SIAM J. DISCRETE MATH.
Vol. 21, No. 1, pp. 1–15

A NEW MIN-CUT MAX-FLOW RATIO FOR MULTICOMMODITY FLOWS[*]

OKTAY GÜNLÜK[†]

Abstract. In this paper we present a new bound on the min-cut max-flow ratio for multicommodity flow problems with specified demands. For multicommodity flows, this is a generalization of the well-known relationship between the capacity of a minimum cut and the value of the maximum flow of a single commodity flow problem. For multicommodity flows, capacity of a cut is scaled by the demand that has to cross the cut to obtain the numerator of this ratio. In the denominator, the maximum concurrent flow value is used. Currently, the best known bound for this ratio is proportional to $\log(k)$, where k is the number of origin-destination pairs with positive demand. Our new bound is proportional to $\log(k^*)$, where k^* is the cardinality of the minimum cardinality vertex cover of the demand graph. To obtain this bound, we start with a so-called aggregated commodity formulation of the maximum concurrent flow problem with k^* commodities. We also show a similar bound for the maximum multicommodity flow problem. The new bound is proportional to $\min\{\log(k^*),\ k^{**}\}$, where k^{**} denotes the size of the minimum cardinality complete bipartite subgraph cover of the demand graph.

Key words. minimum cut, maximum concurrent flow, aggregated commodity formulation

AMS subject classifications. 90C05, 90B10, 68Q25, 68R10

DOI. 10.1137/S089548010138917X

1. Introduction.

In this paper we study multicommodity flow problems and present new bounds on the associated min-cut max-flow ratio. Starting with the pioneering work of Leighton and Rao [13] there has been ongoing research in the area of "approximate min-cut max-flow theorems" for multicommodity flows. We present a summary of previous work later in section 1.3. We next state the well-known min-cut max-flow theorem and present an interpretation of it for flow problems with specified flow requirements. We then clarify what is meant by "minimum cut" and "maximum flow" for multicommodity flow problems.

Throughout the paper, we assume that the input graph is connected and has positive capacity on all edges.

1.1. Single commodity flows.

Given an undirected graph $G = (V, E)$, edge capacities c_e for $e \in E$, and two special nodes $s,\ v \in V$, the well-known min-cut max-flow theorem [6] states that the value of the maximum flow from the *source* node s to the *sink* node v is equal to the capacity of the minimum cut:

$$\min_{S \subset V : s \in S, v \notin S} \left\{ \sum_{e \in \delta(S)} c_e \right\},$$

where $\delta(S) = \{e \in E : |e \cap S| = 1\}$. Let $t \in R_+$ be a specified flow requirement; then the min-cut max-flow theorem implies that t units of flow can be routed from s to v

[*] Received by the editors May 8, 2001; accepted for publication (in revised form) June 15, 2006; published electronically January 8, 2007.
http://www.siam.org/journals/sidma/21-1/38917.html

[†] Department of Mathematical Sciences, IBM Research, Yorktown Heights, NY (oktay@watson.ibm.com).

if and only if *the minimum cut-capacity to cut-load ratio* ρ^*, where

$$\rho^* = \min_{S \subset V : s \in S, v \notin S} \left\{ \frac{\sum_{e \in \delta(S)} \quad c_e}{t} \right\}$$

is at least 1.

We generalize this result to flow requirements with a single common source node and several sink nodes as follows: Given a source node s and a collection of sink nodes $v_q \in V \setminus \{s\}$ for $q \in Q$, it is possible to simultaneously route $t_q \in R_+$ units of flow from s to v_q for all $q \in Q$ if and only if $\rho^* \geq 1$, where

$$\rho^* = \min_{S \subset V : s \in S} \left\{ \frac{\sum_{e \in \delta(S)} \quad c_e}{\sum_{q \in Q \,:\, v_q \notin S} t_q} \right\}.$$

This observation is the main motivation behind our study, as it shows that a min-cut max-flow relationship holds tight for network flow problems (with specified flow requirements) as long as the sink nodes share a common source node. Note that, since G is undirected, the min-cut max-flow relationship also holds tight when there is a single sink node and multiple source nodes.

1.2. Multicommodity flows. A natural extension of this observation is to consider multicommodity flows, where a collection of pairs of vertices $\{s_q, v_q\}$, $q \in Q$, together with a flow requirement t_q for each pair is provided. Let the minimum cut-capacity to cut-load ratio for multicommodity flows be similarly defined as

$$\rho^* = \min_{S \subset V} \left\{ \frac{\sum_{e \in \delta(S)} \quad c_e}{\sum_{q \in Q : |S \cap \{s_q, v_q\}| = 1} t_q} \right\}.$$

In the remainder of the paper we refer to ρ^* as the minimum cut ratio. Clearly, it is possible to simultaneously route t_q units of flow from s_q to v_q for all $q \in Q$, only if $\rho^* \geq 1$. But the converse is not true [16], [18], and a simple counterexample is the complete bipartite graph $K_{2,3}$ with unit capacity edges and unit flow requirements between every pair of nodes that are not connected by an edge.

For multicommodity flows, *metric inequalities* provide the necessary and sufficient conditions for feasibility (see [9] and [19]). More precisely, it is possible to simultaneously route t_q units of flow from s_q to v_q for all $q \in Q$, if and only if the edge capacities satisfy

$$\sum_{e \in E} w_e c_e \;\; \geq \;\; \sum_{q \in Q} dist(s_q, v_q) t_q$$

for all $w \geq 0$, where $dist(u, v)$ denotes the shortest path distance from u to v using w as edge weights. The set of all important edge weights form a well-defined polyhedral cone. Notice that the above example with $K_{2,3}$ does not satisfy the metric inequality "generated" by $w_e = 1$ for all $e \in E$. It is easy to show that the condition $\rho^* \geq 1$ is implied by metric inequalities.

The *maximum concurrent flow* problem is the optimization version of the multicommodity flow feasibility problem (see [22] and [16]). For a given collection of flow requirements and edge capacities, the objective here is to find the maximum value of κ such that $\kappa\, t_q$ units of flow can be simultaneously routed from s_q to v_q for all $q \in Q$. Note that κ can be greater than one.

For a given instance of the multicommodity flow problem, let κ^* denote the value of the maximum concurrent flow. In other words, it is possible to simultaneously route $\kappa\, t_q$ units of flow from s_q to v_q for all $q \in Q$ if and only if $\kappa \leq \kappa^*$. Clearly the maximum concurrent flow value cannot exceed the minimum cut ratio:

$$\rho^* \geq \kappa^*. \tag{1}$$

Our main result in this paper establishes the following reverse relationship between the minimum cut ratio and the maximum concurrent flow value:

$$\kappa^* \geq \frac{1}{c\,\lceil \log k^* \rceil}\,\rho^*, \tag{2}$$

where c is a constant and k^* is the cardinality of the minimum cardinality vertex cover for the demand graph. In other words, k^* is the size of the smallest set $K^* \subseteq V$ such that K^* contains at least one of s_q or v_q for all $q \in Q$. Throughout the paper, we assume that $k^* > 1$.

Combining (1) and (2) we can bound the min-cut max-flow ratio as follows:

$$c\,\lceil \log k^* \rceil \;\; \geq \;\; \frac{\rho^*}{\kappa^*} \;\; \geq \;\; 1. \tag{3}$$

In literature, these bounds are often called "approximate min-cut max-flow theorems," as they relate the maximum (concurrent) flow of a multicommodity flow problem to the (scaled) capacity of the minimum cut. As discussed above, this bound is tight, i.e., $\rho^* = \kappa^*$, when $k^* = 1$.

1.3. Related work. Starting with the pioneering work of Leighton and Rao [13] there has been ongoing interest in the area of approximate min-cut max-flow theorems. The first such result in [13] shows that the upper bound in (3) is at most $O(\log |V|)$ when $t_q = 1$ for all $q \in Q$. Later Klein et al. [12] extend this result to general t_q and show that the bound is $O(\log C \log D)$, where D is the sum of (integral) demands (i.e., $D = \sum_{q\in Q} t_q$) and C is the sum of (integral) capacities (i.e., $C = \sum_{e\in E} c_e$). Tragoudas [23] has later improved this bound to $O(\log |V| \log D)$ and Garg, Vazirani, and Yannakakis [8] have further improved it to $O(\log k \log D)$, where $k = |Q|$.

Plotkin and Tardos [21] present the first bound that does not depend on the input data by showing that the upper bound in (3) is at most $O(\log^2 k)$. Finally, Linial, London, and Rabinovich [14] and Aumann and Rabani [1] independently show that the bound is at most $O(\log k)$.

Our result improves this best known bound to $O(\log k^*)$. To emphasize the difference between $O(\log k)$ and $O(\log k^*)$, we note that for an instance of the multicommodity flow problem with a single source node and $|V|-1$ sink nodes, $k = |V|-1$, whereas $k^* = 1$. In general, $k \geq k^* \geq k/|V|$.

The paper is organized as follows: In section 2, we present a linear programming formulation of the maximum concurrent flow problem using aggregate commodities. A commodity in this formulation combines all demand requirements with a common source node. In section 3, we show the $O(\log k^*)$ bound using this formulation. In section 4, we discuss geometric implications of this result. Finally, in section 5, we show similar bounds for the so-called maximum multicommodity flow problem. In the linear programming formulation of this problem, an aggregate commodity combines all demand requirements that form a complete bipartite subgraph of the demand graph.

2. Formulation. When formulating a multicommodity problem as a linear program, what is meant by a "commodity" can affect the size of the formulation significantly. Even though this has been noticed and exploited by researchers interested in solving these linear programs (see, for example, [2] and [15]), it has been overlooked by researchers interested in the theoretical aspects of multicommodity flows. We next present a formulation for the concurrent flow problem where each commodity aggregates flow requirements with a common source node. We note that the original linear programming formulation of the maximum concurrent flow problem presented in Shahrokhi and Matula [22] also uses aggregate commodities.

2.1. The concurrent flow problem. Given an undirected graph $G = (V, E)$, edge capacities c_e for $e \in E$, and flow requirements t_q for given pairs of vertices $\{s_q, v_q\}$, for all $q \in Q$, let T denote the corresponding flow requirement matrix. More precisely, $T_{[k,u]} = \sum_{q \in Q \,:\, s_q = k,\, v_q = u} t_q$ for all $k, u \in V$. We then define the set of "source nodes" $K \subseteq V$ to be $K = \{k \in V \ : \ \sum_{u \in V} T_{[k,u]} > 0\}$ and formulate the maximum concurrent flow problem as follows:

$$\mathit{Maximize} \qquad \kappa$$

$$\mathit{Subject\ to}$$

$$\begin{aligned}
\sum_{v:\{v,u\}\in E} f^k_{vu} \; - \sum_{v:\{u,v\}\in E} f^k_{uv} \quad &= \quad \kappa\, T_{[k,u]} && \forall\ u \in V,\ k \in K \ \mathit{with}\ u \neq k,\\
\sum_{v:\{v,k\}\in E} f^k_{vk} \; - \sum_{v:\{k,v\}\in E} f^k_{kv} \quad &= \quad -\ \kappa \sum_{u \in V} T_{[k,u]} && \forall\ \ k \in K,\\
\sum_{k \in K} \left(f^k_{uv} + f^k_{vu}\right) \quad &\leq \quad c_{\{u,v\}} && \forall\ \{u,v\} \in E,\\
\kappa \ \geq \ 0, \qquad f^k_{uv} \quad &\geq \quad 0 && \forall\ k \in K,\ \mathit{and}\ \{u,v\} \in E,
\end{aligned}$$

where variable f^k_{vu} denotes the flow of commodity k from node v to node u, and variable κ denotes the value of the concurrent flow. Note that as the commodities are defined with respect to source nodes, f^k_{vu} gives the total flow on edge (v, u) that has originated at node k. The destination of the flow, however, is not specified by the commodity (unlike the "natural formulation" where each source-sink pair defines a commodity).

Given a flow vector f, it is easy to find disaggregated flows for node pairs (k, u) with $T_{[k,u]} > 0$. More precisely, for an aggregate commodity k, if $T_{[k,u]} > 0$ for some node u, then it is possible to trace $\kappa T_{[k,u]}$ units of commodity k entering node u back to its origin k. Reducing aggregate flows as disaggregate flows are identified, and repeating this process iteratively gives a routing of source-sink flows. The disaggregation, however, is not necessarily unique.

2.2. A reformulation of the concurrent flow problem. To find the smallest set of commodities that would model the problem instance correctly, we do the following: Let $G^D = (V, E^D)$ denote the (undirected) demand graph, where $E^D = \{\{s_q, v_q\} \ : \ q \in Q\}$. We first find a minimum cardinality vertex cover $K^* \subseteq V$ of G^D. In other words, K^* is a smallest cardinality set that satisfies $\{s_q, v_q\} \cap K^* \neq \emptyset$ for all $q \in Q$. We then inspect all $q \in Q$, and if $s_q \notin K^*$ and $v_q \in K^*$ for some q, we rename the source node to be v_q so that we have $s_q \in K^*$ for all $q \in Q$. Note that this can be done without loss of generality since the capacity constraints in the formulation do not depend on the orientation of the flow. After this change, the corresponding flow requirement matrix T has the property that $T_{[k,u]} > 0$ only if $k \in K^*$.

We therefore obtain a formulation with $|K^*|$ commodities. In the remainder of the paper we assume that $K = K^*$. Next, we present a slightly modified version of this formulation:

$$\begin{array}{ll} \textit{Maximize} & \kappa \\ \textit{Subject to} & \end{array}$$

$$\begin{array}{rcll} \displaystyle\sum_{v:\{u,v\}\in E} f^k_{uv} - \sum_{v:\{v,u\}\in E} f^k_{vu} + \kappa\, T_{[k,u]} & \leq & 0 & \forall\, u \in V,\ k \in K^* \textit{ with } u \neq k, \\ \displaystyle\sum_{k\in K^*} \left(f^k_{uv} + f^k_{vu}\right) & \leq & c_{\{u,v\}} & \forall\, \{u,v\} \in E, \\ \kappa \ \textit{free},\ f^k_{uv} & \geq & 0 & \forall\, k \in K^*,\ \textit{and}\ \{u,v\} \in E, \end{array}$$

where (i) we have deleted the flow balance equalities for the source nodes $k \in K^*$, (ii) changed the flow balance equalities for the remaining nodes to inequality, and (iii) relaxed the nonnegativity requirement for κ. Note that these modifications do not affect the value of the optimal solution.

The dual of this formulation is

$$\begin{array}{ll} \textit{Minimize} & \displaystyle\sum_{\{u,v\}\in E} c_{\{u,v\}}\ w_{\{u,v\}} \\ \textit{Subject to} & \end{array}$$

$$\begin{array}{rcll} \displaystyle\sum_{k\in K^*}\sum_{u\in V} T_{[k,u]}\ y^k_u & = & 1, & \\ y^k_v - y^k_u + w_{\{u,v\}} & \geq & 0 & \multirow{2}{*}{$\left.\right\}\ \forall\, k \in K^*,\ \textit{and}\ \{u,v\} \in E,$} \\ y^k_u - y^k_v + w_{\{u,v\}} & \geq & 0 & \\ y^k_k & = & 0 & \forall\, k \in K^*, \\ y^k_u & \geq & 0 & \forall\, u \in V,\ k \in K^* \textit{ with } u \neq k, \\ w_{\{u,v\}} & \geq & 0 & \forall\, \{u,v\} \in E, \end{array}$$

where dual variables y^k_k for $k \in K^*$ are included in the formulation even though there are no corresponding primal constraints. These variables are set the zero in a separate constraint. The main reason behind reformulating the primal problem and using redundant variables in the dual problem is to obtain a dual formulation that would have an optimal solution that satisfies the following properties.

Proposition 1. *Let $[\bar{y}, \bar{w}]$ be an optimal solution to the dual problem, and let $\hat{y} \in R^{|V|\times|V|}$ be the vector of shortest path distances (using $\bar{w}$ as edge weights) with $\hat{y}^k_u$ denoting distance from node k to u.*

(i) *For any $k \in K^*$ and $u \in V$, with $T_{[k,u]} > 0$, $\bar{y}^k_u$ is equal to $\hat{y}^k_u$.*

(ii) *For any $\{u,v\} \in E$, $\bar{w}_{\{u,v\}}$ is equal to $\hat{y}^u_v$.*

Proof. (i) For any $k \to u$ path $P = \{\{k, v_1\}, \{v_1, v_2\}, \ldots, \{v_{|P|-1}, u\}\}$ we have $\sum_{e \in P} w_e \geq \bar{y}^k_u$, implying $\hat{y}^k_u \geq \bar{y}^k_u$. If $\hat{y}^k_u > \bar{y}^k_u$ for some $k \in K^*$, $u \in V$ with $T_{[k,u]} > 0$, we can write $\sum_{k \in K^*} \sum_{u \in V} T_{[k,u]}\, \hat{y}^k_u = \sigma > \sum_{k \in K^*} \sum_{u \in V} T_{[k,u]}\, \bar{y}^k_u = 1$. This implies that a new solution, with an improved objective function value, can be constructed by scaling $[\hat{y}, \bar{w}]$ by $1/\sigma$, a contradiction.

(ii) Clearly, $\bar{w}_{\{u,v\}} \geq \hat{y}^u_v$. If $\bar{w}_{\{u,v\}} > \hat{y}^u_v$, replacing $\bar{w}_{\{u,v\}}$ by $\hat{y}^u_v$ in the solution improves the objective function value, a contradiction (remember that $c_{\{u,v\}} > 0$ for all $\{u, v\} \in E$). □

As a side remark, we note that it is therefore possible to substitute some of the dual variables, and consequently it is possible to combine some of the constraints in the primal formulation.

We next express the maximum concurrent flow value as a ratio of a weighted sum of edge capacities and a weighted sum of flow requirements. Notice that this expression resembles the minimum cut ratio ρ^*.

COROLLARY 2. *Let κ^* be the optimal value of the primal* (*or the dual*) *problem. Then,*

$$\kappa^* \quad = \quad \frac{\sum_{\{u,v\} \in E} c_{\{u,v\}}\ dist(u, v)}{\sum_{k \in K^*} \sum_{v \in V} T_{[k,v]}\ dist(k, v)}, \tag{4}$$

where $dist(u, v)$ denotes the shortest path distance from node u to node v with respect to the edge weight vector $\bar{w}$ of Proposition 1.

3. The min-cut max-flow ratio. We next argue that there exists a mapping $\Phi : V \to R^p_+$ for some p, such that $||\Phi(u) - \Phi(v)||_1$ is not very different from $dist(u, v)$ for node pairs $\{u, v\}$ that are of interest. We then substitute $||\Phi(u) - \Phi(v)||_1$ in place of $dist(u, v)$ in (4) and relate the new right-hand side of (4) to the minimum cut ratio. More precisely, we show that

$$\kappa^* \geq \frac{1}{\alpha} \times \frac{\sum_{\{u,v\} \in E} c_{\{u,v\}}\ ||\Phi(u) - \Phi(v)||_1}{\sum_{k \in K^*} \sum_{v \in V} T_{[k,v]}\ ||\Phi(k) - \Phi(v)||_1} \ \geq\ \frac{1}{\alpha}\ \rho^*.$$

3.1. Mapping the nodes of the graph with small distortion. Our approach follows the general structure of the proof of a related result by Bourgain [3] that shows that any n-point metric space can be embedded into l_1 with logarithmic distortion. We state this result more precisely in section 4.

Given an undirected graph $G = (V, E)$ with edge weights $w_e \geq 0$, for $e \in E$, let $d(u, v)$ denote the shortest path distance from $u \in V$ to $v \in V$ using w as edge weights. For $v \in V$ and $S \subseteq V$ let $d(v, S) = \min_{k \in S}\{d(v, k)\}$ and define $d(v, \emptyset) = \sigma = \sum_{u \in V} \sum_{v \in V} d(u, v)$. Furthermore, let $K \subseteq V$ with $|K| > 1$ also be given.

For any $l, t \geq 1$, let Q^t_l be a random subset of K such that members of Q^t_l are chosen independently and with equal probability $P(k \in Q^t_l) = 1/2^t$ for all $k \in K$. Note that for all $l \geq 1$, Q^t_l has an identical probability distribution and $E[|Q^t_l|] = |K|/2^t$. For $m = \lceil \log(|K|) \rceil$ and $L = 300 \cdot \lceil \log(|V|) \rceil$, define the following (random) mapping $\Phi^R : V \to R^{mL}_+$:

$$\Phi^R(v) = \frac{1}{L \cdot m} \begin{bmatrix} d(v, Q^1_1) & d(v, Q^2_1) & \ldots & d(v, Q^m_1) \\ d(v, Q^1_2) & d(v, Q^2_2) & \ldots & d(v, Q^m_2) \\ \vdots & \vdots & \ddots & \vdots \\ d(v, Q^1_L) & d(v, Q^2_L) & \ldots & d(v, Q^m_L) \end{bmatrix}.$$

Note that $|d(u,S) - d(v,S)| \leq d(u,v)$ for any $S \subseteq V$, and therefore

$$\begin{aligned} ||\Phi^R(u) - \Phi^R(v)||_1 &= \frac{1}{L \cdot m} \sum_{t=1}^{m} \sum_{l=1}^{L} \left| d(u, Q_l^t) - d(v, Q_l^t) \right| \\ &\leq \frac{1}{L \cdot m} \cdot L \cdot m \cdot d(u,v) \;=\; d(u,v) \qquad (5) \end{aligned}$$

for all $u, v \in V$.

Notice that as all $Q_l^t \subseteq K$, the expression $||\Phi^R(u) - \Phi^R(v)||_1$ does not give a good approximation of $d(u,v)$ when $u, v \notin K$. To see this consider an example where for some $u, v \in V \setminus K$, we have $d(u,k) = d(v,k)$ for all $k \in K$. In this case $||\Phi^R(u) - \Phi^R(v)||_1 = 0$, whereas the actual distance $d(u,v)$ between the two nodes can be strictly positive. We next bound $||\Phi^R(u) - \Phi^R(v)||_1$ from below when $|\{u,v\} \cap K| \geq 1$. Note that the bound on the distortion of the mapping depends on the size of the set K from which Q_l^t's are chosen.

Lemma 3. *For all $u \in K$ and $v \in V$ and for some $\alpha = O(\log |K|)$ the property*

$$||\Phi^R(u) - \Phi^R(v)||_1 \geq \frac{1}{\alpha} \cdot d(u,v)$$

holds simultaneously with positive probability.

Proof. For any $v \in V$ let $B(v,\delta) = \{k \in K : d(v,k) \leq \delta\}$ and $B^o(v,\delta) = \{k \in K : d(v,k) < \delta\}$, respectively, denote the collection of members of K that lie within the closed and open balls around v. We next define a sequence of δ's for pairs of nodes.

For any fixed $u \in K$ and $v \in V$ let

$$t^*_{uv} = \max\Big\{1, \Big\lceil \log\Big(\max\Big\{|B(u, d(u,v)/2)|, |B(v, d(u,v)/2)|\Big\}\Big)\Big\rceil\Big\}$$

and define

$$\delta^t_{uv} = \begin{cases} 0, & t = 0, \\ \max\{\delta \geq 0 \;:\; |B^o(u,\delta)| < 2^t \text{ and } |B^o(v,\delta)| < 2^t\}, & t^*_{uv} > t > 0, \\ d(u,v)/2, & t = t^*_{uv}. \end{cases}$$

We use the following three observations in the the proof:

1. $m = \lceil \log(|K|) \rceil \geq t^*_{uv} > 0$,
2. $\max\{|B(u, \delta^t_{uv})|, |B(v, \delta^t_{uv})|\} \geq 2^t \quad \forall\, t < t^*_{uv}$, and
3. $\min\{|B^o(u, \delta^t_{uv})|, |B^o(v, \delta^t_{uv})|\} < 2^t \quad \forall\, t \leq t^*_{uv}$.

For a fixed $u, v \in V$, and $t \geq 0$ such that $t < t^*_{uv}$, rename u and v as z_{max} and z_{other} so that $|B(z_{max}, \delta^t_{uv})| \geq |B(z_{other}, \delta^t_{uv})|$. Using $\frac{1}{e} \geq (1 - \frac{1}{x})^x \geq \frac{1}{4}$, for any $x \geq 2$, we can write the following for any Q_l^{t+1} for $L \geq l \geq 1$:

$$\begin{aligned} P\Big(Q_l^{t+1} \cap B(z_{max}, \delta^t_{uv}) = \emptyset\Big) &= \Big(1 - 2^{-(t+1)}\Big)^{|B(z_{max}, \delta^t_{uv})|} \\ &\leq \big(1 - 2^{-(t+1)}\big)^{2^t} \quad \leq \quad e^{-\frac{1}{2}}, \\ P\Big(Q_l^{t+1} \cap B^o(z_{other}, \delta^{t+1}_{uv}) = \emptyset\Big) &= \Big(1 - 2^{-(t+1)}\Big)^{|B^o(z_{other}, \delta^{t+1}_{uv})|} \\ &\geq \big(1 - 2^{-(t+1)}\big)^{2^{t+1}} \quad \geq \quad \tfrac{1}{4}. \end{aligned}$$

Notice that $Q_l^{t+1} \cap B(z_{max}, \delta_{uv}^t) \neq \emptyset$ implies that $d(z_{max}, Q_l^{t+1}) \leq \delta_{uv}^t$, and similarly, $Q_l^{t+1} \cap B^o(z_{other}, \delta_{uv}^{t+1}) = \emptyset$ implies that $d(z_{other}, Q_l^{t+1}) \geq \delta_{uv}^{t+1}$. Using the independence of the two events (since the two balls are disjoint) we can now write

$$P\Big(Q_l^{t+1} \cap B(z_{max}, \delta_{uv}^t) \neq \emptyset \text{ and } Q_l^{t+1} \cap B^o(z_{other}, \delta_{uv}^{t+1}) = \emptyset\Big) \geq \left(1 - e^{-\frac{1}{2}}\right) \times \frac{1}{4} \geq \frac{1}{11},$$

and therefore

$$P\Big(\left| d(z_{other}, Q_l^{t+1}) - d(z_{max}, Q_l^{t+1}) \right| \geq \delta_{uv}^{t+1} - \delta_{uv}^t \Big) \geq \frac{1}{11}$$

or, equivalently,

$$P\Big(\left| d(u, Q_l^{t+1}) - d(v, Q_l^{t+1}) \right| \geq \delta_{uv}^{t+1} - \delta_{uv}^t \Big) \geq \frac{1}{11}$$

for all $t < t_{uv}^*$.

Let X_{uv}^{tl} be a random variable taking value 1 if $\left| d(u, Q_l^{t+1}) - d(v, Q_l^{t+1}) \right| \geq \delta_{uv}^{t+1} - \delta_{uv}^t$, and 0 otherwise. Note that for any fixed $u \in K$ and $v \in V$ if $\sum_{l=1}^{L} X_{uv}^{tl} \geq L/22$ (that is, at least one-half the expected number) for all $t < t_{uv}^*$, then we can write

$$\begin{aligned} ||\Phi^R(u) - \Phi^R(v)||_1 &= \frac{1}{L \cdot m} \quad \sum_{t=1}^{m} \sum_{l=1}^{L} \left| d(u, Q_l^t) - d(v, Q_l^t) \right| \\ &\geq \frac{1}{L \cdot m} \quad \sum_{t=1}^{t_{uv}^*} \frac{L}{22} \quad \left(\delta_{uv}^t - \delta_{uv}^{t-1}\right) \\ &= \frac{1}{22m} \quad \left(\delta_{uv}^{t_{uv}^*} - \delta_{uv}^0\right) = \frac{d(u,v)}{44m}. \end{aligned}$$

To this end, we first use the Chernoff bound (see, for example, [17, Chapter 4]) to claim that

$$P\left(\sum_{l=1}^{L} X_{uv}^{tl} < \frac{1}{2} \times \frac{L}{11} \right) < e^{-\frac{1}{2} \times \frac{1}{4} \times \frac{L}{11}} \;=\; e^{-\frac{L}{88}}$$

for any $u \in K$, $v \in V$, and $t < t_{uv}^*$, which, in turn, implies that

$$P\left(\sum_{l=1}^{L} X_{uv}^{tl} < \frac{L}{22} \quad \textit{for some } u \in K,\ v \in V,\ \textit{and } t < t_{uv}^* \right) < |K||V| \lceil \log(|K|) \rceil \, e^{-L/88},$$

where the right-hand side of the inequality is less than 1 for $L \geq 88(3 \cdot \log(|V|)$. Therefore, with positive probability, $\sum_{l=1}^{L} X_{uv}^{tl} \geq \frac{L}{22}$ for all $u \in K,\ v \in V,$ and $t < t_{uv}^*$, which implies that, with positive probability,

$$||\Phi^R(u) - \Phi^R(v)||_1 \geq \frac{d(u,v)}{44m}$$

for all $u \in K,\ v \in V$. □

An immediate corollary of this result is the existence of a (deterministic) mapping with at most $\log(|K|)$ distortion.

COROLLARY 4. *There exists a collection of sets $\bar{Q}_l^t \subseteq K$ for $m \geq t \geq 1$ and $L \geq l \geq 1$ such that the corresponding mapping $\Phi^D : V \to R_+^{mL}$ satisfies the following two properties:*

(i) $d(u,v) \geq ||\Phi^D(u) - \Phi^D(v)||_1 \quad \forall\ u, v \in V;$

(ii) $d(u,v) \leq \alpha\ ||\Phi^D(u) - \Phi^D(v)||_1 \quad \forall\ u \in K$ *and* $v \in V,$
where $\alpha = c\ \log |K|$ *for some constant c.*

3.2. Bounding the maximum concurrent flow value. Combining Corollaries 2 and 4, we now bound the maximum concurrent flow value as follows:

$$\kappa^* = \frac{\sum_{\{u,v\}\in E} c_{\{u,v\}}\ dist(u,v)}{\sum_{k\in K^*}\sum_{v\in V} T_{[k,v]}\ dist(k,v)} \geq \frac{\sum_{\{u,v\}\in E} c_{\{u,v\}}\ ||\Phi^D(u) - \Phi^D(v)||_1}{\sum_{k\in K^*}\sum_{v\in V} T_{[k,v]}\ \alpha\ ||\Phi^D(k) - \Phi^D(v)||_1}$$

$$= \frac{1}{\alpha} \times \frac{\sum_{t=1}^m \sum_{l=1}^L \left(\sum_{\{u,v\}\in E} c_{\{u,v\}}\ |d(u,\bar{Q}_l^t) - d(v,\bar{Q}_l^t)|\right)}{\sum_{t=1}^m \sum_{l=1}^L \left(\sum_{k\in K^*}\sum_{v\in V} T_{[k,v]}\ |d(k,\bar{Q}_l^t) - d(v,\bar{Q}_l^t)|\right)}$$

$$\geq \frac{1}{\alpha} \times \frac{\sum_{\{u,v\}\in E} c_{\{u,v\}}\ |d(u,Q^*) - d(v,Q^*)|}{\sum_{k\in K^*}\sum_{v\in V} T_{[k,v]}\ |d(k,Q^*) - d(v,Q^*)|} \tag{6}$$

for $Q^* = Q_{l*}^{t*}$ for some $m \geq t^* \geq 1$ and $L \geq l^* \geq 1$. Note that we have essentially bounded maximum concurrent flow value (from below) by a collection of cut ratios. We next bound it by the minimum cut ratio.

First, we assign indices $\{1, 2, \ldots, |V|\}$ to nodes in V so that $d(v_p, Q^*) \geq d(v_{p-1}, Q^*)$ for all $|V| \geq p \geq 2$, and let $x_p = d(v_p, Q^*)$. Next, we define $|V|$ nested sets $S_p = \{v_j \in V : j \leq p\}$ and the associated cuts $C_p = \{\{u,v\} \in E : |\{u,v\} \cap S_p| = 1\}$ and $T_p = \{(k,v) \in K^* \times V : |\{k,v\} \cap S_p| = 1\}$. We can now rewrite the summations in (6) as follows:

$$\frac{1}{\alpha} \times \frac{\sum_{\{v_i,v_j\}\in E} c_{\{v_i,v_j\}}\ |x_i - x_j|}{\sum_{v_i\in K^*}\sum_{v_j\in V} T_{[v_i,v_j]}\ |x_i - x_j|} = \frac{1}{\alpha} \times \frac{\sum_{p=2}^{|V|} (x_p - x_{p-1}) \sum_{\{u,v\}\in C_p} c_{\{u,v\}}}{\sum_{p=2}^{|V|} (x_p - x_{p-1}) \sum_{(k,v)\in T_p} T_{[k,v]}}$$

$$\geq \frac{1}{\alpha} \times \frac{\sum_{\{u,v\}\in C_{p^*}} c_{\{u,v\}}}{\sum_{(k,v)\in T_{p^*}} T_{[k,v]}} \geq \frac{1}{\alpha}\ \rho^*$$

for some $p^* \in \{1, \ldots, |V|\}$. We have therefore shown the following theorem.

THEOREM 5. *Given a multicommodity problem, let κ^* denote the maximum concurrent flow value, ρ^* denote the minimum cut ratio, and k^* denote the cardinality of the minimum cardinality vertex cover of the associated demand graph. If $k^* > 1$, then*

$$c\ \lceil \log\ k^* \rceil \geq \frac{\rho^*}{\kappa^*} \geq 1$$

for some constant c.

3.3. A tight example. We next show that there are problem instances for which the above bound on the min-cut max-flow ratio is tight, up to a constant. This result is a relatively straightforward extension of a similar result by Leighton and Rao [13].

Lemma 6. *For any given $n, k^* \in Z_+$ with $n \geq k^*$, it is possible to construct an instance of the multicommodity problem with n nodes and k^* (minimal) aggregate commodities such that*

$$\frac{\rho^*}{\kappa^*} \geq c \lceil \log k^* \rceil$$

for some constant c.

Proof. We start with constructing a bounded-degree expander graph G^{k^*} with k^* nodes and $O(k^*)$ edges. See, for example, [17] for a definition, and existence of constant degree expander graphs. As discussed in [13], these graphs (with unit capacity for all edges and unit flow requirement between all pairs of vertices) provide examples with $\rho^*/\kappa^* \geq c \lceil \log k^* \rceil$ for some constant c. Note that the demand graph is complete and therefore the minimum cardinality vertex cover has size k^*.

We next augment G^{k^*} by adding $n - k^*$ new vertices and $n - k^*$ edges. Each new vertex has degree one and is connected to an arbitrary vertex of G^{k^*}. The new edges are assigned arbitrary capacities. The augmented graph, with the original flow requirements, has n nodes and satisfies $\rho^*/\kappa^* \geq c \lceil \log k^* \rceil$. □

4. Geometric interpretation.

Both of the more recent studies (namely Linial, London, and Rabinovich [14] and Aumann and Rabani [1]) that relate the min-cut max-flow ratio to the number of origin-destination pairs in the problem instance take a geometric approach and base their results on the fact that a finite metric space can be mapped into a Euclidean space with logarithmic distortion. More precisely, they base their analysis on the following result that shows that n points can be mapped from l_∞^n to l_1^p with $O(\log n)$ distortion (where l_b^a denotes R^a equipped with the norm $||x||_b = (\sum_{i=1}^a |x_i|^b)^{1/b}$).

Lemma 7 (Bourgain [3]; also see [14]). *Given n points $x_1, \ldots, x_n \in R^n$, there exists a mapping $\Phi : R^n \to R^p$, with $p = O(\log n)$, that satisfies the following two properties:*

$$\text{(i)} \quad ||x_i - x_j||_\infty \quad \geq \quad ||\Phi(x_i) - \Phi(x_j)||_1 \quad \forall\, i, j \leq n;$$

$$\text{(ii)} \quad ||x_i - x_j||_\infty \quad \leq \quad \alpha \quad ||\Phi(x_i) - \Phi(x_j)||_1 \quad \forall\, i, j \leq n,$$

where $\alpha = c \log n$ for some constant c. □

Using this result, it is possible to map the optimal dual solution of the disaggregated (one commodity for each source-sink pair) formulation to l_1^p with logarithmic distortion; see [14] and [1]. One can then show a $O(\log k)$ bound by using arguments similar to the ones presented in section 3.2.

We next give a geometric interpretation of Corollary 4 in terms of mapping n points from l_∞^m to l_1^p with logarithmic distortion with respect to a collection of "seed" points.

Lemma 8. *Given n points $x_1, \ldots, x_n \in R^m$, the first $t \leq n$ of which are special, $t > 1$, there exists a mapping $\Phi : R^m \to R^p$ with $p = O(\log n)$ that satisfies the following two properties:*

$$\text{(i)} \quad ||x_i - x_j||_\infty \quad \geq \quad ||\Phi(x_i) - \Phi(x_j)||_1 \quad \forall\, i, j \leq n;$$

$$\text{(ii)} \quad ||x_i - x_j||_\infty \quad \leq \quad \alpha \quad ||\Phi(x_i) - \Phi(x_j)||_1 \quad \forall\, i \leq t, \quad j \leq n,$$

where $\alpha = c \log t$ for some constant c.

Proof. Let $G = (V, E)$ be a complete graph with n nodes where each node v_i is associated with point x_i for $i = 1, \ldots, n$. For $e = \{v_i, v_j\} \in E$, let $w_e = ||x_i - x_j||_\infty$ be the edge weight. Furthermore, let $d(v_i, v_j)$ denote the shortest path length between nodes $v_i, v_j \in V$ using w as edge weights. Note that

$$||x_i - x_j||_\infty \leq ||x_i - x_k||_\infty + ||x_k - x_j||_\infty$$

for any $i, j, k \leq n$, and therefore $d(v_i, v_j) = ||x_i - x_j||_\infty$ for all $i, j \leq n$. We can now use Corollary 4 to show the existence of a mapping $\Phi' : R^m \to R^q$ with $q = O(\log\, n\, \log\, t)$ that satisfies the desired properties.

To decrease the dimension of the image space, we scale Φ' by $\sqrt{Lm}$ to map the points $x_1, \ldots, x_n$ to l_2^q with $c'\ \log\, t$ distortion. More precisely, we use $\Phi'' : R^m \to R^q$, where $\Phi''(x) = \sqrt{Lm}\ \Phi'(x)$. It is easy to see that

$$\begin{aligned} \text{(i)} \quad ||\Phi''(x_i) - \Phi''(x_j)||_2 &\leq \sqrt{(1/Lm) \sum_{k=1}^{m} \sum_{q=1}^{L} d(v_i, v_j)^2} \\ &= d(v_i, v_j) = ||x_i - x_j||_\infty, \\ \text{(ii)} \quad ||\Phi''(x_i) - \Phi''(x_j)||_2 &\geq ||\Phi'(x_i) - \Phi'(x_j)||_1 \\ &\geq c'\ \log\, t\ d(v_i, v_j) = c'\ \log\, t\ ||x_i - x_j||_\infty. \end{aligned}$$

We can now use the following two facts (also used in [14]) to reduce the dimension of the image space to $O(\log\, n)$: (i) for any $q \in Z_+$, n points can be mapped from l_2^q to l_2^p, where $p = O(\log\, n)$ with constant distortion (see [10]), and (ii) for any $p \in Z_+$, l_2^p can be embedded in l_1^{2p} with constant distortion(see [20, Chapter 6]). □

We also note that in Lemma 8 (and Lemma 7), mapping Φ actually satisfies $||x' - x''||_\infty \ \geq \ ||\Phi(x') - \Phi(x'')||_1$ for all $x', x'' \in R^m$ (R^n).

5. Maximum multicommodity flows. The *"maximum multicommodity flow"* problem is a generalization of the (single commodity) maximum flow problem. Given an undirected graph $G = (V, E)$ with edge capacities c_e for $e \in E$, the objective here is to maximize the sum of flows that can be simultaneously sent between given pairs of vertices $\{s_q, v_q\},\ q \in Q$. For this problem, the generalization of the minimum cut is the so-called minimum multicut, which is a collection of edges (of minimum total capacity) that separates s_q from v_q for all $q \in Q$. We denote the capacity of a multicut $\Delta \subseteq E$ by $C(\Delta) = \sum_{e \in \Delta} c_e$.

We next present two formulations for this problem and describe new bounds on the ratio of the minimum multicut capacity to the maximum multicommodity flow.

5.1. The maximum multicommodity flow problem. As in section 2.2, let $K^* \subseteq V$ be a minimum cardinality vertex cover of the demand graph $G^T = (V, E^T)$, where $E^T = \{\{s_q, v_q\} \in V \times V\ :\ q \in Q\}$, and let $T_k = \{v \in V\ :\ \{k, v\} \in E^T\}$ denote the set of sink nodes for $k \in K^*$. The problem can be formulated as follows:

$$\textit{Maximize} \qquad \sum_{k \in K^*} \sum_{u \in T_k} x_u^k$$

Subject to

$$\begin{array}{rcll} \displaystyle\sum_{v:\{u,v\} \in E} f_{uv}^k - \sum_{v:\{v,u\} \in E} f_{vu}^k \ + \quad x_u^k & \leq & 0 & \forall\ u \in V,\ k \in K^*\ \textit{with}\ k \neq u, \\ \displaystyle\sum_{k \in K^*} \left(f_{uv}^k + f_{vu}^k\right) & \leq & c_{\{u,v\}} & \forall\ \{u, v\} \in E, \\ x_u^k \geq 0, \quad f_{uv}^k & \geq & 0 & \forall\ k \in K^*,\ \textit{and}\ \{u, v\} \in E, \end{array}$$

where variable f^k_{uv} denotes the flow of commodity k from node u to node v and x^k_u denotes the total flow of commodity k that terminates at node u. The dual of this formulation is

$$\begin{array}{lll}
\textit{Minimize} & \displaystyle\sum_{\{u,v\}\in E} c_{\{u,v\}}\ w_{\{u,v\}} & \\
\textit{Subject to} & & \\
& \left.\begin{array}{lcl} y^k_v - y^k_u + w_{\{u,v\}} & \geq & 0 \\ y^k_u - y^k_v + w_{\{u,v\}} & \geq & 0 \end{array}\right\} & \forall\ k \in K^*,\ \textit{and}\ \{u,v\} \in E, \\
& y^k_u \quad \geq \begin{cases} 1 & \forall\ k \in K^*,\ u \in T_k, \\ 0 & \forall\ k \in K^*,\ u \in V \setminus T_k, \end{cases} & \\
& y^k_k \quad = \quad 0 \quad \forall\ k \in K^*, & \\
& w_{\{u,v\}} \quad \geq \quad 0 \quad \forall\ \{u,v\} \in E, &
\end{array}$$

where variable y^k_u can be interpreted as the shortest path distance from node k to node u using w as edge weights. Note that any feasible solution to the dual problem assigns weights to the edges in such a way that the shortest path distance from any $k \in K^*$ to any one of its sink nodes is at least 1.

We next state a $O(\log\ k^*)$ bound on the associated min-cut max-flow ratio. This improves the previous best known bound of $O(\log\ k)$ (where k denotes the number of origin-destination pairs) presented in Garg, Vazirani, and Yannakakis [8].

Lemma 9. *Given a maximum multicommodity flow problem, let F^* denote the maximum total flow, $C(\Delta^*)$ denote the capacity of the minimum multicut, and k^* denote the cardinality of the minimum cardinality vertex cover of the associated demand graph. If $k^* > 1$, then*

$$c\ \lceil \log\ k^* \rceil \geq \frac{C(\Delta^*)}{F^*} \geq 1$$

for some constant c.

Proof. Clearly, capacity of any multicut is an upper bound on the total flow implying $C(\Delta^*)/F^* \ \geq\ 1$. For the upper bound, we use the algorithm presented in Garg, Vazirani, and Yannakakis [8] with the input set $V' = K^*$ and an optimal dual solution vector w^*. Given edge weights w, this (constructive) algorithm produces a multicut that separates any $k \in V'$ from vertices that have a shortest path distance of 1 or more from k. The multicut is guaranteed to have a capacity of at most $c\ \lceil \log |V'| \rceil\ (\sum_{\{u,v\}\in E} c_{\{u,v\}}\ w_{\{u,v\}})$ for some constant c. In [8], the authors use this algorithm with $V' = \{s_q\ :\ q \in Q\}$ to prove a $\log(k)$ bound. □

We note that in [8], the authors use their algorithm with $V' = \{s_q\ :\ q \in Q\}$ to prove a $\log(k)$ bound. Garg, in his unpublished thesis [7], however, also observes that their algorithm can be used with the minimum cardinality vertex cover. This (independent) result has been brought to our attention by Chekuri [4].

Also note that if the w variables in the dual linear program are required to be integral, any feasible (integral) solution to the dual problem gives a multicut for the maximum multicommodity flow problem and the optimal solution gives a minimum multicut. Therefore, Lemma 9 implies that the integrality gap of this formulation of the minimum multicut problem is bounded by a factor of $O(\log\ k^*)$.

5.2. A reformulation of the maximum multicommodity flow problem. A more compact formulation of the maximum multicommodity flow problem (i.e., a formulation with fewer variables) can be obtained by allowing a commodity to have multiple source nodes in addition to multiple sink nodes.

Let a *complete bipartite subgraph cover* of a graph be a collection of subgraphs of the graph that satisfy the following two properties: (i) each subgraph is a complete bipartite graph, and (ii) the edges of the subgraphs cover the edges of the graph. See Fishburn and Hammer [5] for a detailed study of these covers. Notice that the complete bipartite subgraph (CBS) cover is a generalization of the vertex cover in the sense that given a vertex cover K, one can construct a (CBS) cover $\mathcal{B}$ with $|K| = |\mathcal{B}|$. We next formulate the problem using a (CBS) cover of the demand graph.

As in section 5.1, let $G^T = (V, E^T)$ be the demand graph where $E^T = \{\{s_q, v_q\} \in V \times V \ : \ q \in Q\}$. Let $\mathcal{B} = \{B_1, B_2, \ldots, B_{|\mathcal{B}|}\}$ be a (CBS) cover of G^T where $B_k = (S_k, T_k, E_k)$ is a complete bipartite graph with $S_k, T_k \subseteq V,\ E_k = \{\{u,v\} \in V \times V : u \in S_k,\ v \in T_k\} \subseteq E^T$, and $\cup_k E_k = E^T$.

In the following reformulation, source nodes of a "commodity" k are denoted by S_k and sink nodes by T_k. Let $B = \{1, 2, \ldots, |\mathcal{B}|\}$ be the index set for commodities.

$$\begin{array}{ll} \mathit{Maximize} & \displaystyle\sum_{k \in B} \sum_{u \in T_k} x_u^k \\ \mathit{Subject\ to} & \end{array}$$

$$\begin{array}{rcll} \displaystyle\sum_{v:\{u,v\}\in E} f_{uv}^k - \sum_{v:\{v,u\}\in E} f_{vu}^k \ + \ x_u^k & \leq & 0 & \forall\ k \in B,\ u \in V \setminus S_k, \\ \displaystyle\sum_{B_k \in B} \left(f_{uv}^k + f_{vu}^k\right) & \leq & c_{\{u,v\}} & \forall\ \{u,v\} \in E, \\ x_u^k \geq 0, \quad f_{uv}^k & \geq & 0 & \forall\ k \in B,\ \mathit{and}\ \{u,v\} \in E, \end{array}$$

where variable f_{uv}^k denotes the flow of commodity k from node u to node v and x_u^k denotes the total flow of commodity k that terminates at node u. Given an aggregate flow vector f, it is easy to find disaggregated flows by tracing each unit of x_u^k from node $u \in T_k$ to some $v \in S_k$. The disaggregation is not necessarily unique.

The dual of this formulation is

$$\begin{array}{ll} \mathit{Minimize} & \displaystyle\sum_{\{u,v\}\in E} c_{\{u,v\}}\ w_{\{u,v\}} \\ \mathit{Subject\ to} & \end{array}$$

$$\begin{array}{rcl} \left.\begin{array}{rcl} y_v^k - y_u^k + w_{\{u,v\}} & \geq & 0 \\ y_u^k - y_v^k + w_{\{u,v\}} & \geq & 0 \end{array}\right\} & & \forall\ k \in B,\ \mathit{and}\ \{u,v\} \in E, \\ y_u^k \quad \geq \begin{cases} 1 & \forall\ k \in B,\ u \in T_k, \\ 0 & \forall\ k \in B,\ u \in V \setminus T_k, \end{cases} & & \\ y_u^k \quad = \quad 0 & & \forall\ k \in B,\ u \in S_k, \\ w_{\{u,v\}} \quad \geq \quad 0 & & \forall\ \{u,v\} \in E, \end{array}$$

where variable y_v^k can be interpreted as the least shortest path distance between v and a member of S_k using w as edge weights.

If $|B| = 1$, the dual feasible set is integral (see Karzanov [11], for example) and an optimal dual solution corresponds to a multicut of capacity equal to the maximum flow. It is also possible to see this by noticing that the maximum multicommodity flow problem can easily be transformed into a maximum flow problem with a single source node and a single sink node.

Based on this observation, we now relate the min-cut max-flow ratio to the size of a minimum cardinality CBS cover of the demand graph. The size of this minimum cardinality cover is called the *bipartite dimension* [5] of the graph.

LEMMA 10. *Let F^* and $C(\Delta^*)$ be defined as in Lemma 9, let $\mathcal{B}^*$ be a minimum cardinality CBS cover of the demand graph, and let $k^{**} = |\mathcal{B}^*|$ denote the bipartite dimension of the demand graph. Then*

$$k^{**} \geq \frac{C(\Delta^*)}{F^*}.$$

Proof. Let $\mathcal{B}^* = \{B_1, B_2, \dots, B_{k^{**}}\}$. We solve k^{**} maximum multicommodity flow problems, one for each $\mathcal{B}_i = \{B_i\}$, and obtain the maximum flow value F_i^* and the corresponding multicut Δ_i. Clearly, $F^* \geq F_i^* = C(\Delta_i)$, and $\sum_{i=1}^{k^{**}} C(\Delta_i) \geq C(\Delta^*)$. We can therefore write

$$k^{**} \quad \geq \quad \sum_{i=1}^{k^{**}} \frac{F_i^*}{F^*} \quad = \quad \sum_{i=1}^{k^{**}} \frac{C(\Delta_i)}{F^*} \quad \geq \quad \frac{C(\Delta^*)}{F^*}. \qquad \square$$

Depending on the problem instance, Lemma 10 can provide a tighter bound than Lemma 9. For example, consider an instance where $S_1 \subseteq V$, $S_2 = V \setminus S_1$ with $|S_1| = |S_2| = n/2$ and $E^T = \{\{s, v\} \in V \times V : s \in S_1, \ v \in S_2\}$. For this problem instance, the number of source-sink pairs is $k = n^2/4$, the size of the minimum cardinality vertex cover of G^T is $k^* = n/2$, and the size of the minimum cardinality CBS cover of G^T is $k^{**} = 1$.

A remaining open question is whether or not one can show a $O(\log k^{**})$ bound on the min-cut max-flow ratio for the maximum multicommodity flow problem. We were unable to prove or disprove such a bound.

6. Conclusion. In this paper we presented improved bounds on the min-cut max-flow ratio for the multicommodity flow and the maximum multicommodity flow problems. Our bounds are motivated by "compact" linear programming formulations based on covers of the demand graph. For both problems, our results suggest that the quality of the ratio depends on the demand graph in a more structural way than the size of the edge set (i.e., the number of origin-destination pairs).

To extend our approach to directed versions of the (maximum) multicommodity flow problems, one needs to find minimum cardinality covers of the "directed" demand graph in the following sense: The demand graph now has two nodes v' and v'' for each original node $v \in V$, and it has an undirected edge $\{u', v''\}$ if there is a flow requirement from node u to node v. A cover $\mathcal{C}$ of this undirected bipartite graph gives a linear programming formulation with $|\mathcal{C}|$ aggregate commodities and therefore provides a $|\mathcal{C}|$ bound on the min-cut max-flow ratio. Relating this ratio logarithmically to the number of aggregate commodities is an open problem.

Acknowledgments. We are very grateful to Gregory Sorkin for his helpful advice in formulating and presenting Lemma 3. We also thank Baruch Schieber for fruitful discussions.

REFERENCES

[1] Y. Aumann and Y. Rabani, *An $O(\log k)$ approximate min-cut max-flow theorem and approximation algorithm*, SIAM J. Comput., 27 (1998), pp. 291–301.

[2] D. Bienstock and O. Günlük, *Computational experience with a difficult multicommodity flow problem*, Math. Programming, 68 (1995), pp. 213–238.

[3] J. Bourgain, *On Lipschitz embedding of finite metric spaces in Hilbert space*, Israel J. Math., 52 (1985), pp. 46–52.

[4] C. Chekuri, *private communication*, 2004.

[5] P. Fishburn and P. Hammer, *Bipartite dimensions and bipartite degrees of graphs*, Discrete Math., 160 (1996), pp. 127–148.

[6] L. R. Ford, Jr., and D. R. Fulkerson, *Flows in Networks*, Princeton University Press, Princeton, NJ, 1962.

[7] N. Garg, *Multicommodity Flows and Approximation Algorithms*, Doctoral thesis, IIT, Delhi, India, 1994.

[8] N. Garg, V. V. Vazirani, and M. Yannakakis, *Approximate max-flow min-(multi)cut theorems and their applications*, in Proceedings of the 25th Annual ACM Symposium on Theory of Computing, 1993, pp. 698–707.

[9] M. Iri, *On an extension of the max-flow min-cut theorem to multicommodity flows*, J. Oper. Res. Soc. Japan 13 (1971), pp. 129–135.

[10] W. Johnson and J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, in Conference in Modern Analysis and Probability, Contemp. Math. 26, AMS, Providence, RI, 1984, pp. 189–206.

[11] A. V. Karzanov, *Polyhedra related to undirected multicommodity flows*, Linear Algebra Appl., 114/115 (1989), pp. 293–328.

[12] P. Klein, S. Rao, A. Agrawal, and R. Ravi, *An approximate max-flow min-cut relation for undirected multicommodity flow, with applications*, Combinatorica, 15 (1995), pp. 187–202.

[13] F. T. Leighton and S. Rao, *An approximate max-flow min-cut theorem for uniform multicommodity flow problems with applications to approximation algorithms*, in Proceedings of the 29th Annual IEEE Symposium on Foundations of Computer Science, 1998, pp. 422–431.

[14] N. Linial, E. London, and Y. Rabinovich, *The geometry of graphs and some of its algorithmic applications*, Combinatorica, 15 (1995), pp. 215–245.

[15] C. Lund, S. Phillips, and N. Reingold, *private communication*, AT&T Research, 1996.

[16] D. W. Matula, *Concurrent flow and concurrent connectivity in graphs*, in Graph Theory and Its Applications to Algorithms and Computer Science, Wiley, New York, 1985, pp. 543–559.

[17] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, Cambridge, UK, 1995.

[18] H. Okamura and P. Seymour, *Multicommodity flows in planar graphs*, J. Combin. Theory Ser. B, 31 (1981), pp. 75–81.

[19] K. Onaga and O. Kakusho, *On feasibility conditions of multicommodity flows in networks*, IEEE Trans. Circuit Theory, 18 (1971), pp. 425–429.

[20] G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry*, Cambridge University Press, Cambridge, UK, 1989.

[21] S. Plotkin and E. Tardos, *Improved bounds on the max-flow min-cut ratio for multicommodity flows*, Combinatorica, 15 (1995), pp. 425–434.

[22] F. Shahrokhi and D. W. Matula, *The maximum concurrent flow problem*, J. Assoc. Comput. Mach., 37 (1990), pp. 318–334.

[23] S. Tragoudas, *Improved approximations for the min-cut max-flow ratio and the flux*, Math. Systems Theory, 29 (1996), pp. 157–167.

SIAM J. DISCRETE MATH.
Vol. 21, No. 1, pp. 16–25

A SIMPLE GRAY CODE TO LIST ALL MINIMAL SIGNED BINARY REPRESENTATIONS[*]

J. SAWADA[†]

Abstract. A signed binary representation (SBR) of an integer N is a string $a_b \cdots a_2 a_1 a_0$ over the alphabet $\{-1, 0, 1\}$ such that $N = \sum_{i=0}^{b} a_i 2^i$. An SBR of an integer N is said to be *minimal* if the number of nonzero digits is minimum. In this paper, we describe a simple 3-close Gray code for listing all minimal SBRs of an integer N. The algorithm is implemented to run in constant amortized time. In addition, we identify the values for N that have the maximum number of minimal SBRs given the length of the binary representation of N.

Key words. Signed binary representations, Gray code

AMS subject classification. 05A

DOI. 10.1137/050641405

1. Introduction.

A *signed binary representation* (SBR) of an integer N is a string $a_b \cdots a_2 a_1 a_0$ over the alphabet $\{-1, 0, 1\}$ such that $N = \sum_{i=0}^{b} a_i 2^i$. An example of an SBR for $N = 51$ is $10\bar{1}010\bar{1}$ (where for convenience we use $\bar{1}$ for -1), which corresponds to $2^6 - 2^4 + 2^2 - 2^0$.

An SBR of an integer N is said to be *minimal* if the number of nonzero digits is minimum. A minimal SBR for an integer N is not necessarily unique; in fact, we will show that there may be an exponential number of such strings with respect to the length of the binary representation of N. As an example, there are 5 minimal SBRs for $N = 51$ each requiring 4 nonzero bits:

$$0110011, 011010\bar{1}, 100\bar{1}\bar{1}0\bar{1}, 10\bar{1}0011, 10\bar{1}010\bar{1}.$$

Booth [2] first applied the notion of SBRs to a signed binary multiplication technique. A decade later, Reitwiesner [10] gave the first linear time algorithm to find a minimal SBR for a given integer N. Since then, several other researchers have provided similar linear time algorithms, including the following one-line algorithm (based on work by Güntzer and Paul [4]) given by Prodinger [9]: "writing $3N/2$ in binary and subtracting (bitwise) the binary representation of $N/2$." For a more thorough history of SBRs and how they apply to fast exponentiation and cryptography, consult [6, 8, 12, 14].

As there are potentially an exponential number of minimal SBRs for an integer N (with respect to the length of the binary representation of N), it is natural to ask how efficiently we can produce an exhaustive listing of these objects. Ideally, a listing algorithm will run in time proportional to the number of objects (strings) generated. Such algorithms are said to be CAT for constant amortized time. Also, it is often useful for a listing of objects to have the *Gray code property*: successive objects in the listing differ by a constant amount.

In [6], Ganesan and Manku show how minimal SBRs can be used to find optimal routes in a network derived from Chord [5], a peer-to-peer network topology. They

[*] Received by the editors September 28, 2005; accepted for publication (in revised form) July 23, 2006; published electronically January 8, 2007.
http://www.siam.org/journals/sidma/21-1/64140.html

[†] Computing and Information Science, University of Guelph, Guelph N1G 2W1, ON, Canada (sawada@cis.uoguelph.ca).

also present the only previously known algorithm for exhaustively listing all minimal SBRs. Unfortunately, no analysis of the algorithm was provided and the resulting listing does not have the Gray code property. To remedy this situation, we modify their algorithm into one that is a 3-close Gray code listing (successive strings differ in 3 consecutive positions) and provide steps to make the algorithm CAT. This is discussed in section 2. Then in section 3, as a secondary result, we identify precisely the values for N that have the maximum number of minimal SBRs given the length of the binary representation for N. In section 4 we identify two interesting sequences with respect to SBRs and conclude with final remarks in section 5.

For the remainder of this paper we will let SBR(N) denote the set of all minimal signed binary representations of an integer N. It also assumed that N is represented in binary as $a_b \cdots a_2 a_1 a_0$, and as mentioned earlier, we will use $\bar{1}$ to represent -1 for convenience.

2. Listing minimal SBRs.

2. Listing minimal SBRs. The following is a recursive description of Ganesan and Manku's algorithm [6] where $\mathbf{B}(N)$ denotes a listing of all the strings in SBR(N). The notation $\mathbf{B}(N) \cdot 1$ denotes the listing $\mathbf{B}(N)$ with an additional 1 appended to each string. The notation $\mathbf{B}(N), \mathbf{B}(M)$ indicates the list of strings $\mathbf{B}(N)$ followed by the list of strings $\mathbf{B}(M)$:

$$\mathbf{B}(N) = \begin{cases} 0 & \text{if } N = 0, \\ \mathbf{B}(\frac{N}{2}) \cdot 0 & \text{if } \mathbf{suffix}(N, 0) \text{ and } N > 0, \\ \mathbf{B}(\frac{N-1}{2}) \cdot 1 & \text{if } \mathbf{suffix}(N, 0(01)^*01), \\ \mathbf{B}(\frac{N+1}{2}) \cdot \bar{1} & \text{if } \mathbf{suffix}(N, 1(10)^*11)\ , \\ \mathbf{B}(\frac{N+1}{2}) \cdot \bar{1},\ \mathbf{B}(\frac{N-1}{2}) \cdot 1 & \text{if } \mathbf{suffix}(N, 11(01)^*01) \text{ or} \\ & \text{if } \mathbf{suffix}(N, 00(10)^*11). \end{cases}$$

The predicate $\mathbf{suffix}(N, \mathit{expr})$ returns true if a suffix of N, represented in binary (and padded with 0's on the left), matches the regular expression expr. An example computation tree for $\mathbf{B}(51)$ is given in Figure 1. The nodes in the tree represent the input strings, and the labels on the edges represent the character to be prepended to the output string as specified by $\mathbf{B}(N)$. Thus, each minimal SBR can be found by tracing a path from a leaf back to the root while recording the labels on the edges. Observe that since 0011 is a suffix of 110011, it satisfies the last case in the recursive description. Thus, the root node in Figure 1 has 2 children producing strings that end with $\bar{1}$ and 1, respectively.

2.1. A Gray code. In general the listing $\mathbf{B}(N)$ is not a Gray code since successive strings in the listing may differ by up to a linear amount $\Omega(b)$. However, by studying the listings for a variety of input values and focusing on the parities of the repeated terms in the regular expressions, we discover a 3-close Gray code description for SBR(N). This new listing is obtained by reversing the order of particular subtrees within the computation tree of $\mathbf{B}(N)$. The result is a listing that will produce the same strings but in a different order. The overline in the description of this new listing $\mathbf{L}(N)$ indicates that the listing of strings is reversed:

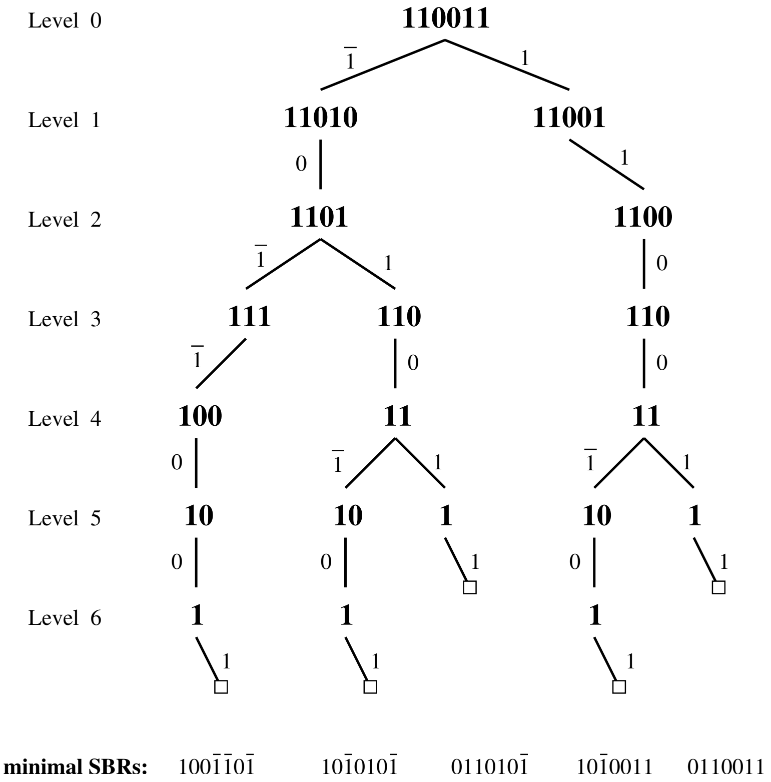


FIG. 1. *Computation tree for* $\mathbf{B}((110011)_2) = \mathbf{B}(51)$.

$$\mathbf{L}(N) = \begin{cases} 0 & \text{if } N = 0, \\ \mathbf{L}(\frac{N}{2}) \cdot 0 & \text{if } \mathbf{suffix}(N, 0) \text{ and } N > 0, \\ \mathbf{L}(\frac{N-1}{2}) \cdot 1 & \text{if } \mathbf{suffix}(N, 0(01)^*01), \\ \mathbf{L}(\frac{N+1}{2}) \cdot \bar{1} & \text{if } \mathbf{suffix}(N, 1(10)^*11), \\ \overline{\mathbf{L}(\frac{N+1}{2})} \cdot \bar{1},\ \mathbf{L}(\frac{N-1}{2}) \cdot 1 & \text{if } \mathbf{suffix}(N, 11(01)^t01) \text{ and } t \text{ even} \qquad (1), \\ \mathbf{L}(\frac{N+1}{2}) \cdot \bar{1},\ \overline{\mathbf{L}(\frac{N-1}{2})} \cdot 1 & \text{if } \mathbf{suffix}(N, 00(10)^t11) \text{ and } t \text{ even} \qquad (2), \\ \mathbf{L}(\frac{N+1}{2}) \cdot \bar{1},\ \mathbf{L}(\frac{N-1}{2}) \cdot 1 & \text{if } \mathbf{suffix}(N, 11(01)^t01) \text{ and } t \text{ odd or} \qquad (3), \\ & \text{if } \mathbf{suffix}(N, 00(10)^t11) \text{ and } t \text{ odd} \qquad (4). \end{cases}$$

THEOREM 1. *The listing* $\mathbf{L}(N)$ *of all strings in SBR(N) where* $N > 0$ *is a* 3*-close Gray code.*

Proof. Assume that N is represented in binary. Let $\mathbf{first}(N)$ and $\mathbf{last}(N)$ denote the first and last strings in the listing of $\mathbf{L}(N)$. To prove that the listing $\mathbf{L}(N)$ is a 3-close Gray code we show that the interface strings for Cases (1), (2), (3), and (4) differ in *exactly the last three positions*. Applying induction completes the proof.

Case (1). N is of the form $x11(01)^t01$, where x is some binary string and t is even. Here we must compare the last string in $\overline{\mathbf{L}(\frac{N+1}{2})} \cdot \bar{1} = \mathbf{first}(x11(01)^t1) \cdot \bar{1}$ and the first string in $\mathbf{L}(\frac{N-1}{2}) \cdot 1 = \mathbf{first}(x11(01)^t0) \cdot 1$. First consider $t > 0$:

$$\begin{aligned}
\mathbf{first}(x11(01)^t1) \cdot \bar{1} &= \mathbf{first}(x1(10)^t11) \cdot \bar{1} \\
&= \mathbf{first}(x1(10)^{t-1}110) \cdot \bar{1}\bar{1} \\
&= \mathbf{first}(x1(10)^{t-1}11) \cdot 0\bar{1}\bar{1}, \\
\mathbf{first}(x11(01)^t0) \cdot 1 &= \mathbf{first}(x11(01)^t) \cdot 01 \\
&= \mathbf{first}(x11(01)^{t-1}01) \cdot 01 \\
&= \mathbf{first}(x11(01)^{t-1}1) \cdot \bar{1}01 \\
&= \mathbf{first}(x1(10)^{t-1}11) \cdot \bar{1}01.
\end{aligned}$$

If $t = 0$, let $x = y01^r$:

$$\begin{aligned}
\mathbf{first}(y01^r111) \cdot \bar{1} &= \mathbf{first}(y10^r00) \cdot \bar{1}\bar{1} \\
&= \mathbf{first}(y10^r0) \cdot 0\bar{1}\bar{1}, \\
\mathbf{first}(y01^r110) \cdot 1 &= \mathbf{first}(y01^r11) \cdot 01 \\
&= \mathbf{first}(y10^r0) \cdot \bar{1}01.
\end{aligned}$$

Case (2). N is of the form $x00(10)^t11$, where x is some binary string and t is even. Again we consider two subcases depending on the value for t. If $t > 0$, then we must compare the last string in $\mathbf{L}(\frac{N+1}{2}) \cdot \bar{1} = \mathbf{last}(x00(10)^{t-1}110) \cdot \bar{1}$ and the first string in $\overline{\mathbf{L}(\frac{N-1}{2})} \cdot 1 = \mathbf{last}(x00(10)^t1) \cdot 1$:

$$\begin{aligned}
\mathbf{last}(x00(10)^{t-1}110) \cdot \bar{1} &= \mathbf{last}(x00(10)^{t-1}11) \cdot 0\bar{1} \\
&= \mathbf{last}(x00(10)^{t-1}1) \cdot 10\bar{1} \\
&= \mathbf{last}(x0(01)^t) \cdot 10\bar{1}, \\
\mathbf{last}(x00(10)^t1) \cdot 1 &= \mathbf{last}(x0(01)^t01) \cdot 1 \\
&= \mathbf{last}(x0(01)^t0) \cdot 11 \\
&= \mathbf{last}(x0(01)^t) \cdot 011.
\end{aligned}$$

If $t = 0$, then the two interface strings are $\mathbf{last}(x010) \cdot \bar{1}$ and $\mathbf{last}(x001) \cdot 1$, respectively:

$$\begin{aligned}
\mathbf{last}(x010) \cdot \bar{1} &= \mathbf{last}(x01) \cdot 0\bar{1} \\
&= \mathbf{last}(x0) \cdot 10\bar{1}, \\
\mathbf{last}(x001) \cdot 1 &= \mathbf{last}(x001) \cdot 1 \\
&= \mathbf{last}(x0) \cdot 011.
\end{aligned}$$

Case (3). N is of the form $x11(01)^t01$, where x is some binary string and t is odd. Here we must compare the last string in $\mathbf{L}(\frac{N+1}{2}) \cdot \bar{1} = \mathbf{last}(x11(01)^t1) \cdot \bar{1}$ and the first string in $\mathbf{L}(\frac{N-1}{2}) \cdot 1 = \mathbf{first}(x11(01)^t0) \cdot 1$:

$$\begin{aligned}
\mathbf{last}(x11(01)^t1) \cdot \bar{1} &= \mathbf{last}(x1(10)^t11) \cdot \bar{1} \\
&= \mathbf{last}(x1(10)^{t-1}110) \cdot \bar{1}\bar{1} \\
&= \mathbf{last}(x1(10)^{t-1}11) \cdot 0\bar{1}\bar{1}, \\
\mathbf{first}(x11(01)^t0) \cdot 1 &= \mathbf{first}(x11(01)^{t-1}01) \cdot 01 \\
&= \mathbf{last}(x11(01)^{t-1}1) \cdot \bar{1}01 \\
&= \mathbf{last}(x1(10)^{t-1}11) \cdot \bar{1}01.
\end{aligned}$$

Case (4). N is of the form $x00(10)^t11$, where x is some binary string and t is odd. Here we must compare the last string in $\mathbf{L}(\frac{N+1}{2}) \cdot \bar{1} = \mathbf{last}(x00(10)^{t-1}110) \cdot \bar{1}$ and the first string in $\mathbf{L}(\frac{N-1}{2}) \cdot 1 = \mathbf{first}(x00(10)^t1) \cdot 1$:

$$\begin{aligned}
\mathbf{last}(x00(10)^{t-1}110) \cdot \bar{1} &= \mathbf{last}(x00(10)^{t-1}11) \cdot 0\bar{1} \\
&= \mathbf{first}(x00(10)^{t-1}1) \cdot 10\bar{1} \\
&= \mathbf{first}(x0(01)^t) \cdot 10\bar{1}, \\
\mathbf{first}(x00(10)^t1) \cdot 1 &= \mathbf{first}(x0(01)^t01) \cdot 1 \\
&= \mathbf{first}(x0(01)^t0) \cdot 11 \\
&= \mathbf{first}(x0(01)^t) \cdot 011.
\end{aligned}$$

In all cases we have shown that the interface strings differ in exactly the last 3 positions. By applying induction, this proves that $\mathbf{L}(N)$ is a 3-close Gray code for the strings in $\mathrm{SBR}(N)$. □

As an example, Figure 2 displays the computation tree for $\mathbf{L}(110011)$. As before, the nodes represent the input strings, but now if a subtree is to be reversed, this information is additionally passed down via the edges and represented by R. Naturally, a reversal of a reversed subtree produces the regular ordering (see the 11 node furthest to the right in Figure 2).

The following is the final listing generated for $\mathbf{L}(110011)$:

$$\begin{array}{cccccccccc}
0 & 1 & 0 & 0 & \bar{1} & \bar{1} & 0 & \bar{1} & & \\
0 & 1 & 0 & \bar{1} & 0 & 1 & 0 & \bar{1} & & (2) \\
0 & 0 & 1 & 1 & 0 & 1 & 0 & \bar{1} & & (4) \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & & (0) \\
0 & 1 & 0 & \bar{1} & 0 & 0 & 1 & 1 & & (4)
\end{array}$$

Observe that each successive string differs in exactly 3 *consecutive* positions by either the transformation $011 \leftrightarrow 10\bar{1}$ or $0\bar{1}\bar{1} \leftrightarrow \bar{1}01$. Also observe that the rightmost of these positions (indicated in parentheses) corresponds to the levels of the degree 2 nodes in the computation tree when visited in order. These properties can also be inferred from Theorem 1 and its proof. Therefore given the in-order sequence of levels of the degree 2 nodes along with the first output string in the listing, we can generate the Gray code listing $\mathbf{L}(N)$ in constant amortized time. In the next subsection, we describe how we can efficiently generate this sequence.

It is also interesting to note from this example that a 2-close listing is impossible in general. This is because the first string is the only one that contains $0\bar{1}\bar{1}$ in positions 4, 5, and 6.

Also of interest is the underlying graph with vertices corresponding to the minimal SBRs (for a given integer N) and edges between two vertices if and only if their corresponding SBRs differ in exactly 3 adjacent positions. Clark and Liang [3] showed that this graph is connected. The result in this paper shows that this graph contains a Hamilton path. In general there is no Hamilton cycle since the underlying graph from our example has a vertex with degree 1: $0100\bar{1}\bar{1}0\bar{1}$.

2.2. Efficiency considerations. If we apply the algorithm for $\mathbf{L}(N)$ or $\mathbf{B}(N)$ directly, we may require a linear amount of work to process each node: computing the suffix and modifying the input string for the next recursive call. This amount of computation is not desirable; however, because there are repeated subtrees, we can precompute the parent child relationships for each node. In fact, given that

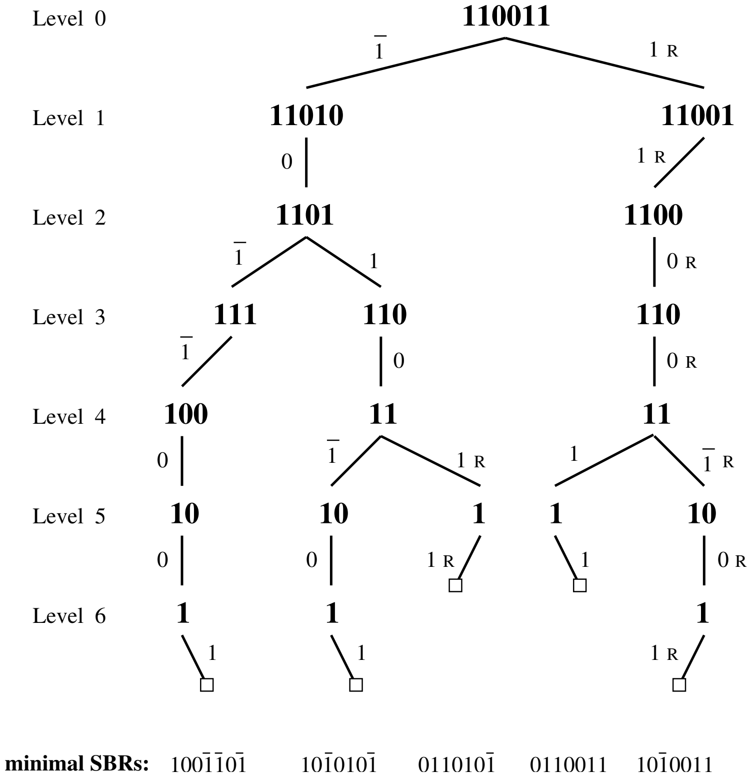


FIG. 2. *Computation tree for* $\mathbf{L}(110011)$.

$N = a_b \cdots a_2 a_1 a_0$ and that α_i denotes the prefix $a_b \cdots a_i$ and $\beta_i = \alpha_i + 1$, the following lemma shows that the number of different possible nodes in the computation tree for $\mathbf{L}(N)$ or $\mathbf{B}(N)$ is at most $2(b+1)$.

LEMMA 2. *At level i in the computation tree of $\boldsymbol{L}(N)$ or $\boldsymbol{B}(N)$ the input string is either α_i or β_i.*

Proof. The proof is by induction on i. When $i = 0$, we are at the root of the computation tree and the input string is α_0. For the inductive hypothesis, suppose that the input string for a node at level $i \geq 0$ is either α_i or β_i. By using the rules of the listing $\mathbf{L}(N)$ or $\mathbf{B}(N)$, observe that the input string for the child of a node is obtained either by trimming off the least significant bit or by adding one first and then trimming off the final bit. Thus, if the input string of a node at level i is α_i, then the input for its children must be either α_{i+1} or β_{i+1}. In the case where the input string is β_i we consider two subcases depending on the last bit. If β_i ends with 0, then its only child is $(\beta_i)/2 = \beta_{i+1}$. Otherwise, if β_i ends with 1, then trimming off the last bit will result in α_{i+1}. If we add one first and then trim the last bit, we will obtain β_{i+1}. □

Since there at most $2(b+1)$ different nodes in the computation tree for $\mathbf{L}(N)$, we can precompute the parent child relationships for all nodes in $O(b^2)$ time. (In fact, if we reuse the suffix details starting at node α_0, we could perform this precomputation in linear time $O(b)$.) After performing this precomputation, we can determine the children of a node in constant time, allowing us to construct the computation tree for $\mathbf{L}(N)$ in constant time per node. Applying this method, the overall running time of the algorithm will be proportional to the number of nodes in the computation tree. If this number is proportional to the number of strings generated (the leaves in the tree), then the algorithm will be CAT. However, in general, this will not be the case due to the large number of degree 1 nodes.

Fortunately, as discussed at the end of the previous subsection, we need only visit the degree 2 nodes (in order) from the computation tree to obtain the Gray code listing. Again, since there are only a linear number of nodes, we can precompute the nearest (left and right) descendants that have degree 2 for each potential node in the computation tree. This can be done in $O(b)$ time, since there are only $O(b)$ nodes to visit. Now, for each node, we can find the nearest left and/or right descendant that has degree 2 in constant time. All that remains is to compute the initial minimal SBR by following the leftmost path in the computation tree and then traversing the degree 2 nodes in order, outputting the original level of each node. When traversing this tree we must be careful to maintain information about subtree reversal so that we know which child branch to visit first.

The following is a detailed summary of the steps required to produce the listing $\mathbf{L}(N)$ in constant amortized time:

1. For $0 \leq i \leq n$ determine the child or children of each node α_i and β_i. The level of these nodes will be i, and from the recursive description of $\mathbf{L}(N)$ we can determine whether or not the subtrees for each child should be reversed. This will take $O(b^2)$ time.
2. For $0 \leq i \leq n$ determine the nearest left and/or right descendant of α_i and β_i that has degree 2. This can be computed in linear time $O(b)$ by starting with $i = b$ and working back to $i = 0$. Details about whether or not the subtrees are to be reversed must be maintained for each degree 2 node.
3. Determine the initial minimal SBR of the listing $\mathbf{L}(N)$ by tracing a path through the virtual computation tree rooted by α_0. This will take linear time $O(b)$.
4. Visit the degree 2 nodes in order, being careful to consider when subtrees are to be reversed. For each level i that is output, modify the current minimal SBR in positions $i+2, i+1, i$. This is done by scanning these 3 characters and applying the appropriate transformation rule: $011 \leftrightarrow 10\bar{1}$ or $0\bar{1}\bar{1} \leftrightarrow \bar{1}01$. The degree 2 nodes can be traversed in constant amortized time; thus the running time of this step will be proportional to the number of minimal SBRs generated.

Observe that the original computation tree is never actually constructed.

THEOREM 3. *The Gray code listing $\mathbf{L}(N)$ can be generated in constant amortized time with $O(b^2)$ initialization.*

3. Maximizing the number of minimal SBRs. If N is represented by the binary string $a_b \cdots a_2 a_1 a_0$, where $a_b = 1$, then there may be only one string in SBR(N) or there could potentially be an exponential number with respect to b. Thus given b, we are interested in finding a tight upper bound on the number of strings in SBR(N), denoted $Max(b)$, as well as a characterization of the bitstrings that obtain this upper bound. Note that the actual length of the binary representation of N is $b+1$. When $b = 0, 1, 2$, the binary representations that produce the maximum number of minimal SBRs are 1, 11, and 110, respectively. The values $Max(0) = 1$ and $Max(1) = Max(2) = 2$. For $3 \leq b \leq 10$ we apply a generation algorithm to determine which binary representations of N have $Max(b)$ strings in SBR(N):

b	Binary representations of N		$Max(b)$
3	1011	1101	3
4	10110	11010	3
5	101101	110011	5
6	1011010	1100110	5
7	10110011	11001101	8
8	101100110	110011010	8
9	1011001101	1100110011	13
10	10110011010	11001100110	13

THEOREM 4. *$Max(b) = f_{\lceil b/2 \rceil + 2}$, the $\lceil b/2 \rceil$+2^{nd} Fibonacci number. Moreover, the SBRs of the two values of N that have $Max(b)$ minimal SBRs where $b \geq 3$ are*

$$\begin{array}{llll}
10(1100)^t11 & \text{and} & 11(0011)^t01 & \text{if } b = 4t{+}3, \\
10(1100)^t110 & \text{and} & 11(0011)^t010 & \text{if } b = 4t{+}4, \\
10(1100)^t1101 & \text{and} & 11(0011)^t0011 & \text{if } b = 4t{+}5, \\
10(1100)^t11010 & \text{and} & 11(0011)^t00110 & \text{if } b = 4t{+}6.
\end{array}$$

Proof. Applying a generation algorithm, it is trivial to verify the theorem for $3 \leq b \leq 6$. For $b > 6$ we assume that $Max(i) = f_{\lceil \frac{i}{2} \rceil + 2}$ for $3 \leq i < b$ (inductive hypothesis) and consider $a_b \cdots a_2 a_1 a_0$, the binary representation for an integer N. Using the recursive listing $\mathbf{B}(N)$, we will examine each possible suffix of N to determine restrictions on the strings in SBR(N). In particular, we will show that the strings in SBR(N) must end with particular bit sequences for a given suffix.

suffix$(N, 0)$. All strings end with 0. Thus, the maximum number of strings is bounded by $Max(b-1)$. This result implies that $Max(b)$ does not decrease as b increases.

suffix$(N,\ 0(01)^*01)$. Applying the recursive rules twice, all strings must end with 01. Thus, the maximum number of strings is bounded by $Max(b-2)$.

suffix$(N,\ 1(10)^*11)$. Applying the recursive rules twice, all strings must end with $0\bar{1}$. Thus, the maximum number of strings is bounded by $Max(b-2)$.

For the remaining two suffixes, the strings may end with either 1 or $\bar{1}$.

suffix$(N, 11(01)^t01)$. Applying the recursive rules, all strings ending with 1 will end with 01. Otherwise if a string ends with $\bar{1}$, then if $t = 0$, it must end with $00\bar{1}\bar{1}$; if $t = 1$, it must end with $00\bar{1}0\bar{1}\bar{1}$; if $t > 1$, it must end with $\bar{1}0\bar{1}0\bar{1}\bar{1}$. Thus, when the suffix of N is 1101, an upper bound on the maximum number of minimal SBRs is $Max(b-2) + Max(b-4)$. Otherwise if $t > 1$, the upper bound is $Max(b-2) + Max(b-6)$.

suffix$(N, 00(10)^t11)$. Applying the recursive rules, all strings ending with $\bar{1}$ will end with $0\bar{1}$. Otherwise if a string ends with 1, then if $t = 0$, it must end with 0011; if $t = 1$, it must end with 001011; if $t > 1$, it must end with 101011. Thus, when the suffix of N is 0011, an upper bound on the maximum number of minimal SBRs is $Max(b-2) + Max(b-4)$. Otherwise if $t > 1$, the upper bound is $Max(b-2) + Max(b-6)$.

Observe that from our inductive hypotheses that when b is even, $Max(b-1) = Max(b-2) + Max(b-4) = f_{\lceil \frac{b}{2} \rceil + 2}$, and when b is odd, $Max(b-1) = f_{\lceil \frac{b}{2} \rceil + 2 - 1}$. Thus an overall upper bound on $Max(b)$ is $f_{\lceil \frac{b}{2} \rceil + 2}$. We complete the proof by showing the two strings that obtain this bound, thus proving $Max(b) = f_{\lceil \frac{b}{2} \rceil + 2}$. We examine four cases depending on b. Since $b > 6$, we must have $t \geq 1$.

$b = 4t - 1$. In this case b is odd, so if $Max(b)$ is to obtain the upper bound of $Max(b-2) + Max(b-4)$, N must have the suffix 1101 or 0011. If it ends with 1101, then in order for the number of strings in SBR(N) that end with 01 to meet the maximum of $Max(b-2)$, the first $b-2$ bits must be either $10(1100)^{t-1}1101$ or $11(0011)^{t-1}0011$ (by induction). Additionally, in order for the number of strings that end with $00\bar{1}\bar{1}$ to meet the maximum of $Max(b-4)$, the first $b-4$ bits must be either $10(1100)^{t-1}11$ or $11(0011)^{t-1}01$ with 1 subtracted as a result of the recursive definition applied to the final 4 bits. Thus taking the union of these criteria over all binary strings, we are left with $N = 11(0011)^t01$. A similar examination will show that when N has a suffix 0011, the only string that will obtain the upper bound of $Max(b-2) + Max(b-4)$ strings is $N = 10(1100)^t11$.

$b = 4t$. Using induction, if N ends with 0, then $Max(b) = Max(b-1)$ if and only if $N = 10(1100)^t110$ or $N = 11(0011)^t010$. Otherwise, in order for the size of SBR(N) to meet the upper bound, it must have the suffix 1101 or 0011. If it ends with 1101, the first $b-2$ bits must be either $10(1100)^t11010$ or $11(0011)^t00110$. However, since neither of these strings ends with 11, no value for N ending with 1101 will meet the upper bound in this case. If it ends with 0011, the first $b-2$ bits must be either $10(1100)^t11010$ or $11(0011)^t00110$, but this time with 1 subtracted since the strings must end with $0\bar{1}$ (from the recursive definition). However, since neither of the resulting strings ends with 00, no value for N ending with 0011 will meet the upper bound.

$b = 4t + 1$. This is similar to the case $b = 4t - 1$.

$b = 4t + 2$. This is similar to the case $b = 4t$. □

4. Related sequences. If we consider the number of bits required to represent each string in SBR(N) for each value of N starting from $N = 0$, we obtain the following sequence:

$$A = 0, 1, 1, 2, 1, 2, 2, 2, 1, 2, 2, 3, 2, 3, 2, 2, \ldots .$$

For example, the minimum number of bits required to represent the integer 3 as the difference of 2 binary numbers is 2 (given by the fourth element in the sequence). This sequence corresponds to sequence A007302 in Sloane's *The On-Line Encyclopedia of Integer Sequences* [13]. Interestingly, these values also correspond to the cost of grid communications on the Connection Machine [15]. This sequence is also discussed with respect to k-regular sequences in [1].

Another interesting sequence is the one obtained from the number of strings in SBR(N) for each value of N starting from 0:

$$B = 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 1, 3, 2, 3, 1, 1, \ldots .$$

For example, the fourth element in this sequence is 2 since there are two strings in SBR(3). This sequence corresponds to sequence A110955 in Sloane's *The On-Line Encyclopedia of Integer Sequences* [13].

5. Final remarks. In this paper we have presented a 3-close Gray code algorithm to generate all minimal SBRs of an integer N. After some initialization, the algorithm can be implemented to run in constant amortized time. A CAT implementation is available from the author upon request or at http://www.cis.uoguelph.ca/~sawada/prog.html. As a secondary result, we have precisely identified the values for N that produce the maximum number of minimal SBRs given the length of the binary representation of N.

A preliminary version of this work appears in the proceedings of GRACO 2005 [11]. Since this manuscript was submitted, a related result by Manku and Sawada appeared in the proceedings of ESA 2005 [7]. In that work, a loopless algorithm to list all minimal SBRs of an integer N is provided. The loopless algorithm is based on the binary reflected Gray code and is significantly more complex than the simple recursive description given in this paper.

Acknowledgments. The author would like to thank the anonymous referees and Frank Ruskey for helpful comments to improve the final version of this paper.

REFERENCES

[1] J. P. Allouche and J. Shallit, *The ring of k-regular sequences*, II, Theoret. Comput. Sci., 307 (2003), pp. 3–29.

[2] A. D. Booth, *A signed binary multiplication technique*, Quart. J. Mech. Appl. Math., 4 (1951), pp. 236–240.

[3] W. E. Clark and J. J. Liang, *On arithmetic weight for a general radix representation of integers*, IEEE Trans. Inform. Theory, 19 (1973), pp. 823–826.

[4] U. Güntzer and M. Paul, *Jump interpolation search trees and symmetric binary numbers*, Inform. Process. Lett., 26 (1987), pp. 193–204.

[5] I. Stoica, R. Morris, D. Liben-Lowell, D. R. Karger, M. F. Kaashoek, F. Dabek, and H. Balakrishnan, *Chord: A scalable peer-to-peer lookup protocol for Internet applications*, IEEE/ACM Trans. Networking, 11 (2003), pp. 17–32.

[6] P. Ganesan and G. S. Manku, *Optimal routing in Chord*, in Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), New Orleans, 2004, pp. 169–178.

[7] G. S. Manku and J. Sawada, *A loopless Gray code for minimal signed-binary representations*, in Proceedings of the 13th Annual European Symposium on Algorithms (ESA), Lecture Notes in Comput. Sci. 3669, Springer-Verlag, Berlin, 2005, pp. 438–447.

[8] K. Okeya, K. Schmidt-Samoa, C. Spahn, and T. Takagi, *Signed binary representations revisited*, in Advances in Cryptology—CRYPTO 2004, Lecture Notes in Comput. Sci. 3152, Springer-Verlag, Berlin, 2004, pp. 123–139.

[9] H. Prodinger, *On binary representations of integers with digits* -1, 0, 1, Integers (2000), A8, 14 pp.

[10] G. W. Reitwiesner, *Binary arithmetic*, in Advances in Computers, Vol. 1, Academic Press, New York, 1960, pp. 231–308.

[11] J. Sawada, *A Gray code for binary subtraction*, in Proceedings of GRACO 2005, Electron. Notes Discrete Math., 19 (2005), pp. 125–131.

[12] K. Schmidt-Samoa, O. Semay, and T. Takagi, *Analysis of fractional window recoding methods and their application to elliptic curve cryptosystems*, IEEE Trans. Comput., 55 (2006), pp. 48–57.

[13] N. Sloane, *The on-line encyclopedia of integer sequences*: ID A007302, A110955, http://www.research.att.com/~njas/sequences/index.html (2006).

[14] T. Takagi, D. Reis, Jr., S. Yen, and B. Wu, *Radix-r nonadjacent form and its application to pairing-based cryptosystem*, IEICE Trans. Fund. Elec., Comm., & Comp. Sci., E89-A (2006), pp. 115–123.

[15] A. Weitzman, *Transformation of parallel programs guided by micro-analysis*, in Algorithms Seminar (1992–1993), B. Salvy, ed., INRIA, Rocquencourt, France, 1993, Rapport de Recherche 2130, pp. 155–159.

SIAM J. Discrete Math.
Vol. 21, No. 1, pp. 26–32

CROSSING GRAPHS AS JOINS OF GRAPHS AND CARTESIAN PRODUCTS OF MEDIAN GRAPHS[*]

BOŠTJAN BREŠAR[†] AND SANDI KLAVŽAR[‡]

Abstract. For a partial cube G its crossing graph $G^{\#}$ is the graph whose vertices are the Θ-classes of G, two classes being adjacent if they cross on some cycle in G. The following problem posed in [S. Klavžar and H. M. Mulder, *SIAM J. Discrete Math.*, 15 (2002), pp. 235–251, Problem 7.1] is considered: What can be said about the partial cube G if $G^{\#}$ is the join $A \oplus B$ of graphs A and B with at least one edge? It is proved that for arbitrary graphs A and B, where at least one of them contains an edge, there exists a Cartesian prime partial cube G such that $G^{\#} = A \oplus B$. On the other hand, if G is a median graph, then $G^{\#} = A \oplus B$ if and only if $G = H \,\square\, K$, where $H^{\#} = A$ and $K^{\#} = B$. Along the way some new facts about partial cubes are obtained; for instance, a bipartite graph of radius 2 is a partial cube if and only if it is $K_{2,3}$-free.

Key words. intersection graph, partial cube, median graph, Cartesian product of graphs, join of graphs

AMS subject classifications. 05C75, 05C12

DOI. 10.1137/050622997

1. Introduction.

Intersection concepts in graph theory have been extensively studied [16]. Although some of the intersection operations yield all graphs (for instance, every graph is the intersection graph of some set system), their importance is due to their usefulness in the characterization of particular classes of graphs, thus leading to a deeper structural understanding. Here we study a nonstandard intersection operation where vertices of the intersection graph (called *crossing graph*) are equivalence classes of a certain equivalence relation Θ defined on the edge-set of a graph. Hence the edges of the crossing graph are not defined in the standard way (by intersections of subsets). The graphs that we are interested in are isometric subgraphs of hypercubes, and the relation Θ is of great importance for understanding the structure of these graphs. So before presenting the preliminary work on these graphs and the crossing graph operation, let us recall necessary definitions.

The distance $d_G(u, v)$ between vertices u and v of a graph G is the length of a shortest u, v-path in G. A subgraph U of G is *isometric* if $d_U(u, v) = d_G(u, v)$ for all $u, v \in U$. The *interval* $I_G(u, v)$ is the set of vertices that lie on shortest paths between u and v in G. A subgraph U is *convex* if $I_G(u, v) \subseteq U$ for all $u, v \in U$. (Indices in the above definitions are omitted when the graph is understood from the context.) Recall that the *hypercube* Q_k, or *k-cube*, is the graph with the vertex set $\{0, 1\}^k$, where two vertices are adjacent whenever they differ in exactly one position.

Partial cubes are isometric subgraphs of hypercubes. This class of graphs has been extensively investigated; see, for instance, [3, 5, 6, 7, 8, 21]. A well-known

[*] Received by the editors January 20, 2005; accepted for publication (in revised form) April 25, 2006; published electronically January 8, 2007. This work was supported in part by the Ministry of Science of Slovenia under grant P1-0297.

http://www.siam.org/journals/sidma/21-1/62299.html

[†] Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia, and the Institute of Mathematics, Physics and Mechanics, Jadranska 19, 1000 Ljubljana, Slovenia (bostjan.bresar@uni-mb.si).

[‡] Department of Mathematics and Computer Science, PeF, University of Maribor, Koroška cesta 160, 2000 Maribor, Slovenia, and the Institute of Mathematics, Physics and Mechanics, Jadranska 19, 1000 Ljubljana, Slovenia (sandi.klavzar@uni-mb.si).

characterization of partial cubes is by the relation Θ on the edge-set of a graph. Two edges $e = xy$ and $f = uv$ of a graph G are in the Djoković–Winkler [7, 21] relation Θ_G, Θ for short, if $d_G(x,u)+d_G(y,v) \neq d_G(x,v)+d_G(y,u)$. Winkler [21] proved that a bipartite graph is a partial cube if and only if Θ is transitive. Letting R^* denote the transitive closure of a relation R, Winkler's result reads as follows: A connected bipartite graph G is a partial cube if and only if $\Theta = \Theta^*$. Hence in partial cubes the relation Θ is an equivalence relation on $E(G)$, and the classes of the corresponding partition will be called Θ-*classes*.

For a partial cube G its *crossing graph* $G^{\#}$ was introduced in [15] as follows. The vertices of $G^{\#}$ are the Θ-classes of G, two vertices being adjacent if the respective Θ-classes meet (or cross) on some cycle (that is, there is a cycle C that contains edges of both Θ-classes). In fact, in the class of median graphs the same concept was introduced earlier by Bandelt and Chepoi under the name *incompatibility graph* [1].

In this paper we address the problem of what can be said about the partial cube G if $G^{\#} = A \oplus B$, where A and B have at least one edge. Here $A \oplus B$ denotes the *join* of graphs A and B, that is, the graph obtained from the disjoint union of A and B by joining every vertex of A with every vertex of B by an edge. In the next section we state important properties of the Cartesian product of graphs and median graphs that are needed later. In section 3 we prove that for arbitrary graphs A and B, where at least one of them contains an edge, there exists a Cartesian prime partial cube G such that $G^{\#} = A \oplus B$. Then we restrict our attention to median graphs and prove that the crossing graph of a median graph G is the join of two graphs A and B if and only if G is a Cartesian product graph. In due course we also characterize partial cubes of radius 2 and observe that a partial cube contains no nontrivial convex subgraph that meets all of its Θ-classes.

2. Cartesian products and median graphs.

The *Cartesian product* $G \,\square\, H$ of the graphs G and H is the graph with the vertex set $V(G) \times V(H)$ in which two vertices (a,x) and (b,y) are adjacent whenever $ab \in E(G)$ and $x = y$, or $a = b$ and $xy \in E(H)$. The Cartesian product is associative and commutative with K_1 as its unit. It is easy to see that the Cartesian product of k copies of K_2 is the *hypercube* Q_k. A graph G is called *prime* (with respect to the Cartesian product) if it cannot be represented as the product of two nontrivial graphs; that is, $G = G_1 \,\square\, G_2$ implies that G_1 or G_2 is the one-vertex graph K_1.

The well-known prime factorization theorem, proved by Sabidussi [19] and independently by Vizing [20], states that every connected graph has a unique prime factor decomposition with respect to the Cartesian product. This decomposition can be made explicit in the following way: Edges uv and uw are said to be in relation τ_G, or τ for short, if u is the unique common neighbor of v and w. Feder [9] proved (cf. also [11, Theorem 4.8] and [13]) that $(\Theta \cup \tau)^*$ is the Cartesian product relation of a connected graph. This actually means that the equivalence classes of the relation $(\Theta \cup \tau)^*$ determine the prime factor decomposition of a graph—every equivalence class yields one factor of the decomposition. The following consequence of this theorem will be useful for us.

Corollary 1. *A connected graph G is prime if and only if $(\Theta_G \cup \tau_G)^* = E(G)$.*

We will also need the following result (in a way part of the folklore) on the Cartesian product; see [4].

Lemma 2. *A subgraph C of the Cartesian product $G_1 \,\square \cdots \square\, G_m$ of connected graphs is convex if and only if $C = p_1(C) \,\square \cdots \square\, p_m(C)$, where $p_i(C)$ is convex in G_i, $1 \le i \le m$. (Here p_i is the projection map from G onto G_i.)*

The most important subclass of partial cubes are median graphs. They have been rediscovered several times, and a rich theory of these graphs and related structures has been developed; cf. the survey [14]. The most common definition is the following: G is a *median graph* if for every triple of vertices $u, v, w \in V(G) : I(u, v) \cap I(u, w) \cap I(v, w)$ consists of precisely one vertex (which is called the *median* of the triple u, v, w). One of the most well-known characterizations of median graphs involves a certain expansion procedure, a result due to Mulder [17]. (By the way, it inspired Chepoi [5] to prove a similar characterization of partial cubes.) In this note we will make use of a variation of the expansion procedure that involves peripheral subgraphs of a median graph [18]; see also [2].

Let G be a connected graph and G_0 a convex subgraph. Then the *peripheral expansion* of G is the graph G' obtained as follows. Take the disjoint union of a copy of G and a copy of G_0. Join each vertex u in the copy of G_0 with the vertex that corresponds to u in the copy of G (actually in the subgraph G_0 of G). We say that the resulting graph G' is *obtained by a (peripheral) expansion from G along G_0*. We also say that we *expand* G_0 in G to obtain G'. Note that in a peripheral expansion one new Θ-class appears. It is easy to prove that expanding a convex subgraph of a median graph yields again a median graph. It is more surprising that the converse is also true, as proved by Mulder in [18].

Theorem 3. *A graph G is a median graph if and only if it can be obtained from K_1 by a sequence of peripheral expansions.*

Hence each median graph contains a *peripheral subgraph*, that is, a subgraph H whose vertices are all incident with a particular Θ-class F in G, such that H is a connected component of $G - F$ (the graph obtained from G by removal of edges from F). Even more is known [18], as stated in the following proposition.

Proposition 4. *Let G be a median graph and F any Θ-class in G. Then both connected components of $G - F$ contain a peripheral subgraph of G.*

It is easy to see that median graphs are closed under Cartesian multiplication and that, conversely, if a median graph is not prime, all of the factors also must be median graphs.

3. Partial cubes whose crossing graphs are joins. Crossing graphs of Cartesian products have a simple structure [15, Proposition 6.1].

Proposition 5. *Let H and K be partial cubes. Then $(H \,\Box\, K)^{\#} = H^{\#} \oplus K^{\#}$.*

Let A and B be graphs. Clearly, $A \oplus B$ is a complete bipartite graph if and only if both A and B have no edges. In [15] it has also been proved that $G^{\#}$ is a complete bipartite graph if and only if G is the Cartesian product of two trees. In this section we show, a bit surprisingly, that any other join of graphs can be realized as the crossing graph of a partial cube that is prime with respect to the Cartesian product.

Recall that the *radius* of a connected graph G is $\min_{u \in V(G)} \max_{v \in V(G)} d_G(u, v)$ and that G is called $K_{2,3}$-free if it contains no induced subgraph isomorphic to $K_{2,3}$. Note that partial cubes are $K_{2,3}$-free, as follows readily from the fact that Θ is not transitive on $K_{2,3}$.

For the main result of this section we first state the following lemma, which might be of independent interest.

Lemma 6. *Let G be a bipartite graph of radius 2. Then G is a partial cube if and only if G is $K_{2,3}$-free.*

Proof. We only need to show that if G is bipartite of radius 2 and $K_{2,3}$-free, then G is a partial cube. Let u be a vertex that realizes the radius of G and let $v_1, \ldots, v_k$ be its neighbors. As G is bipartite, $v_1, \dots, v_k$ is an independent set of G. Let $w_1, \dots, w_r$ be the remaining vertices of G; then they are all at distance 2 from u. Again, there is no edge between w_i and w_j.

Note that a graph is a partial cube if and only if the graph obtained from it by removing a pendant vertex is a partial cube. Hence we may without loss of generality assume that G has no pendant vertex. Since G is $K_{2,3}$-free, it follows that every vertex w_i is of degree 2. Moreover, no two vertices w_i and w_j, $i \neq j$, have the same pair of neighbors. Therefore every edge of the form $w_i v_j$ lies in precisely one square.

No two edges uv_i and uv_j, $i \neq j$, are in relation Θ. We claim that G isometrically embeds into Q_k and construct edge-subsets $E_1, \dots, E_k$ of $E(G)$ as follows. For $i = 1, \dots, k$ put uv_i in E_i. Consider an edge $w_i v_j$ and let $w_i v_j u v_\ell$ be the unique square containing this edge. Then $w_i v_j$ is in relation Θ with uv_ℓ. Put $w_i v_j \in E_\ell$. We claim that $E_1, \dots, E_k$ form the $\Theta = \Theta^*$-classes of G.

Clearly, $E_1, \dots, E_k$ is a partition of $E(G)$. Suppose $w_i v_j$ and $w_{i'} v_{j'}$ are two distinct edges of E_ℓ. Note first that $i \neq i'$, for otherwise w_i would have three neighbors at distance 1 from u; see Figure 1(i). The case $j = j'$ leads to another $K_{2,3}$; see Figure 1(ii). Hence $i \neq i'$ and $j \neq j'$ and we have the situation as shown in Figure 1(iii).

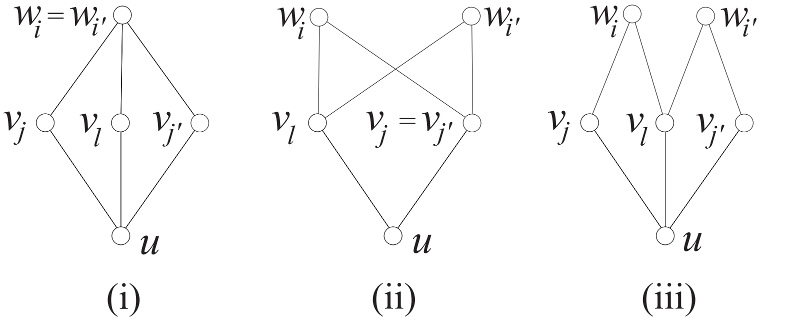


FIG. 1. *Cases in the proof of Lemma* 6.

Then $w_i v_\ell \in E(G)$ and $w_{i'} v_\ell \in E(G)$, which implies that $w_i v_j$ is in relation Θ with $w_{i'} v_{j'}$. Thus all pairs of edges from E_ℓ are in relation Θ. Now assume $w_i v_j \in E_\ell$ and $w_{i'} v_{j'} \in E_{\ell'}$, where $\ell \neq \ell'$. If $i = i'$ or $j = j'$, then clearly $w_i v_j$ and $w_{i'} v_{j'}$ are not in relation Θ. Next, if $\ell = j'$, then $d(w_i, w_{i'}) + d(v_j, v_{j'}) = 2 + 2$ is equal to $d(w_i, v_{j'}) + d(w_{i'}, v_j) = 1 + 3$; hence they are again not in relation Θ (the case $\ell' = j$ is analogous). Otherwise we get $d(w_i, w_{i'}) + d(v_j, v_{j'}) = 4 + 2 = 3 + 3 = d(w_i, v_{j'}) + d(w_{i'}, v_j)$. Hence we conclude that $\Theta = \Theta^*$ and thus G is a partial cube by Winkler's theorem. ∎

THEOREM 7. *Let A and B be arbitrary graphs, where at least one of them contains an edge. Then there exists a Cartesian prime partial cube G such that $G^{\#} = A \oplus B$.*

Proof. For a graph H let $\widetilde{H}$ be the graph obtained from H by subdividing all edges of H and adding a new vertex u joined to all the original vertices of H. (This construction has been introduced in [12] to establish a connection between median graphs and triangle-free graphs.) We claim that $G = \widetilde{A \oplus B}$ does the job.

Let $V(A) = \{a_1, \dots, a_n\}$ and $V(B) = \{b_1, \dots, b_m\}$, so that in G the vertex u is adjacent to $a_1, \dots, a_n$ and to $b_1, \dots, b_m$. Let x_{ij} be the vertex of G obtained by subdividing the edge $a_i b_j$, $1 \leq i \leq n$, $1 \leq j \leq m$.

We first observe that G is a partial cube by Lemma 6. Let E_i be the Θ-classes of G with the representative ua_i, $1 \leq i \leq n$, and let F_i be the Θ-classes of G with the representative ub_i, $1 \le i \le m$. Consider the square $ua_ix_{ij}b_j$ to infer that E_i and F_j cross. Similarly, E_i and E_j (resp., F_i and F_j) cross if and only if $a_ia_j \in E(A)$ (resp., $b_ib_j \in E(B)$). Hence $G^{\#} = A \oplus B$.

It remains to show that G is prime with respect to the Cartesian product. Assume without loss of generality that $n \ge 2$ and that $a_1a_2 \in E(A)$. Let a_i, a_j, $i \ne j$, be arbitrary vertices of A and b_k a vertex of B. Then we have $x_{ik}b_k \in E_i$ and $x_{jk}b_k \in E_j$. By the construction of G (recall that x_{ik} and x_{jk} are of degree 2) we infer that the edges $x_{ik}b_k$ and $x_{jk}b_k$ are in relation τ. As i and j were arbitrary, it follows that $E_1, \ldots, E_n$ belong to the same equivalence class of $(\Theta_G \cup \tau_G)^*$. Analogously, $F_1, \ldots, F_m$ belong to the same equivalence class of $(\Theta_G \cup \tau_G)^*$. Let y be the vertex of G obtained by subdividing the edge a_1a_2. Then we have $a_1y \in E_2$ and $a_1x_{11} \in F_1$. Moreover, a_1y is in relation τ with a_1x_{11}, which implies that $(\Theta_G \cup \tau_G)^*$ consists of a single equivalence class. By Corollary 1 we conclude that G is a Cartesian prime graph. $\square$

Other constructions that yield joins of graphs as crossing graphs can also be obtained. Let A be a graph and let G be the graph that is obtained from $\widetilde{A}$ by the Chepoi expansion (cf. [5]) with covering sets A and the star induced by u and its neighbors. Then G is a partial cube with $G^{\#} = K_1 \oplus A$. This construction is illustrated in Figure 2 for the case when A is the graph on four vertices and five edges. The new Θ-class of G that yields the K_1 in the join decomposition is denoted with thick lines.

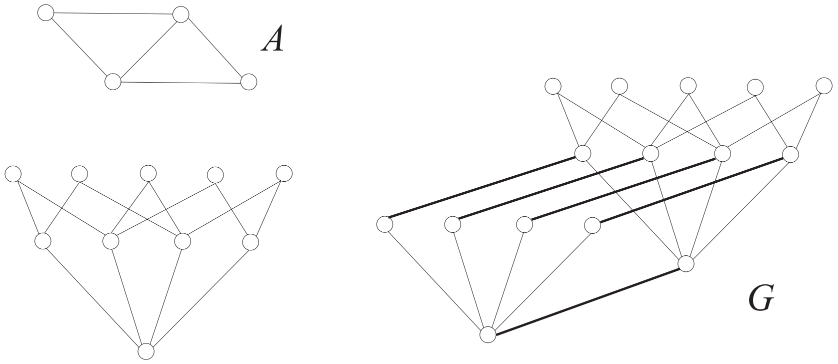


FIG. 2. *Expanding $\widetilde{A}$ into G, so that $G^{\#} = K_1 \oplus A$.*

4. The case of median graphs. Crossing graphs of median graphs are easier to study than those of general partial cubes, since if two Θ-classes of a median graph cross on some cycle, then there exists a square in which they cross. This fact can be easily seen by using the expansion procedure and induction.

In [15] it is proved that every graph is the crossing graph of some median graph. However, it was erroneously mentioned that there are prime median graphs whose crossing graphs are joins of two graphs. The graph presented in Figure 7.2 of [15] is a Cartesian product graph, namely $P_3 \,\square\, G$, where G is the graph obtained from C_4 and another vertex joined to one of the vertices of C_4. In this section we prove that the above remark is indeed wrong by proving that a median graph whose crossing graph is the join of two graphs is necessarily the Cartesian product of two graphs. Note that this is in surprising contrast to the situation from the previous section. We will need the following lemma that might be of independent interest. It follows from the

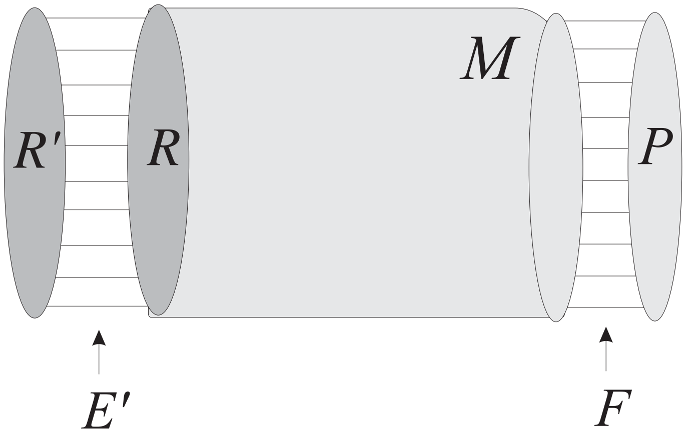


Fig. 3. *Case* $|A| = 1$ *in the proof of Theorem* 9.

Convexity Lemma from [10], which asserts that an induced connected subgraph H of a bipartite graph G is convex if and only if no edge with one endvertex in H and the other not in H is in relation Θ to an edge in H.

Lemma 8. *Let G be a partial cube and H a convex subgraph of G. If H intersects all Θ-classes of G, then $H = G$.*

Proof. Suppose H is a proper subgraph of G. Then, since H is convex and hence induced, there exists an edge uv of G such that $u \in H$ and $v \notin H$. By the Convexity Lemma, uv is in relation Θ to no edge of H. But then H does not intersect the Θ-class of uv, a contradiction. □

We can now state the main result of this section.

Theorem 9. *Let G be a median graph. Then $G^{\#} = A \oplus B$ if and only if $G = H \,\square\, K$, where $H^{\#} = A$ and $K^{\#} = B$.*

Proof. By Proposition 5 one direction is proved: The crossing graph of the Cartesian product of median graphs is the join of the crossing graphs of the factors. Hence it remains to prove the converse of this statement, for which we will use induction on the number of Θ-classes of a median graph G. Clearly the smallest graph that is the join of two graphs and the crossing graph of a median graph is K_2. It is obvious that the only median graph with exactly two Θ-classes that cross is C_4, and $C_4 = K_2 \,\square\, K_2$, providing the basis of the induction.

Assume the statement holds for median graphs with fewer than k Θ-classes. Let G be a median graph with k Θ-classes and $G^{\#} = A \oplus B$. By Theorem 3, G can be obtained by the peripheral expansion from a median graph M along its convex subgraph R. Denote by R' the corresponding peripheral subgraph (isomorphic to R), that is, $R' = G - M$. As M has one Θ-class less than G, $M^{\#}$ is an induced subgraph of $G^{\#}$. More precisely $M^{\#} = G^{\#} - u$, where u corresponds to the peripheral Θ-class E' of G. Without loss of generality we may assume that $u \in A$.

Assume first that $|A| = 1$. By Proposition 4 both connected components of $G - E'$ contain a peripheral subgraph. One component clearly induces the peripheral subgraph R'. Let P be a peripheral subgraph in the other component of $G - E'$. Denote by F the Θ-class such that P is a component of $G - F$ and denote by v the vertex of $G^{\#}$ that corresponds to F (see Figure 3). If $F \neq E'$, then F and E' do not cross, for otherwise P would lie in both components of $G - E'$. Hence, in $G^{\#}$ vertices u and v are not adjacent, which means that they must both be in A, but this is a contradiction with $|A| = 1$. The remaining case is $E' = F$, which implies $P = R$. Hence $G = K_2 \,\square\, R$, where $R^{\#} = B$.

Now, let $|A| > 1$. Then $M^{\#} = (A - u) \oplus B$, and by the induction hypothesis, $M = U \,\square\, K$, where $U^{\#} = A - u$ and $K^{\#} = B$. Note that Θ-classes of M consist of Θ-classes of U and of Θ-classes of K. More precisely, if F is a Θ-class of U (resp., K), then $F \times V(K)$ (resp., $V(U) \times F$) is a Θ-class of $U \,\square\, K$; cf. [11, Lemma 4.3]. Denote by $u_1, \ldots, u_p$ the vertices of $A - u$ that correspond to Θ-classes of U, and by $v_1, \ldots, v_r$ the vertices of B that correspond to Θ-classes of K. By Lemma 2, $R = U' \,\square\, K'$, where U' is a convex subgraph of U and K' is a convex subgraph of K.

Suppose K' is a proper subgraph of K. By Lemma 8 there exists a Θ-class of K that does not intersect with R, and thus it does not cross with E'. This implies that there is a vertex $v_i \in B$ which is not adjacent to $u \in A$, a contradiction. Hence $K' = K$ and $R = U' \,\square\, K$, where U' is a convex subgraph of U. We deduce that $G = H \,\square\, K$, where H is the graph obtained from U by expanding U'. Clearly $H^{\#} = A$ and $K^{\#} = B$, which completes the proof. □

REFERENCES

[1] H.-J. Bandelt and V. Chepoi, *Graphs of acyclic cubical complexes*, European J. Combin., 17 (1996), pp. 113–120.

[2] B. Brešar, *Arboreal structure and regular graphs of median-like classes*, Discuss. Math. Graph Theory, 23 (2003), pp. 215–225.

[3] B. Brešar, S. Klavžar, A. Lipovec, and B. Mohar, *Cubic inflation, mirror graphs, regular maps, and partial cubes*, European J. Combin., 25 (2004), pp. 55–64.

[4] S. R. Canoy, Jr., and I. J. L. Garces, *Convex sets under some graph operations*, Graphs Combin., 18 (2002), pp. 787–793.

[5] V. D. Chepoĭ, *Isometric subgraphs of Hamming graphs and d-convexity*, Kibernetika (Kiev) (1988), pp. 6–9, 15, 133 (in Russian); Cybernetics, 24 (1988), pp. 6–11 (in English).

[6] M. Deza, M. Dutour-Sikiric, and S. Shpectorov, *Graphs 4_n that are isometrically embeddable in hypercubes*, Southeast Asian Bull. Math., 29 (2005), pp. 469–484.

[7] D. Djoković, *Distance preserving subgraphs of hypercubes*, J. Combin. Theory Ser. B, 14 (1973), pp. 263–267.

[8] D. Eppstein, *The lattice dimension of a graph*, European J. Combin., 26 (2005), pp. 585–592.

[9] T. Feder, *Product graph representations*, J. Graph Theory, 16 (1992), pp. 467–488.

[10] W. Imrich and S. Klavžar, *A convexity lemma and expansion procedures for bipartite graphs*, European J. Combin., 19 (1998), pp. 677–685.

[11] W. Imrich and S. Klavžar, *Product Graphs: Structure and Recognition*, Wiley Interscience, New York, 2000.

[12] W. Imrich, S. Klavžar, and H. M. Mulder, *Median graphs and triangle-free graphs*, SIAM J. Discrete Math., 12 (1999), pp. 111–118.

[13] W. Imrich and J. Žerovnik, *Factoring Cartesian-product graphs*, J. Graph Theory, 18 (1994), pp. 557–567.

[14] S. Klavžar and H. M. Mulder, *Median graphs: Characterizations, location theory, and related structures*, J. Combin. Math. Combin. Comput., 30 (1999), pp. 103–127.

[15] S. Klavžar and H. M. Mulder, *Partial cubes and crossing graphs*, SIAM J. Discrete Math., 15 (2002), pp. 235–251.

[16] T. A. McKee and F. R. McMorris, *Topics in Intersection Graph Theory*, SIAM Monogr. Discrete Math. Appl. 2, SIAM, Philadelphia, 1999.

[17] H. M. Mulder, *The structure of median graphs*, Discrete Math., 24 (1978), pp. 197–204.

[18] H. M. Mulder, *The expansion procedure for graphs*, in Contemporary Methods in Graph Theory, R. Bodendiek, ed., B.I.-Wissenschaftsverlag, Mannheim, Wien, Zürich, 1990, pp. 459–477.

[19] G. Sabidussi, *Graph multiplication*, Math. Z., 72 (1960), pp. 446–457.

[20] V. G. Vizing, *On an estimate of the chromatic class of a p-graph*, Diskret. Analiz No., 3 (1964), pp. 25–30 (in Russian).

[21] P. Winkler, *Isometric embeddings in products of complete graphs*, Discrete Appl. Math, 7 (1984), pp. 221–225.

SIAM J. Discrete Math.
Vol. 21, No. 1, pp. 33–46

CORRELATION OF GRAPH-THEORETICAL INDICES[*]

STEPHAN G. WAGNER[†]

Abstract. The correlation of graph characteristics, such as the number of independent vertex or edge subsets, the number of connected subsets, or the sum of distances, which also play a role in combinatorial chemistry, is studied by a generating function approach and asymptotic analysis. It is shown how an asymptotic formula for the correlation coefficient can be obtained when simply generated families of trees are investigated. For rooted ordered trees, the calculations are done explicitly. Further feasible correlation measures are discussed.

Key words. trees, correlation, graph-theoretical index

AMS subject classifications. 05A15, 05A16, 05C05, 05C30

DOI. 10.1137/050631446

1. Introduction.

In combinatorial chemistry, so-called topological indices are used for the description of the structural properties of molecular graphs. Formally, such an index is a map from the set of graphs into the real numbers (usually integer-valued). Typically, for a fixed number of vertices, the trees of maximal and minimal indices are the path and the star, respectively (or vice versa). A variety of graph-theoretical indices has been proposed for this purpose, and their connection to the physico-chemical properties of the corresponding molecules has been studied (cf. [19, 23]).

The isomer-discriminating power, a measure of the ability of an index to distinguish between isomeric compounds, has been considered in the paper [15], and there is also a large amount of literature on extremal and asymptotic properties of various indices; we refer to [3, 4, 11, 12, 16, 20, 22].

However, it seems that there is yet no theoretical result on the correlation between the different indices. It should be quite natural to claim some strong correlation between them, since they all reflect the structural properties of graphs in some way. This paper tries to fill this gap a little by proposing and discussing measures for the correlation of two indices.

The main part of this paper will deal with the asymptotic behavior of the classical correlation coefficient given by

$$r(X_n, Y_n) = \frac{E(X_nY_n) - E(X_n)E(Y_n)}{\sqrt{\operatorname{Var}(X_n)\operatorname{Var}(Y_n)}}. \tag{1.1}$$

Here, $X_n = X(T_n)$ and $Y_n = Y(T_n)$ are the X-index and Y-index of a tree T_n on n vertices taken uniformly at random from some family of trees—for simplicity, we will consider only rooted ordered trees in detail; however, the methods can be extended to other families of simply generated trees (such as binary trees; cf. [4, 17]) quite easily.

The asymptotic behavior of the correlation coefficient will give us a measure of the linear correlation of the indices X and Y. Other possible ways to define such

[*] Received by the editors May 12, 2005; accepted for publication (in revised form) September 5, 2006; published electronically January 12, 2007. This work was supported by Austrian Science Fund project S-8307-MAT.

http://www.siam.org/journals/sidma/21-1/63144.html

[†] Department of Mathematics, Graz University of Technology, Steyrergasse 30, A-8010 Graz, Austria (wagner@finanz.math.tugraz.at).

a measure are discussed in the last section, but it seems that a similar asymptotic analysis is not practicable in these cases.

The indices that will be taken into consideration are the following:

(1) The *Merrifield–Simmons-index* (or σ-index) is defined to be the number of independent vertex subsets of a graph, i.e., the number of vertex subsets in which no two vertices are adjacent, including the empty set. Merrifield and Simmons investigated the σ-index in their work [19] and pointed out its correlation to boiling points of molecules.

(2) The *Hosoya-index* (or Z-index) [8] is defined as the number of independent edge subsets (also referred to as "matchings"), i.e., the number of edge subsets in which no two edges are adjacent, again including the empty set.

(3) The *number of subtrees* is called the ρ-index in [19] and was discussed recently in a paper of Székely and Wang [22].

(4) The *Wiener-index* is probably the most popular topological index (see [3, 4, 26]). It is defined as the sum of all the distances between pairs of vertices, i.e.,

$$W(G) = \sum_{v,w \in V(G)} d_G(v,w). \tag{1.2}$$

Section 2 will deal with the correlation of (1), (2), and (3). The Wiener-index has a different growth structure than the other three, so we need a different approach, which will be presented in section 3. Finally, we will take a look at some other statistical measures in section 4.

2. σ-index, Z-index, and ρ-index. The method for determining the expected values of these indices for rooted ordered trees on n vertices has been given in several papers [11, 12, 13]. However, for the sake of completeness, it is repeated here. It is well known that the generating function for the number of rooted ordered trees is given by the functional equation

$$T(z) = \frac{z}{1 - T(z)}, \tag{2.1}$$

which is an immediate consequence of the recursive structure of this family of trees. Now, consider the σ-index, for instance. We want to determine the function

$$S(z) = \sum_T \sigma(T) z^{|T|},$$

where the sum goes over all trees T, and $|T|$ denotes the number of vertices. Now, we distinguish between independent sets containing the root and those not containing it and denote the corresponding quantities by $\sigma_1(T), \sigma_2(T)$. If $T_1, \ldots, T_k$ are the branches of the rooted tree T, it is easy to see that the recursive relations

$$\sigma_1(T) = \prod_{i=1}^{k} \sigma_2(T_i),$$
$$\sigma_2(T) = \prod_{i=1}^{k} (\sigma_1(T_i) + \sigma_2(T_i))$$

hold. These relations can be translated into equations for the corresponding generating functions: if $S_1(z)$ is the generating function for the number of subsets of the first type and $S_2(z)$ the generating function for the number of subsets of the second type, we obtain

$$\begin{aligned} S_1(z) &= \sum_T \sigma_1(T) z^{|T|} \\ &= \sum_{k\geq 0} \sum_{T_1} \sum_{T_2} \cdots \sum_{T_k} \left(\prod_{i=1}^{k} \sigma_2(T_i) \right) z^{|T_1|+\cdots+|T_k|+1} \\ &= z \sum_{k\geq 0} \left(\sum_T \sigma_2(T) z^{|T|} \right)^k \\ &= z \sum_{k\geq 0} S_2(z)^k = \frac{z}{1-S_2(z)}, \end{aligned} \tag{2.2}$$

and in exactly the same way,

$$S_2(z) = \frac{z}{1-S_1(z)-S_2(z)}. \tag{2.3}$$

The asymptotic growth of the coefficients of functions satisfying algebraical equations of this kind can be determined by a standard application of the Flajolet–Odlyzko singularity analysis, which is discussed in several papers, such as [1, 2, 5, 18] (sometimes, one can even find exact expressions by means of Lagrange's inversion formula; this is the case for this example (see [11, 12]), but we won't need the exact solution, which can be given as a hypergeometric sum). However, the details can be intricate, as will be explained in the following. Here, using (2.2) in (2.3) yields

$$S_2(z) = \frac{z}{1 - \frac{z}{1-S_2(z)} - S_2(z)}$$

or

$$S_2(z)^3 - 2S_2(z)^2 + S_2(z) - z = 0.$$

Bender [1] gives a general theorem dealing with functional equations of the type $F(z, w(z)) = 0$. His theorem states that, given a minimal solution (with respect to absolute value) (α, β) of the system

$$F(z,w) = 0, F_w(z,w) = 0,$$

which lies within the region of analyticity of F and satisfies $F_z(\alpha,\beta), F_{ww}(\alpha,\beta) \neq 0$, the asymptotic behavior of the coefficients a_n of $w(z)$ is determined by

$$a_n \sim \sqrt{\frac{\alpha F_z(\alpha,\beta)}{2\pi F_{ww}(\alpha,\beta)}} n^{-3/2} \alpha^{-n}.$$

However, there is a slight mistake in this theorem, as was pointed out by Canfield [2], and the method might give erroneous results. The theorem holds only if α is indeed the radius of convergence of $w(z)$ and the only singularity on the circle of convergence.

In the present case, we know from [7, Thm. 12.2.1] (see also [2]) that a singularity of an algebraic function $w(z)$ given by a polynomial equation of the form

$$F(z,w) = \sum_{j=0}^{k} p_{k-j}(z) w^j = 0$$

is either a zero of $p_0(z)$ (here, there is no such zero) or given by a solution of the system $F(z,w) = 0$, $F_w(z,w) = 0$.

Therefore, the common singularity z_0 of $S_1(z), S_2(z)$, and $S(z) = S_1(z) + S_2(z)$ nearest to the origin is given by the system of equations

$$\begin{aligned} F(s,z) &= s^3 - 2s^2 + s - z = 0, \\ \frac{\partial}{\partial s} F(s,z) &= 3s^2 - 4s + 1 = 0, \end{aligned}$$

yielding $z_0 = \frac{4}{27}$. Using the formula for the number of rooted ordered trees on n vertices,

$$t_n = \frac{1}{n}\binom{2n-2}{n-1} \sim \frac{1}{4\sqrt{\pi}} n^{-3/2} 4^n,$$

it is easy now to find the asymptotics for the expected σ-index:

$$E(\sigma_n) \sim \sqrt{3}\left(\frac{27}{16}\right)^{n-1} \approx (1.02640) \cdot (1.6875)^n.$$

Similarly, for the Z-index, we have

$$\begin{aligned} Z_1(T) &= \sum_{j=1}^{k} Z_2(T_j) \prod_{\substack{i=1 \\ i \neq j}}^{k} (Z_1(T_i) + Z_2(T_i), \\ Z_2(T) &= \prod_{i=1}^{k} (Z_1(T_i) + Z_2(T_i)), \end{aligned}$$

where $Z_1(T)$ and $Z_2(T)$ denote the number of independent edge subsets containing (resp., not containing) an edge incident to the root. From this, we obtain the equations

$$\begin{aligned} Z_1(z) &= \frac{zZ_2(z)}{(1 - Z_1(z) - Z_2(z))^2}, \\ Z_2(z) &= \frac{z}{1 - Z_1(z) - Z_2(z)} \end{aligned} \tag{2.4}$$

for the respective generating functions. This system gives us the asymptotic expression for the average Z-index:

$$E(Z_n) \sim \sqrt{\frac{65 - \sqrt{13}}{78}} \left(\frac{35 + 13\sqrt{13}}{54}\right)^n \approx (0.88719) \cdot (1.51615)^n.$$

Finally, for the ρ-index,

$$\begin{aligned} \rho_1(T) &= \prod_{i=1}^{k} (1 + \rho_1(T_i)), \\ \rho_2(T) &= \sum_{i=1}^{k} (\rho_1(T_i) + \rho_2(T_i)), \end{aligned}$$

where $\rho_1(T)$ and $\rho_2(T)$ denote the number of subtrees containing (resp., not containing) an edge incident to the root. Here, the system of equations for the corresponding generating functions is

$$\begin{aligned} R_1(z) &= \frac{z}{1-R_1(z)-T(z)}, \\ R_2(z) &= \frac{z}{(1-T(z))^2}(R_1(z)+R_2(z)), \end{aligned} \tag{2.5}$$

yielding

$$E(\rho_n) \sim \frac{16}{3\sqrt{15}}\left(\frac{25}{16}\right)^n \approx (1.37706)\cdot(1.5625)^n.$$

All these results have already been given in a paper of Klazar [13]. Now, to find the covariances, one needs four generating functions connected by a system of equations. For the covariance of the σ-index and Z-index, for example, we take $\mathrm{SZ}_{11},\dots,\mathrm{SZ}_{22}$ to be the generating functions for the product of the number of independent vertex subsets and independent edge subsets such that the root is contained in

- the vertex and the edge subset;
- the vertex, but not the edge subset;
- the edge, but not the vertex subset;
- neither,

respectively. The functional equations can be seen to be a combination of those for S_1 and S_2 (resp., Z_1 and Z_2):

$$\begin{aligned} \mathrm{SZ}_{11}(z) &= \frac{z\,\mathrm{SZ}_{22}(z)}{(1-\mathrm{SZ}_{21}(z)-\mathrm{SZ}_{22}(z))^2}, \\ \mathrm{SZ}_{12}(z) &= \frac{z}{1-\mathrm{SZ}_{21}(z)-\mathrm{SZ}_{22}(z)}, \\ \mathrm{SZ}_{21}(z) &= \frac{z(\mathrm{SZ}_{12}(z)+\mathrm{SZ}_{22}(z))}{(1-\mathrm{SZ}_{11}(z)-\mathrm{SZ}_{12}(z)-\mathrm{SZ}_{21}(z)-\mathrm{SZ}_{22}(z))^2}, \\ \mathrm{SZ}_{22}(z) &= \frac{z}{1-\mathrm{SZ}_{11}(z)-\mathrm{SZ}_{12}(z)-\mathrm{SZ}_{21}(z)-\mathrm{SZ}_{22}(z)}. \end{aligned} \tag{2.6}$$

For instance, the functional equation for SZ_{11} is derived as follows:

$$\begin{aligned} \mathrm{SZ}_{11}(z) &= \sum_T \sigma_1(T)Z_1(T)z^{|T|} \\ &= \sum_{k\ge 0}\sum_{j=1}^{k}\sum_{T_1}\sum_{T_2}\cdots\sum_{T_k}\left(\sigma_2(T_j)Z_2(T_j)\prod_{i\ne j}\sigma_2(T_i)(Z_1(T_i)+Z_2(T_i))\right) \\ &\quad\cdot z^{|T_1|+\cdots+|T_k|+1} \\ &= z\sum_{k\ge 0} k\,\mathrm{SZ}_{22}(z)(\mathrm{SZ}_{21}(z)+\mathrm{SZ}_{22}(z))^{k-1} \\ &= \frac{z\,\mathrm{SZ}_{22}(z)}{(1-SZ_{21}(z)-\mathrm{SZ}_{22}(z))^2}. \end{aligned}$$

Since all the functional equations can be written in polynomial form, it is possible to employ the method of Gröbner bases (cf. [6]) and a computer algebra package, such as Mathematica (for details, see [24]), to obtain a single polynomial equation from the system. In this case, we find that $s = \mathrm{SZ}_{22}(z)$ satisfies the polynomial equation

$$F(z,s) = s^{10} + 2zs^8 - 3zs^7 + z^2s^6 - 4z^2s^5 + 3z^2s^4 - z^3s^3 + 2z^3s^2 - z^3s + z^4 = 0.$$

Since $\mathrm{SZ}(z) = \mathrm{SZ}_{11}(z) + \mathrm{SZ}_{12}(z) + \mathrm{SZ}_{21}(z) + \mathrm{SZ}_{22}(z) = 1 - \frac{z}{\mathrm{SZ}_{22}(z)}$, the smallest singularity of SZ is either a singularity of SZ_{22} or a zero of SZ_{22}. However, from the functional equation we know that SZ_{22} has only one zero at $z = 0$, where the zero cancels out with the numerator. Therefore, we have only to find the smallest singularity of SZ_{22} to apply Bender's theorem. Fortunately, things are still comparatively simple since we can bound the range of the singularity by an a priori estimate.

Again, the leading coefficient of the polynomial equation is 1, so it has no zeroes. Therefore, the dominating singularity is a solution of the system $F(z,w) = 0$, $F_w(z,w) = 0$ again. The solutions of this system can be found by the method of Gröbner bases as well—it turns out that a singularity z_0 of SZ must be a solution of

$$5038848z^4 - 221833728z^3 + 5017360096z^2 + 3451610880z - 387420489 = 0.$$

Now we note that, for trivial reasons, $1 \leq \sigma(T), Z(T), \rho(T) \leq 2^{|T|}$ for all trees T. This shows that the coefficients c_n of SZ are bounded by

$$\frac{1}{n}\binom{2n-2}{n-1} \leq c_n \leq \frac{1}{n}\binom{2n-2}{n-1} \cdot 4^n,$$

so the radius of convergence of SZ lies in the interval $\left[\frac{1}{16}, \frac{1}{4}\right]$. Thus we have only to search for a solution whose absolute value lies within this interval. There is only one such solution in this case, which is given by $z_0 \approx 0.0982673$. Expanding SZ_{22} and SZ around this singularity and applying Bender's formula yields an asymptotic expression for the expected product of the σ-index and Z-index:

$$E(\sigma_n Z_n) \sim (0.92565) \cdot (2.54408)^n.$$

Of course, the same reasoning can also be used to determine the other expected values $E(\sigma_n \rho_n)$ and $E(Z_n \rho_n)$, as well as the variances of all our random variables. All details (which are mostly analogous to the example) are given in [24]. Therefore, we list all the asymptotics only in Table 2.1.

Table 2.1
Asymptotic formulas for expected values and variances.

$E(\sigma_n)$	$\sqrt{3}\left(\frac{27}{16}\right)^{n-1} \sim (1.02640) \cdot (1.6875)^n$
$E(Z_n)$	$\sqrt{\frac{65-\sqrt{13}}{78}}\left(\frac{35+13\sqrt{13}}{54}\right)^n \sim (0.88719) \cdot (1.51615)^n$
$E(\rho_n)$	$\frac{16}{3\sqrt{15}}\left(\frac{25}{16}\right)^n \sim (1.37706) \cdot (1.5625)^n$
$E(\sigma_n Z_n)$	$(0.92565) \cdot (2.54408)^n$
$E(\sigma_n \rho_n)$	$(1.36653) \cdot (2.66477)^n$
$E(Z_n \rho_n)$	$\frac{1}{116}\sqrt{\frac{5(128985+57683\sqrt{5})}{58}} \cdot \left(8(7-3\sqrt{5})\right)^n \sim (1.28557) \cdot (2.33437)^n$
$\mathrm{Var}(\sigma_n)$	$(1.03802) \cdot (2.86096)^n$
$\mathrm{Var}(Z_n)$	$(0.77227) \cdot (2.31549)^n$
$\mathrm{Var}(\rho_n)$	$\frac{64\sqrt{14}}{147} \cdot \left(\frac{81}{32}\right)^n \sim (1.79509) \cdot (2.53125)^n$

Now we can turn to the correlation coefficients. We see that

$$\begin{aligned} r(\sigma_n, Z_n) &\sim (-1.01706) \cdot (0.99405)^n, \\ r(\sigma_n, \rho_n) &\sim (1.05088) \cdot (0.99023)^n, \\ r(Z_n, \rho_n) &\sim (-1.08924) \cdot (0.97853)^n \end{aligned}$$

and conclude that the σ-index and ρ-index are positively correlated, whereas they are both negatively correlated to the Z-index. The correlation coefficient tends to zero as $n \to \infty$, but very slowly. The constant factor as well as the basis of the exponential term can be used as a measure for the correlation. So we may claim that the closest correlation of the three is between the σ-index and the Z-index.

3. Correlation to the Wiener-index. The Wiener-index has a different recursive structure than the indices discussed in the preceding chapter, and its growth is not exponential. Entringer et al. [4] were able to show that the average Wiener-index is asymptotically $K \cdot n^{5/2}$ for a simply generated family of trees, where K is a constant depending on the specific family. For rooted ordered trees, the constant K is $\frac{\sqrt{\pi}}{4}$. We repeat the argument of [4] here since it will be needed for the computation of the covariances.

We are first going to consider an auxiliary value, $D(T)$, denoting the sum of the distances of all vertices from the root. This is also known as the *total height* [21] or *internal path length* [9] of the tree T. Then, we set

$$D(z) := \sum_T D(T) z^{|T|},$$

where the sum again runs over all rooted ordered trees T. The value $D(T)$ can be calculated recursively from the branches of T: in fact, if $T_1, \dots, T_k$ are the branches of T, we have

$$D(T) = \sum_{i=1}^{k} D(T_i) + |T| - 1, \tag{3.1}$$

where $|T|$ is the size (number of vertices) of T. In terms of $D(z)$, this gives

$$\begin{aligned} D(z) &= \sum_T D(T) z^{|T|} \\ &= \sum_{k\geq 0} \sum_{i=1}^{k} \sum_{T_1} \sum_{T_2} \dots \sum_{T_k} D(T_i) z^{|T_1|+\cdots+|T_k|+1} + \sum_T (|T|-1) z^{|T|} \\ &= z \sum_{k \geq 0} k D(z) T(z)^{k-1} + zT'(z) - T(z) \\ &= \frac{zD(z)}{(1-T(z))^2} + zT'(z) - T(z). \end{aligned} \tag{3.2}$$

Now, the Wiener-index of a tree can also be determined recursively from its branches:

$$W(T) = D(T) + \sum_{i=1}^{k} W(T_i) + \sum_{i \neq j} \Big(D(T_i) + |T_i| \Big) |T_j|, \tag{3.3}$$

where the last sum goes over all $k(k-1)$ pairs of different branches. Thus, if

$$W(z) := \sum_T W(T) z^{|T|},$$

we have

$$W(z) = D(z) + \frac{zW(z)}{(1-T(z))^2} + \frac{2z^2T'(z)(D(z)+zT'(z))}{(1-T(z))^3}. \tag{3.4}$$

It turns out that $W(z) = \frac{z^2}{(1-4z)^2}$, giving an average Wiener-index of asymptotically $\frac{\sqrt{\pi}}{4}n^{5/2}$. Now, we introduce various generating functions for the correlation of $D(T), W(T)$, and $\sigma(T)$: let $\mathrm{DS}_1, \mathrm{DS}_2, \mathrm{WS}_1$, and WS_2 be the generating functions for the product of $D(T)$ (resp., $W(T)$) with the number of independent vertex subsets containing (resp., not containing) the root. In analogy to the functional equations for $D(z)$ and $W(z)$, we obtain a system of linear equations—for example, we have

$$\begin{aligned}
\mathrm{DS}_1(z) &= \sum_T D(T)\sigma_1(T)z^{|T|} \\
&= \sum_{k\ge 0}\sum_{T_1}\cdots\sum_{T_k}\left(\sum_{i=1}^k D(T_i)\prod_{j=1}^k \sigma_2(T_j)\right) z^{|T_1|+\cdots+|T_k|+1} \\
&\quad + \sum_T (|T|-1)\sigma_1(T)z^{|T|} \\
&= \sum_{k\ge 0}\sum_{T_1}\cdots\sum_{T_k}\left(\sum_{i=1}^k D(T_i)\sigma_2(T_i)\prod_{j\ne i} \sigma_2(T_j)\right) z^{|T_1|+\cdots+|T_k|+1} \\
&\quad + zS_1'(z) - S_1(z) \\
&= z\sum_{k\ge 0} k\,\mathrm{DS}_2(z)S_2(z)^{k-1} + zS_1'(z) - S_1(z) \\
&= \frac{z\,\mathrm{DS}_2(z)}{(1-S_2(z))^2} + zS_1'(z) - S_1(z).
\end{aligned}$$

Altogether, we obtain

$$\begin{aligned}
\mathrm{DS}_1(z) &= \frac{z\,\mathrm{DS}_2(z)}{(1-S_2(z))^2} + zS_1'(z) - S_1(z), \\
\mathrm{DS}_2(z) &= \frac{z(\mathrm{DS}_1(z)+\mathrm{DS}_2(z))}{(1-S_1(z)-S_2(z))^2} + zS_2'(z) - S_2(z), \\
\mathrm{WS}_1(z) &= \mathrm{DS}_1(z) + \frac{z\,\mathrm{WS}_2(z)}{(1-S_2(z))^2} + \frac{2z^2S_2'(z)(\mathrm{DS}_2(z)+zS_2'(z))}{(1-S_2(z))^3}, \\
\mathrm{WS}_2(z) &= \mathrm{DS}_2(z) + \frac{z(\mathrm{WS}_1(z)+\mathrm{WS}_2(z))}{(1-S_1(z)-S_2(z))^2} \\
&\quad + \frac{2z(zS_1'(z)+zS_2'(z))(\mathrm{DS}_1(z)+\mathrm{DS}_2(z)+zS_1'(z)+zS_2'(z))}{(1-S_1(z)-S_2(z))^3}.
\end{aligned} \tag{3.5}$$

We solve this system for WS_1 and WS_2 (which can be done explicitly in terms of S_1 and S_2 since the system is linear) and write the total generating function $\mathrm{WS}(z) = \mathrm{WS}_1(z) + \mathrm{WS}_2(z)$ in terms of S_1, S_2, S_1', S_2'. Then we make use of the functional equations for S_1 and S_2 and replace $S_1(z)$ with $\frac{z}{1-S_2(z)}$. Implicit differentiation of the equation $S_2(z)^3 - 2S_2(z)^2 + S_2(z) - z = 0$ yields

$$S_2'(z) = \frac{1}{3S_2(z)^2 - 4S_2(z) + 1},$$

so WS can be written in terms of only S_2 and z. In fact, we have

$$\mathrm{WS}(z) = \frac{N}{(1-3S_2(z))^2(1-S_2(z))^3(S_2(z)^2+S_2(z)^3-z)^2},$$

where N is a polynomial in S_2 and z. The denominator vanishes only at 0 and at the dominating singularity $\frac{4}{27}$ of S_2. Therefore, we have only to expand WS around $\frac{4}{27}$:

$$\mathrm{WS}(z) \sim \frac{5}{81\left(1-\frac{27z}{4}\right)^2},$$

which once again gives us the expected value $E(W_n\sigma_n)$ by means of the Flajolet–Odlyzko singularity analysis [5]:

$$E(W_n\sigma_n) \sim \frac{20\sqrt{\pi}}{81} n^{5/2} \left(\frac{27}{16}\right)^n.$$

It was shown by Janson [9] that the variance of the Wiener-index for rooted ordered trees is given asymptotically by

$$\mathrm{Var}(W_n) \sim \frac{16-5\pi}{80} n^5,$$

and thus the correlation coefficient of W_n and σ_n is

$$r(W_n, \sigma_n) \sim (-0.27891) \cdot (0.99767)^n.$$

Similarly, we obtain

$$\begin{aligned} r(W_n, Z_n) &\sim (0.40351) \cdot (0.99637)^n, \\ r(W_n, \rho_n) &\sim (-1.78357) \cdot (0.98209)^n. \end{aligned}$$

Again, the calculational details are given in [24].

4. Some numerical values and their interpretation. We have seen that in all the considered cases, the correlation coefficient was asymptotically of the form

$$\alpha \cdot \beta^n$$

for some constants α and β. The significance of these constants can be roughly described as follows:

- A large value of α usually means a higher correlation for trees with few vertices.
- A large value of β means that the correlation decreases very slowly—thus, it is a measure of the correlation of the indices when the number of vertices is large.

When the pairwise correlation of σ, Z, and ρ was considered, β depended on the growth of both indices. If the correlation was negative in these cases (which it was, except for the correlation of the σ-index and ρ-index), the exact asymptotics of the

Table 4.1
$E(X_nY_n)$ and $E(X_n)E(Y_n)$ separated.

Indices	$\frac{E(X_nY_n)}{\sqrt{\operatorname{Var}(X_n)\operatorname{Var}(Y_n)}}$	$\frac{E(X_n)E(Y_n)}{\sqrt{\operatorname{Var}(X_n)\operatorname{Var}(Y_n)}}$	$\frac{E(X_nY_n)}{E(X_n)E(Y_n)}$
$\sigma - Z$	$(1.03386) \cdot (0.988448)^n$	$(1.01706) \cdot (0.99405)^n$	$(1.01652) \cdot (0.99436)^n$
$\sigma - \rho$	$(1.05088) \cdot (0.99023)^n$	$(1.08694) \cdot (0.97981)^n$	$(0.96683) \cdot (1.01064)^n$
$Z - \rho$	$(1.14617) \cdot (0.96423)^n$	$(1.08924) \cdot (0.97853)^n$	$(1.05227) \cdot (0.98539)^n$
$\sigma - W$	$(7.10957) \cdot (0.99767)^n$	$(7.38848) \cdot (0.99767)^n$	0.96225
$Z - W$	$(7.80764) \cdot (0.99637)^n$	$(7.40413) \cdot (0.99637)^n$	1.05450
$\rho - W$	$(6.12924) \cdot (0.98209)^n$	$(7.91281) \cdot (0.98209)^n$	0.77460

Table 4.2
Correlation coefficients for rooted ordered trees, $n \leq 25$.

n	$r(\sigma_n, Z_n)$	$r(\sigma_n, \rho_n)$	$r(Z_n, \rho_n)$	$r(\sigma_n, W_n)$	$r(Z_n, W_n)$	$r(\rho_n, W_n)$
4	-1.000000	1.000000	-1.000000	-1.000000	1.000000	-1.000000
5	-0.991189	0.971494	-0.994334	-0.923381	0.966092	-0.988064
6	-0.970054	0.947369	-0.955649	-0.870581	0.918482	-0.977131
7	-0.959741	0.926080	-0.926321	-0.829908	0.883867	-0.966673
8	-0.950801	0.907123	-0.898558	-0.796570	0.853248	-0.956356
9	-0.943296	0.890225	-0.873371	-0.768197	0.826459	-0.945962
10	-0.936479	0.875159	-0.850213	-0.743446	0.802492	-0.935353
11	-0.930116	0.861703	-0.828817	-0.721477	0.780828	-0.924449
12	-0.924048	0.849641	-0.808906	-0.701723	0.761060	-0.913214
13	-0.918187	0.838772	-0.790246	-0.683782	0.742891	-0.901641
14	-0.912479	0.828909	-0.772640	-0.667357	0.726088	-0.889750
15	-0.906888	0.819890	-0.755923	-0.652218	0.710467	-0.877574
20	-0.880077	0.783214	-0.681768	-0.590624	0.645700	-0.814057
25	-0.854498	0.753917	-0.617683	-0.544547	0.596088	-0.750155

expected value of their product would be redundant for the asymptotics of the correlation coefficient. So, in order to exploit this piece of information as well, one should separately consider normalized values of the form

$$\frac{E(X_nY_n)}{\sqrt{\operatorname{Var}(X_n)\operatorname{Var}(Y_n)}} \text{ and } \frac{E(X_n)E(Y_n)}{\sqrt{\operatorname{Var}(X_n)\operatorname{Var}(Y_n)}},$$

where X_n and Y_n are the X-index and Y-index, respectively, of random trees.

Further problems arise in the study of the Wiener-index. Since the Wiener-index grows only polynomially, β depends only on the expected value and variance of the second index. Again, one should also consider separately the coefficients given above. We have seen that they are of the same asymptotic order except from the constant factors, so one might use their quotient as a correlation measure as well. Table 4.1 gives the asymptotic behavior of these coefficients and their quotient. In any case, our approach will yield us only quantitative correlation measures; qualitative information on the correlation structure is not provided.

One can calculate the exact correlation coefficients for small values of n quite easily from the functional equations. In Table 4.2, some numerical examples are given; note that the correlation coefficient makes sense only for $n \geq 4$: for $n \leq 3$, all trees are isomorphic.

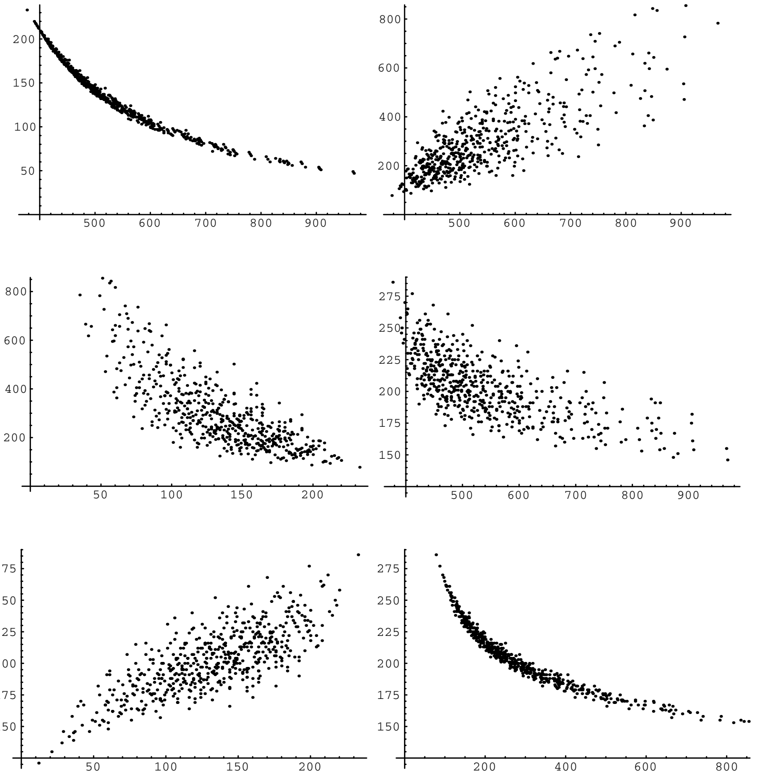


FIG. 4.1. *Top, left to right: σ- and Z-index, σ- and ρ-index. Middle, left to right: Z- and ρ-index, σ- and Wiener-index. Bottom, left to right: Z- and Wiener-index, ρ- and Wiener-index.*

We see that the correlation coefficient between the σ-index and Z-index is largest among those investigated in section 2. Likewise, the correlation to the Wiener-index is highest for the ρ-index. This observation agrees with the asymptotic results of the preceding sections. The plots in Figure 4.1 suggest that the correlation is in fact very strong in both cases (much stronger than for the other pairs, which is quite remarkable), but not entirely linear, which is clear from the exponential growth of the σ-index, Z-index, and ρ-index (this phenomenon will be discussed in detail in the following section). The plots show the values of all trees with 12 vertices.

5. Other correlation measures. Unfortunately, there are some drawbacks in our approach. Apart from the obvious fact that asymptotic correlations might hold only for a considerably large number of vertices, the correlation coefficient principally measures linear dependence. But since the σ-, Z-, and ρ- indices grow exponentially with different growth rates, the dependence cannot be completely linear. Thus, it might be reasonable to instead study the correlation of their logarithms. The problem with that approach is the fact that generating function methods as presented in this paper will no longer be applicable. The corresponding plot for the correlation of $\log \sigma_n$ and $\log Z_n$ (the random variables are rescaled in such a way that they are of equal order now!) suggests that it is reasonable to use a logarithmic transformation—it shows an almost linear correspondence (Figure 5.1). This suggests that a sharp inequality of the form

$$f_n(\sigma(T)) \leq Z(T) \leq g_n(\sigma(T)) \tag{5.1}$$

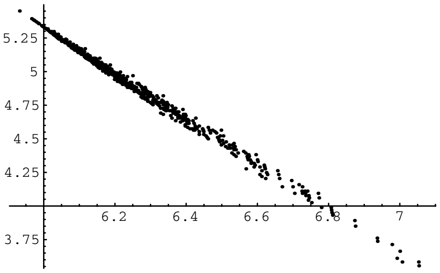


FIG. 5.1. *σ- and Z-index after logarithmic transformation.*

should hold for all trees T on n vertices, where $f_n(x), g_n(x)$ behave like negative powers of x, i.e., $f_n(x) \sim a_1(n)x^{-c_1}$, $g_n(x) \sim a_2(n)x^{-c_2}$. However, it is not difficult to construct discordant pairs of trees, i.e., two trees T_1, T_2 such that $Z(T_1) > Z(T_2)$ and $\sigma(T_1) > \sigma(T_2)$.

This leads us to an alternative method of measuring correlation: the use of rank statistics (cf. [10, 14]). Given two indices X and Y, we assign ranks x_i and y_i to all trees $T_1, \ldots, T_s$ on n vertices such that x_i and y_i range from 1 to s and such that $x_i < x_j$ if $X(T_i) < X(T_j)$ (resp., $y_i < y_j$ if $Y(T_i) < Y(T_j)$). Then, a correlation measure is given by Spearman's ρ,

$$\rho_S(X_n, Y_n) = 1 - \frac{6\sum_{i=1}^{s}(x_i - y_i)^2}{s^3 - s}, \tag{5.2}$$

which ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation). Unfortunately, even though rank statistics are an interesting means of measuring the statistical dependence of random variables, it seems virtually impossible to apply them to our problem, since generating function methods are not applicable in the treatment of ranks. It seems that rank statistics can be applied to our problem only if the number of vertices is considerably small, so that everything can be calculated explicitly.

Another problem with rank statistics is the occurrence of ties—all the random variables under consideration are discrete, and the number of trees grows larger than the maximal index in all our cases, so ties (i.e., several nonisomorphic trees of the same index) are inevitable. There are statistical methods for coping with this problem (cf. [10, 14]); usually, if ties occur, the average rank is allotted to all tied elements. This method is used in the examples at the end of this section.

The problem of ties leads us to our final remark. The methods of this paper easily generalize to all simply generated families of trees. However, one would like to apply them to unordered rooted trees or trees (so one can take isomorphisms into account). This should be doable (in essentially the same way as in [25]), but it certainly requires very lengthy calculations.

In Table 5.1, correlation coefficients for trees with ≤ 14 vertices are given. If we compare them to the values of Table 4.2, we see that the correlation coefficients for rooted ordered trees provide suitable estimates.

Finally, we examine the rank correlation. Table 5.2 shows the numerical values of Spearman's ρ for all trees with ≤ 14 vertices.

Again, we observe the striking correspondence between the σ-index and Z-index (resp., ρ-index and Wiener-index). It seems to be a challenging graph-theoretical problem to explain this phenomenon.

Table 5.1
Correlation coefficients for trees, $n \leq 14$.

n	$r(\sigma_n, Z_n)$	$r(\sigma_n, \rho_n)$	$r(Z_n, \rho_n)$	$r(\sigma_n, W_n)$	$r(Z_n, W_n)$	$r(\rho_n, W_n)$
4	-1.000000	1.000000	-1.000000	-1.000000	1.000000	-1.000000
5	-0.995871	0.986241	-0.997176	-0.960769	0.981981	-0.993399
6	-0.977051	0.969611	-0.982970	-0.901473	0.953231	-0.977255
7	-0.955329	0.959254	-0.943865	-0.863896	0.911843	-0.959471
8	-0.930868	0.947142	-0.918181	-0.819996	0.886845	-0.940935
9	-0.908594	0.932074	-0.869200	-0.778345	0.841803	-0.91815
10	-0.890714	0.920543	-0.836300	-0.748034	0.816189	-0.899454
11	-0.877343	0.903475	-0.797497	-0.714065	0.782806	-0.879018
12	-0.869047	0.889422	-0.767693	-0.689129	0.758290	-0.860836
13	-0.862946	0.872456	-0.739304	-0.663493	0.732342	-0.843721
14	-0.859211	0.857532	-0.715078	-0.642464	0.710476	-0.827013

Table 5.2
Spearman's ρ for $n \leq 14$.

n	$\rho_S(\sigma_n, Z_n)$	$\rho_S(\sigma_n, \rho_n)$	$\rho_S(Z_n, \rho_n)$	$\rho_S(\sigma_n, W_n)$	$\rho_S(Z_n, W_n)$	$\rho_S(\rho_n, W_n)$
4	-1.000000	1.000000	-1.000000	-1.000000	1.000000	-1.000000
5	-1.000000	1.000000	-1.000000	-1.000000	1.000000	-1.000000
6	-1.000000	0.942857	-0.942857	-0.942857	0.942857	-1.000000
7	-1.000000	0.918182	-0.918182	-0.877273	0.886364	-0.986364
8	-0.994071	0.881670	-0.876729	-0.867836	0.870800	-0.996789
9	-0.996126	0.854591	-0.852798	-0.805273	0.809349	-0.990171
10	-0.997048	0.832577	-0.834320	-0.774514	0.777381	-0.992314
11	-0.997392	0.811737	-0.814267	-0.746093	0.749423	-0.990921
12	-0.997471	0.796388	-0.801514	-0.724382	0.729450	-0.990146
13	-0.997421	0.781437	-0.787808	-0.697123	0.703244	-0.987169
14	-0.997383	0.770002	-0.777472	-0.675956	0.682617	-0.984820

Acknowledgments. The author thanks the referees for their valuable help in improving this paper and in particular for bringing references [2, 18] to his attention.

REFERENCES

[1] E. A. Bender, *Asymptotic methods in enumeration*, SIAM Rev., 16 (1974), pp. 485–515.

[2] E. R. Canfield, *Remarks on an asymptotic method in combinatorics*, J. Combin. Theory Ser. A, 37 (1984), pp. 348–352.

[3] A. A. Dobrynin, R. Entringer, and I. Gutman, *Wiener index of trees: Theory and applications*, Acta Appl. Math., 66 (2001), pp. 211–249.

[4] R. C. Entringer, A. Meir, J. W. Moon, and L. A. Székely, *The Wiener index of trees from certain families*, Australas. J. Combin., 10 (1994), pp. 211–224.

[5] P. Flajolet and A. Odlyzko, *Singularity analysis of generating functions*, SIAM J. Discrete Math., 3 (1990), pp. 216–240.

[6] R. Fröberg, *An Introduction to Gröbner Bases*, Pure Appl. Math. (NY), John Wiley & Sons, Ltd., Chichester, 1997.

[7] E. Hille, *Analytic Function Theory. Vol.* II, Introductions to Higher Mathematics, Ginn and Co., Boston, MA–New York–Toronto, ON, 1962.

[8] H. Hosoya, *Topological index as a common tool for quantum chemistry, statistical mechanics, and graph theory*, in Mathematical and Computational Concepts in Chemistry (Dubrovnik, 1985), Ellis Horwood Ser. Math. Appl., Horwood, Chichester, 1986, pp. 110–123.

[9] S. Janson, *The Wiener index of simply generated random trees*, Random Structures Algorithms, 22 (2003), pp. 337–358.

[10] M. Kendall and J. D. Gibbons, *Rank Correlation Methods*, 5th ed., A Charles Griffin Title, Edward Arnold, London, 1990.

[11] P. Kirschenhofer, H. Prodinger, and R. F. Tichy, *Fibonacci numbers of graphs.* II, Fibonacci Quart., 21 (1983), pp. 219–229.

[12] P. Kirschenhofer, H. Prodinger, and R. F. Tichy, *Fibonacci numbers of graphs.* III. *Planted plane trees*, in Fibonacci Numbers and Their Applications (Patras, 1984), Math. Appl. 28, D. Reidel, Dordrecht, 1986, pp. 105–120.

[13] M. Klazar, *Twelve countings with rooted plane trees*, European J. Combin., 18 (1997), pp. 195–210.

[14] E. L. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day Inc., San Francisco, CA, 1975.

[15] M. Lepović and I. Gutman, *A collective property of trees and chemical trees*, J. Chem. Inf. Comput. Sci., 38 (1998), pp. 823–826.

[16] X. Li, Z. Li, and L. Wang, *The inverse problems for some topological indices in combinatorial chemistry*, J. Comput. Biol., 10 (2003), pp. 47–55.

[17] A. Meir and J. W. Moon, *On the altitude of nodes in random trees*, Canad. J. Math., 30 (1978), pp. 997–1015.

[18] A. Meir and J. W. Moon, *On an asymptotic method in enumeration*, J. Combin. Theory Ser. A, 51 (1989), pp. 77–89.

[19] R. E. Merrifield and H. E. Simmons, *Topological Methods in Chemistry*, Wiley, New York, 1989.

[20] H. Prodinger and R. F. Tichy, *Fibonacci numbers of graphs*, Fibonacci Quart., 20 (1982), pp. 16–21.

[21] J. Riordan and N. J. A. Sloane, *The enumeration of rooted trees by total height*, J. Austral. Math. Soc., 10 (1969), pp. 278–282.

[22] L. A. Székely and H. Wang, *On subtrees of trees*, Adv. in Appl. Math., 34 (2005), pp. 138–155.

[23] N. Trinajstić, *Chemical Graph Theory*, CRC Press, Boca Raton, FL, 1992.

[24] S. Wagner, *Calculating the Correlation Coefficients of Graph-Theoretical Indices*, http://www.arxiv.org/abs/math.CO/0608753 (2006). Also available at http://finanz.math.tugraz.at/~wagner/Correlation, 2006.

[25] S. Wagner, *Subset counting in trees*, Ars Combin., to appear.

[26] H. Wiener, *Structural determination of paraffin boiling points*, J. Amer. Chem. Soc., 69 (1947), pp. 17–20.

SIAM J. Discrete Math.
Vol. 21, No. 1, pp. 47–56

THE SPECTRUM OF THE CORONA OF TWO GRAPHS*

S. BARIK†, S. PATI‡, AND B. K. SARMA†

Abstract. We consider only simple graphs. Given two graphs G with vertices $1, \dots, n$ and H, the corona $G \circ H$ is defined as the graph obtained by taking n copies of H and for each i inserting edges between the ith vertex of G and each vertex of the ith copy of H. For a connected graph G and any r-regular graph H we provide complete information about the spectrum of $G \circ H$ using the spectrum of G and spectrum of H. Complete information about the Laplacian spectrum of $G \circ H$ is also provided even when H is not regular. A graph G is said to have the property (R) if $\frac{1}{\lambda}$ is an eigenvalue of G whenever λ is an eigenvalue of G. Further, if λ and $\frac{1}{\lambda}$ have the same multiplicity, for each eigenvalue λ, then it is said to have the property (SR). We characterize all trees with property (SR) and show that such a tree is the corona product of some tree and an isolated vertex. We supply a family of bipartite graphs with property (R). As an application we construct infinitely many pairs of nonisomorphic graphs with the same spectrum and the same Laplacian spectrum. We prove some results about the eigenvector related to the second smallest eigenvalue of the Laplacian matrix of $G \circ H$ and give an application.

Key words. adjacency matrix, Laplacian matrix, property (R), property (SR), corona

AMS subject classifications. 05C05, 05C50, 15A18

DOI. 10.1137/050624029

1. Introduction.

Throughout this article we consider only simple graphs. Let $G = (V, E)$ be a graph with vertex set $V = \{1, 2, \dots, n\}$. The *adjacency matrix* of G, denoted by $A(G)$, is defined as $A(G) = [a_{ij}]_n$, where

$$a_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are adjacent in } G, \\ 0 & \text{otherwise.} \end{cases}$$

The *spectrum of* G is defined throughout as

$$\sigma(G) = (\lambda_1(G), \lambda_2(G), \dots, \lambda_n(G)),$$

where $\lambda_1(G) \leq \lambda_2(G) \leq \cdots \leq \lambda_n(G)$ are the eigenvalues of $A(G)$. The largest eigenvalue of $A(G)$ is called the *spectral radius* of G and is denoted by $\rho(G)$. If G is connected, then $A(G)$ is irreducible, thus the spectral radius of G is of multiplicity one and is afforded by a positive eigenvector called the *Perron vector*. A graph G is said to be singular if $A(G)$ is singular.

It is well known [4] that a graph G is bipartite if and only if the negative of each eigenvalue of G is also an eigenvalue of G. Let us say that *a graph G has property (R)*[1] if $\frac{1}{\lambda}$ is an eigenvalue of G whenever λ is an eigenvalue of G. Further, if λ and $\frac{1}{\lambda}$ have the same multiplicity for each eigenvalue λ, then we say that the *graph has property (SR)*.[2]

*Received by the editors February 8, 2005; accepted for publication (in revised form) May 23, 2006; published electronically January 12, 2007.
http://www.siam.org/journals/sidma/21-1/62402.html

†Department of Mathematics, IIT Guwahati, Guwahati-781039, India (sasmitab@iitg.ernet.in, bks@iitg.ernet.in). The first author's research was supported by CSIR.

‡Corresponding author. Department of Mathematics, IIT Guwahati, Guwahati-781039, India (pati@iitg.ernet.in).

[1] (R) for reciprocal.

[2] (SR) for strong reciprocal.

In section 2, we supply a class of graphs satisfying property (R), using the corona of a bipartite graph and a single vertex. We characterize the trees satisfying property (SR). We show that this is the class of trees on $2n$ vertices with n matchings which are leaves. We refer to such trees as *corona trees.* We supply suitable examples to show that a graph with property (R) is not necessarily the corona of two graphs and not necessarily bipartite.

The *Laplacian matrix* of G, denoted by $L(G)$ is defined as $D(G) - A(G)$, where $D(G)$ is the diagonal degree matrix of G. The *Laplacian spectrum* of G is defined as

$$S(G) = (\gamma_1(G), \gamma_2(G), \ldots, \gamma_n(G)),$$

where $\gamma_1(G) \leq \gamma_2(G) \leq \cdots \leq \gamma_n(G)$ are the eigenvalues of $L(G)$. For any graph G, $\gamma_1(G) = 0$ afforded by the all ones eigenvector $\mathbb{1}$. It is well known that $L(G)$ is a positive semidefinite matrix and there is extensive literature available on works related to Laplacian matrices. We refer the interested reader to a survey article [13] and the references therein to know more. Fiedler [5] showed that the second smallest eigenvalue of $L(G)$ is 0 if and only if the graph is disconnected. Thus the second smallest eigenvalue of $L(G)$ is popularly known as the *algebraic connectivity* of G and is denoted by $a(G)$. The eigenvectors corresponding to $a(G)$ are called *Fiedler vectors* of the graph G.

In section 3, we give a complete description of the spectrum of $G \circ H$ using the spectrum of G and the spectrum of H when H is r-regular. We also give a complete description of the Laplacian spectrum of $G \circ H$ (here H is not necessarily regular). As an application we show how to construct infinitely many pairs of nonisomorphic graphs which have the same spectrum and same Laplacian spectrum. We study the algebraic connectivity and the characteristic set of the corona of graphs and prove some structural results. An application is indicated.

The complete graph of order n is denoted by K_n and the *star graph* of order n is denoted by $K_{1,n-1}$. Let $R = [r_{ij}]$, S be matrices. Then the *Kronecker product* of R and S is defined to be the partitioned matrix $[r_{ij}S]$ and is denoted by $R \otimes S$. The vector with ith entry equal to one and all other entries zero is denoted by e_i.

Definition 1.1 (see [9]). *Let G_1 and G_2 be two graphs on disjoint sets of n and m vertices, respectively. The corona $G_1 \circ G_2$ of G_1 and G_2 is defined as the graph obtained by taking one copy of G_1 and n copies of G_2, and then joining the ith vertex of G_1 to every vertex in the ith copy of G_2.*

Note that the corona $G_1 \circ G_2$ has $n(m+1)$ vertices and $|E(G_1)| + n(|E(G_2)| + m)$ edges. There has been some research on the corona of two graphs; see, for example, [7].

Example 1.2. *Let $G_1 = C_4$, the cycle of order 4 and $G_2 = K_2$. The two different coronas $G_1 \circ G_2$ and $G_2 \circ G_1$ are shown in Figure 1.1.*

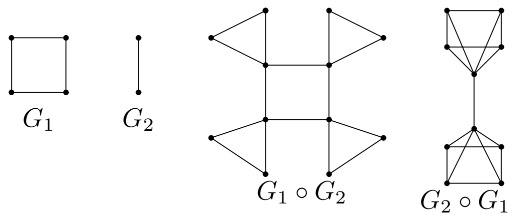


Fig. 1.1. *Coronas of two graphs.*

We close this section by citing some well known results which will be used later. The results can be found in [3, 4].

LEMMA 1.3. *Let $P(x) = x^n + C_1x^{n-1} + C_2x^{n-2} + \cdots + C_{n-2}x^2 + C_{n-1}x + C_n$ be the characteristic polynomial of a tree T on n vertices. Then $C_{2i+1} = 0$, and*

$$C_{2i} = (-1)^i(\textit{the number of pairwise disjoint edge subsets of size } i).$$

LEMMA 1.4. *Let T be a tree with vertex set $\{1, 2, \ldots, n\}$. If two or more pendant vertices have a common neighbor, then T is singular.*

LEMMA 1.5. *A graph G with diameter d has at least $d + 1$ distinct eigenvalues.*

2. Trees with property (SR).

In this section we first investigate graphs with property (R). The first examples are paths P_2 and P_4. For P_2 the eigenvalues are $1, -1$, where as for P_4 the eigenvalues are

$$\frac{1 \pm \sqrt{5}}{2}, \frac{-1 \pm \sqrt{5}}{2}.$$

A careful examination of P_4 leads us to the following result.

LEMMA 2.1. *Let G_1 be any graph and G be obtained by adding a new pendant to every vertex of G_1. Then λ is an eigenvalue of G if and only if $\frac{-1}{\lambda}$ is an eigenvalue of G. Further, if G_1 is bipartite, then G has property (R).*

Proof. Let G_1 be on n vertices. It is clear that $G = G_1 \circ K_1$. Thus

$$A(G) = \left[\begin{array}{c|c} A(G_1) & I_n \\ \hline I_n & \mathbf{0} \end{array}\right].$$

Let $\mu_1, \ldots, \mu_n$ be the eigenvalues of $A(G_1)$ corresponding to the eigenvectors $x_1, \ldots, x_n$, respectively, where the set $\{x_1, \ldots, x_n\}$ is orthonormal. Then the vectors

$$\left[\begin{array}{c} x_1 \\ \frac{2}{\mu_1+\sqrt{\mu_1^2+4}}x_1 \end{array}\right], \left[\begin{array}{c} x_1 \\ \frac{2}{\mu_1-\sqrt{\mu_1^2+4}}x_1 \end{array}\right], \ldots, \left[\begin{array}{c} x_n \\ \frac{2}{\mu_n+\sqrt{\mu_n^2+4}}x_n \end{array}\right], \left[\begin{array}{c} x_n \\ \frac{2}{\mu_n-\sqrt{\mu_n^2+4}}x_n \end{array}\right]$$

are all eigenvectors of $A(G)$ corresponding to the eigenvalues

$$\frac{\mu_1 + \sqrt{\mu_1^2 + 4}}{2}, \frac{\mu_1 - \sqrt{\mu_1^2 + 4}}{2}, \ldots, \frac{\mu_n + \sqrt{\mu_n^2 + 4}}{2}, \frac{\mu_n - \sqrt{\mu_n^2 + 4}}{2},$$

respectively.

We observe that $\frac{\mu_i + \sqrt{\mu_i^2 + 4}}{2}\frac{\mu_i - \sqrt{\mu_i^2 + 4}}{2} = -1$ and the first conclusion follows. Note that if G_1 is bipartite, then G is bipartite. Thus if $\lambda \in \sigma(G)$, then by the above $\frac{-1}{\lambda} \in \sigma(G)$ and as G is bipartite $\frac{1}{\lambda} \in \sigma(G)$. □

The following is an immediate corollary.

COROLLARY 2.2. *Let $G = G_1 \circ K_1$. Then*

(a) *G is nonsingular and the determinant of $A(G) = (-1)^n$, where n is the number of vertices in G_1.*

(b) *There are n positive and n negative eigenvalues of G. If λ_i are the positive eigenvalues of G, then $\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} \frac{1}{\lambda_i}$.*

(c) $\rho(G) = \frac{\rho(G_1) + \sqrt{\rho(G_1)^2 + 4}}{2}$, *where $\rho(H)$ is the spectral radius of a graph H.*

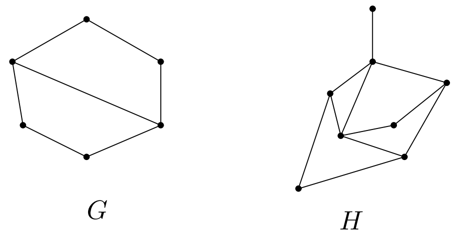


FIG. 2.1. *Graphs with property (R) which are not corona of two graphs.*

It is natural to ask whether the converse of Lemma 2.1 is true, that is, if G is any graph which has property (R), is it necessarily the corona of a bipartite graph and K_1? The answer is in the negative, in general, as can be seen from the following example.

EXAMPLE 2.3. *The graphs G, H in Figure 2.1 satisfy property (R). The eigenvalues of the graph on the left are $1, -1, \pm 1 + \sqrt{2}, \pm 1 - \sqrt{2}$. The eigenvalues of the graph on the right are*

$$1, 1, \frac{-3 \pm \sqrt{5}}{2}, \frac{1 + \sqrt{33} + \sqrt{18 + 2\sqrt{33}}}{4}, \frac{1 + \sqrt{33} - \sqrt{18 + 2\sqrt{33}}}{4},$$

$$\frac{1 - \sqrt{33} + \sqrt{18 - 2\sqrt{33}}}{4}, \frac{1 - \sqrt{33} - \sqrt{18 - 2\sqrt{33}}}{4}.$$

We can argue that G is not the corona of two graphs. Suppose that $G = G_1 \circ G_2$. Thus $6 = |G| = (|G_2| + 1)|G_1|$, where $|G|$ means the number of vertices of G. Note that $|G_1|$ cannot be 1 because in that case G would have a vertex of degree 5 and $|G_1| \neq 6$, otherwise G_2 has to be empty. If $|G_1| = 2$, then $|G_2| = 2$ and thus $G_1 \circ G_2$ cannot have a 6-cycle. If $|G_1| = 3$, then $|G_2| = 1$ in which case $G_1 \circ G_2$ should have 3 pendants. One can argue in a similar way that H is not a corona of two graphs.

Thus we have two immediate questions.

1. Characterize the trees with property (R).
2. Characterize the trees with property (SR).

In this section we supply an answer to question 2.

It turns out that any tree with property (SR) is of the form $T \circ K_1$, for some tree T. Before we prove that we need the following observation.

LEMMA 2.4. *Let G be a graph on n vertices with property (SR) and $P(x)$ be the characteristic polynomial of $A(G)$. Then $|C_r| = |C_{n-r}|$, where C_r is the coefficient of x^{n-r} in $P(x)$.*

Proof. Since G satisfies property (SR), G is nonsingular. Moreover $P(x)$ and $x^n P\left(\frac{1}{x}\right)$ have the same roots. Since $P(x)$ is monic and the leading coefficient of $x^n P\left(\frac{1}{x}\right)$ is ± 1, it follows that $P(x) = \pm\, x^n P\left(\frac{1}{x}\right)$ and the conclusion follows. □

The following is the main result of this section.

THEOREM 2.5. *Let T be a tree on n vertices. Then T has property (SR) if and only if $T = T_1 \circ K_1$, for some tree T_1.*

Proof. We prove the only if part here. Let T have property (SR). Then $n = 2k$, for some k. If $k = 1, 2$, then the only nonsingular trees of size $2k$ are the paths which are $K_1 \circ K_1$, $K_2 \circ K_1$. Assume that $k \geq 3$. Further, T has a perfect matching. Let

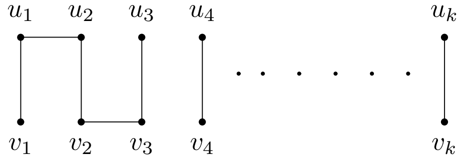


FIG. 2.2.

$\mathcal{M} = \{f_i = u_iv_i,\ i = 1, \ldots, k\}$ be those edges of T (see Figure 2.2). Note here that if we put back the remaining $k-1$ edges we get Figure 2.2.

We claim that for each edge f_i at least one of u_i, v_i is of degree 1 in T. Suppose this is not the case. Thus there is an edge, say f_2, such that u_2, v_2 both have degrees greater than 1, say u_1u_2, v_2v_3 are present.

Let

$$P(x) = x^{2k} + C_1x^{2k-1} + C_2x^{2k-2} + \cdots + C_{2k-2}x^2 + C_{2k-1}x + C_{2k}$$

be the characteristic polynomial of the tree T. By Lemma 2.4, $|C_2| = |C_{2k-2}|$.

But by Lemma 1.3, $|C_2| =$ the number of edges in $T = 2k-1$ and $|C_{2k-2}| =$ the number of pairwise disjoint edge subsets of size $k-1$. Hence the number of pairwise disjoint edge subsets of size $k-1$ is $2k-1$.

There are k pairwise edge disjoint subsets of size $k-1$ of T of the form

$$\{f_1, \ldots, f_k\} \setminus \{f_1\}, \{f_1, \ldots, f_k\} \setminus \{f_2\}, \ldots, \{f_1, \ldots, f_k\} \setminus \{f_k\}.$$

Any edge e of T which is not in $\mathcal{M}$ is incident to exactly two edges in $\mathcal{M}$, say f_i, f_j and gives us a pairwise edge disjoint subset of size $k-1$ of T of the form

$$\{e\} \cup \mathcal{M} \setminus \{f_i, f_j\}.$$

We will have $k-1$ such pairwise edge disjoint subsets of size $k-1$ of T.

Further, the set $\{u_1u_2, v_2v_3, f_4, \ldots f_k\}$ is also a pairwise edge disjoint subset of size $k-1$ of T. Thus the number of pairwise edge disjoint subsets of size $k-1$ of T exceeds $2k-1$, which is a contradiction and the claim is justified. Assume that the k pendants of T are $\{u_1, \ldots, u_k\}$ and let T_1 be the subtree of T induced by $\{v_1, \ldots, v_k\}$. Then $T = T_1 \circ K_1$ and the proof is complete. □

Let $\mathcal{F}$ be the class of all trees with property (R) and $\mathcal{S}$ be the class of all trees with property (SR). Then it is clear that $\mathcal{S} \subset \mathcal{F}$. It has been shown in [2] that these two classes are in fact the same.

The questions of characterizing all bipartite graphs/nonbipartite graphs with property (R)/(SR) remain open.

3. The corona $G_1 \circ G_2$. Throughout this section G_1 is assumed to be connected with $n > 1$ vertices and G_2 any graph on $m \geq 1$ vertices. In this section we give a complete description of the eigenvalues and the corresponding eigenvectors of the adjacency matrix of $G_1 \circ G_2$, where G_2 is regular and also give a complete description of the eigenvalues and the corresponding eigenvectors of the Laplacian matrix of $G_1 \circ G_2$, for any graph G_2. As a consequence we obtain some interesting results.

Let G_1 be a graph with vertex set $V = \{1, 2, \ldots, n\}$ and G_2 be a regular graph of order m and of regularity r (say), $r \leq m-1$. Let $G = G_1 \circ G_2$. Thus

$$|V(G)| = (m+1)n, \text{ and } |E(G)| = |E(G_1)| + mn + \frac{nmr}{2},$$

and the adjacency matrix of G is

$$A(G) = \left[\begin{array}{c|ccc} A(G_1) & I_n & \dots & I_n \\ \hline I_n & & & \\ \vdots & & A(G_2) \otimes I_n & \\ I_n & & & \end{array}\right],$$

where $A(G_1)$ and $A(G_2)$ are the adjacency matrices of the graphs G_1 and G_2, respectively. The following gives a complete characterization of the eigenvalues and the eigenvectors of $G_1 \circ G_2$.

THEOREM 3.1. *Let G_1 be any graph, G_2 be an r-regular graph, and $G = G_1 \circ G_2$. Let $\sigma(G_1) = (\mu_1, \mu_2, \dots, \mu_n)$ and $\sigma(G_2) = (\eta_1, \eta_2, \dots, \eta_m = r)$. Then*

(a) $\dfrac{\mu_i + r \pm \sqrt{(r-\mu_i)^2 + 4m}}{2} \in \sigma(G)$ *with multiplicity* 1 *for* $i = 1, \dots, n$ *and*

(b) $\eta_j \in \sigma(G)$ *with multiplicity* n *for* $j = 1, \dots, m-1$.

Thus $\rho(G) = \dfrac{\rho(G_1) + r + \sqrt{(r - \rho(G_1))^2 + 4m}}{2}$.

Proof. Let $X_1, \dots, X_n$ be the orthonormal eigenvectors of $A(G_1)$ corresponding to the eigenvalues $\mu_1, \mu_2, \dots, \mu_n$, respectively. For $i = 1, \dots, n$, let

$$\lambda_i = \frac{\mu_i + r + \sqrt{(r-\mu_i)^2 + 4m}}{2}, \ \hat{\lambda}_i = \frac{\mu_i + r - \sqrt{(r-\mu_i)^2 + 4m}}{2}.$$

Note that $\dfrac{\mu_i + r \pm \sqrt{(r-\mu_i)^2 + 4m}}{2} = r$ implies $m = 0$, so that $\lambda_i, \hat{\lambda}_i$ are never r.

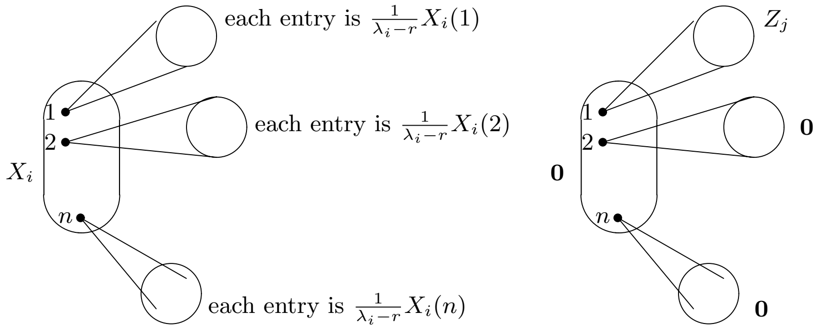

each entry is $\frac{1}{\lambda_i - r} X_i(1)$
1
2
each entry is $\frac{1}{\lambda_i - r} X_i(2)$
X_i
n
each entry is $\frac{1}{\lambda_i - r} X_i(n)$
Z_j
1
2
$\mathbf{0}$
$\mathbf{0}$
n
$\mathbf{0}$

FIG. 3.1. *Left: eigenvector corresponding to λ_i. Right: one of the eigenvectors corresponding to η_j.*

Observe that $\lambda_i, \hat{\lambda}_i$ are eigenvalues of $A(G)$ corresponding to the eigenvectors

$$\begin{pmatrix} X_i \\ \frac{1}{\lambda_i - r} X_i \\ \vdots \\ \frac{1}{\lambda_i - r} X_i \end{pmatrix}, \begin{pmatrix} X_i \\ \frac{1}{\hat{\lambda}_i - r} X_i \\ \vdots \\ \frac{1}{\hat{\lambda}_i - r} X_i \end{pmatrix},$$

respectively (see Figure 3.1, picture on left).

Further, for $1 \le j \le m-1$, let Z_j be the eigenvector corresponding to the eigenvalue η_j of $A(G_2)$. Then for $i = 1, \dots, n$ we have (see Figure 3.1, picture on right, for $i = 1$)

$$A(G)\begin{pmatrix} \mathbf{0} \\ Z_j \otimes e_i \end{pmatrix} = \eta_j \begin{pmatrix} \mathbf{0} \\ Z_j \otimes e_i \end{pmatrix}.$$

In the previous equation we use that G_2 is r-regular and hence $Z_j \perp \mathbb{1}$, for $j = 1, 2, \dots, m-1$. Hence the proof. □

Next we talk about the Laplacian matrix of $G_1 \circ G_2$. Let $L(G_1)$ and $L(G_2)$ be the Laplacian matrices of the graphs G_1 and G_2, respectively. Thus the Laplacian matrix of G is

$$L(G) = \left[\begin{array}{c|ccc} L(G_1) + mI_n & -I_n & \dots & -I_n \\ \hline -I_n & & & \\ \vdots & & (L(G_2) + I_m) \otimes I_n & \\ -I_n & & & \end{array}\right].$$

THEOREM 3.2. *Let G_1, G_2 be any graphs, not necessarily regular and $G = G_1 \circ G_2$. Let $S(G_1) = (0 = \nu_1, \nu_2, \dots, \nu_n)$ and $S(G_2) = (0 = \delta_1, \delta_2, \dots, \delta_m)$. Then*

(a) $\dfrac{\nu_i + m + 1 \pm \sqrt{(m+1)^2 - 4\nu_i}}{2} \in S(G)$ *with multiplicity* 1 *for* $i = 1, \dots, n$ *and*

(b) $\delta_j + 1 \in S(G)$ *with multiplicity* n *for* $j = 2, \dots, m$.

Thus

(i) $1 \notin S(G)$ *if and only if* G_2 *is connected.*

(ii) $m + 1 \in S(G)$ *always.*

(iii) $a(G) = \dfrac{a(G_1) + m + 1 - \sqrt{(a(G_1) + m + 1)^2 - 4a(G_1)}}{2} < 1.$

Proof. Suppose that $\mathbb{1} = Y_1, Y_2, \dots, Y_n$, are the eigenvectors of $L(G_1)$ corresponding to the eigenvalues $0 = \nu_1, \nu_2, \dots, \nu_n$, respectively. For $i = 1, \dots, n$, let

$$\gamma_i = \frac{\nu_i + m + 1 + \sqrt{(\nu_i + m + 1)^2 - 4\nu_i}}{2} = \frac{\nu_i + m + 1 + \sqrt{(\nu_i + m - 1)^2 + 4m}}{2},$$

$$\hat{\gamma}_i = \frac{\nu_i + m + 1 - \sqrt{(\nu_i + m + 1)^2 - 4\nu_i}}{2} = \frac{\nu_i + m + 1 - \sqrt{(\nu_i + m - 1)^2 + 4m}}{2}.$$

Notice that $\dfrac{\nu_i + m + 1 \pm \sqrt{(\nu_i + m - 1)^2 + 4m}}{2} = 1$ implies $m = 0$, so that $\gamma_i, \hat{\gamma}_i$ are never 1.

Observe that γ_i, and $\hat{\gamma}_i$ are eigenvalues of $L(G)$ afforded by the eigenvectors

$$\begin{pmatrix} Y_i \\ \frac{1}{1-\gamma_i} Y_i \\ \vdots \\ \frac{1}{1-\gamma_i} Y_i \end{pmatrix}, \begin{pmatrix} Y_i \\ \frac{1}{1-\hat{\gamma}_i} Y_i \\ \vdots \\ \frac{1}{1-\hat{\gamma}_i} Y_i \end{pmatrix},$$

respectively.

Also if the eigenvalues $\delta_1 (= 0), \delta_2, \dots, \delta_{m-1}, \delta_m$ of $L(G_2)$ are afforded by the eigenvectors $Z_1, Z_2, \dots, Z_m$, respectively, then for $j = 2, \dots, m$,

$$\begin{pmatrix} \mathbf{0} \\ Z_j \otimes e_1 \end{pmatrix}, \begin{pmatrix} \mathbf{0} \\ Z_j \otimes e_2 \end{pmatrix}, \dots, \begin{pmatrix} \mathbf{0} \\ Z_j \otimes e_n \end{pmatrix}$$

are the n eigenvectors corresponding to the eigenvalue $\delta_j + 1$ of $L(G)$. Hence the first statement follows. Items (i), (ii) are routine. Item (iii) follows from the fact that $\frac{\nu_i + m + 1 - \sqrt{(\nu_i + m - 1)^2 + 4m}}{2} < 1$ and if $\nu_i \leq \nu_j$, then

$$\frac{\nu_i + m + 1 - \sqrt{(\nu_i + m - 1)^2 + 4m}}{2} \leq \frac{\nu_j + m + 1 - \sqrt{(\nu_j + m - 1)^2 + 4m}}{2}.$$

Hence the proof. □

In order to show an application of Theorem 3.2 we need the following setup. Let G be a connected graph. Define a relation R on the edge set as: $e_1 R e_2$ if and only if either $e_1 = e_2$ or there is a simple cycle containing both of them. Then R is an equivalence relation. Let $E_1 \bigcup E_2 \bigcup \cdots \bigcup E_k$ be the decomposition of the edge set into equivalence classes. The subgraphs G_i, $i = 1, \ldots, k$ of G consisting of all edges in E_i and all vertices adjacent to them is called a *block* of G. A vertex v is called a *point of articulation* if v is common to more than one block. Let Y be a Fiedler vector of G. A vertex v of G is called a *characteristic vertex* of G if $Y(v) = 0$ and if there is a vertex w, adjacent to v, such that $Y(w) \neq 0$. An edge e with end vertices u, w is called a *characteristic edge* if $Y(u)Y(w) < 0$. By $C(G, Y)$ we denote the characteristic set of G which is defined as the collection of all characteristic vertices and characteristic edges of G (keeping the notations from [1]). The following is essentially contained in [6].

Proposition 3.3. *Let G be a connected graph and Y a Fiedler vector. Then exactly one of the following holds.*
Case 1. *$C(G, Y) = \{v\}$, where v is a point of articulation.*
Case 2. *Or Case* 1 *does not hold and there is a unique block B (called characteristic block) of G which contains all the characteristic vertices and edges.*

The following was shown in [12].

Proposition 3.4. *Let G be connected. If Case* 1 *of Proposition* 3.3 *holds, then for any Fiedler vector Z of G, $C(G, Z) = \{v\}$. If Case* 2 *holds, then for any Fiedler vector z of G the characteristic block of G is B.*

The following is an easy consequence of Theorem 3.2.

Corollary 3.5. *Let G_1 be a graph with vertex set $V = \{1, 2, \ldots, n\}$ and G_2 be any graph of order m and $G = G_1 \circ G_2$. Then exactly one of the following holds.*
Case 1. *For some Fiedler vector Y of G_1, $C(G_1, Y) = \{v\}$. Then for each Fiedler vector Z of G we have $C(G, Z) = \{v\}$.*
Case 2. *For some Fiedler vector Y of G_1 there is a unique characteristic block B. Then for each Fiedler vector Z of G the characteristic block is also B.*

Proof. We know that

$$a(G) = \frac{a(G_1) + m + 1 - \sqrt{(a(G_1) + m + 1)^2 - 4a(G_1)}}{2},$$

and the vector

$$\begin{pmatrix} Y \\ \frac{1}{1-a(G)} Y \\ \vdots \\ \frac{1}{1-a(G)} Y \end{pmatrix}$$

is a Fiedler vector of G, where Y is a Fiedler vector of the graph G_1. Hence the proof. □

A graph T which satisfies Case 1 of Proposition 3.3 is called a Type I graph. Constructions of an infinite class of Type I trees with nonisomorphic Perron branches has been discussed in [11]. The following result which is immediate from previous results, helps in the construction of Type I graphs with nonisomorphic Perron branches.

COROLLARY 3.6. *Let $G_1 = T$ be a tree and G_2 be any graph. Then the characteristic set $C(G, Y)$ of $G = T \circ G_2$ (with respect to any Fiedler vector Y) is completely determined by the nature of T. The set $C(G, Y)$ always has only one element, either a vertex or an edge. Further, $C(G, Y) = C(G, Z) = C(T, X)$, where Z, X are any Fiedler vectors of G, T, respectively.*

In particular, if T is Type I *with characteristic vertex v, then G is Type* I *with characteristic vertex v.*

In view of this result, to construct Type I graphs with nonisomorphic Perron branches all we need is to take a Type I tree on more than 2 vertices, say T and any graph H. Then $G = T \circ H$ is an example; note that $G \circ H$ is also an example. In particular, considering the tree T in Figure 3.2, which is known to be Type I with nonisomorphic Perron branches (see [8]), and taking H to be an isolated vertex, we see that $T, T \circ H, (T \circ H) \circ H, \ldots$ gives us a different infinite class of Type I trees with nonisomorphic Perron branches.

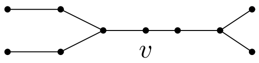


FIG. 3.2. *A Type I tree with characteristic vertex v.*

Below we discuss another application of the results in this article. Two graphs G and H are called *cospectral* if the spectrum of $A(H)$ and $A(G)$ are the same. Two graphs are called *Laplacian cospectral* if $L(G)$ and $L(H)$ have the same spectrum. This topic has been an area of interest for many researchers. We refer the reader to [14] and the references therein to learn more. Our aim here is to construct infinite pairs of nonisomorphic graphs G, H which are cospectral and Laplacian cospectral.

Let G, H be two nonisomorphic cospectral and Laplacian cospectral graphs (such a pair can be found in [14]). Let B be the graph of an isolated vertex. Let $G_1 = G \circ B$, $H_1 = H \circ B$, and for $i = 2, \ldots$ define $G_i = G_{i-1} \circ B$, $H_i = H_{i-1} \circ B$. By Theorems 3.1 and 3.2, we see that the spectrum and the Laplacian spectrum of G_1, H_1 is completely determined by the spectrum of G, H and they are the same. Use of induction leads us to the following conclusion.

COROLLARY 3.7. *Let G_i, H_i be defined as above, for $i \in \mathbb{N}$. Then for each i the graphs G_i, H_i are nonisomorphic cospectral and Laplacian cospectral nonregular graphs.*

Acknowledgment. We sincerely thank the referee for many valuable suggestions which improved the presentation of the article.

REFERENCES

[1] R. B. BAPAT AND S. PATI, *Algebraic connectivity and the characteristic set of a graph*, Linear Multilinear Algebra, 45 (1998), pp. 247–273.

[2] S. BARIK, M. NEUMANN, AND S. PATI, *On nonsingular trees and a reciprocal eigenvalue property*, Linear Multilinear Algebra, 54 (2006), pp. 453–465.

[3] R. A. BRUALDI AND H. J. RYSER, *Combinatorial Matrix Theory*, Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge, UK, 1991.

[4] D. M. Cvetkovic, M. Doob, and H. Sachs, *Spectra of Graphs*, Academic Press, New York, 1980.

[5] M. Fiedler, *Algebraic connectivity of graphs*, Czechoslovak Math. J., 23 (1973), pp. 298–305.

[6] M. Fiedler, *A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory*, Czechoslovak Math. J., 23 (1975), pp. 619–633.

[7] R. Frucht and F. Harary, *On the corona of two graphs*, Aequationes Math., 4 (1970), pp. 322–325.

[8] R. Grone and R. Merris, *Algebraic connectivity of trees*, Czechoslovak Math. J., 37 (1987), pp. 660–670.

[9] F. Harary, *Graph Theory*, Addison-Wesley, Reading, MA, 1969.

[10] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, MA, 1987.

[11] S. Kirkland, *Constructions for type* I *trees with nonisomorphic Perron branches*, Czechoslovak Math. J., 49 (1999), pp. 617–632.

[12] S. Kirkland and S. Fallat, *Perron components and algebraic connectivity for weighted graphs*, Linear Multilinear Algebra, 44 (1998), pp. 131–148.

[13] R. Merris, *Laplacian matrices of graphs: A survey*, Linear Algebra Appl., 197/198 (1994), pp. 143–176.

[14] E. R. van Dam and W. H. Haemers, *Which graphs are determined by their spectrum?*, Linear Algebra Appl., 373 (2003), pp. 241–272.

SIAM J. DISCRETE MATH.
Vol. 21, No. 1, pp. 57–65

SHARP THRESHOLD FOR HAMILTONICITY OF RANDOM GEOMETRIC GRAPHS*

JOSEP DÍAZ[†], DIETER MITSCHE[‡], AND XAVIER PÉREZ[†]

Abstract. We show for an arbitrary ℓ_p norm that the property that a random geometric graph $\mathcal{G}(n,r)$ contains a Hamiltonian cycle exhibits a sharp threshold at $r = r(n) = \sqrt{\frac{\log n}{\alpha_p n}}$, where α_p is the area of the unit disk in the ℓ_p norm. The proof is constructive and yields a linear time algorithm for finding a Hamiltonian cycle of $\mathcal{G}(n,r)$ asymptotically almost surely, provided $r = r(n) \geq \sqrt{\frac{\log n}{(\alpha_p - \epsilon)n}}$ for some fixed $\epsilon > 0$.

Key words. random geometric graphs, Hamilton cycles

AMS subject classifications. Primary, 05C80, 60D05; Secondary, 05C85

DOI. 10.1137/060665300

1. Introduction.

Given a graph G on n vertices, a *Hamiltonian cycle* is a simple cycle that visits each vertex of G exactly once. A graph is said to be *Hamiltonian* if it contains a Hamiltonian cycle. The problem of deciding if a given graph is Hamiltonian is known to be NP-complete [5]. Two known facts for the Hamiltonicity of random graphs are that almost all d-regular graphs ($d \geq 3$) are Hamiltonian [14], and that in the $\mathcal{G}_{n,p}$ model if $p(n) = (\log n + \log\log n + \omega(n))/n$, then a.a.s. $\mathcal{G}_{n,p}$ is Hamiltonian [9] (see also Chapter 8 of [3]). Throughout this paper, "a.a.s." means *asymptotically almost surely*, that is, with probability tending to 1 as n goes to ∞.

A *random geometric graph* $\mathcal{G}(n,r)$ [6] is a graph resulting from placing a set of n vertices uniformly at random and independently on the unit square $[0,1]^2$, and connecting two vertices if and only if their *distance* is at most the given radius r, the distance depending on the type of metric being used. The two metrics more often used are the ℓ_2 and the ℓ_∞ norms. In recent times, random geometric graphs have received quite a bit of attention in the modeling of sensor networks, and in general ad hoc wireless networks (see, e.g., [1]).

Random geometric graphs are the randomized version of unit disk graphs. An undirected graph is a *unit disk graph* if its vertices can be put into one-to-one correspondence with circles of equal radius in the plane in such a way that two vertices are joined by an edge if and only if their corresponding circles intersect. W.l.o.g. it can be assumed that the radius of the circles is 1 [4]. The problem of deciding if a given unit disk graph is Hamiltonian is known to be NP-complete [8].

Many properties of random geometric graphs have been intensively studied, from both the theoretical and the empirical points of view. It is known (see [7]) that all monotone properties of $\mathcal{G}(n,r)$ exhibit a sharp threshold. For the present paper, the

*Received by the editors July 17, 2006; accepted for publication (in revised form) September 8, 2006; published electronically January 26, 2007. This research was partially supported by the 6th Framework under contract 001907 (DELIS) and the Spanish CYCIT: TIN2004-07925-C03-01 (GRAMMARS). The first author was also supported by La distinció per a la promació de la recerca de la Generalitat de Catalunya, 2002.

http://www.siam.org/journals/sidma/21-1/66530.html

[†]Llenguatges i Sistemes Informàtics, UPC, 08034 Barcelona, Spain (diaz@lsi.upc.edu, xperez@lsi.upc.edu).

[‡]Institut für Theoretische Informatik, ETH Zürich, 8092 Zürich, Switzerland (dmitsche@inf.ethz.ch).

most relevant result on random geometric graphs is the connectivity threshold: in [10] it is proved that $r = r(n) = \sqrt{\log n/(\pi n)}$ is the sharp threshold for the connectivity of $\mathcal{G}(n, r)$ in the ℓ_2 norm. For the ℓ_∞ norm, the sharp threshold for connectivity occurs at $r = r(n) = \sqrt{\log n/(4n)}$ (see [2]). In general, for an arbitrary ℓ_p norm, for some fixed p, $1 \leq p \leq \infty$, the sharp threshold is known to be $r = r(n) = \sqrt{\log n/(\alpha_p n)}$, where α_p is the area of the unit disk in the ℓ_p norm (see [11] and [12]).

A natural issue to study is the existence of Hamiltonian cycles in $\mathcal{G}(n, r)$. Penrose in his book [12] poses as an open problem whether, exactly at the point where $\mathcal{G}(n, r)$ becomes 2-connected, the graph also becomes Hamiltonian a.a.s., Petit in [13] proved that for $r = \omega(\sqrt{\log n/n})$, $\mathcal{G}(n, r)$ is Hamiltonian a.a.s., and he also gave a distributed algorithm for finding a Hamiltonian cycle in $\mathcal{G}(n, r)$ with his choice of radius. In the present paper, we find the sharp threshold for this property in any ℓ_p metric. In fact, let p $(1 \leq p \leq \infty)$ be arbitrary but fixed throughout the paper, and let $\mathcal{G} = \mathcal{G}(n, r)$ be a random geometric graph with respect to ℓ_p. We first show the following

Theorem 1. *The property that a random geometric graph $\mathcal{G} = \mathcal{G}(n, r)$ contains a Hamiltonian cycle exhibits a sharp threshold at $r = \sqrt{\frac{\log n}{\alpha_p n}}$, where α_p is the area of the unit disk in the ℓ_p norm.*

More precisely, for any $\epsilon > 0$,

- *if $r = \sqrt{\frac{\log n}{(\alpha_p+\epsilon)n}}$, then a.a.s. $\mathcal{G}$ contains no Hamiltonian cycle;*
- *if $r = \sqrt{\frac{\log n}{(\alpha_p-\epsilon)n}}$, then a.a.s. $\mathcal{G}$ contains a Hamiltonian cycle.*

As a corollary of the proof, we describe a linear time algorithm that finds a Hamiltonian cycle in $\mathcal{G}(n, r)$ a.a.s., provided that $r \geq \sqrt{\frac{\log n}{(\alpha_p-\epsilon)n}}$ for some fixed $\epsilon > 0$.

2. Proof of Theorem 1. To prove Theorem 1, note that the lower bound of the threshold is trivial. In fact, if $r = \sqrt{\frac{\log n}{(\alpha_p+\epsilon)n}}$, then a.a.s. $\mathcal{G}$ is disconnected [11], and hence it cannot contain any Hamiltonian cycle. To simplify the proof of the upper bound, we need some auxiliary definitions and lemmas. In the remainder of the section, we assume that $r = \sqrt{\frac{\log n}{(\alpha_p-\epsilon)n}}$ for some fixed $\epsilon > 0$, and we show that a.a.s. $\mathcal{G}$ contains a Hamiltonian cycle.

Let us take $y = \left\lceil \frac{2}{r} \right\rceil^{-1}$. Intuitively, y is close to $r/2$ but slightly smaller. We divide $[0, 1]^2$ into squares of side length y. Call this the *initial tessellation* of $[0, 1]^2$. Two different squares R and S are defined to be *friends* if they are either adjacent (i.e., they share at least one corner) or there exists at least one other square T adjacent to both R and S. Thus, each square has at most 24 friends. Then we create a second and finer tessellation of $[0, 1]^2$ by dividing each square into k^2 new squares of side length $y/k \sim r/(2k)$, for some large enough but fixed $k = k(\epsilon) \in \mathbb{N}$. We call this the *fine tessellation* of $[0, 1]^2$, and we refer to these smaller squares as *cells.* We note that the total numbers of squares and cells are both $\Theta(1/r^2)$. Note that with probability 1, for every fixed n, any vertex will be contained in exactly one cell (and exactly one square). In the following we always assume this.

We say that a cell is *dense* if it contains at least 48 vertices of $\mathcal{G}$. If the cell contains at least one vertex but less than 48 vertices, we say the cell is *sparse.* If the cell contains no vertex, the cell is *empty.* Furthermore we define an *animal* to be a union of cells which is topologically connected. The *size* of an animal is the number of different cells it contains. In particular, the squares of the initial tessellation of $[0, 1]^2$ are animals of size k^2. An animal is called *dense* if it contains at least one dense cell. If an animal contains no dense cell, but it contains at least one vertex of $\mathcal{G}$, it is called *sparse.*

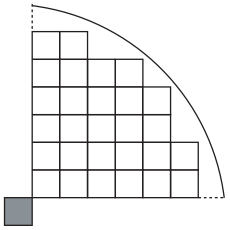


FIG. 1. *Set of cells close to, above, and to the right of the shaded cell.*

From hereinafter, all distances in $[0,1]^2$ will be taken in the ℓ_p metric. As usual, the distance between two sets of points P_1 and P_2 in $[0,1]^2$ is the infimum of the distances between any pair of points in P_1 and P_2. Two cells c_1 and c_2 are said to be *close* to each other if

$$\sup_{p_1 \in c_1, p_2 \in c_2} \{\text{distance}(p_1, p_2)\} \leq r.$$

For an arbitrary cell c at distance at least r from the boundary of $[0,1]^2$, let $K = K(n)$ be the number of cells which are close to c and also above and to the right of c (see Figure 1). Obviously, K does not depend on the particular cell we chose.

LEMMA 1. *For any $\eta > 0$, we can choose k sufficiently large such that $K > (\alpha_p - \eta)k^2$ for n large enough.*

Proof. Let c be a cell at distance at least r from the boundary of $[0,1]^2$. Call A the union of the cells which are close to c and also above and to the right of c. Let p be the top right corner point of c. Define the set

$$B = \{q \in [0,1]^2 \cap R : \text{distance}(p,q) \leq r - 4y/k\},$$

where R is the set of points which are above and to the right of p. Observe that $B \subseteq A$. Moreover, if k is chosen large enough, the area of B is at least $\frac{1}{4}(\alpha_p - \eta)r^2$. Thus, A contains at least $\frac{1}{4}(\alpha_p - \eta)r^2/(\frac{y}{k})^2 > (\alpha_p - \eta)k^2$ cells. □

LEMMA 2. *The following statements are true a.a.s:*

(i) *All animals of size* $4K$ *are dense.*

(ii) *All animals of size* $2K$ *which touch any of the four sides of* $[0,1]^2$ *are dense.*

(iii) *All cells at distance less than* $4y$ *from two sides of* $[0,1]^2$ *are dense.*

Proof. Let $0 < \delta < \epsilon$. Taking into account that the side length of each cell is $(\frac{y}{k}) \geq \frac{1}{2k}\sqrt{\frac{\log n}{(\alpha_p - \delta)n}}$ (but also $(\frac{y}{k}) \leq c\sqrt{\log n/n}$ for some $c > 0$), the probability that any given cell is not dense (i.e., it contains at most 47 vertices) is

$$\sum_{i=0}^{47} \binom{n}{i} \left(\frac{y^2}{k^2}\right)^i \left(1 - \frac{y^2}{k^2}\right)^{n-i} = \Theta(1)n^{47} \left(\frac{y^2}{k^2}\right)^{47} \left(1 - \frac{y^2}{k^2}\right)^n,$$

since the weight of this sum is concentrated in the last term. Then, plugging in the bounds for y/k, we get that the probability above is

$$O(1)\left(\frac{ny^2}{k^2}\right)^{47} e^{-y^2 n/k^2} = O(1)(\log n)^{47} n^{-\frac{1}{4k^2(\alpha_p - \delta)}}.$$

For each one of the cells of a given animal, we can consider the event that this particular cell is not dense. Notice that these events are negatively correlated (i.e., the probability that any particular cell is not dense conditional upon having some other cells with at most 47 vertices is not greater than the unconditional probability). Thus, the probability that a given animal of size $4K$ contains no dense cell is at most

$$\left(O(1)(\log n)^{47} n^{-\frac{1}{4k^2(\alpha_p-\delta)}}\right)^{4K} = O(1)(\log n)^{C} n^{-\frac{K}{k^2(\alpha_p-\delta)}},$$

for some constant C. Let $\rho = \frac{K}{k^2(\alpha_p-\delta)}$. From Lemma 1 applied with any $0 < \eta < \delta$, by choosing k sufficiently large, we can guarantee that $\rho > 1$. Now note that the number of animals of size $4K$ is $O(1/r^2)$ since for each fixed shape of an animal there are $O(1/r^2)$ many choices and there is only a constant number of shapes. Thus, by taking a union bound over all animals and plugging in the value of r, we get that the probability of having an animal without any dense cell is

$$O(1)(\log n)^{C-1}/n^{\rho-1} = o(1),$$

and (i) holds.

An analogous argument shows that any given animal of size $2K$ is not dense with probability

$$O(1)(\log n)^{C/2} n^{-\rho/2}.$$

Observe that there exist only $O(1/r)$ animals touching any of the four sides of $[0,1]^2$. Hence, the probability that one of these is not dense is

$$O(1)(\log n)^{(C-1)/2}/n^{(\rho-1)/2} = o(1),$$

and (ii) is proved.

To prove (iii), we simply recall that the probability that a given cell is not dense is $o(1)$. By taking a union bound, the same argument holds for a constant number of cells. $\square$

LEMMA 3. *A.a.s., for any cell c_1, there exists a cell c_2 which is dense and close to c_1.*

Proof. Let S be the square of the initial tessellation of $[0,1]^2$, where c_1 is contained, and let A be the animal containing all the cells which are close to c_1 but different from c_1. Suppose that S is at distance at least $2y$ from all sides of $[0,1]^2$. Then A has size greater than $4K$, and it must contain some dense cell by Lemma 2(i) a.a.s.

Otherwise, suppose that S is at distance less than $2y$ from just one side of $[0,1]^2$. Then, A has size greater than $2K$ and it touches one side of $[0,1]^2$, and thus it must contain some dense cell by Lemma 2(ii) a.a.s.

Finally, if S is at distance less than $2y$ from two sides of $[0,1]^2$, then all cells in that square must be dense by Lemma 2(iii) a.a.s. $\square$

We now consider the following auxiliary graph $\mathcal{G}'$: The vertices of $\mathcal{G}'$ are all those squares belonging to the initial tessellation of $[0,1]^2$ which are dense, and there is an edge between two dense squares R and S if they are friends and there exist cells $c_1 \subset R$ and $c_2 \subset S$ which are dense and close to each other. We observe that the maximal degree of $\mathcal{G}'$ is 24.

LEMMA 4. *A.a.s., $\mathcal{G}'$ is connected.*

Proof. Suppose for contradiction that $\mathcal{G}'$ contains at least two connected components C_1 and C_2. We denote by D the union of all dense cells which are contained in some vertex (i.e., dense square) of C_1, and let $H \supseteq D$ be the union of all cells which are close to some cell contained in D. Note that H is topologically connected, and let the closed curve γ be the outer boundary of H with respect to $\mathbb{R}^2$. Each connected part obtained by removing from γ the intersection with the sides of $[0,1]^2$ is called a *piece* of γ. Define by E the union of all cells in H but not in D. In general, E might have several connected components (animals). Moreover, all cells in E must be not dense, by construction. Note that any cell in D cannot touch any piece of γ. Hence, each piece of γ is touched by exactly one connected component $A \subseteq E$. Observe that, if γ touches some side of $[0,1]^2$, then all connected components of E touching some piece of γ must also touch some side of $[0,1]^2$.

Given any of the four sides s of $[0,1]^2$, the distance between s and C_1 is understood to be the distance between s and the dense square of C_1 which has the smallest distance to s. We now distinguish between a few cases depending on whether C_1 is at distance less than $2y$ from one (or more) side(s) of $[0,1]^2$ or not.

Case 1. C_1 is at distance at least $2y$ from any side of $[0,1]^2$.

In this case, let A be the only connected component of E which touches γ. Consider the uppermost dense cell $c \subset D$ (if there are several ones, choose an arbitrary one) and the lowermost dense cell $d \subset D$ (possibly equal to c). Then all cells which are close to c and above c and all cells which are close to d and below d belong to A. Since there are at least as many as $4K$ of these, we have an animal A of size at least $4K$ without any dense cell, which by Lemma 2(i) does not happen a.a.s.

Case 2. C_1 is at distance less than $2y$ from exactly one side of $[0,1]^2$.

W.l.o.g. we can assume that C_1 is at distance less than $2y$ from the bottom side of $[0,1]^2$. Consider the uppermost dense cell $c \subset D$ (if there are several, choose an arbitrary one). Let A be the connected component of E which contains all cells which are close to c and above c. Note that there are at least as many as $2K$ of these cells. Moreover, A touches one of the pieces of γ. Hence, we have an animal A of size at least $2K$ without any dense cell and that touches some side of $[0,1]^2$. By Lemma 2(ii) this does not happen a.a.s.

Case 3. C_1 is at distance less than $2y$ from two opposite sides of $[0,1]^2$.

W.l.o.g. we can assume that C_1 is at distance less than $2y$ from the top and bottom sides of $[0,1]^2$. From among all cells contained in squares of C_1 that are at distance less than $4y$ from the top side of $[0,1]^2$, consider the rightmost dense cell c. If c is at distance less than $2y$ from that side, consider all K cells which are close to c and below and to the right of c. Otherwise, if c is at distance at least $2y$ from that side, consider all K cells which are close to c and above and to the right of c. Let A be the connected component of E containing these cells. Similarly, from among all cells contained in squares of C_1 that are at distance less than $4y$ from the bottom side of $[0,1]^2$, consider the rightmost dense cell d. Again, if d is at distance less than $2y$ from that side, consider all K cells which are close to d and above and to the right of d. Otherwise, if d is at distance at least $2y$ from that side, consider all K cells which are close to d and below and to the right of d. Thus, in either case, we obtain K cells pairwise different from the K previously described ones, and let A' be the connected component containing them. A and A' must be the same, since they touch the same piece of γ. Hence, we have an animal A of size at least $2K$ touching at least one side of $[0,1]^2$ and without any dense cell. By Lemma 2(ii) this does not happen a.a.s.

Case 4. C_1 is at distance less than $2y$ from one vertical side and one horizontal side of $[0,1]^2$.

W.l.o.g. we can assume that C_1 is at distance less than $2y$ from the left and top sides of $[0,1]^2$. From among all cells contained in squares of C_1 that are at distance less than $4y$ from the top side of $[0,1]^2$, consider the rightmost dense cell c. If c is at distance less than $2y$ from that side, consider all K cells which are close to c and below and to the right of c. Otherwise, if c is at distance at least $2y$ from that side, consider all K cells which are close to c and above and to the right of c. Let A be the connected component of E containing all these K cells. By construction, all these K cells are at distance less than $4y$ from the top side of $[0, 1]^2$. Then, by Lemma 2(iii), they must be a.a.s. at distance at least $4y$ from the left side of $[0, 1]^2$, since otherwise they would be all dense. Similarly, from among all cells contained in squares of C_1 that are at distance less than $4y$ from the left side of $[0, 1]^2$, consider the lowermost dense cell d. Again, if d is at distance less than $2y$ from that side, consider all K cells which are close to d and below and to the right of d. Otherwise, if d is at distance at least $2y$ from that side, consider all K cells which are close to d and below and to the left of d. Let A' be the connected component of E containing these K cells. By construction, all these K cells are at distance less than $4y$ from the left side of $[0, 1]^2$, and hence they must be pairwise different from the K ones previously described a.a.s. (note that we used Lemma 2(iii) to prove that the K cells contained in A described above must be at distance at least $4y$ from the top side of $[0, 1]^2$). Moreover, A and A' must be the same, since they touch the same piece of γ. Then we have an animal A of size at least $2K$ touching at least one side of $[0, 1]^2$ without any dense cell. By Lemma 2(ii) this does not happen a.a.s.

Case 5. C_1 is at distance less than $2y$ from three sides of $[0, 1]^2$.

W.l.o.g. we can assume that C_1 is at distance less than $2y$ from the left, top, and bottom sides of $[0, 1]^2$. The argument is exactly the same as in Case 3, and hence this case does not occur a.a.s.

In the case when C_2 is at distance at least $2y$ from some side of $[0, 1]^2$, we can apply one of the above cases with C_2 instead of C_1. Thus, it suffices to consider the following case.

Case 6. Both C_1 and C_2 are at distance less than $2y$ from all four sides of $[0, 1]^2$.

Let Q be the union of all those cells at distance less than $4y$ from both the bottom and left sides of $[0, 1]^2$. By Lemma 2, all the cells in Q must be dense, and thus must belong to squares of the same connected component of $\mathcal{G}'$. W.l.o.g., we can assume that they are not in D (i.e., are not contained in squares of C_1). Moreover, A must touch the left (and the bottom) side of $[0, 1]^2$. From among all cells contained in squares of C_1 that are at distance less than $4y$ from the bottom side of $[0, 1]^2$, consider the leftmost dense cell c. If c is at distance less than $2y$ from that side, consider all K cells which are close to c and above and to the left of c. Otherwise, if c is at distance at least $2y$ from that side, consider all K cells which are close to c and below and to the left of c. Let A be the connected component of E containing all these K cells. By construction, all these K cells are at distance less than $4y$ from the bottom side of $[0, 1]^2$. Then, by Lemma 2(iii), they must be a.a.s. at distance at least $4y$ from the left side of $[0, 1]^2$, since otherwise they all would be dense. Similarly, from among all cells contained in squares of C_1 that are at distance less than $4y$ from the left side of $[0, 1]^2$, consider the lowermost dense cell d. Again, if d is at distance less than $2y$ from that side, consider all K cells which are close to d and below and to the right of d. Otherwise, if d is at distance at least $2y$ from that side, consider all K cells which are close to d and below and to the left of d. Let A' be the connected component of E containing all these K cells. By construction, all these K cells are at distance less than $4y$ from the left side of $[0, 1]^2$, and hence they must be pairwise different from the K ones previously described a.a.s. Moreover, A and A' must be the same, since they touch the same piece of γ. Then we have an animal A of size at least $2K$ touching at least one side of $[0, 1]^2$ without any dense cell. By Lemma 2(ii) this does not happen a.a.s. $\square$

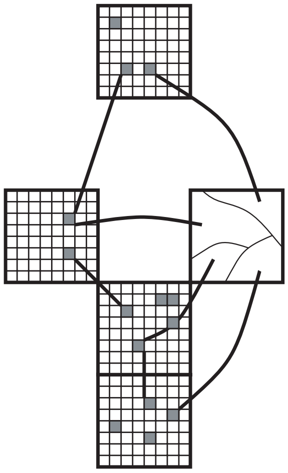


FIG. 2. *Illustration of $\mathcal{G}''$.*

Proof of the upper bound of Theorem 1. Starting from $\mathcal{G}'$ we construct a new graph $\mathcal{G}''$ by adding some new vertices and edges as follows. Let us consider one fixed sparse square S of the initial tessellation of $[0,1]^2$. For each sparse cell c contained in S, we can a.a.s. find at least one dense cell close to it (by Lemma 3) which we call the *hook cell* of c (if this cell is not unique, or even if the square containing these cell(s) is not unique, take an arbitrary one). This hook cell must lie inside some dense square R, which is a friend of S. Then that sparse cell c gets the label R. By grouping those ones sharing the same label, we partition the sparse cells of S into at most 24 groups. Each of these groups of sparse cells will be thought of as a new vertex, added to graph $\mathcal{G}'$ and connected by an edge to the vertex of $\mathcal{G}'$ described by the common label. By performing this same procedure for all remaining sparse squares, we obtain the desired graph $\mathcal{G}''$ (see Figure 2). Those vertices in $\mathcal{G}''$ which already existed in $\mathcal{G}'$ (i.e., dense squares) are called *old*, and those newly added ones are called *new*. Notice that by construction of $\mathcal{G}''$ and by Lemma 4, $\mathcal{G}''$ must be connected a.a.s.

Now, consider an arbitrary spanning tree $\mathcal{T}$ of $\mathcal{G}''$. Observe that the maximal degree of $\mathcal{T}$ is 24, and that all new vertices of $\mathcal{T}$ have degree one and are connected to old vertices. We use capital letters U, V to denote vertices of $\mathcal{T}$ and reserve the lowercase u, v, w for vertices of $\mathcal{G}$. Fix an arbitrary traversal of $\mathcal{T}$ which, starting at an arbitrary vertex, traverses each edge of $\mathcal{T}$ exactly twice and returns to the starting vertex. Note that such a traversal always exists: Fix an arbitrary vertex of $\mathcal{T}$ to be the root vertex, and always follow the edge going to the leftmost neighbor of that vertex (the vertex with the smallest x-coordinate; if there are more, the one among them with the smallest y-coordinate) which was not yet visited. Do this recursively for each vertex. When all neighbors of a vertex are visited, we go back to the vertex from which we came. We iterate this procedure until all vertices are visited and we are back at the root vertex. This traversal gives an ordering in which we construct our Hamiltonian cycle in $\mathcal{G}$ (i.e., as the Hamiltonian cycle travels along the vertices of $\mathcal{G}$, it will visit the vertices of $\mathcal{T}$ according to this traversal).

Let us give a constructive description of our Hamiltonian cycle. Suppose that at some time we visit an old vertex U of $\mathcal{T}$ and that the next vertex V (w.r.t. the traversal) is also old. Then there must exist a pair of dense cells $c_1 \subset U$, $c_2 \subset V$ close to each other, and let $u \in c_1$ and $v \in c_2$ be vertices not used so far. In case this is not the last time we visit U (w.r.t. the traversal), immediately after entering vertex w inside U we connect w to u, and then u is connected to v. If U is visited for the last time (w.r.t. the traversal), we connect from the entering vertex w all vertices inside U not yet used by an arbitrary Hamiltonian path (note that they form a clique in $\mathcal{G}$) before leaving U via u, and subsequently we connect u to v.

Otherwise, suppose that at some time we visit an old vertex U of $\mathcal{T}$ and that the next vertex V (w.r.t. the traversal) is new. We connect all the vertices inside V (possibly just one) by an arbitrary Hamiltonian path, whose endpoints lie inside the sparse cells $d_1 \subset V$ and $d_2 \subset V$ (possibly d_1 equals d_2). Again this is possible since these vertices form a clique in $\mathcal{G}$. Let $c_1 \subset U$ and $c_2 \subset U$ (possibly c_1 equals c_2) be the hook cells of d_1 and d_2 (i.e., c_i is a dense cell in U close to the sparse cell d_i in V). Let $u \in c_1$ and $v \in c_2$ be vertices not used so far. Then immediately after entering vertex w inside U we connect w to u, and then u is joined to the corresponding endpoint of the Hamiltonian path connecting the vertices inside V. The other endpoint is connected to v, and so we again visit U.

We observe that at some steps of the above construction we request unused vertices of $\mathcal{G}$. This is always possible; in fact, each vertex of $\mathcal{T}$ is visited as many times as its degree (at most 24); for each visit of an old vertex U our construction requires exactly two unused vertices $v \in c$, $w \in c$ inside some dense cell $c \subset U$; and c contains at least 48 vertices. By construction, the described cycle is Hamiltonian and the result holds. ∎

In the following corollary, we give an informal definition of a linear time algorithm that constructs a Hamiltonian cycle for a specific instance of $\mathcal{G}(n, r)$. The procedure is based on the previous constructive proof. We assume that real arithmetic can be done in constant time.

COROLLARY 1. *Let* $r \geq \sqrt{\frac{\log n}{(\alpha_p - \epsilon)n}}$ *for some fixed* $\epsilon > 0$. *The proof of Theorem* 1 *yields an algorithm that a.a.s. produces a Hamiltonian cycle in* $\mathcal{G}(n, r)$ *in linear time with respect to* n.

Proof. Assume that the input graph satisfies all the conditions required in the proof of Theorem 1, which happens a.a.s. Assume also that each vertex of the input graph is represented by a pair of coordinates. Observe that the total number of squares is $O(n/\log n)$, and since the number of cells per square is constant, the same holds for the total number of cells. First, we compute in linear time the label of the cell and the square where each vertex is contained. At the same time, we can find for each cell (and square) the set of vertices it contains, and mark those cells (squares) which are dense. Now, for the construction of $\mathcal{G}'$, note that each dense square has at most a constant number of friends to which it can be connected. Thus, the edges of $\mathcal{G}'$ can be obtained in time $O(n/\log n)$. In order to construct $\mathcal{G}''$, for each of the $O(n/\log n)$ cells in sparse squares, we compute in constant time its hook cell and the dense square containing it. Since both the number of vertices and the number of edges of $\mathcal{G}''$ are $O(n/\log n)$, we can compute in time $O(n)$ (e.g., by Kruskal's algorithm) an arbitrary spanning tree $\mathcal{T}$ of $\mathcal{G}''$. The traversal and construction of the Hamiltonian cycle is proportional to the number of edges in $\mathcal{T}$ plus the number of vertices in $\mathcal{G}$ and thus can be done in linear time. ∎

3. Conclusion and outlook.

We believe that the above construction can be generalized to obtain sharp thresholds for Hamiltonicity for random geometric graphs in $[0, 1]^d$ (d being fixed). However, it seems much more difficult to generalize the results to arbitrary distributions of the vertices. The problem posed by Penrose [12], whether or not the graph also becomes Hamiltonian a.a.s. exactly at the point where $\mathcal{G}(n, r)$ gets 2-connected, still remains open.

REFERENCES

[1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, *Wireless sensor networks: A survey*, Computer Networks, 38 (2002), pp. 393–422.

[2] M. Appel and R. P. Russo, *The connectivity of a graph on uniform points on* $[0, 1]^d$, Statist. Probab. Lett., 60 (2002), pp. 351–357.

[3] B. Bollobás, *Random Graphs*, 2nd ed., Cambridge University Press, Cambridge, UK, 2001.

[4] N. B. Clark, C. J. Colbourn, and D. S. Johnson, *Unit disk graphs*, Discrete Math., 86 (1990), pp. 165–177.

[5] M. Garey and D. Johnson, *Computers and Intractability*, W. H. Freeman, NY, 1979.

[6] E. N. Gilbert, *Random plane networks*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 533–543.

[7] A. Goel, S. Rai, and V. Krishnamachari, *Sharp thresholds for monotone properties in random geometric graphs*, in Proceedings of the 36th ACM Symposium (Theory of Computing), ACM, New York, 2004, pp. 580–586.

[8] A. Itai, C. H. Papadimitriou, and J. L. Szwaracfiter, *Hamilton paths in grid graphs*, SIAM J. Comput., 11 (1982), pp. 676–686.

[9] J. Komlós and E. Szemerédi, *Limit distribution for the existence of Hamiltonian cycles in a random graph*, Discrete Math., 43 (1983), pp. 55–63.

[10] M. D. Penrose, *The longest edge of the random minimal spanning tree*, Ann. Appl. Probab., 7 (1997), pp. 340–361.

[11] M. D. Penrose, *On k-connectivity for a geometric random graph*, Random Structures Algorithms, 15 (1999), pp. 145–164.

[12] M. D. Penrose, *Random Geometric Graphs*, Oxford Stud. Probab., Oxford University Press, Oxford, UK, 2003.

[13] J. Petit, *Layout Problems*, Ph.D. Thesis, Universitat Politècnica de Catalunya, Barcelona, Spain.

[14] M. W. Robinson and N. Wormald, *Almost all regular graphs are Hamiltonian*, Random Structures Algorithms, 5 (1994), pp. 363–374.

SIAM J. Discrete Math.
Vol. 21, No. 1, pp. 66–72

TURÁN'S THEOREM IN THE HYPERCUBE*

NOGA ALON[†], ANJA KRECH[‡], AND TIBOR SZABÓ[§]

Abstract. We are motivated by the analogue of Turán's theorem in the hypercube Q_n: How many edges can a Q_d-free subgraph of Q_n have? We study this question through its Ramsey-type variant and obtain asymptotic results. We show that for every odd d it is possible to color the edges of Q_n with $\frac{(d+1)^2}{4}$ colors such that each subcube Q_d is polychromatic, that is, contains an edge of each color. The number of colors is tight up to a constant factor, as it turns out that a similar coloring with $\binom{d+1}{2}+1$ colors is not possible. The corresponding question for vertices is also considered. It is not possible to color the vertices of Q_n with $d+2$ colors such that any Q_d is polychromatic, but there is a simple $d+1$ coloring with this property. A relationship to anti-Ramsey colorings is also discussed. We discover much less about the Turán-type question which motivated our investigations. Numerous problems and conjectures are raised.

Key words. hypercube, Turán-type problems, Ramsey's theorem, anti-Ramsey problems

AMS subject classifications. Primary, 05D05; Secondary, 05D10, 05C15, 05C35

DOI. 10.1137/060649422

1. Introduction.

For graphs G and H, let $ex(G, H)$ denote the maximum number of edges in a subgraph of G which does not contain a copy of H. The quantity $ex(G, H)$ was first investigated in the case when G is a clique. Turán's theorem resolves the problem precisely when H is a clique as well.

In this paper, we study these Turán-type problems for the case when the base graph G is the n-dimensional hypercube Q_n. This setting was initiated by Erdős [8] who posed the problem of determining the largest number of edges in a C_4-free subgraph of the hypercube. He conjectured that the answer is $(\frac{1}{2}+o(1))e(Q_n)$ and offered \$100 for a solution. The current best upper bound, due to Chung [6], stands at $\approx .623e(Q_n)$. The best known lower bound is $\frac{1}{2}\left(n+\sqrt{n}\right)2^{n-1}$ (for $n=4^r$) due to Brass, Harborth, and Nienborg [5].

Erdős [8] also raised the extremal question for even cycles. Chung [6] obtained that $\frac{ex(Q_n,C_{4k})}{e(Q_n))} \to 0$ for every $k \geq 2$, i.e., cycles with length divisible by 4, starting from 8 are harder to avoid than the 4-cycle. She also showed that

$$\frac{1}{4}e(Q_n) \leq ex(Q_n, C_6) \leq (\sqrt{2}-1+o(1))e(Q_n).$$

Later Conder [7] improved the lower bound to $\frac{1}{3}e(Q_n)$ by defining a 3-coloring of the edges of the n-cube such that every color class is C_6-free. On the other hand, it is shown in [1] that for any fixed k, in any k-coloring of the edges of a sufficiently

[*] Received by the editors January 9, 2006; accepted for publication (in revised form) August 21, 2006; published electronically February 5, 2007.
http://www.siam.org/journals/sidma/21-1/64942.html

[†] Schools of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel, and IAS, Princeton, NJ 08540 (nogaa@tau.ac.il). This author's research was supported in part by a USA-Israel BSF grant, by the Israel Science Foundation, by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University, and by the Von Neumann Fund.

[‡] Department of Mathematics and Theoretical Computer Science, Free University Berlin, D-14195 Berlin, Germany (krech@math.fu-berlin.de).

[§] Institute of Theoretical Computer Science, ETH Zürich, CH-8092 Zürich, Switzerland (szabo@inf.ethz.ch).

large cube there are monochromatic cycles of every even length greater than 6. Note, however, that the Turán problem for cycles of length $4k+2$ is still wide open. For $k \geq 2$, it is not even known whether $ex(Q_n, C_{4k+2}) = o(e(Q_n))$.

In the present paper we consider a generalization of the C_4-free subgraph problem in a different direction, which we feel is the true analogue of Turán's theorem in the hypercube. For arbitrary d we give bounds on $ex(Q_n, Q_d)$. For convenience we will talk about the complementary problem; i.e., let $f(n,d)$ denote the minimum number of edges one must delete from the n-cube to make it d-cube-free. Obviously $f(n,d) = e(Q_n) - ex(Q_n, Q_d)$. By a simple averaging argument one can see that for any fixed d the function $f(n,d)/e(Q_n)$ is nondecreasing in n, so a limit c_d exists. (In fact this limit exists for an arbitrary forbidden subgraph H in the place of Q_d). Erdős's conjecture then could be stated as $c_2 = \frac{1}{2}$.

Trivially $f(d,d) = 1$, so by the above, $c_d \geq \frac{1}{d2^{d-1}}$. On the other hand, if one deletes edges of the hypercube on every dth level, one obtains a Q_d-free subgraph. For this, observe that every d-dimensional subcube spans edges on d consecutive levels. Thus $c_d \leq \frac{1}{d}$.

In the present paper we improve on these trivial bounds.

Theorem 1.

$$\Omega\left(\frac{\log d}{d2^d}\right) = c_d \leq \begin{cases} \frac{4}{(d+1)^2} & \textit{if } d \textit{ is odd,} \\ \frac{4}{d(d+2)} & \textit{if } d \textit{ is even.} \end{cases}$$

We conjecture that our construction is essentially optimal for $d = 3$.

Conjecture 2.

$$c_3 = \frac{1}{4}.$$

The best known lower bound on c_3 is $1 - \left(\frac{5}{8}\right)^{1/4} \approx 0.11$ and follows from some property of the four-dimensional cube. (A Q_3-free subgraph of Q_4 cannot contain more than 10 vertices of degree 4; see the paper of Graham et al. [10]).

For arbitrary d we are less confident; it would certainly be very interesting to determine how fast c_d tends to 0, when d tends to infinity.

Problem 3. *Determine the order of magnitude of c_d.*

We tend to think that c_d is larger than inverse exponential, but feel that we are very far from understanding the truth. In fact all our arguments are set in the related Ramsey-type framework, rather than the original Turán-type. A coloring of the edges of Q_n is called *d-polychromatic* if every subcube of dimension d is *polychromatic* (i.e., it has all the colors represented on its edges). Let $pc(n,d)$ be the largest integer p such that there exists a d-polychromatic coloring of the edges of Q_n in p colors. Clearly, $pc(n,d) \leq d2^{d-1}$ and $f(n,d) \leq e(Q_n)/pc(n,d)$. Since $pc(n,d)$ is a nonincreasing function in n, it stabilizes for large n. Let p_d be this limit; then we have $c_d \leq 1/p_d$. We can determine p_d up to a factor of 2.

Theorem 4.

$$\binom{d+1}{2} \geq p_d \geq \begin{cases} \frac{(d+1)^2}{4} & \textit{if } d \textit{ is odd,} \\ \frac{d(d+2)}{4} & \textit{if } d \textit{ is even.} \end{cases}$$

The lower bound implies the upper bound in Theorem 1. It would be interesting to resolve the following problem.

Problem 5. *Determine the asymptotic behavior of p_d.*

The lower bound in Theorem 1 is a consequence of some known results on the analogous problem for vertices of the cube. Let $g(n,d)$ be the minimum number of vertices one must delete from the n-cube to make it d-cube-free. Clearly $g(n,d) \leq f(n,d)$. Again, simple averaging shows that for any fixed d the function $g(n,d)/2^n$ is nondecreasing in n, so a limit c_d^0 exists.

The problem of determining $g(n,d)$ was investigated early and widely by several research communities, mostly in a dual formulation under the different names of t-independent sets [12], qualitatively t-independent 2-partitions [13], and (n,t)-universal vector sets [15], where $t = n - d$. These investigations mostly deal with the case when d is *large*, i.e., very close to n. The lone result we are aware of about $g(n,d)$ for d small compared to n is due to Johnson and Entringer [11] who prove that $c_2^0 = 1/3$. Even more, they show that the unique smallest set breaking all copies of Q_2 is in the form of every third level of the cube. In general we know very little.

Proposition 6.

$$\frac{1}{d+1} \geq c_d^0 \geq \frac{\log d}{2^{d+2}}.$$

Again, the Ramsey analogue of the problem is more clear. In fact we have here a precise result. A coloring of the vertices of Q_n is called *d-polychromatic* if every subcube of dimension d has all the colors represented on its vertices. Let $pc^0(n,d)$ be the largest integer p such that there exists a d-polychromatic coloring of the vertices of Q_n in p colors. Clearly, $pc^0(n,d) \leq 2^d$ and $g(n,d) \leq 2^n/pc^0(n,d)$. Since $pc^0(n,d)$ is a nonincreasing function of n, it stabilizes for large n. Let p_d^0 be this limit; then we have $c_d^0 \leq 1/p_d^0$. We can determine p_d^0 for every d.

Theorem 7.

$$p_d^0 = d + 1.$$

1.1. Relation to rainbow colorings. In this subsection we point out a relation between the established notion of anti-Ramsey coloring and the one of polychromatic coloring introduced in this paper. We also note how Theorem 4 could be applied to improve a result of [2].

An edge-coloring $r : E(H) \to \{1, 2, \dots\}$ of a graph H is called *rainbow* if no two edges of H receive the same color. A coloring c of the edges of graph G is called *H-anti-Ramsey* if the restriction of c to any subgraph $H_0 \subseteq G$, $H_0 \cong H$, is *not* rainbow. Let $ar(G,H)$ be the largest number of colors used in an H-anti-Ramsey coloring of G. The function $ar(G,H)$ was introduced by Erdős, Simonovits, and Sós [9]. It is well known that $ar(G,H) \leq ex(G,H)$ since taking one arbitrary edge from each color class of an H-anti-Ramsey coloring, one must obtain an H-free subgraph of G.

For any graph G and H, we call a p-coloring $c : E(G) \to \{1, \dots, p\}$ of the edges of G *H-polychromatic* if every subgraph $H_0 \subseteq G$, $H_0 \cong H$, has *all* the p colors represented on its edges. Let $pc(G,H)$ be the largest number p such that there is an H-polychromatic coloring of the edges of G. The following proposition establishes a relationship between H-anti-Ramsey and H-polychromatic colorings.

Proposition 8.

$$ar(G,H) \geq \left(1 - \frac{2}{pc(G,H)}\right) e(G).$$

Proof. Given an H-polychromatic coloring c of G with $p = pc(G,H)$ colors, we define an H-anti-Ramsey coloring r of G with at least $(1 - 2/p)e(G)$ colors. Let F

be the set of edges formed by the union of the two smallest color classes of c. The coloring r will be chosen constant on F; say, all edges in F receive color 1. All other edges of G will receive distinct colors. Then we used at least $(1-\frac{2}{p})e(G)+1$ colors. Also, the coloring r defined this way is H-anti-Ramsey since each copy of H in G contains at least two edges of F, and thus at least two edges receive the color 1 in every copy of H. □

In a recent paper [2] Axenovich et al. investigated Q_d-anti-Ramsey colorings of Q_n. Lower and upper bounds for $ar(Q_n, Q_d)$ are found. In particular for fixed d, the leading terms of their bounds amount to

$$\left(1-\frac{4}{d2^d}\right)e(Q_n) \geq ar(Q_n, Q_d) \geq \left(1-\frac{1}{d}\right)e(Q_n).$$

One can improve the upper bound by applying Theorem 1, and the lower bound by using the polychromatic coloring of Theorem 4.

Corollary 9.

$$\left(1-\Omega\left(\frac{\log d}{d2^d}\right)\right)e(Q_n) \geq ar(Q_n, Q_d) \geq \left(1-\frac{8}{d^2}-O\left(\frac{1}{d^3}\right)\right)e(Q_n).$$

Notation. We consider the cube as a set of n-dimensional 0/1-vectors, where the coordinates are labeled by the first n positive integers, $[n]=\{1,\dots,n\}$. A d-dimensional subcube of the n-dimensional cube is denoted by a vector from $\{0,1,\star\}^n$ which contains d $\star$-entries; the stars represent the nonconstant coordinates of the subcube. For a subcube D of the n-dimensional cube we denote by $ONE(D)$, $ZERO(D)$, and $STAR(D)$ the set of labels of those coordinates which are 1, 0, and $\star$, respectively.

2. Q_d-free subgraphs of Q_n.

In this section we give a proof of the lower bound in Theorem 4.

Proof. First assume that d is odd. We define a $\frac{(d+1)^2}{4}$-coloring of the edges of Q_n, which is d-polychromatic.

We color the edges of Q_n with elements of $\mathbb{Z}_{\frac{d+1}{2}} \times \mathbb{Z}_{\frac{d+1}{2}}$ in the following way. The edge e with a star at coordinate a is colored with the vector whose first coordinate is $|\{x \in ONE(e) : x < a\}| \pmod{\frac{d+1}{2}}$ and whose second coordinate is $|\{x \in ONE(e) : x > a\}| \pmod{\frac{d+1}{2}}$.

Now consider a d-dimensional subcube C of Q_n with $STAR(C)=\{a_1,\dots,a_d\}$, where $a_1 < a_2 < \dots < a_d$. Let s be the vertex of C with the least number of ones. So for each vertex x of C we have that $ONE(s) \subseteq ONE(x) \subseteq ONE(s) \cup \{a_1,\dots,a_d\}$.

We will show that all $\frac{(d+1)^2}{4}$ colors appear on edges of C whose star is at position $a_{\frac{d+1}{2}}$. Let (u,v) be an arbitrary element of $\mathbb{Z}_{\frac{d+1}{2}} \times \mathbb{Z}_{\frac{d+1}{2}}$.

Let $l := |\{x \in ONE(s) : x < a_{\frac{d+1}{2}}\}| \pmod{\frac{d+1}{2}}$ and $r := |\{x \in ONE(s) : x > a_{\frac{d+1}{2}}\}| \pmod{\frac{d+1}{2}}$. Choose any $k \equiv u-l \pmod{\frac{d+1}{2}}$ elements K from $\{a_1,\dots,a_{\frac{d+1}{2}-1}\}$ and any $p \equiv v-r \pmod{\frac{d+1}{2}}$ elements L from $\{a_{\frac{d+1}{2}+1},\dots,a_d\}$. Define s' by $ONE(s') = ONE(s) \cup K \cup L$. Then the edge incident to s' and having a star at position $a_{\frac{d+1}{2}}$ has color (u,v).

For even d a similar construction works; the only difference is that we take the number of ones to the left of the label of the edge modulo $\frac{d}{2}$ and the number of ones to the right modulo $\frac{d+2}{2}$. Then one can prove that among the edges with label $\frac{d}{2}$ all colors appear. □

3. Upper bound in the Ramsey problems. First, we prove the upper bound in Theorem 4.

Proof of Theorem 4. Suppose we have a d-polychromatic p-edge-coloring c of Q_n where n is huge. We will use Ramsey's theorem for d-uniform hypergraphs with $p^{d2^{d-1}}$ colors. We define a $p^{d2^{d-1}}$-coloring of the d-subsets of $[n]$. Fix an arbitrary ordering of the edges of Q_d. For an arbitrary subset S of the coordinates, define *cube*(S) to be the subcube whose $\star$ coordinates are at the positions of S and all its other coordinates are 0, i.e., $STAR(cube(S)) = S$ and $ZERO(cube(S)) = [n] \setminus S$. Let S be a d-subset of $[n]$ and define the color of S to be the vector whose coordinates are the c-values of the edges of the d-dimensional subcube *cube*(S) (according to the fixed ordering of the edges of Q_d). By Ramsey's theorem, if n is large enough, there is a set $T \subseteq [n]$ of $d^2 + d - 1$ coordinates such that the color-vector is the same for any d-subset of T. Let us now fix a set S of d particular coordinates from T: those ones which are the (id)th elements of T for some $i = 1, \dots, d$. Hence any two elements of S have at least $d - 1$ elements of T in between.

Claim 10. *The c-value of an edge e of cube(S) depends only on the number of ones to the left of the $\star$ of e and the number of ones to the right of this $\star$.*

Proof. Let e_1 and e_2 be two edges of *cube*(S) such that they have the same number of ones to the left of their respective star and the same number of ones to the right as well. We can find d coordinates S' from T such that $STAR(e_2) \cup ONE(e_2) \subseteq S'$ (i.e., e_2 is an edge of *cube*(S')), and the vector e_2 restricted to S' is *equal* to the vector e_1 restricted to S. Indeed, there are enough unused 0-coordinates of e_2 in T between any two elements of S.

Now, since every d-subset of T has the same color-vector, the corresponding edges of the cubes *cube*(S) and *cube*(S') have the same c-value. In particular the colors of e_1 and e_2 are equal. The claim is proved. □

To finish the proof of the upper bound in Theorem 4 we just note that there are exactly $1 + \cdots + d = \binom{d+1}{2}$ many ways to separate at most $d - 1$ ones by a $\star$. By Claim 10 a d-polychromatic edge-coloring is not possible with more colors. □

With a very similar argument one can prove the matching upper bound in the analogous question for vertices.

Proof of Theorem 7. Assume we have a d-polychromatic coloring of the vertices of Q_n. Let us define a d^{2^d}-coloring of the d-tuples of $[n]$. For a d-subset S let the color be determined by the vector of the 2^d colors of the vertices of the subcube *cube*(S) with $STAR(cube(S)) = S$ and $ZERO(cube(S)) = [n] \setminus S$ (according to some fixed ordering of the vertex set of Q_d). By Ramsey's theorem there is a set T of $d^2 + d - 1$ coordinates such that the color-vector is the same for any d-subset of T. Let us again fix d coordinates S in T such that any two elements of S have at least $d - 1$ elements of T in between (as in the proof of Theorem 4).

Claim 11. *The color of a vertex in cube(S) depends only on its number of ones.*

Proof. Let v_1 and v_2 be two vectors from *cube*(S) such that $|ONE(v_1)| = |ONE(v_2)|$. We can find d coordinates S' from T such that $ONE(v_2) \subseteq S'$, and the vector v_2 restricted to S' is *equal* to the vector v_1 restricted to S. Indeed, there are enough unused 0 coordinates in T between any two elements of S to do this. Now, since T is monochromatic according to our color-vectors, the color of v_1 and v_2 is the same as well. The claim is proved. □

To finish the proof of the upper bound in Theorem 7 we just note that there are exactly $d + 1$ possible values for the number of ones on d coordinates. By Claim 11 a d-polychromatic coloring is not possible with more colors.

For the lower bound in Theorem 7 one can color each vertex of the cube with the number of its nonzero coordinates modulo $d+1$. This gives a d-polychromatic vertex coloring in $d+1$ colors. ∎

4. A lower bound on c_d. The lower bound in Proposition 6 can be deduced from earlier results on the d-independent set problem and is essentially stated (implicitly) in [10]. For completeness we sketch the proof.

Let G be a set of g vertices which intersects all d-cubes of the n-cube. This happens if and only if, interpreting these vertices as subsets of an n-element base set X, G shatters all $(n-d)$-element subsets of X. (A family $\mathcal{F}$ of subsets *shatters* a given subset K if all $2^{|K|}$ subsets of K can be represented as $K \cap F$ for some $F \in \mathcal{F}$.) Now let M_G be the $g \times n$ 0/1-matrix whose rows correspond to the elements of G. Then the columns of M_G can be interpreted as a family L of n subsets of a g-element base set Y such that all the 2^{n-d} parts of the Venn diagram of any $n-d$ members of L are nonempty. (A family L satisfying this property is usually called $(n-d)$*-independent.*)

Thus determining $g(d+t,d)$ is the same problem as determining the largest size of a t-independent family. This was first done by Schönheim [14] and Brace and Daykin [4] for $t=2$ and later reproved and generalized by many others, e.g., Kleitman and Spencer [12].

It is known that $g(d+2,d) \geq \log d$, and thus the lower bound on c_d^0 follows by the monotonicity of $g(n,d)/2^n$. The lower bound in Theorem 1 also follows since $f(d+2,d) \geq g(d+2,d)$ and $f(n,d)/e(Q_n)$ is nondecreasing.

5. Remarks and more open problems.

Remark. The following claim shows that if $1/c_d$ is indeed subexponential, then one has to search for the evidence in very large, i.e., doubly exponential, dimensions.

For simplicity we write here the proof for c_d^0 (the vertex version); the argument for c_d follows along similar lines.

Claim. *For any $p \leq \frac{2^d}{2d}$, there is a d-polychromatic p-coloring of the n-cube, with $n = \frac{1}{2}\exp\left\{\frac{2^d}{2dp}\right\}$. In particular, for any $\epsilon > 0$ and $n \leq \frac{1}{2}\exp\left\{2^{(1-\epsilon)d}\right\}$,*

$$g(n,d) \leq \frac{2d}{2^{\epsilon d}} \cdot 2^n.$$

Proof. We randomly color the vertices of Q_n with p colors. For each vertex v select a color uniformly at random from $\{1,\ldots,p\}$, with choices being independent from the choices on all other vertices. For a d-cube D, let A_D be the event that there is a color which does not appear on the vertices of D. The probability of A_D is at most $p\left(1-1/p\right)^{2^d}$. Each d-cube intersects less than $2^d\binom{n}{d}$ other d-cubes. Obviously A_D is independent from the set of all events $A_{D'}$, where D' is disjoint from D.

For $p \leq \frac{2^d}{2d}$ and $n = \frac{1}{2}\exp\left\{\frac{2^d}{2dp}\right\}$,

$$e \cdot p\left(1-\frac{1}{p}\right)^{2^d} 2^d\binom{n}{d} \leq e^{1+\log p - \frac{2^d}{p} + d\log 2n} = o_d(1).$$

Hence the Lovász Local Lemma implies that with nonzero probability all p colors are represented on all d-cubes.

For the second part of the claim, choose $p = 2^{\epsilon d}/2d$ and leave out the vertices of the sparsest color class in a d-polychromatic p-coloring of the n-cube. ∎

Open problems. Since $f(n,2)$ is known to be strictly larger than one-third of the number of edges in Q_n for large n [6], it is clear that $p_2 = 2$. Bialostocki [3] proved that in any 2-polychromatic edge-2-coloring of Q_n the color classes are asymptotically equal. The next natural question is the determination of p_3, which is either $4, 5$, or 6. Once p_3 is known, it would be interesting to generalize Bialostocki's theorem and decide whether in any 3-polychromatic p_3-edge-coloring of Q_n, each color class contains approximately $\frac{1}{p_3}e(Q_n)$ edges.

Everything above could be generalized, quite straightforwardly, but would not solve the following problems:

Turán-type: Let $f^{(l)}(n,d)$ be the smallest integer f such that there is a family of f l-faces of Q_n such that every d-face contains at least one member of this family. Again, $f^{(l)}(n,d)/\binom{n}{l}2^{n-l}$ is nondecreasing, so there is a limit $c_d^{(l)}$. Determine it!

Ramsey-type: A coloring of the l-faces of Q_n is *d-polychromatic* if for every d-face S and color s there is an l-face of S with color s. Let $pc^{(l)}(n,d)$ be the largest number of colors with which there is a d-polychromatic coloring of the l-faces of Q_n. Again, the limit $p_d^{(l)}$ of $pc^{(l)}(n,d)$ exists. Determine it!

Note added in proof. Problem 5 was recently solved by David Offner. He showed that the lower bound in Theorem 4 is tight for every d.

Acknowledgment. We would like to thank an anonymous referee for pointing out reference [2] to us.

REFERENCES

[1] N. Alon, R. Radoičić, B. Sudakov, and J. Vondrák, *A Ramsey-type result for the hypercube*, J. Graph Theory, 53 (2006), pp. 196–208.

[2] M. Axenovich, H. Harborth, A. Kemnitz, M. Möller, and I. Schiermeyer, *Rainbows in the hypercube*, Graphs Combin., to appear.

[3] A. Bialostocki, *Some Ramsey type results regarding the graph of the n-cube*, Ars Combin., 16-A (1983), pp. 39–48.

[4] A. Brace and D. E. Daykin, *Sperner type theorems for finite sets*, in Proceedings of the British Combinatorial Conference, Oxford, 1972, pp. 18–37.

[5] P. Brass, H. Harborth, and H. Nienborg, *On the maximum number of edges in a C_4-free subgraph of Q_n*, J. Graph Theory, 19 (1995), pp. 17–23.

[6] F. Chung, *Subgraphs of a hypercube containing no small even cycles*, J. Graph Theory, 16 (1992), pp. 273–286.

[7] M. Conder, *Hexagon-free subgraphs of hypercubes*, J. Graph Theory, 17 (1993), pp. 477–479.

[8] P. Erdős, *Some problems in graph theory, combinatorial analysis and combinatorial number theory*, in Graph Theory and Combinatorics, B. Bollobás, ed., Academic Press, New York, 1984, pp. 1–17.

[9] P. Erdős, M. Simonovits, and V. T. Sós, *Anti-Ramsey theorems*, in Infinite and finite sets (Colloquium held at Keszthely, 1973; dedicated to P. Erdős on his 60th birthday), Vol. II, Colloq. Math. Soc. János Bolyai, 10, North–Holland, Amsterdam, 1975, pp. 633–643.

[10] N. Graham, F. Harary, M. Livingston, and Q. Stout, *Subcube fault-tolerance in hypercubes*, Inform. Comput., 102 (1993), pp. 280–314.

[11] K. A. Johnson and R. Entringer, *Largest induced subgraphs of the n-cube that contain no 4-cycles*, J. Combin. Theory Ser. B, 46 (1989), pp. 346–355.

[12] D. Kleitman and J. Spencer, *Families of k-independent sets*, Discrete Math., 6 (1973), pp. 255–262.

[13] A. Rényi, *Foundations of Probability*, Wiley New York, 1971.

[14] J. Schönheim, *A generalization of results of P. Erdős, G. Katona, and D. J. Kleitman concerning Sperner's theorem*, J. Combin. Theory Ser. A, 11 (1971), pp. 111–117.

[15] G. Seroussi and N. H. Bshouty, *Vector sets for exhaustive testing of logic circuits*, IEEE Trans. Inform. Theory, 34 (1988), pp. 513–522.

SIAM J. Discrete Math.
Vol. 21, No. 1, pp. 73–92

EVERY MONOTONE 3-GRAPH PROPERTY IS TESTABLE*

CHRISTIAN AVART†, VOJTĚCH RÖDL†, AND MATHIAS SCHACHT‡

Abstract. Recently Alon and Shapira [*Every monotone graph property is testable*, New York, Proceedings of the 37th Annual ACM Symposium on Theory of Computing, Baltimore, MD, ACM Press, 2005, pp. 128–137] have established that every monotone graph property is testable. They raised the question whether their results can be extended to hypergraphs. The aim of this paper is to address this problem. Based on the recent regularity lemma of Rödl and Schacht [*Regular partitions of hypergraphs*, Combin. Probab. Comput., to appear], we prove that any monotone property of 3-uniform hypergraphs is testable answering in part the question of Alon and Shapira. Our approach is similar to the one developed by Alon and Shapira for graphs. We believe that based on the general version of the hypergraph regularity lemma the proof presented in this article extends to k-uniform hypergraphs.

Key words. hypergraph, monotone properties, property testing, hypergraph regularity lemma

AMS subject classifications. 68R10, 05C65, 05C35, 68Q25, 68W20

DOI. 10.1137/060652294

1. Introduction.

1.1. Basic definitions.

Let $k \geq 2$ be an integer and $\mathscr{A}$ be a property of k-uniform hypergraphs. In other words, $\mathscr{A}$ is a (possibly infinite) family of k-uniform hypergraphs, and we say that a given hypergraph $\mathcal{H}$ satisfies $\mathscr{A}$ if $\mathcal{H} \in \mathscr{A}$. In this paper we consider only *decidable properties* $\mathscr{A}$, which are those for which there is an algorithm that decides if $\mathcal{H} \in \mathscr{A}$ or $\mathcal{H} \notin \mathscr{A}$ in finite time (depending on the size of $\mathcal{H}$) for every k-uniform hypergraph $\mathcal{H}$.

For a given constant $\eta > 0$, we say a k-uniform hypergraph $\mathcal{H}$ on n vertices is *η-far* from $\mathscr{A}$ if no k-uniform hypergraph $\mathcal{G}$ on the same vertex set with $|E(\mathcal{G})\triangle E(\mathcal{H})| \leq \eta n^k$ satisfies $\mathscr{A}$. This is a natural measure of how far the given hypergraph $\mathcal{H}$ is from satisfying the property $\mathscr{A}$.

We consider randomized algorithms which for an input hypergraph $\mathcal{H}$ on the vertex set $\{1, 2, \ldots, n\} = [n]$ are able to make queries whether a given k-tuple of vertices spans an edge in $\mathcal{H}$. For a property $\mathscr{A}$ and a constant $\eta > 0$, such an algorithm will be called a *tester* for $\mathscr{A}$ if it can distinguish with, say, probability 2/3 whether $\mathcal{H}$ satisfies $\mathscr{A}$ or is η-far from it. If a property $\mathscr{A}$ has for every $\eta > 0$ a tester whose *query complexity* (i.e., the number of queries) is bounded by a function of η and $\mathscr{A}$ but is independent of the number of vertices of the input hypergraph $\mathcal{H}$, the property is called *testable*.

One can observe that some simple properties such as connectivity or containing a copy of some fixed hypergraph $\mathcal{F}$ are not testable. Perhaps surprisingly, many other properties, e.g., being $\mathcal{F}$-free, are testable.

*Received by the editors February 17, 2006; accepted for publication (in revised form) September 19, 2006; published electronically February 5, 2007.

http://www.siam.org/journals/sidma/21-1/65229.html

†Department of Mathematics and Computer Science, Emory University, Atlanta, GA 30322 (cavart@mathcs.emory.edu, rodl@mathcs.emory.edu). The research of the second author was partially supported by NSF grant DMS 0300529.

‡Institut für Informatik, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany (schacht@informatik.hu-berlin.de). The research of this author was supported by DFG grant SCHA 1263/1-1.

1.2. Testable graph properties. The general notion of property testing was introduced by Rubinfeld and Sudan in [26]. In [14], Goldreich, Goldwasser, and Ron initiated the study of property testing for combinatorial structures. In the present paper the combinatorial structures we focus on are hypergraphs. Our work builds on some of the earlier work of Alon et al. In a series of papers [1, 2, 3, 6, 4, 5] Alon and his co-authors investigated testability of graph properties. This line of research culminated in the recent result of Alon and Shapira [4] asserting that every hereditary property $\mathscr{A}$, i.e., $\mathscr{A}$ is closed under taking induced subgraphs, is testable (see also Lovász and Szegedy [19] for an alternative proof). A central tool in the work for graphs is Szemerédi's regularity lemma (see Theorem 4) for graphs [28].

Some ideas of property testing for graphs were already present before the notion of a tester was developed. For example if $\mathscr{A}$ consists of all graphs not containing a fixed graph F (as a not necessarily induced subgraph), then the existence of a tester for $\mathscr{A}$ follows from the so-called *removal lemma* for graphs. The removal lemma asserts that for every graph F and every $\eta > 0$ there exists a $c > 0$ such that if G is an n-vertex graph which is η-far from being F-free, then G contains at least $cn^{|V(F)|}$ copies of F. This result was first obtained for F being the triangle K_3 by Ruzsa and Szemerédi [27] and later extended to arbitrary graphs F by Erdős, Frankl, and Rödl [12]. Those results can straightforwardly be generalized to prove the testability of properties $\mathscr{A}$, which can be defined by a *finite* collection $\mathscr{F}$ of forbidden subgraphs, i.e.,

$$\mathscr{A} = \mathrm{Forb}(\mathscr{F}) := \{G\colon\ F \nsubseteq G \text{ for every } F \in \mathscr{F}\}, \tag{1}$$

with $|\mathscr{F}| < \infty$ (see the discussion in section 3.1).

For *infinite* families $\mathscr{F}$ it follows for example from a result of Bollobás et al. [7] that being bipartite is a testable graph property. In [10], answering a question of Erdős (see, e.g., [11]), Duke and Rödl generalized the result from [7] and proved that being h-colorable is testable for any $h \geq 2$. The proof in [10] is also based on Szemerédi's regularity lemma. Later this result and related results were established by Goldreich, Goldwasser, and Ron [14] and subsequently improved by Alon and Krivelevich [3]. The authors of [14] and [3] could avoid using Szemerédi's regularity lemma and, consequently, obtained much better bounds on the query complexity for the testers.

The general problem for *monotone* graph properties, which are those properties as described in (1) with a possibly infinite forbidden family $\mathscr{F}$, was solved by Alon and Shapira [5]. They showed that every monotone graph property is testable and asked if the same holds for hypergraphs. In this paper we answer their question positively for 3-uniform hypergraphs (see Theorem 2 below). Our proof uses the ideas of Alon and Shapira and is based on the recent hypergraph extensions of Szemerédi's regularity lemma [13, 15, 21, 23, 24, 29]. The transition from graphs to hypergraphs leads, however, to some technical difficulties. In this paper we restrict ourselves to 3-uniform hypergraphs. This case already reflects the main differences between the graphs and the general case of k-uniform hypergraphs but allows us to simplify the notation and to improve the presentation of the proof. We believe that the argument can be extended with no major conceptual modification to k-uniform hypergraphs (see also section 5).

The main result of the present paper is the first general result for 3-uniform hypergraphs which establishes testability for a fairly natural and general class of properties. A few other hypergraph results were already known before, e.g., h-colorability [9], not containing one fixed *induced* subhypergraph [17] and not containing one fixed *non-induced* subhypergraph [20, 21].

1.3. Main result. We now state the main result of the paper. A 3-uniform hypergraph $\mathcal{H}$ on the vertex set V is some family of 3-element subsets of V, i.e., $\mathcal{H} \subseteq \binom{V}{3}$. Note that we identify hypergraphs with its edge set and we write $V(\mathcal{H})$ for the vertex set. We recall that a property $\mathscr{A}$ of 3-uniform hypergraphs is *monotone* if $\mathcal{H} \in \mathscr{A}$ implies that every (not necessarily induced) subhypergraph $\mathcal{G} \subseteq \mathcal{H}$ exhibits property $\mathscr{A}$ as well. In other words, $\mathscr{A}$ is closed under removal of vertices and edges. Note that if $\mathscr{A}$ is a monotone property and the hypergraph $\mathcal{H}$ does not satisfy $\mathscr{A}$, then no hypergraph obtained by adding edges to $\mathcal{H}$ will satisfy $\mathscr{A}$. Consequently, for monotontone properties the definition of η-far given earlier is equivalent to the following.

Definition 1. *For a monotone property $\mathscr{A}$ we say an n-vertex 3-uniform hypergraph $\mathcal{H}$ is* η-far from $\mathscr{A}$ *if every subhypergraph $\mathcal{G}$ of $\mathcal{H}$ with $|\mathcal{H} \setminus \mathcal{G}| \leq \eta n^3$ satisfies $\mathcal{G} \notin \mathscr{A}$.*

We say a tester has *one-sided error* if it confirms with probability 1 that $\mathcal{H} \in \mathscr{A}$. In other words, whenever $\mathcal{H}$ satisfies $\mathscr{A}$, the algorithm will be correct with probability equal to 1. Moreover, a property $\mathscr{A}$ is testable with one-sided error if for every $\eta > 0$ there exists a tester with one-sided error.

In [5] Alon and Shapira proved that for any (decidable) monotone graph property $\mathscr{A}$ and any $\eta > 0$ there exists a tester which after a bounded number of random edge queries comes to the following conclusion:

- If $\mathcal{H} \in \mathcal{P}$, then the tester confirms it with probability 1.
- If $\mathcal{H}$ is η-far from $\mathscr{A}$, then the tester outputs with probability 2/3 that $\mathcal{H} \notin \mathcal{P}$.
- Otherwise, if $\mathcal{H} \notin \mathscr{A}$ and $\mathcal{H}$ is not η-far from $\mathscr{A}$, then there are no guarantees for the output of the tester.

In this paper we generalize this result from graphs to 3-uniform hypergraphs.

Theorem 2. *Every decidable and monotone property $\mathscr{A}$ of 3-uniform hypergraphs is testable with one-sided error.*

As discussed earlier, monotone properties can be described by a (possibly infinite) family of forbidden hypergraphs, i.e, for every monotone property $\mathscr{A}$ there exists a family of hypergraphs $\mathscr{F}$ such that $\mathscr{A} = \mathrm{Forb}(\mathscr{F})$, where $\mathrm{Forb}(\mathscr{F})$ is the family of those hypergraphs not containing any element of $\mathscr{F}$ as a (not necessarily induced) subhypergraph. Theorem 2 is then a consequence of the following result, as we will show momentarily.

Theorem 3. *Let $\mathscr{F}$ be a family of 3-uniform hypergraphs and $\mathscr{A} = \mathrm{Forb}(\mathscr{F})$. For all $\eta > 0$ there exists $c = c(\mathscr{A}, \eta) > 0$, and there are positive integers $C = C(\mathscr{A}, \eta)$ and $n_0 = n_0(\mathscr{A}, \eta)$ such that the following holds.*

If $\mathcal{H}$ is a 3-uniform hypergraph on $n \geq n_0$ vertices which is η-far from satisfying $\mathscr{A}$, then there exists a hypergraph $\mathcal{F}_0 \in \mathscr{F}$ on $f_0 \leq C$ vertices such that the number of copies of $\mathcal{F}_0$ in $\mathcal{H}$ is at least cn^{f_0}.

Theorem 1 easily follows from Theorem 2.

Proof (Theorem 3 is a direct consequence of Theorem 2). Let a decidable and monotone property $\mathscr{A} = \mathrm{Forb}(\mathscr{F})$ and some $\eta > 0$ be given. By Theorem 2, there is some $c > 0$ and there are integers C and $n_0 \in \mathbb{N}$ such that any 3-uniform hypergraph on $n \geq n_0$ vertices, which is η-far from exhibiting $\mathscr{A}$, contains at least $cn^{|V(\mathcal{F}_0)|}$ copies of some $\mathcal{F}_0 \in \mathcal{F}$ with $|V(\mathcal{F}_0)| \leq C$.

Let $s \in \mathbb{N}$ be such that $(1-c)^{s/C} < 1/3$, and set $m_0 = \max\{s, n_0\}$. We claim that there exists a one-sided tester with query complexity $\binom{m_0}{3}$ for $\mathscr{A}$. For that let $\mathcal{H}$ be a 3-uniform hypergraph on n vertices. If $n \leq m_0$, then the tester simply queries all edges of $\mathcal{H}$, and since $\mathscr{A}$ is decidable, there is an exact algorithm with running time depending only on the fixed m_0, which determines correctly if $\mathcal{H} \in \mathscr{A}$ or not.

Consequently, let $n > m_0$. Then we choose uniformly at random a set S of s vertices from $\mathcal{H}$. Consider the hypergraph $\mathcal{H}[S] = \mathcal{H} \cap \binom{S}{3}$ induced on S. If $\mathcal{H}[S]$ has $\mathscr{A}$, then the tester says "yes" and otherwise "no." Since $\mathscr{A}$ is decidable and s is fixed, the algorithm decides whether or not $\mathcal{H}[S]$ is in $\mathscr{A}$ in constant time (constant depending only on s and $\mathscr{A}$).

Clearly, if $\mathcal{H} \in \mathscr{A}$ or $n \leq m_0$, then this tester outputs correctly, and hence it is one-sided. On the other hand, if $\mathcal{H}$ is η-far from $\mathscr{A}$ and $n > m_0$, then due to Theorem 3 the random set S spans a copy of $\mathcal{F}_0$ for some $\mathcal{F}_0 \in \mathscr{F}$ on $f_0 \leq C$ vertices, with probability at least

$$cn^{f_0}/\tbinom{n}{f_0} \geq c. \tag{2}$$

Hence the probability that S does not span any copy of $\mathcal{F}_0$ is at most $(1-c)^{s/f_0} \leq (1-c)^{s/C} < 1/3$. In other words, S spans a copy of $\mathcal{F}_0$ with probability at least $2/3$, which shows that the tester works as specified. □

From now on we are concerned only with the proof of Theorem 3. The main philosophy of the proof of Theorem 3 is similar to the corresponding statement for graphs in [5], which was originally obtained by Alon et al. in [2]. The proof requires a strengthening of the hypergraph regularity lemma analogous to the modification of Szemerédi's regularity lemma proved in [2]. A similar lemma for 3-uniform hypergraphs was already proved by Kohayakawa, Nagle, and Rödl [17] based on the regularity lemma for 3-uniform hypergraphs of Frankl and Rödl [13] (see also [24]). We give here a different (and simpler) proof, based on a "cleaner" version of the regularity lemma from [13], which was obtained for general k-uniform hypergraphs by Rödl and Schacht [23]. We call this auxiliary result the *representative lemma* (see Lemma 16 below).

This paper is organized as follows: In section 2, we develop the necessary definitions for the regularity method of 3-uniform hypergraphs. In particular, we state the *hypergraph regularity lemma* (Theorem 13), the corresponding *counting lemma* (Theorem 8), and the *representative lemma* (Lemma 16). Section 3 is devoted to the proof of Theorem 3, and in section 4 we prove the representative lemma.

2. Regularity method for hypergraphs.

In this section we recall some definitions of the hypergraph regularity method following the approach from [13].

2.1. Szemerédi's regularity lemma.

We start the discussion with graphs. Given a graph G and disjoint subsets $X, Y \subseteq V(G)$, the *density* of the pair (X, Y) is

$$d_G(X,Y) = \frac{e_G(X,Y)}{|X||Y|}\,, \tag{3}$$

where $e_G(X, Y)$ denotes the number of edges in G with one vertex in X and one vertex in Y. The pair (X, Y) will be called (ε, d)*-regular* if for every $X' \subseteq X$ and $Y' \subseteq Y$ such that $|X'| \geq \varepsilon|X|$ and $|Y'| \geq \varepsilon|Y|$ we have

$$|d_G(X', Y') - d| < \varepsilon\,. \tag{4}$$

We also say (X, Y) is ε-regular if it is (ε, d)-regular for some d. Roughly speaking, an (ε, d)-regular pair (X, Y) behaves in a similar way to a random bipartite graph on the same vertex sets, where each edge appears with probability d. Szemerédi's regularity lemma [28] states that, for every $\varepsilon > 0$, we can partition the vertex set of any large graph into a bounded number (depending only on ε) of sets such that almost all bipartite graphs between the partition classes are ε-regular.

THEOREM 4 (Szemerédi's regularity lemma [28]). *For any $\varepsilon > 0$ and any integer t_0, there are positive integers $T_0 = T_0(\varepsilon, t_0) > t_0$ and $n_0 = n_0(\varepsilon, t_0)$ such that for every graph $G = (V, E)$ with $|V| = n \geq n_0$ there exists a partition $\mathscr{P}^{(1)} = \{V_1, V_2, \dots, V_t\}$ of V such that*

(i) $t_0 \leq t \leq T_0$,

(ii) $||V_i| - |V_j|| \leq 1$ *for all* $1 \leq i < j \leq t$, *and*

(iii) *all but* εt^2 *pairs* (V_i, V_j) *are* ε*-regular, where* $1 \leq i < j \leq t$.

This lemma is a powerful tool in extremal graph theory (see [18] for a survey or many of its applications). It is often used in conjunction with the so-called counting lemma for graphs. We will later need the simplest form of that lemma for triangles.

LEMMA 5 (triangle counting lemma [18]). *For all constants $\gamma > 0$ and $d > 0$ there exists $\varepsilon_{\mathrm{tcl}} = \varepsilon_{\mathrm{tcl}}(\gamma, d) > 0$ and $m_{\mathrm{tcl}} = m_{\mathrm{tcl}}(f, \gamma, d) \in \mathbb{N}$ such that the following holds. If $\mathcal{P}$ is a tripartite graph with vertex classes V_1, V_2, and V_3 of size $|V_1| = |V_2| = |V_3| = m \geq m_{\mathrm{tcl}}$ and if, moreover, (V_i, V_j) is $(\varepsilon_{\mathrm{tcl}}, d)$-regular for all $1 \leq i < j \leq 3$, then the number of triangles K_3 in $\mathcal{P}$ is in the interval $(1 \pm \gamma) d^3 m^3$.*

An extension of Szemerédi's regularity lemma for 3-uniform hypergraphs has been developed in [13]. More recently extensions to k-uniform hypergraphs were obtained by several authors in [15, 16, 24] and subsequently in [23, 29]. The key feature of all those extensions of Theorem 4 to hypergraphs mentioned above is that it allows one to prove a corresponding extension of the counting lemma, Lemma 5, as shown in [13, 15, 16, 20, 21, 23, 29].

In our proof we will use the regularity lemma and the counting lemma for hypergraphs from [23]. Since, in this paper, we focus only on 3-uniform hypergraphs, we develop the definitions only for that case, following the approach from [13]. Moreover, from now on, by a hypergraph we mean a 3-uniform hypergraph.

2.2. Regular hypergraphs and the counting lemma for hypergraphs. Let V_1, V_2, and V_3 be mutually disjoint subsets of some vertex set V. We call a triple $\hat{\mathcal{Q}} = (Q^{12}, Q^{13}, Q^{23})$ of bipartite graphs with vertex sets $V_1 \cup V_2$, $V_2 \cup V_3$ and $V_1 \cup V_3$ a *triad*. Usually, we will think of a triad $\hat{\mathcal{Q}} = (Q^{12}, Q^{13}, Q^{23})$ as a tripartite graph with vertex set $V_1 \cup V_2 \cup V_3$ and edge set $E(Q^{12}) \cup E(Q^{13}) \cup E(Q^{23})$. For the regularity of hypergraphs, triads play the same role as pairs of vertex sets in Szemerédi's regularity lemma.

For a triad $\hat{\mathcal{Q}} = (Q^{12}, Q^{13}, Q^{23})$ with vertex set $V_1 \cup V_2 \cup V_3$ we define $\mathrm{Tr}(\hat{\mathcal{Q}})$ as the set of triples of vertices of $\hat{\mathcal{Q}}$ each inducing a triangle in $\hat{\mathcal{Q}}$:

$$\mathrm{Tr}(\hat{\mathcal{Q}}) = \left|\left\{\{v_1, v_2, v_3\}\colon\ v_i \in V_i \text{ and } v_i v_j \in E(Q^{ij}) \text{ for all } 1 \leq i < j \leq 3\right\}\right|.$$

For a hypergraph $\mathcal{H}$ on some vertex set V and a triad $\hat{\mathcal{Q}}$ with vertex classes V_1, V_2, and $V_3 \subset V$, we define the *density of $\mathcal{H}$ on the triad $\hat{\mathcal{Q}}$* as

$$d_{\mathcal{H}}(\hat{\mathcal{Q}}) = \begin{cases} \frac{|\mathcal{H} \cap \mathrm{Tr}(\hat{\mathcal{Q}})|}{|\mathrm{Tr}(\hat{\mathcal{Q}})|} & \text{if } |\mathrm{Tr}(\hat{\mathcal{Q}})| > 0\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{5}$$

This is a natural extension of the notion of density from graphs w.r.t. pairs (see (3)) to hypergraphs w.r.t. triads. We generalize the last definition to the density of an r-tuple of subtriads of a given triad. We say a tripartite graph $\hat{\mathcal{X}} = (X^{12}, X^{13}, X^{23})$ with vertex sets W_1, W_2, and W_3 is a subtriad of a triad $\hat{\mathcal{Q}} = (Q^{12}, Q^{13}, Q^{23})$ with vertex sets $V_1 \supseteq W_1$, $V_2 \supseteq W_2$, and $V_3 \supseteq W_3$ if for every $1 \leq i < j \leq 3$ we have $E(X^{ij}) \subseteq E(Q^{ij})$. For a given triad $\hat{\mathcal{Q}} = (Q^{12}, Q^{13}, Q^{23})$ and a family of (not necessarily disjoint) subtriads $\hat{\boldsymbol{\mathcal{X}}} = \{\hat{\mathcal{X}}_s = (X_s^{12}, X_s^{13}, X_s^{23})\colon\ s = 1, \dots, r\}$ we define

$$\operatorname{Tr}(\hat{\boldsymbol{\mathcal{X}}}) = \bigcup_{i=1}^{r} \operatorname{Tr}(\hat{\mathcal{X}}_i)$$

and extend (5) by setting

$$d_{\mathcal{H}}(\hat{\boldsymbol{\mathcal{X}}}) = \begin{cases} \frac{|\mathcal{H} \cap \operatorname{Tr}(\hat{\boldsymbol{\mathcal{X}}})|}{|\operatorname{Tr}(\hat{\boldsymbol{\mathcal{X}}})|} & \text{if } |\operatorname{Tr}(\hat{\boldsymbol{\mathcal{X}}})| > 0\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

We now proceed to a central definition and extend the notion of a regular pair to a regular triad.

Definition 6 ((δ, d, r)-regularity). *Let $\delta > 0$, $d > 0$, and $r \in \mathbb{N}$. We say a hypergraph $\mathcal{H}$ is (δ, d, r)-regular with respect to a triad $\hat{\mathcal{Q}} = (Q^{12}, Q^{13}, Q^{23})$ on the vertex sets V_1, V_2, and $V_3 \subseteq V(\mathcal{H})$ if for any family of r subtriads $\hat{\boldsymbol{\mathcal{X}}} = \{\hat{\mathcal{X}}_s\colon\ s = 1, \dots, r\}$ satisfying*

$$|\operatorname{Tr}(\hat{\boldsymbol{\mathcal{X}}})| > \delta |\operatorname{Tr}(\hat{\mathcal{Q}})|$$

we have

$$|d_{\mathcal{H}}(\hat{\boldsymbol{\mathcal{X}}}) - d| < \delta\,.$$

This notion was introduced in [13] and, similar to Szemerédi's regularity lemma, decomposes every graph in a bounded number of "mostly" regular pairs; the hypergraph regularity lemma (Theorem 13) will partition the edge set of any given hypergraph into "triads" in such a way that most of them are regular in the sense of Definition 6. In order to simplify the notation we sometimes do not specify the density d. We will say a hypergraph is $(\delta, *, r)$*-regular* if it is (δ, d, r)-regular for some density d.

The counting lemma for hypergraphs is a crucial tool in our proof of Theorem 3. It ensures the existence of many copies of a fixed small hypergraph inside a larger, dense and "sufficiently regular" hypergraph $\mathcal{H}$. We need a few more definitions before we give the precise statement.

Let $V_1 \cup V_2 \cup \dots \cup V_f$ be a partition of some vertex set V. We denote by $K_f(V_1, \dots, V_f)$ the complete f-partite graph on that partition. Let R be any f-partite subgraph of $K_f(V_1, \dots, V_f)$ on the same vertex partition, and as above, let $\operatorname{Tr}(R)$ be the set of those 3-element subsets of V, which span a K_3 in R. We say R *underlies a hypergraph* $\mathcal{H}$ on the same vertex set V if $\mathcal{H} \subseteq \operatorname{Tr}(R)$. This leads to the notion of a *regular complex*.

Definition 7 (regular complex). *Let positive integers f, m, and $r \in \mathbb{N}$ and positive constants δ_2, δ_3, d_2, $d_3 > 0$ be given. Suppose $\mathcal{F}$ is a hypergraph with vertex set $[f] = \{1, 2, \dots, f\}$, $V_1 \cup V_2 \cup \dots \cup V_f$ is a partition of some vertex set V, $R \subseteq K_f(V_1, \dots, V_f)$, and R underlies a hypergraph $\mathcal{H}$ with vertex set $V(\mathcal{H}) = V$. We say the pair $(R, \mathcal{H})$ is a $(\delta_2, \delta_3, d_2, d_3, r)$-regular $(m, \mathcal{F})$-complex if the following holds:*

(i) *$|V_i| = m$ for all $i = 1, \dots, f$;*

(ii) *for every $1 \le i < j \le f$ such that $\{i, j, k\} \in \mathcal{F}$ for some $k \in [f]$, the induced subgraph $R^{ij} = R[V_i, V_j]$ of R on the vertex sets V_i and V_j is (δ_2, d_2)-regular; and*

(iii) *for every $\{i, j, k\} \in \mathcal{F}$ the hypergraph $\mathcal{H}$ is (δ_3, d_{ijk}, r)-regular w.r.t. the triad $\hat{\mathcal{R}} = (R^{ij}, R^{ik}, R^{jk})$ for some $d_{ijk} \ge d_3$.*

The counting lemma for hypergraphs extends Lemma 5 and gives a bound on the number of copies of a fixed hypergraph $\mathcal{F}$ in $\mathcal{H}$ for sufficiently $(\delta_2, \delta_3, d_2, d_3, r)$-regular $(m, \mathcal{F})$-complexes $(R, \mathcal{H})$.

THEOREM 8 (counting lemma for 3-uniform hypergraphs [20]). *For every $f \in \mathbb{N}$ and constants $\gamma > 0$ and $d_3 > 0$, there exist $\delta_3 = \delta_3(f, \gamma, d_3) > 0$ such that for every $d_2 > 0$ there exist $\delta_2 = \delta_2(f, \gamma, d_3, d_2) > 0$, and positive integers $r = r(f, \gamma, d_3, d_2)$ and $m_0 = m_0(f, \gamma, d_3, d_2) \in \mathbb{N}$ such that the following holds.*

Suppose $\mathcal{F}$ is a hypergraph with vertex set $[f] = \{1, \dots, f\}$, $V_1 \cup V_2 \cup \dots \cup V_f$ is a partition of some vertex set V, $R \subseteq K_f(V_1, \dots, V_f)$, and R underlies a hypergraph $\mathcal{H}$ with vertex set $V(\mathcal{H}) = V$. If, moreover, $(R, \mathcal{H})$ is a $(\delta_2, \delta_3, d_2, d_3, r)$-regular $(m, \mathcal{F})$-complex with $m \geq m_0$, then the number of copies of $\mathcal{F}$ in $\mathcal{H}$ is at least

$$(1 - \gamma) d_2^{|\Delta(\mathcal{F})|} d_3^{|\mathcal{F}|} m^f \,, \tag{6}$$

where $\Delta(\mathcal{F})$ is the shadow of $\mathcal{F}$, i.e.,

$$\Delta(\mathcal{F}) = \big\{\{i, j\} \colon\ 1 \leq i < j \leq f \text{ so that there exists } k \in [f] \text{ with } \{i, j, k\} \in \mathcal{F}\big\} .$$

A generalization of this counting lemma to k-uniform hypergraphs can be found in [21] and [23].

2.3. Regularity lemma for hypergraphs. In this section we state a variant of the regularity for 3-uniform hypergraphs [13], which was obtained by the Rödl and Schacht for general k-uniform hypergraphs in [23]. First we generalize the concept of vertex partition present in Szemerédi's regularity lemma.

DEFINITION 9 ((t, ℓ)-partition). *Let V be a vertex set, $\mathscr{P}^{(1)} = \{V_1, V_2, \dots, V_t\}$ be a partition of V, and $\mathscr{P}^{(2)} = \{P^{ij}_\alpha \colon\ 1 \leq i < j \leq t \text{ and } 1 \leq \alpha \leq \ell\}$ be a family of $\binom{t}{2}\ell$ bipartite graphs. We say the pair $\mathscr{P} = \{\mathscr{P}^{(1)}, \mathscr{P}^{(2)}\}$ is a (t, ℓ)-partition*[1] *on V if for every $1 \leq i < j \leq t$ the family $\{E(P^{ij}_1), E(P^{ij}_2), \dots, E(P^{ij}_\ell)\}$ is a partition of the edge set of the complete bipartite graph $K_2(V_i, V_j)$.*

We say a (t, ℓ)-partition is T-bounded if $\max\{t, \ell\} \leq T$. *Moreover, for a (t, ℓ)-partition $\mathscr{P}$, we denote by $\hat{\mathscr{P}}$ the set of all triads of the form $(P^{ij}_\alpha, P^{ik}_\beta, P^{jk}_\gamma)$ with $1 \leq \alpha, \beta, \gamma \leq \ell$ and $1 \leq i < j < k \leq t$.*

We consider such (t, ℓ)-partitions for which the bipartite graphs P^{ij}_α are μ-regular. Moreover, as in Szemerédi's regularity lemma we will require the vertex partition classes to have almost the same size. This leads us to the following definition.

DEFINITION 10 ((μ, t, ℓ)-equitable). *We say a (t, ℓ)-partition $\mathscr{P} = \{\mathscr{P}^{(1)}, \mathscr{P}^{(2)}\}$ is (μ, t, ℓ)-equitable if*

(i) *$\mathscr{P}^{(1)} = \{V_1, V_2, \dots, V_t\}$ is equitable, i.e., for all $1 \leq i < j \leq t$ we have $||V_i| - |V_j|| \leq 1$, and*

(ii) *for every $1 \leq i < j \leq t$ and $1 \leq \alpha \leq \ell$ the bipartite graph $P^{ij}_\alpha \in \mathscr{P}^{(2)}$ is $(\mu, 1/\ell)$-regular on the pair (V_i, V_j).*

The regularity lemma from [23] guarantees the existence of a T-bounded (μ, t, ℓ)-equitable partition $\mathscr{P}$ (where $\mu = \mu(t, \ell)$ is any function of t and ℓ) for any hypergraph $\mathcal{H}$, where T is independent of the number of vertices of $\mathcal{H}$. Moreover, the hypergraph $\mathcal{H}$ will be $(\delta, *, r)$-regular w.r.t. almost all triads of $\hat{\mathscr{P}}$.

THEOREM 11 (regularity lemma for 3-uniform hypergraphs [23]). *For every integer $t_0 \in \mathbb{N}$, every constant $\delta_{\mathscr{P}} > 0$, and all functions $\mu_{\mathscr{P}} \colon \mathbb{N}^2 \to (0, 1]$ and $r_{\mathscr{P}} \colon \mathbb{N}^2 \to$*

[1] Note that while $\mathscr{P}^{(1)}$ is a partition of V, the family of bipartite graphs $\mathscr{P}^{(2)}$ is not a partition of $\binom{V}{2}$ but a partition of the edge set of the complete t-partite graph $K_t(V_1, \dots, V_t)$.

$\mathbb{N}$, *there exist positive integers* $T_0 = T_0(t_0, \delta_{\mathscr{P}}, \mu_{\mathscr{P}}, r_{\mathscr{P}})$ *and* $n_0 = n_0(t_0, \delta_{\mathscr{P}}, \mu_{\mathscr{P}}, r_{\mathscr{P}})$ *such that for every hypergraph* $\mathcal{H}$ *with* $n \geq n_0$ *vertices* V *there exists a partition* $\mathscr{P}$*, and there are positive integers* $t_{\mathscr{P}}$ *and* $\ell_{\mathscr{P}}$ *such that for* $\mu_{\mathscr{P}} = \mu_{\mathscr{P}}(t_{\mathscr{P}}, \ell_{\mathscr{P}})$ *and* $r_{\mathscr{P}} = r_{\mathscr{P}}(t_{\mathscr{P}}, \ell_{\mathscr{P}})$ *the following holds:*

(i) $\mathscr{P}$ *is* $(\mu_{\mathscr{P}}, t_{\mathscr{P}}, \ell_{\mathscr{P}})$*-equitable and a* T_0*-bounded partition on* V*; and*

(ii) $\mathcal{H}$ *is* $(\delta_{\mathscr{P}}, *, r_{\mathscr{P}})$*-regular w.r.t. all but at most* $\delta_{\mathscr{P}} t_{\mathscr{P}}^3 \ell_{\mathscr{P}}^3$ *triads* $\hat{\mathcal{P}} \in \hat{\mathscr{P}}$.

In our proof we will use the regularity lemma twice. First we use it in the form as stated above, and in the second application we will refine the given partition $\mathscr{P}$ to obtain a partition $\mathscr{Q}$ w.r.t. which $\mathcal{H}$ will be "more regular." To state that version we need the notion of a refinement of a partition.

Definition 12 (refinement). *We say a partition* $\mathscr{Q} = \{\mathscr{Q}^{(1)}, \mathscr{Q}^{(2)}\}$ *on* V *refines a partition* $\mathscr{P} = \{\mathscr{P}^{(1)}, \mathscr{P}^{(2)}\}$ *on* V *and write* $\mathscr{Q} \prec \mathscr{P}$ *if*

(i) *for every vertex set* $U \in \mathscr{Q}^{(1)}$ *there exists* $W \in \mathscr{P}^{(1)}$ *such that* $U \subseteq W$*, and*

(ii) *for every bipartite graph* $Q \in \mathscr{Q}^{(2)}$ *there exists* $P \in \mathscr{P}^{(2)}$ *such that* Q *is a subgraph of* P.

We now state that *refinement version* of Theorem 11. In fact, Theorem 11 is a simple corollary of the refinement version, and a proof of that stronger version can be found in [23]. The lemma roughly states that given a $(\mu, t_{\mathscr{P}}, \ell_{\mathscr{P}})$-equitable partition $\mathscr{P}$ (with $(\mu, 1/\ell_{\mathscr{P}})$-regular auxiliary graphs $P_\alpha^{ij} \in \mathscr{P}^{(2)}$ for sufficiently small μ) any hypergraph $\mathcal{H}$ admits a partition $\mathscr{Q} \prec \mathscr{P}$ for which $\mathcal{H}$ is $(\delta, *, r)$-regular on most triads $\hat{\mathcal{Q}} \in \hat{\mathscr{Q}}$.

Theorem 13 (refinement version of the regularity lemma [23]). *For all positive integers* $t_{\mathscr{P}}$*,* $\ell_{\mathscr{P}} \in \mathbb{N}$*, every constant* $\delta_{\mathscr{Q}} > 0$*, and all functions* $\varepsilon_{\mathscr{Q}}\colon \mathbb{N}^2 \to (0,1]$ *and* $r_{\mathscr{Q}}\colon \mathbb{N}^2 \to \mathbb{N}$*, there exist* $\mu_{\mathrm{hrl}} = \mu_{\mathrm{hrl}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}, \delta_{\mathscr{Q}}, \varepsilon_{\mathscr{Q}}, r_{\mathscr{Q}}) > 0$ *and positive integers* $T_{\mathrm{hrl}} = T_{\mathrm{hrl}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}, \delta_{\mathscr{Q}}, \varepsilon_{\mathscr{Q}}, r_{\mathscr{Q}})$ *and* $n_{\mathrm{hrl}} = n_{\mathrm{hrl}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}, \delta_{\mathscr{Q}}, \varepsilon_{\mathscr{Q}}, r_{\mathscr{Q}})$ *such that the following holds. If*

(a) $\mathcal{H}$ *is a hypergraph with* $n \geq n_{\mathrm{hrl}}$ *vertices* V*, and*

(b) $\mathscr{P}$ *is a* $(\mu_{\mathrm{hrl}}, t_{\mathscr{P}}, \ell_{\mathscr{P}})$*-equitable (and hence* $\max\{t_{\mathscr{P}}, \ell_{\mathscr{P}}\}$*-bounded) partition on* V*,*

then there exists a partition $\mathscr{Q}$ *and there are positive integers* $t_{\mathscr{Q}}$ *and* $\ell_{\mathscr{Q}}$ *such that the following holds for* $t_{\mathscr{P}\mathscr{Q}} = t_{\mathscr{P}} t_{\mathscr{Q}}$*,* $\ell_{\mathscr{P}\mathscr{Q}} = \ell_{\mathscr{P}} \ell_{\mathscr{Q}}$*,* $\varepsilon_{\mathscr{Q}} = \varepsilon_{\mathscr{Q}}(t_{\mathscr{P}\mathscr{Q}}, \ell_{\mathscr{P}\mathscr{Q}})$*, and* $r_{\mathscr{Q}} = r_{\mathscr{Q}}(t_{\mathscr{P}\mathscr{Q}}, \ell_{\mathscr{P}\mathscr{Q}})$*:*

(i) $\mathscr{Q}$ *is* $(\varepsilon_{\mathscr{Q}}, t_{\mathscr{P}\mathscr{Q}}, \ell_{\mathscr{P}\mathscr{Q}})$*-equitable and* T_{hrl}*-bounded partition on* V*;*

(ii) $\mathscr{Q} \prec \mathscr{P}$*; and*

(iii) $\mathcal{H}$ *is* $(\delta_{\mathscr{Q}}, *, r_{\mathscr{Q}})$*-regular w.r.t. all but at most* $\delta_{\mathscr{Q}} t_{\mathscr{P}\mathscr{Q}}^3 \ell_{\mathscr{P}\mathscr{Q}}^3$ *triads* $\hat{\mathcal{Q}} \in \hat{\mathscr{Q}}$.

2.4. Statement of the representative lemma for hypergraphs. We now turn to the key definition of a *representative* of a (t, ℓ)-partition $\mathscr{P}$. Roughly speaking, a representative is a subobject of a (t, ℓ)-partition $\mathscr{P}$ reflecting the structure of $\mathscr{P}$.

Definition 14 (representative). *Let* $\mathscr{P} = \{\mathscr{P}^{(1)}, \mathscr{P}^{(2)}\}$ *be a* (t, ℓ)*-partition on* V *with vertex partition* $\mathscr{P}^{(1)} = \{V_1, V_2, \ldots, V_t\}$ *and* $\mathscr{P}^{(2)} = \{P_\alpha^{ij}\colon 1 \leq i < j \leq t \text{ and } 1 \leq \alpha \leq \ell\}$*. We say* $\mathscr{R} = \{\mathscr{R}^{(1)}, \mathscr{R}^{(2)}\}$*, where* $\mathscr{R}^{(1)} = \{W_1, W_2, \ldots, W_t\}$ *and* $\mathscr{R}^{(2)} = \{R_\alpha^{ij}\colon 1 \leq i < j \leq t \text{ and } 1 \leq \alpha \leq \ell\}$ *is a representative of* $\mathscr{P}$ *(or* $\mathscr{R}$ *represents* $\mathscr{P}$*) if*

(i) $W_i \subseteq V_i$ *for every* $1 \leq i \leq t$*, and*

(ii) R_α^{ij} *is a (bipartite) subgraph of* P_α^{ij} *with vertex classes* W_i *and* W_j *for every* $1 \leq i < j \leq t$ *and* $1 \leq \alpha \leq \ell$.

Moreover, we define for every triad $\hat{\mathcal{P}} = (P_\alpha^{ij}, P_\beta^{ik}, P_\gamma^{jk}) \in \hat{\mathscr{P}}$ *the corresponding triad* $\hat{\mathcal{R}}(\hat{\mathcal{P}}) = (R_\alpha^{ij}, R_\beta^{ik}, R_\gamma^{jk})$*, and we let* $\hat{\mathscr{R}} = \{\hat{\mathcal{R}}(\hat{\mathcal{P}})\colon \hat{\mathcal{P}} \in \hat{\mathscr{P}}\}$ *be the family of triads of the representative.*

In our proof of Theorem 3 the representative $\mathscr{R}$ of the partition $\mathscr{P}$ will be appropriately chosen from an equitable refinement $\mathscr{Q}$ of $\mathscr{P}$ (cf. Theorem 13), and hence, $\mathscr{R}$ will be equitable in the following sense.

Definition 15 ($(\varepsilon_{\mathscr{R}}, t_{\mathscr{R}}, \ell_{\mathscr{R}})$-representative). *Let $\varepsilon_{\mathscr{R}} > 0$ and positive integers $t_{\mathscr{R}}$ and $\ell_{\mathscr{R}} \in \mathbb{N}$ be given. We say a representative $\mathscr{R} = \{\mathscr{R}^{(1)}, \mathscr{R}^{(2)}\}$ of a $(t_{\mathscr{P}}, \ell_{\mathscr{P}})$-partition $\mathscr{P}$ on n vertices is an $(\varepsilon_{\mathscr{R}}, t_{\mathscr{R}}, \ell_{\mathscr{R}})$-representative if*

(i) *$|W| = n/(t_{\mathscr{P}} t_{\mathscr{R}})$ for every $W \in \mathscr{R}^{(1)}$, and*

(ii) *R is $(\varepsilon_{\mathscr{R}}, 1/(\ell_{\mathscr{P}} \ell_{\mathscr{R}}))$-regular for every $R \in \mathscr{R}^{(2)}$.*

We say the $(\varepsilon_{\mathscr{R}}, t_{\mathscr{R}}, \ell_{\mathscr{R}})$-representative $\mathscr{R}$ of a $(t_{\mathscr{P}}, \ell_{\mathscr{P}})$-partition $\mathscr{P}$ is T-bounded for some $T \in \mathbb{N}$ if $\max\{t_{\mathscr{P}}, t_{\mathscr{R}}, \ell_{\mathscr{P}}, \ell_{\mathscr{R}}\} \le T$.

Recall that due to the quantification of the regularity lemma (Theorem 11), which states that for every $\delta_{\mathscr{P}}$ there exists T_0, the resulting T_0-bounded partition $\mathscr{P}$ may satisfy $t_{\mathscr{P}}, \ell_{\mathscr{P}} \gg 1/\delta_{\mathscr{P}}$. This would, however, not suffice to count hypergraphs $\mathcal{F}$ of a size comparable to $t_{\mathscr{P}}$ or $\ell_{\mathscr{P}}$, the number of the blocks in the partition $\mathscr{P}$. That is due to the quantification of the counting lemma (Theorem 8), which for a given hypergraph $\mathcal{F}$ of size f ensures the existence of sufficiently small $\delta \ll 1/f$ (δ_3 in the statement).

To circumvent a similar problem arising in the graph case, Alon and Shapira [5] used an iterated version of Szemerédi's regularity lemma, which was first obtained and used by Alon et al. in [2]. This iterated regularity lemma yields for a given graph G a vertex partition $\mathscr{P}^{(1)} = \{V_1, V_2, \dots, V_t\}$ and a representative $\mathscr{R}^{(1)} = \{W_1, W_2, \dots, W_t\}$ with $W_i \subseteq V_i$ for every $i = 1, 2, \dots, t$. In that lemma the representative $\mathscr{R}^{(1)}$ resembles *typically* the density of G w.r.t. $\mathscr{P}$, i.e., $d_G(W_i, W_j) \sim d_G(V_i, V_j)$ for *most* pairs $1 \le i < j \le t$. Moreover, the graph G is ε-regular on *every* pair (W_i, W_j) of the representative, and (most importantly) ε can be chosen as an arbitrary function of t, e.g., on the representative one can count graphs of order t, i.e., the size of the partition $\mathscr{P}^{(1)}$.

The representative lemma, Lemma 16 below, is an analogous statement for 3-uniform hypergraphs. For a given hypergraph $\mathcal{H}$ it asserts the existence of a partition $\mathscr{P} = \{\mathscr{P}^{(1)}, \mathscr{P}^{(2)}\}$ and of a representative $\mathscr{R} = \{\mathscr{R}^{(1)}, \mathscr{R}^{(2)}\}$ of $\mathscr{P}$ so that $\mathcal{H}$ is $(\delta_{\mathscr{R}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}), *, r_{\mathscr{R}}(t_{\mathscr{P}}, t_{\mathscr{R}}, \ell_{\mathscr{P}}, \ell_{\mathscr{R}}))$-regular on *every* triad of the representative (see (iv) in Lemma 16). Note that the number of *partition blocks* in $\mathscr{R}$, which is the same as that in the partition $\mathscr{P}$, depends on $t_{\mathscr{P}}$ and $\ell_{\mathscr{P}}$ only, and here $\delta_{\mathscr{R}}(t_{\mathscr{P}}, \ell_{\mathscr{P}})$ is a function of those parameters. On the other hand, the functions $r_{\mathscr{R}}(t_{\mathscr{P}}, t_{\mathscr{R}}, \ell_{\mathscr{P}}, \ell_{\mathscr{R}})$ and $\varepsilon_{\mathscr{R}}(t_{\mathscr{P}}, t_{\mathscr{R}}, \ell_{\mathscr{P}}, \ell_{\mathscr{R}})$ can depend on $t_{\mathscr{P}}$, $t_{\mathscr{R}}$, $\ell_{\mathscr{P}}$, and $\ell_{\mathscr{R}}$, so in particular, they can depend on $\ell_{\mathscr{P}} \ell_{\mathscr{R}}$, which is the reciprocal of the densities of $R \in \mathscr{R}^{(2)}$. Choosing $\delta_{\mathscr{R}}(t_{\mathscr{P}}, \ell_{\mathscr{P}})$ and $r_{\mathscr{R}}(t_{\mathscr{P}}, t_{\mathscr{R}}, \ell_{\mathscr{P}}, \ell_{\mathscr{R}})$ appropriately as functions of $t_{\mathscr{P}}$, $\ell_{\mathscr{P}}$, $t_{\mathscr{R}}$, and $\ell_{\mathscr{R}}$ will allow us to satisfy the quantification of the counting lemma, Theorem 8, for counting hypergraphs $\mathcal{F}$ whose size depends on $t_{\mathscr{P}}$ and $\ell_{\mathscr{P}}$. Additionally, we will also ensure that $d_{\mathcal{H}}(\hat{\mathcal{R}}(\hat{\mathcal{P}})) \sim d_{\mathcal{H}}(\hat{\mathcal{P}})$ for *most* triads $\hat{\mathcal{P}} \in \hat{\mathscr{P}}$ (see (iii) in Lemma 16).

Lemma 16 (representative lemma). *For every $t_1 \in \mathbb{N}$ and $\delta > 0$ and all functions $\varepsilon_{\mathscr{P}} : \mathbb{N}^2 \to (0,1]$ $\delta_{\mathscr{R}} : \mathbb{N}^2 \to (0,1]$, $\varepsilon_{\mathscr{R}} : \mathbb{N}^4 \to (0,1]$, and $r_{\mathscr{R}} : \mathbb{N}^4 \to \mathbb{N}$, there exist positive integers $T_{\mathscr{P}}$, $T_{\mathscr{P}\mathscr{R}}$, and $n_1 \in \mathbb{N}$ such that the following holds.*

If $\mathcal{H}$ is a hypergraph on at least n_1 vertices, then there exist positive integers $t_{\mathscr{P}}$ and $\ell_{\mathscr{P}}$ and a $T_{\mathscr{P}}$-bounded $(t_{\mathscr{P}}, \ell_{\mathscr{P}})$-partition $\mathscr{P} = \{\mathscr{P}^{(1)}, \mathscr{P}^{(2)}\}$ with $t_1 \le t_{\mathscr{P}}$, and there is representative $\mathscr{R} = \{\mathscr{R}^{(1)}, \mathscr{R}^{(2)}\}$ of $\mathscr{P}$ such that

(i) *every graph in $P \in \mathscr{P}^{(2)}$ is $(\varepsilon_{\mathscr{P}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}), 1/\ell_{\mathscr{P}})$-regular;*

(ii) *$\mathscr{R}$ is a $T_{\mathscr{P}\mathscr{R}}$-bounded $(\varepsilon_{\mathscr{R}}(t_{\mathscr{P}}, t_{\mathscr{R}}, \ell_{\mathscr{P}}, \ell_{\mathscr{R}}), t_{\mathscr{R}}, \ell_{\mathscr{R}})$-representative of $\mathscr{P}$ for some positive integers $t_{\mathscr{R}}$ and $\ell_{\mathscr{R}}$;*

(iii) *for all but at most $\delta t_{\mathscr{P}}^3 \ell_{\mathscr{P}}^3$ triads $\hat{\mathcal{P}} \in \hat{\mathscr{P}}$ we have $|d_{\mathcal{H}}(\hat{\mathcal{P}}) - d_{\mathcal{H}}(\hat{\mathcal{R}}(\hat{\mathcal{P}}))| \le \delta$; and*

(iv) *$\mathcal{H}$ is $(\delta_{\mathscr{R}}(t_{\mathscr{P}},\ell_{\mathscr{P}}),*,r_{\mathscr{R}}(t_{\mathscr{P}},t_{\mathscr{R}},\ell_{\mathscr{P}},\ell_{\mathscr{R}}))$-regular w.r.t. $\hat{\mathcal{R}}(\hat{\mathcal{P}})\in\hat{\mathscr{R}}$ for every triad $\hat{\mathcal{P}}\in\hat{\mathscr{P}}$.*

A similar lemma was proved by Kohayakawa, Nagle, and Rödl [17]. In section 4 we give a different proof of the representative lemma based on Theorem 13.

3. Proof of Theorem 3. A weakened version of Theorem 3 is obtained by restricting the theorem to *finite* forbidden families $\mathscr{F}$. In that case the corresponding statement of Theorem 3 has essentially been proved for k-uniform hypergraphs in [15, 21, 25]. We briefly outline that proof in section 3.1 and discuss its limitations w.r.t. *infinite* families $\mathscr{F}$. The representative lemma, Lemma 16, will allow us to overcome those difficulties, and in section 3.2 we give a proof of Theorem 3 based on Lemma 16 and the counting lemma, Theorem 8.

3.1. The finite case. We now sketch a (straightforwardly adjusted) proof of the removal lemma [15, 21, 25], based on the regularity and the counting lemma for hypergraphs, which yields the restricted version of Theorem 3 for finite forbidden families $\mathscr{F}$.

Let $\mathscr{F}$ be a finite family of hypergraphs and $\eta>0$ be given, and consider an n-vertex hypergraph $\mathcal{H}$ which is η-far from $\mathscr{A}=\text{Forb}(\mathscr{F})$. We apply Theorem 11 with appropriately chosen parameters $\delta_{\mathscr{P}}$ and functions $\mu_{\mathscr{P}}$ and $r_{\mathscr{P}}$ (discussed below). This way we obtain a partition $\mathscr{P}=\{\mathscr{P}^{(1)},\mathscr{P}^{(2)}\}$. We then delete those edges H of $\mathcal{H}$ which satisfy one of the following conditions:

(a) H is *noncrossing* in $\mathscr{P}$, i.e., there exist $V_i\in\mathscr{P}^{(1)}$ so that $|H\cap V_i|\geq 2$;

(b) H belongs to a *sparse* triad, i.e., $d_{\mathcal{H}}(\hat{\mathcal{P}})<\eta/3$ for the unique $\hat{\mathcal{P}}\in\hat{\mathscr{P}}$, with $H\in\text{Tr}(\hat{\mathcal{P}})$; or

(c) H belongs to a $(\delta_{\mathscr{P}},*,r_{\mathscr{P}}(t_{\mathscr{P}},\ell_{\mathscr{P}}))$-*irregular* triad, i.e., the hypergraph $\mathcal{H}$ is not $(\delta_{\mathscr{P}},*,r_{\mathscr{P}}(t_{\mathscr{P}},\ell_{\mathscr{P}}))$-regular w.r.t. the unique $\hat{\mathcal{P}}\in\hat{\mathscr{P}}$, with $H\in\text{Tr}(\hat{\mathcal{P}})$.

We call the resulting hypergraph $\mathcal{H}'$. Choosing $\delta_{\mathscr{P}}<\eta/3$ and provided the regular partition has sufficiently many vertex classes (which implies that only a few tuples, e.g., less than $\eta n^3/3$, are deleted because of (a)), one can show that at most ηn^3 edges of $\mathcal{H}$ were deleted. Since by assumption $\mathcal{H}$ is η-far from $\mathscr{A}$, the hypergraph $\mathcal{H}'$ still does not satisfy $\mathscr{A}$. Consequently, $\mathcal{H}'$ contains a subhypergraph $\mathcal{F}_0$ isomorphic to some forbidden hypergraph from $\mathscr{F}$. Due to the construction of $\mathcal{H}'$ all edges of $\mathcal{F}_0$ are crossing w.r.t. the vertex partition $\mathscr{P}^{(1)}$ and belong to dense and regular triads from $\hat{\mathscr{P}}$. Suppose now that the entire copy $\mathcal{F}_0$ is crossing w.r.t. $\mathscr{P}^{(1)}$, i.e., $V(\mathcal{F}_0)$ intersects each vertex partition class $V_i\in\mathscr{P}^{(1)}$ in at most one vertex. (The case when $\mathcal{F}_0$ is not crossing can be handled similarly, as we will show in the general proof for not necessarily finite families $\mathscr{F}$.)

Since each edge of $\mathcal{F}_0$ belongs to a dense and regular triad, the union of those triads defines a dense and regular $(m,\mathcal{F}_0)$-complex (see Definition 7) with $m=n/t_{\mathscr{P}}$. Moreover, $\max_{\mathcal{F}\in\mathscr{F}}|V(\mathcal{F})|$ exists since $|\mathscr{F}|<\infty$. In other words, we can forecast the maximum possible size of $\mathcal{F}_0$ we may encounter, and we can choose $\delta_{\mathscr{P}}$ and functions $\mu_{\mathscr{P}}$ and $r_{\mathscr{P}}$ at the beginning of the proof appropriately so that the $(m,\mathcal{F}_0)$-complex from above is ready for an application of the counting lemma, Theorem 8. The counting lemma then guarantees $\Omega(n^{|V(\mathcal{F}_0)|})$ copies of $\mathcal{F}_0$ in the $(m,\mathcal{F}_0)$-complex and, consequently, in $\mathcal{H}'\subseteq\mathcal{H}$, which concludes the proof.

Clearly, this argument breaks down for infinite families $\mathscr{F}$, as we cannot forecast an upper bound on the size of $\mathcal{F}_0$. The representative lemma, Lemma 16, allows us to get around this issue. Given a hypergraph $\mathcal{H}$ we apply Lemma 16 and delete noncrossing edges and edges belonging to triads $\hat{\mathcal{P}}$ for which $\hat{\mathcal{R}}(\hat{\mathcal{P}})$ is sparse; similarly, as in the discussion above (and also using (iii) of Lemma 16), we will be left with a hypergraph $\mathcal{H}'$ which again contains a hypergraph $\mathcal{F}_0$ from the forbidden family $\mathscr{F}$. In this (infinite) case we have no upper bound on the size of $\mathcal{F}_0$. However, since all edges of $\mathcal{F}_0$ belong to triads of the $(t_{\mathscr{P}}, \ell_{\mathscr{P}})$-partition $\mathscr{P}$, we will argue that $\mathcal{H}'$ also contains some other forbidden hypergraph $\mathcal{F}_1$ of $\mathscr{F}$, whose edges belong to the same triads as the edges of $\mathcal{F}_0$ and, more importantly, the size of $\mathcal{F}_1$ will be bounded by a function depending only on $t_{\mathscr{P}}$ and $\ell_{\mathscr{P}}$. (Roughly speaking, the $\mathcal{F}_0$ and $\mathcal{F}_1$ can both be homomorphically mapped in the "cluster-structure" of the partition $\mathscr{P}$, and the size of $\mathcal{F}_1$ depends only on the number of partition blocks of $\mathscr{P}$.) This will allow us, with appropriately chosen functions $\delta_{\mathscr{R}}(t_{\mathscr{P}}, \ell_{\mathscr{P}})$ and $r_{\mathscr{R}}(t_{\mathscr{P}}, t_{\mathscr{R}}, \ell_{\mathscr{P}}, \ell_{\mathscr{R}})$ for the regularity of the representative, to find (using the counting lemma) $\Omega(n^{|V(\mathcal{F}_1)|})$ copies of $\mathcal{F}_1$ in $\mathcal{H}' \subseteq \mathcal{H}$. We now give the details of this outline.

3.2. The general case. In this section we make the ideas presented in the outline above precise. For that we need a few more definitions. For positive integers t and ℓ we denote by $M(t, \ell)$ the complete multigraph with vertex set $[t] = \{1, \dots, t\}$ and edge multiplicity ℓ. We can view edges as ordered pairs $(\{i, j\}, \alpha)$, where $1 \le i < j \le t$ and $\alpha \in [l]$. We denote by $\mathrm{Tr}(M(t, \ell))$ the set of all $\binom{t}{3}\ell^3$ triangles of $M(t, \ell)$. We will identify a triangle on the vertices $1 \le i < j < k \le t$ and edges $(\{i, j\}, \alpha)$, $(\{i, k\}, \beta)$, $(\{j, k\}, \gamma)$ with the 6-tuple $((i, j, k), (\alpha, \beta, \gamma))$ and set

$$\mathrm{Tr}\left(M(t, \ell)\right) = \left\{\left((i, j, k),\, (\alpha, \beta, \gamma)\right) \colon\, 1 \le i < j < k \le t \text{ and } \alpha, \beta, \gamma \in [l]\right\}.$$

We also consider homomorphisms into sub-multi-hypergraphs of $\mathrm{Tr}(M(t, \ell))$. Recall that for a hypergraph $\mathcal{F}$ we denote by $\Delta(\mathcal{F})$ the shadow of $\mathcal{F}$, i.e., the family of all pairs of vertices contained in an edge of $\mathcal{F}$.

Definition 17. *Let t and ℓ be integers, and let $\mathcal{S} \subseteq \mathrm{Tr}(M(t, \ell))$ be a multihypergraph. For a 3-uniform hypergraph $\mathcal{F}$ on f vertices, we say a pair of mappings (φ, ψ)*

$$\varphi \colon\, V(\mathcal{F}) \to V(\mathcal{S}) \subseteq [t] \qquad \textit{and} \qquad \psi \colon\, \Delta(\mathcal{F}) \to [\ell]$$

is a homomorphism from $\mathcal{F}$ to the multihypergraph $\mathcal{S}$ if φ is onto and if there exists a labeling of $V(\mathcal{F}) = \{v_1, \dots, v_f\}$ such that for every edge $\{v_x, v_y, v_z\} \in \mathcal{F}$, with $1 \le x < y < z \le f$, we have $\varphi(v_x) < \varphi(v_y) < \varphi(v_z)$ and

$$\Big(\big(\varphi(v_x), \varphi(v_y), \varphi(v_z)\big),\ \big(\psi(v_x, v_y), \psi(v_x, v_z), \psi(v_y, v_z)\big)\Big) \in \mathcal{S}.$$

We will abbreviate the existence of a homomorphism from $\mathcal{F}$ to $\mathcal{S}$ as $\mathcal{F} \twoheadrightarrow \mathcal{S}$.

Proof of Theorem 3. Let $\mathscr{A} = \mathrm{Forb}(\mathscr{F})$ for a (possibly infinite) family of forbidden hypergraphs $\mathscr{F}$, and let $\eta > 0$ be a positive constant. We need a few auxiliary functions before we reveal the promised constants $c > 0$, C, and n_0 (see (13) below). Given two positive integers t and ℓ and a multihypergraph $\mathcal{S} \subseteq \mathrm{Tr}(M(t, \ell))$, we set

$$\mathscr{F}_{\mathcal{S}} = \{\mathcal{F} \in \mathscr{F} \colon\, \mathcal{F} \twoheadrightarrow \mathcal{S}\}$$

and

$$C_{\mathcal{S}} = \begin{cases} \min\{|V(\mathcal{F})| \colon\, \mathcal{F} \in \mathscr{F}_{\mathcal{S}}\} & \text{if } \mathscr{F}_{\mathcal{S}} \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Let $\Psi \colon\, \mathbb{N}^2 \to \mathbb{N} \cup \{0\}$ be defined for two positive integers t and ℓ as

$$\Psi(t, l) = \max\left\{C_{\mathcal{S}} \colon\, \mathcal{S} \subseteq \mathrm{Tr}(M(t, \ell))\right\}.$$

The function $\Psi(t,\ell)$ is designed to forecast the maximal size of a witness from $\mathscr{F}$ we may encounter after application of the representative lemma, Lemma 16. Next we introduce the parameters δ and t_1 and the functions $\varepsilon_{\mathscr{P}}$, $\delta_{\mathscr{R}}$, $\varepsilon_{\mathscr{R}}$, and $r_{\mathscr{R}}$ with which we are going to apply Lemma 16. Recall the functions $\delta_3(f,\gamma,d_3)$, $\delta_2(f,\gamma,d_3,d_2)$, $r(f,\gamma,d_3,d_2)$, and $m_0(f,\gamma,d_3,d_2)$ from Theorem 8 and $\varepsilon_{\mathrm{tcl}}(\gamma,d)$ and $m_{\mathrm{tcl}}(\gamma,d)$ from Lemma 5. For the given η from above we set

$$\delta=\frac{\eta}{3} \qquad \text{and} \qquad t_1=\left\lceil\frac{4}{\eta}\right\rceil \tag{8}$$

and define functions in integer variables t, t', ℓ, and ℓ':

$$\varepsilon_{\mathscr{P}}(t,\ell)=\varepsilon_{\mathrm{tcl}}(\gamma=1/2,d=1/\ell), \tag{9}$$

$$\delta_{\mathscr{R}}(t,\ell)=\delta_3\big(f=\Psi(t,\ell),\gamma=1/2,d_3=\eta/3\big), \tag{10}$$

$$\varepsilon_{\mathscr{R}}(t,t',\ell,\ell')=\delta_2\big(f=\Psi(t,\ell),\gamma=1/2,d_3=\eta/3,d_2=1/(\ell\ell')\big), \tag{11}$$

$$r_{\mathscr{R}}(t,t',\ell,\ell')=r\big(f=\Psi(t,\ell),\gamma=1/2,d_3=\eta/3,d_2=1/(\ell\ell')\big)\,. \tag{12}$$

For that choice of δ, t_1, $\varepsilon_{\mathscr{P}}$, $\delta_{\mathscr{R}}$, $\varepsilon_{\mathscr{R}}$, and $r_{\mathscr{R}}$, Lemma 16 yields constants

$$T_{\mathscr{P}}\,, \qquad T_{\mathscr{P}\mathscr{R}}\,, \quad \text{and} \quad n_1\,.$$

Now we define the constants c, C, and n_0 promised by Theorem 3, and we set

$$\begin{aligned} &C=\Psi(T_{\mathscr{P}},T_{\mathscr{P}})\,, \qquad c=\frac{1}{4C!}\times\left(\frac{1}{T_{\mathscr{P}}T_{\mathscr{P}\mathscr{R}}}\right)^{\binom{C}{2}}\times\left(\frac{\eta}{3}\right)^{\binom{C}{3}}\left(\frac{1}{T_{\mathscr{P}}T_{\mathscr{P}\mathscr{R}}}\right)^{C}, \\ &\text{and} \\ &n_0=\max\big\{n_1,\,T_{\mathscr{P}}\times m_{\mathrm{tcl}}(\gamma=1/2,d=1/T_{\mathscr{P}}),\,C^2/c, \\ &\qquad\qquad T_{\mathscr{P}}T_{\mathscr{P}\mathscr{R}}\times m_0(f=C,\gamma=1/2,d_3=\eta/3,d_2=1/(T_{\mathscr{P}}T_{\mathscr{P}\mathscr{R}}))\big\}\,. \end{aligned} \tag{13}$$

This concludes the definition of all constants and functions involved in the proof.

Let $\mathcal{H}$ be a hypergraph on $n\geq n_0$ vertices which is η-far from $\mathscr{A}$. Due to Lemma 16 the hypergraph $\mathcal{H}$ admits a $(t_{\mathscr{P}},\ell_{\mathscr{P}})$-partition $\mathscr{P}$ (where $t_{\mathscr{P}}\geq t_1$) with a representative $\mathscr{R}$, and there are integers $t_{\mathscr{R}}$ and $\ell_{\mathscr{R}}$ such that (i)–(iv) of the lemma hold. We formally fix

$$\begin{aligned} \varepsilon_{\mathscr{P}}=\varepsilon_{\mathscr{P}}(t_{\mathscr{P}},\ell_{\mathscr{P}}), \quad \delta_{\mathscr{R}}=\delta_{\mathscr{R}}(t_{\mathscr{P}},\ell_{\mathscr{P}}), \\ \varepsilon_{\mathscr{R}}=\varepsilon_{\mathscr{R}}(t_{\mathscr{P}},t_{\mathscr{R}},\ell_{\mathscr{P}},\ell_{\mathscr{R}}), \quad \text{and} \quad r_{\mathscr{R}}=r_{\mathscr{R}}(t_{\mathscr{P}},t_{\mathscr{R}},\ell_{\mathscr{P}},\ell_{\mathscr{R}})\,. \end{aligned}$$

Now we delete all edges H of $\mathcal{H}$ which satisfy at least one of the two properties below:

(a) H is noncrossing w.r.t. the vertex partition $\mathscr{P}^{(1)}$; i.e., there is some $V_i\in\mathscr{P}^{(1)}$ $(1\leq i\leq t_{\mathscr{P}})$ such that $|H\cap V_i|\geq 2$; or

(b) $H\in\mathrm{Tr}(\hat{\mathcal{P}})$ for which $d_{\mathcal{H}}(\hat{\mathcal{R}}(\hat{\mathcal{P}}))<2\eta/3$.

We call the resulting subhypergraph $\mathcal{H}'\subseteq\mathcal{H}$. Next we estimate $\mathcal{H}\setminus\mathcal{H}'$. We first consider the edges deleted due to (a). Recalling the definition of t_1 in (8) and $t_{\mathscr{P}}\geq t_1$ we get that the number of noncrossing triples in $\mathcal{H}$ is at most

$$2\binom{t_{\mathscr{P}}}{2}\binom{n/t_{\mathscr{P}}}{2}\frac{n}{t_{\mathscr{P}}}+t_{\mathscr{P}}\binom{n/t_{\mathscr{P}}}{3}\leq\frac{n^3}{t_{\mathscr{P}}}\leq\frac{\eta}{4}n^3\,. \tag{14}$$

Next we estimate the number of edges deleted because of (b). By property (i) of Lemma 16 all graphs $P^{ij}_{\alpha}\in\mathscr{P}^{(2)}$ are $(\varepsilon_{\mathscr{P}},1/\ell_{\mathscr{P}})$-regular, and consequently, Lemma 5 applies to every triad $\hat{\mathcal{P}} \in \hat{\mathscr{P}}$ (with $\gamma = 1/2$ and $d = 1/\ell_{\mathscr{P}}$). We consider two subcases. First, the edge H could belong to a triad $\hat{\mathcal{P}} \in \hat{\mathscr{P}}$ which is exceptional in the sense of (iii) of Lemma 16. However, the number of those edges cannot exceed

$$\begin{aligned}(15)\qquad &\Big|\bigcup\big\{\operatorname{Tr}(\hat{\mathcal{P}})\colon\ \hat{\mathcal{P}} \in \hat{\mathscr{P}} \text{ and } |d_{\mathcal{H}}(\hat{\mathcal{P}}) - d_{\mathcal{H}}(\hat{\mathcal{R}}(\hat{\mathcal{P}}))| > \delta\big\}\Big| \\ &\qquad \leq \delta t_{\mathscr{P}}^3 \ell_{\mathscr{P}}^3 \times \max_{\hat{\mathcal{P}} \in \hat{\mathscr{P}}} \big|\operatorname{Tr}(\hat{\mathcal{P}})\big| \leq \delta t_{\mathscr{P}}^3 \ell_{\mathscr{P}}^3 \times \left(1 + \frac{1}{2}\right)\frac{n^3}{t_{\mathscr{P}}^3 \ell_{\mathscr{P}}^3} \leq \frac{\eta}{2} n^3 \,.\end{aligned}$$

Finally, the number of edges of $\mathcal{H}$ in triads $\hat{\mathcal{P}}$ which are not exceptional in the sense of part (iii) of Lemma 16 but satisfy (b) is at most

$$(16)\qquad \left(\frac{2\eta}{3} + \delta\right) \times \max_{\hat{\mathcal{P}} \in \hat{\mathscr{P}}} \big|\operatorname{Tr}(\hat{\mathcal{P}})\big| \times \binom{t_{\mathscr{P}}}{3} \ell_{\mathscr{P}}^3 \leq \eta \times \frac{3}{2} \frac{n^3}{t_{\mathscr{P}}^3 \ell_{\mathscr{P}}^3} \times \frac{t_{\mathscr{P}}^3}{6} \ell_{\mathscr{P}}^3 \leq \frac{\eta}{4} n^3 \,.$$

It follows from the considerations above and from (14)–(16) that

$$|\mathcal{H} \setminus \mathcal{H}'| \leq \eta n^3 \,.$$

Hence, by assumption on $\mathcal{H}$ the hypergraph $\mathcal{H}' \notin \mathscr{A}$ and contains some copy $\mathcal{F}_0$ of some forbidden hypergraph from $\mathscr{F}$. Note that since $\mathcal{H}'$ contains only crossing edges in $\mathscr{P}$, the existence of $\mathcal{F}_0 \subseteq \mathcal{H}'$ yields the existence of some homomorphism to $\operatorname{Tr}(M(t_{\mathscr{P}}, \ell_{\mathscr{P}}))$. Let $\mathcal{S} \subseteq \operatorname{Tr}(M(t_{\mathscr{P}}, \ell_{\mathscr{P}}))$ be a homomorphic image of such a homomorphism. In particular, $\mathscr{F}_{\mathcal{S}} \neq \emptyset$, since $\mathcal{F}_0 \in \mathscr{F}_{\mathcal{S}}$. We denote by $\mathcal{F}_1$ some hypergraph in $\mathscr{F}_{\mathcal{S}}$ with the minimum number $C_{\mathcal{S}}$ of vertices (see (7)) and let (φ, ψ) be the homomorphism from $\mathcal{F}_1$ to $\mathcal{S}$. Let

$$f_1 = C_{\mathcal{S}} = |V(\mathcal{F}_1)| \leq C,$$

and let $V(\mathcal{F}_1) = \{v_1, \dots, v_{f_1}\}$. We are going to show that the number of copies of $\mathcal{F}_1$ in $\mathcal{H}'$ will satisfy

$$(17)\qquad \#\{\mathcal{F}_1 \subseteq \mathcal{H}'\} \geq c n^{f_1} \,,$$

which clearly implies Theorem 3.

We define the graph

$$R_{\mathcal{S}} = \bigcup_{(\{i,j\},\alpha) \in \mathcal{S}} R_{\alpha}^{ij} \,,$$

where $R_{\alpha}^{ij} \in \mathscr{R}^{(2)}$ are graphs of the representative. Assume without loss of generality that $V(\mathcal{S}) = \{1, 2, \dots, s\}$ for some $s \leq t_{\mathscr{P}}$ and thus $V(R_{\mathcal{S}}) = W_1 \cup W_2 \cup \dots \cup W_s$ with $W_i \in \mathscr{R}^{(1)}$. If $f = s$, then $(R_{\mathcal{S}}, \mathcal{H}' \cap \operatorname{Tr}(R_{\mathcal{S}}))$ is a $(n/(t_{\mathscr{P}} t_{\mathscr{R}}), \mathcal{F}_1)$-complex (since by definition φ is onto), and we could invoke the counting lemma, Theorem 8, which would yield (17). However, since this does not have to be the case, we will define an auxiliary $(n/(t_{\mathscr{P}} t_{\mathscr{R}}), \mathcal{F}_1)$-complex $(G, \mathcal{G})$ which will satisfy the assumptions of Theorem 8, and due to its construction we will infer (17) from it.

We first define the vertex set of $(G, \mathcal{G})$. For each $x = 1, 2, \dots, f_1$ let Y_x be a copy of $W_{\varphi(x)}$ such that for all $1 \leq x < y \leq f_1$ we have $Y_x \cap Y_y = \emptyset$. Moreover, for every $x = 1, 2, \dots, f_1$ let $\vartheta_x\colon W_{\varphi(x)} \to Y_x$ be a bijection. Note that if $\{x, y\} \in \Delta(\mathcal{F}_1)$, then $\varphi(x) \neq \varphi(y)$ and, consequently, $(\{\varphi(x), \varphi(y)\}, \psi(x, y)) \in \mathcal{S}$ and $R_{\psi(x,y)}^{\varphi(x)\varphi(y)} \in \mathscr{R}^2$.

Hence, we can define for every $\{x,y\} \in \Delta(\mathcal{F}_1)$ with $x < y$ a bipartite graph G^{xy} with vertex classes Y_x and Y_y and edge set

$$E(G^{xy}) = \big\{\{\vartheta(w), \vartheta(w')\}\colon\ \{w,w'\} \in E(R^{\varphi(x)\varphi(y)}_{\psi(x,y)})\big\}\,.$$

It follows from that definition that G^{xy} is an isomorphic copy of $R^{\varphi(x)\varphi(y)}_{\psi(x,y)}$ and that $G = (Y, E_G)$ defined by

$$Y = Y_1 \cup \dots \cup Y_{f_1} \quad \text{and} \quad E_G = \bigcup \big\{E(G^{xy})\colon\ \{x,y\} \in \Delta(\mathcal{F}_1)\big\}$$

is an f_1-partite graph satisfying (i) and (ii) of Definition 7 with $m = n/(t_{\mathscr{P}} t_{\mathscr{R}})$, $\delta_2 = \varepsilon_{\mathscr{R}}$, and $d_2 = 1/(\ell_{\mathscr{P}} \ell_{\mathscr{R}})$. Similarly, for every edge $\{x,y,z\} \in \mathcal{F}_1$ there is a triad $\mathcal{R}(x,y,z) \in \hat{\mathscr{R}}$ defined by

$$\mathcal{R}(x,y,z) = (R^{\varphi(x)\varphi(y)}_{\psi(x,y)},\ R^{\varphi(x)\varphi(z)}_{\psi(x,z)},\ R^{\varphi(y)\varphi(z)}_{\psi(y,z)})\,.$$

We set

$$\mathcal{G}^{xyz} = \big\{\{w,w',w''\}\colon\ \{w,w',w''\} \in \mathcal{H}' \cap \mathrm{Tr}(\mathcal{R}(x,y,z))\big\}$$

and

$$\mathcal{G} = \bigcup \big\{\mathcal{G}^{xyz}\colon\ \{x,y,z\} \in \mathcal{F}_1\big\}\,.$$

Again it follows from the definition that $\mathcal{G}$ satisfies (iii) of Definition 7 with $\delta_3 = \delta_{\mathscr{R}}$, $d_3 \geq \eta/3$, and $r = r_{\mathscr{R}}$. Summarizing, $(G, \mathcal{G})$ is an $(\varepsilon_{\mathscr{R}}, \delta_{\mathscr{R}}, 1/(\ell_{\mathscr{P}}\ell_{\mathscr{R}}), \eta/3, r_{\mathscr{R}})$-regular $(n/(t_{\mathscr{P}} t_{\mathscr{R}}), \mathcal{F}_1)$-complex. By the choices in (10)–(12) and (13) we can apply the counting lemma, Theorem 8, and hence,

$$\#\{\mathcal{F}_1 \subseteq \mathcal{G}\} \geq \left(1 - \frac{1}{2}\right)\left(\frac{1}{\ell_{\mathscr{P}}\ell_{\mathscr{R}}}\right)^{|\Delta(\mathcal{F}_1)|}\left(\frac{\eta}{3}\right)^{|\mathcal{F}_1|}\left(\frac{n}{t_{\mathscr{P}} t_{\mathscr{R}}}\right)^{f_1} \overset{(13)}{\geq} 2 f_1! c n^{f_1}\,.$$

Observe that almost every copy of $\mathcal{F}_1$ in $\mathcal{G}$ corresponds to a labeled copy of $\mathcal{F}_1$ in $\mathcal{H}'$ (with $\vartheta_x\colon\ Y_x \to W_{\varphi(x)}$ defining the isomorphism). The only possible exceptions are those copies of $\mathcal{F}_1$ with an image of size less then f_1. This may happen since for each $x = 1, \dots, f_1$ the map ϑ_x is a bijection, but $\vartheta_x^{-1}(u) = \vartheta_y^{-1}(w)$ for two different vertices u and w of a copy of $\mathcal{F}_1$ in $\mathcal{G}$ if, e.g., x and y are such that $W_{\varphi(x)} = W_{\varphi(y)}$. The number of those copies of $\mathcal{F}_1$ in $\mathcal{G}$ is however bounded from above by $\binom{f_1}{2}(n/t_{\mathscr{P}} t_{\mathscr{R}})^{f_1 - 1}$. Consequently, we can find $2 f_1! c n^{f_1} - \binom{f_1}{2}(n/t_{\mathscr{P}} t_{\mathscr{R}})^{f_1-1}$ labeled copies of $\mathcal{F}_1$ in $\mathcal{H}'$ and by the choice of $n_0 \geq C^2/c$ at least $c n^{f_1}$ unlabeled copies. Hence, we verified (17), which concludes the proof of Theorem 3. □

4. Proof of the representative lemma.

4. Proof of the representative lemma. The proof of Lemma 16 is based on two successive applications of the regularity lemma (first in the form of Theorem 11 and second in the form of Theorem 13).

Proof of Lemma 16. First we recall the quantification of the representative lemma, Lemma 16. Let constants $t_1 \in \mathbb{N}$ and $\delta > 0$ and functions $\varepsilon_{\mathscr{P}}\colon\ \mathbb{N}^2 \to (0,1]$, $\delta_{\mathscr{R}}\colon\ \mathbb{N}^2 \to (0,1]$, $\varepsilon_{\mathscr{R}}\colon\ \mathbb{N}^4 \to (0,1]$, and $r_{\mathscr{R}}\colon\ \mathbb{N}^4 \to \mathbb{N}$ be given. We are supposed to define positive integers $T_{\mathscr{P}}$, $T_{\mathscr{P}\mathscr{R}}$, and n_1, and we are going to define them in (23). First, however, we need some preparations.

Our proof of Lemma 16 will rely on the regularity lemma for hypergraphs in the form of Theorems 13 and 11 and the counting lemma for graphs, Lemma 5. Below we will use the functions $\mu_{\mathrm{hrl}}(t_{\mathscr{P}},\ell_{\mathscr{P}},\delta_{\mathscr{Q}},\varepsilon_{\mathscr{Q}},r_{\mathscr{Q}})$, $T_{\mathrm{hrl}}(t_{\mathscr{P}},\ell_{\mathscr{P}},\delta_{\mathscr{Q}},\varepsilon_{\mathscr{Q}},r_{\mathscr{Q}})$, and $n_{\mathrm{hrl}}(t_{\mathscr{P}},\ell_{\mathscr{P}},\delta_{\mathscr{Q}},\varepsilon_{\mathscr{Q}},r_{\mathscr{Q}})$ given by Theorem 13; the functions $T_0(t_0,\delta_{\mathscr{P}},\mu_{\mathscr{P}},r_{\mathscr{P}})$ and $n_0(t_0,\delta_{\mathscr{P}},\mu_{\mathscr{P}},r_{\mathscr{P}})$ given by Theorem 11; and the functions $\varepsilon_{\mathrm{tcl}}(\gamma,d)$ and $m_{\mathrm{tcl}}(\gamma,d)$ given by the triangle counting lemma, Lemma 5.

We define auxiliary functions $\mu_{\mathrm{aux}}\colon \mathbb{N}^2\to(0,1]$, $T_{\mathrm{aux}}\colon \mathbb{N}^2\to\mathbb{N}$, and $n_{\mathrm{aux}}\colon \mathbb{N}^2\to\mathbb{N}$, and we set for positive integers t and ℓ and $x\in\{\mu,T,n\}$

$$\begin{aligned} x_{\mathrm{aux}}(t,\ell)=x_{\mathrm{hrl}}\Big(t_{\mathscr{P}}=t,\ \ell_{\mathscr{P}}=\ell,\ \delta_{\mathscr{Q}}=\min\big\{\delta_{\mathscr{R}}(t,\ell),\ (t\ell)^{-3}/3\big\},&\\ \varepsilon_{\mathscr{Q}}(t',\ell')=\min\big\{\varepsilon_{\mathscr{R}}(t,t',\ell,\ell'),\ \varepsilon_{\mathrm{tcl}}\big(\gamma=\tfrac12,\ d=\tfrac{1}{\ell\ell'}\big)\big\},&\\ r_{\mathscr{Q}}(t',\ell')=r_{\mathscr{R}}(t,t',\ell,\ell')\Big)\,.& \end{aligned}\tag{18}$$

In other words, for fixed t and ℓ the values $\mu_{\mathrm{aux}}(t,\ell)$, $T_{\mathrm{aux}}(t,\ell)$, and $n_{\mathrm{aux}}(t,\ell)$ are defined by the corresponding constants μ_{hrl}, T_{hrl}, and n_{hrl} given by Theorem 13 for those parameters $t_{\mathscr{P}}$, $\ell_{\mathscr{P}}$, and $\delta_{\mathscr{Q}}$ and functions $\varepsilon_{\mathscr{Q}}$ and $r_{\mathscr{Q}}$ displayed in (18). Note that the choice in (18) is such that, for fixed integers t and ℓ, the parameters $t_{\mathscr{P}}$, $\ell_{\mathscr{P}}$, and $\delta_{\mathscr{Q}}$ are constants, while $\varepsilon_{\mathscr{Q}}\colon \mathbb{N}^2\to(0,1]$ and $r_{\mathscr{Q}}\colon\mathbb{N}^2\to\mathbb{N}$ are functions in the variables t' and ℓ', which matches the quantification of Theorem 13.

With those auxiliary functions at hand, we define the parameters and constants with which we will apply the “simple” regularity lemma, Theorem 11, later. For that we fix constants

$$t_0=t_1\qquad\text{and}\qquad \delta_{\mathscr{P}}=\frac{\delta}{9}\tag{19}$$

and functions $\mu_{\mathscr{P}}\colon\mathbb{N}^2\to(0,1]$ and $r_{\mathscr{P}}\colon\mathbb{N}^2\to\mathbb{N}$ defined for all positive integers t and ℓ by

$$\mu_{\mathscr{P}}(t,\ell)=\min\big\{\mu_{\mathrm{aux}}(t,\ell),\ \varepsilon_{\mathscr{P}}(t,\ell),\ \varepsilon_{\mathrm{tcl}}(\gamma=1/2,\ d=1/\ell)\big\}\,,\tag{20}$$

$$r_{\mathscr{P}}(t,\ell)=\big(T_{\mathrm{aux}}(t,\ell)\big)^6\,,\tag{21}$$

where t_1, δ, and $\varepsilon_{\mathscr{P}}$ are input parameters of Lemma 16. Given t_0, $\delta_{\mathscr{P}}$, $\mu_{\mathscr{P}}$, and $r_{\mathscr{P}}$ from above, Theorem 11 yields positive integers

$$T_0=T_0(t_0,\delta_{\mathscr{P}},\mu_{\mathscr{P}},r_{\mathscr{P}})\quad\text{and}\quad n_0(t_0,\delta_{\mathscr{P}},\mu_{\mathscr{P}},r_{\mathscr{P}})\,.\tag{22}$$

Now we are able to determine the promised constants $T_{\mathscr{P}}$, $T_{\mathscr{P}\mathscr{R}}$, and n_1 of Lemma 16, and we set

$$\begin{aligned} T_{\mathscr{P}}=T_0,\qquad T_{\mathscr{P}\mathscr{R}}=\max_{1\le t,\ell\le T_0}T_{\mathrm{aux}}(t,\ell),&\\ \text{and}\quad n_1=\max\Big\{\max_{1\le t,\ell\le T_0}n_{\mathrm{aux}}(t,\ell),\ n_0,\ T_{\mathscr{P}}\times m_{\mathrm{tcl}}\big(\gamma=\tfrac12,\ d=\tfrac{1}{T_{\mathscr{P}}}\big),&\\ T_{\mathscr{P}}T_{\mathscr{P}\mathscr{R}}\times m_{\mathrm{tcl}}\big(\gamma=\tfrac12,\ d=\tfrac{1}{T_{\mathscr{P}}T_{\mathscr{P}\mathscr{R}}}\big)\Big\}\,.& \end{aligned}\tag{23}$$

Having defined those constants, let $\mathcal{H}$ be a given hypergraph on $n\ge n_1$ vertices. Since $n_1\ge n_0$ we can apply Theorem 11 with constants t_0 and $\delta_{\mathscr{P}}$ and functions $\mu_{\mathscr{P}}$ and $r_{\mathscr{P}}$ defined in (19)–(21). Theorem 11 yields a partition $\mathscr{P}=\{\mathscr{P}^{(1)},\mathscr{P}^{(2)}\}$ and positive integers $t_{\mathscr{P}}$ and $\ell_{\mathscr{P}}$ such that (i) and (ii) of Theorem 11 hold; i.e., (P1) $\mathscr{P}$ is $(\mu_{\mathscr{P}}(t_{\mathscr{P}},\ell_{\mathscr{P}}),t_{\mathscr{P}},\ell_{\mathscr{P}})$-equitable, T_0-bounded, and $t_{\mathscr{P}}\geq t_0$; and

(P2) $\mathcal{H}$ is $(\delta_{\mathscr{P}},*,r_{\mathscr{P}}(t_{\mathscr{P}},\ell_{\mathscr{P}}))$-regular w.r.t. all but at most $\delta_{\mathscr{P}}t_{\mathscr{P}}^3\ell_{\mathscr{P}}^3$ triads $\hat{\mathcal{P}}\in\hat{\mathscr{P}}$.

Next we will apply the *refining version* of the regularity lemma, Theorem 13, to $\mathcal{H}$ and $\mathscr{P}$, with parameters $t_{\mathscr{P}}$, $\ell_{\mathscr{P}}$, $\delta_{\mathscr{Q}}=\min\{\delta_{\mathscr{R}}(t_{\mathscr{P}},\ell_{\mathscr{P}}),\,(3t_{\mathscr{P}}^3\ell_{\mathscr{P}}^3)^{-1})\}$, $\varepsilon_{\mathscr{Q}}(t',\ell')=\min\{\varepsilon_{\mathscr{R}}(t_{\mathscr{P}},t',\ell_{\mathscr{P}},\ell'),\varepsilon_{\mathrm{tcl}}(\gamma=\frac{1}{2},d=\frac{1}{\ell_{\mathscr{P}}\ell'})\}$, and $r_{\mathscr{Q}}(t',\ell')=r_{\mathscr{R}}(t_{\mathscr{P}},t',\ell_{\mathscr{P}},\ell')$. In other words, we will apply Theorem 13 with precisely the same parameters as those in (18). Therefore, we have to check that $\mathcal{H}$ and $\mathscr{P}$ satisfy the assumptions (a) and (b) of Theorem 13 for $n_{\mathrm{aux}}(t_{\mathscr{P}},\ell_{\mathscr{P}})$ and $\mu_{\mathrm{aux}}(t_{\mathscr{P}},\ell_{\mathscr{P}})$. However, due to (23) and (P1) we have $n\geq n_{\mathrm{aux}}(t_{\mathscr{P}},\ell_{\mathscr{P}})$ and due to the choice in (20) we have $\mu_{\mathscr{P}}(t_{\mathscr{P}},\ell_{\mathscr{P}})\leq\mu_{\mathrm{aux}}(t_{\mathscr{P}},\ell_{\mathscr{P}})$. Consequently, assumptions (a) and (b) of Theorem 13 hold, and Theorem 13 yields a partition $\mathscr{Q}=\{\mathscr{Q}^{(1)},\mathscr{Q}^{(2)}\}$ and positive integers $t_{\mathscr{Q}}$ and $\ell_{\mathscr{Q}}$ such that with

$$\delta_{\mathscr{Q}}=\min\{\delta_{\mathscr{R}}(t_{\mathscr{P}},\ell_{\mathscr{P}}),\,(t_{\mathscr{P}}\ell_{\mathscr{P}})^{-3}/3\}\,,\tag{24}$$

$$\varepsilon_{\mathscr{Q}}=\min\{\varepsilon_{\mathscr{R}}(t_{\mathscr{P}},t_{\mathscr{Q}},\ell_{\mathscr{P}},\ell_{\mathscr{Q}}),\varepsilon_{\mathrm{tcl}}(\gamma=\tfrac{1}{2},d=\tfrac{1}{\ell_{\mathscr{P}}\ell_{\mathscr{Q}}})\}\,,\tag{25}$$

and

$$r_{\mathscr{Q}}=r_{\mathscr{R}}(t_{\mathscr{P}},t_{\mathscr{Q}},\ell_{\mathscr{P}},\ell_{\mathscr{Q}})\,.\tag{26}$$

we have

(Q1) $\mathscr{Q}$ is $(\varepsilon_{\mathscr{Q}},t_{\mathscr{P}}t_{\mathscr{Q}},\ell_{\mathscr{P}}\ell_{\mathscr{Q}})$-equitable and a $T_{\mathrm{aux}}(t_{\mathscr{P}},\ell_{\mathscr{P}})$-bounded partition on V,

(Q2) $\mathscr{Q}\prec\mathscr{P}$, and

(Q3) $\mathcal{H}$ is $(\delta_{\mathscr{Q}},*,r_{\mathscr{Q}})$-regular w.r.t. all but at most $\delta_{\mathscr{Q}}t_{\mathscr{P}}^3t_{\mathscr{Q}}^3\ell_{\mathscr{P}}^3\ell_{\mathscr{Q}}^3$ triads $\hat{\mathcal{Q}}\in\hat{\mathscr{Q}}$.

It follows directly from (P1), (20), and the choice of $T_{\mathscr{P}}$ in (23) that

$$\mathscr{P}\text{ is }T_{\mathscr{P}}\text{-bounded and satisfies (i) of Lemma 16.}\tag{27}$$

Below we will select the representative $\mathscr{R}$ from the finer partition $\mathscr{Q}$. For property (iii) of Lemma 16 the following claim will be useful.

Claim 18. *If $\hat{\mathcal{P}}\in\hat{\mathscr{P}}$ is such that $\mathcal{H}$ is $(\delta_{\mathscr{P}},*,r_{\mathscr{P}}(t_{\mathscr{P}},\ell_{\mathscr{P}}))$-regular w.r.t. $\mathscr{P}$, then all but at most $2\delta_{\mathscr{P}}t_{\mathscr{Q}}^3\ell_{\mathscr{Q}}^3$ triads $\hat{\mathcal{Q}}\in\hat{\mathscr{Q}}$ with $\hat{\mathcal{Q}}\subseteq\hat{\mathcal{P}}$ satisfy*

$$\big|d_{\mathcal{H}}(\hat{\mathcal{P}})-d_{\mathcal{H}}(\hat{\mathcal{Q}})\big|\leq\delta\,.\tag{28}$$

Proof. Let $\hat{\mathcal{P}}\in\hat{\mathscr{P}}$ as in the claim be given. We set

$$\mathcal{B}_+=\{\hat{\mathcal{Q}}\in\hat{\mathscr{Q}}\colon\ \hat{\mathcal{Q}}\subseteq\hat{\mathcal{P}}\text{ and }d_{\mathcal{H}}(\hat{\mathcal{Q}})>d_{\mathcal{H}}(\hat{\mathcal{P}})+\delta\},$$
$$\mathcal{B}_-=\{\hat{\mathcal{Q}}\in\hat{\mathscr{Q}}\colon\ \hat{\mathcal{Q}}\subseteq\hat{\mathcal{P}}\text{ and }d_{\mathcal{H}}(\hat{\mathcal{Q}})<d_{\mathcal{H}}(\hat{\mathcal{P}})-\delta\}\,.$$

We first consider $|\mathcal{B}_+|$. Observe that by definition of $\mathcal{B}_+$

$$d_{\mathcal{H}}\left(\bigcup_{\hat{\mathcal{Q}}\in\mathcal{B}_+}\hat{\mathcal{Q}}\right)>d_{\mathcal{H}}(\hat{\mathcal{P}})+\delta\overset{(19)}{>}d_{\mathcal{H}}(\hat{\mathcal{P}})+\delta_{\mathscr{P}}\,.\tag{29}$$

On the other hand, recalling that by (Q1) every $Q\in\mathscr{Q}^{(2)}$ is an $(\varepsilon_{\mathscr{Q}},1/(\ell_{\mathscr{P}}\ell_{\mathscr{Q}}))$-regular bipartite graph and that by (25) $\varepsilon_{\mathscr{Q}}\leq\varepsilon_{\mathrm{tcl}}(\gamma=\frac{1}{2},d=\frac{1}{\ell_{\mathscr{P}}\ell_{\mathscr{Q}}})$, we infer from Lemma 5 that the total number of triangles contained in some $\hat{\mathcal{Q}}\in\mathcal{B}_+\subseteq\hat{\mathscr{Q}}$ does not exceed

$$|\mathcal{B}_+|\times\frac{3}{2}\left(\frac{1}{\ell_{\mathscr{P}}\ell_{\mathscr{Q}}}\right)^3\left(\frac{n}{t_{\mathscr{P}}t_{\mathscr{Q}}}\right)^3\,.\tag{30}$$

Since $\mathcal{H}$ is $(\delta_{\mathscr{P}}, *, r_{\mathscr{P}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}))$-regular w.r.t. $\hat{\mathcal{P}}$ and by (21)

$$r_{\mathscr{P}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}) \geq \big(T_{\mathrm{aux}}(t_{\mathscr{P}}, \ell_{\mathscr{P}})\big)^6 \overset{\text{(Q1)}}{\geq} t_{\mathscr{Q}}^3 \ell_{\mathscr{Q}}^3 \geq |\mathcal{B}_+|,$$

we infer from (29) that the quantity from (30) is smaller than

$$|\mathcal{B}_+| \times \frac{3}{2}\left(\frac{1}{\ell_{\mathscr{P}}\ell_{\mathscr{Q}}}\right)^3 \left(\frac{n}{t_{\mathscr{P}}t_{\mathscr{Q}}}\right)^3 \leq \delta_{\mathscr{P}} \big|\operatorname{Tr}(\hat{\mathcal{P}})\big| \leq \delta_{\mathscr{P}} \times \frac{3}{2}\left(\frac{1}{\ell_{\mathscr{P}}}\right)^3 \left(\frac{n}{t_{\mathscr{P}}}\right)^3,$$

where we used the $(\mu_{\mathscr{P}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}), 1/\ell_{\mathscr{P}})$-regularity of the bipartite subgraphs of every $\hat{\mathcal{P}} \in \hat{\mathscr{P}}$ (see (P1)), $\mu_{\mathscr{P}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}) \leq \varepsilon_{\mathrm{tcl}}(\gamma = \frac{1}{2}, d = \frac{1}{\ell_{\mathscr{P}}})$ (see (20)), and Lemma 5 for the last inequality. Consequently,

$$|\mathcal{B}_+| \leq \delta_{\mathscr{P}} t_{\mathscr{Q}}^3 \ell_{\mathscr{Q}}^3 .$$

Repeating the same argument for $|\mathcal{B}_-|$ yields $|\mathcal{B}_+| + |\mathcal{B}_-| \leq 2\delta_{\mathscr{P}} t_{\mathscr{Q}}^3 \ell_{\mathscr{Q}}^3$, and the claim follows. □

In what follows we will select the representative $\mathscr{R}$ from $\mathscr{Q}$ randomly and show that with positive probability such an $\mathscr{R}$ satisfies properties (ii)–(iv) of Lemma 16. For that the following notation will be useful. Recall that $\mathscr{P}^{(1)} = \{V_1, V_2, \dots, V_{t_{\mathscr{P}}}\}$ and $\mathscr{P}^{(2)} = \{P_\alpha^{ij} \colon 1 \leq i < j \leq t_{\mathscr{P}}, \alpha \in [\ell_{\mathscr{P}}]\}$, where $V(P_\alpha^{ij}) = V_i \cup V_j$. Let the vertex partition classes of $\mathscr{Q}^{(1)}$ be labeled in such a way that $\mathscr{Q}^{(1)} = \{W_{i,i'} \colon (i,i') \in [t_{\mathscr{P}}] \times [t_{\mathscr{Q}}]\}$ and $V_i = W_{i,1} \cup W_{i,2} \cup \dots \cup W_{i,t_{\mathscr{Q}}}$ for every $i = 1, 2, \dots, t_{\mathscr{P}}$. Furthermore, let the graphs of $\mathscr{Q}_{\mathscr{P}}^{(2)} = \{Q \in \mathscr{Q}^{(2)} \colon Q \subseteq P \text{ for some } P \in \mathscr{P}^{(2)}\}$ be labeled

$$\mathscr{Q}_{\mathscr{P}}^{(2)} = \big\{Q_{\alpha,\alpha'}^{(i,i'),(j,j')} \colon (\alpha,\alpha') \in [\ell_{\mathscr{P}}] \times [\ell_{\mathscr{Q}}], (i,i'), (j,j') \in [t_{\mathscr{P}}] \times [t_{\mathscr{Q}}], \text{ and } i < j\big\}$$

such that for every $(i,i'), (j,j') \in [t_{\mathscr{P}}] \times [t_{\mathscr{Q}}]$ with $i < j$ and $(\alpha,\alpha') \in [\ell_{\mathscr{P}}] \times [\ell_{\mathscr{Q}}]$,

$$\begin{aligned} V(Q_{\alpha,\alpha'}^{(i,i'),(j,j')}) &= W_{i,i'} \cup W_{j,j'} , \\ E(P_\alpha^{ij}[W_{i,i'} \cup W_{j,j'}]) &= \bigcup_{\alpha' \in [\ell_{\mathscr{Q}}]} E(Q_{\alpha,\alpha'}^{(i,i'),(j,j')}) . \end{aligned}$$

Now consider a pair of random mappings

$$\varphi \colon [t_{\mathscr{P}}] \to [t_{\mathscr{Q}}] \quad \text{and} \quad \psi \colon \binom{[t_{\mathscr{P}}]}{2} \times [\ell_{\mathscr{P}}] \to [\ell_{\mathscr{Q}}];$$

both mappings are chosen independently and uniformly at random from the set of all $t_{\mathscr{Q}}^{t_{\mathscr{P}}}$ or $\ell_{\mathscr{Q}}^{\binom{t_{\mathscr{P}}}{2}\ell_{\mathscr{P}}}$ mappings. To each such pair of mappings we associate a random representative $\mathscr{R} = \mathscr{R}(\varphi, \psi) = \{\mathscr{R}^{(1)}, \mathscr{R}^{(2)}\}$ defined by

$$\mathscr{R}^{(1)} = \{W_{i,\varphi(i)} \colon i \in [t_{\mathscr{P}}]\}$$

and

$$\mathscr{R}^{(2)} = \Big\{Q_{\alpha,\psi(\{i,j\},\alpha)}^{(i,\varphi(i)),(j,\varphi(j))} \colon 1 \leq i < j \leq t_{\mathscr{P}}, \alpha \in [\ell_{\mathscr{P}}]\Big\}. \tag{31}$$

It is easy to check that $\mathscr{R}(\varphi, \psi)$ indeed is a representative of $\mathscr{P}$ for every choice of φ and ψ. Moreover, we infer from (Q1), (25), and the choice of $T_{\mathscr{P}\mathscr{R}}$ in (23) that setting

$$t_{\mathscr{R}} = t_{\mathscr{Q}} \qquad \text{and} \qquad \ell_{\mathscr{R}} = \ell_{\mathscr{Q}} \tag{32}$$

yields

$$\mathscr{R}(\varphi,\psi) \text{ satisfies (ii) of Lemma 16 for every choice of } \varphi \text{ and } \psi. \tag{33}$$

We are going to show that there is a choice of mappings φ and ψ such that the representative $\mathscr{R} = \mathscr{R}(\varphi,\psi)$ satisfies properties (iii) and (iv) of Lemma 16 as well.

Due to Claim 18 we have that if $\hat{\mathcal{P}} \in \hat{\mathscr{P}}$ is such that $\mathcal{H}$ is $(\delta_{\mathscr{P}}, *, r_{\mathscr{P}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}))$-regular w.r.t. $\hat{\mathcal{P}}$, then at most $2\delta_{\mathscr{P}} t_{\mathscr{Q}}^3 \ell_{\mathscr{Q}}^3$ subtriads $\hat{\mathcal{Q}} \subseteq \hat{\mathcal{P}}$ from $\hat{\mathscr{Q}}$ violate (28). Moreover, by (P2) the number of $(\delta_{\mathscr{P}}, *, r_{\mathscr{P}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}))$-irregular triads $\hat{\mathcal{P}} \in \hat{\mathscr{P}}$ does not exceed $\delta_{\mathscr{P}} t_{\mathscr{P}}^3 \ell_{\mathscr{P}}^3$, and consequently, the total number of subtriads $\hat{\mathcal{Q}} \in \hat{\mathscr{Q}}$ with $\hat{\mathcal{Q}} \subseteq \hat{\mathcal{P}} \in \hat{\mathscr{P}}$, where $\hat{\mathcal{P}}$ is $(\delta_{\mathscr{P}}, *, r_{\mathscr{P}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}))$-irregular, is at most $\delta_{\mathscr{P}} t_{\mathscr{P}}^3 \ell_{\mathscr{P}}^3 \times t_{\mathscr{Q}}^3 \ell_{\mathscr{Q}}^3$.

We say a triad $\hat{\mathcal{Q}} \in \hat{\mathscr{Q}}$ is *bad* if there exists some $\hat{\mathcal{P}} \in \hat{\mathscr{P}}$ such that $\hat{\mathcal{P}} \supseteq \hat{\mathcal{Q}}$ and either (28) is violated or $\hat{\mathcal{P}}$ is $(\delta_{\mathscr{P}}, *, r_{\mathscr{P}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}))$-irregular. From the discussion above we clearly infer that

$$\left|\{\hat{\mathcal{Q}} \in \hat{\mathscr{Q}} \colon\ \hat{\mathcal{Q}} \text{ is bad}\}\right| \le 2\delta_{\mathscr{P}} t_{\mathscr{Q}}^3 \ell_{\mathscr{Q}}^3 \times |\hat{\mathscr{P}}| + \delta_{\mathscr{P}} t_{\mathscr{P}}^3 \ell_{\mathscr{P}}^3 t_{\mathscr{Q}}^3 \ell_{\mathscr{Q}}^3 \le 3\delta_{\mathscr{P}} (t_{\mathscr{P}} \ell_{\mathscr{P}} t_{\mathscr{Q}} \ell_{\mathscr{Q}})^3 \,.$$

Since each $\hat{\mathcal{Q}} \in \hat{\mathscr{Q}}$ which is a subtriad of some $\hat{\mathcal{P}} \in \hat{\mathscr{P}}$ is contained in the same number $(t_{\mathscr{Q}}^{t_{\mathscr{P}}-3} \ell_{\mathscr{Q}}^{\binom{t_{\mathscr{P}}}{2}\ell_{\mathscr{P}}-3})$ of all representatives $\mathscr{R}(\varphi,\psi)$, the expected number of bad triads contained in a random representative $\mathscr{R}(\varphi,\psi)$ is smaller than $3\delta_{\mathscr{P}} t_{\mathscr{P}}^3 \ell_{\mathscr{P}}^3$. Let B_1 be the event for which the random representative contains more than $9\delta_{\mathscr{P}} t_{\mathscr{P}}^3 \ell_{\mathscr{P}}^3 = \delta t_{\mathscr{P}}^3 \ell_{\mathscr{P}}^3$ (see (19)) bad triads $\hat{\mathcal{Q}} \in \hat{\mathscr{Q}}$. From Markov's inequality we infer $\mathbb{P}(\mathscr{R}(\varphi,\psi) \in B_1) \le 1/3$. In other words,

$$\mathbb{P}\big(\mathscr{R}(\varphi,\psi) \text{ satisfies (iii) of Lemma 16}\big) \ge \frac{2}{3}\,. \tag{34}$$

Similarly, due to (Q3) combined with (24), (26), and (32), we have that the number of $(\delta_{\mathscr{R}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}), *, r_{\mathscr{R}}(t_{\mathscr{P}}, t_{\mathscr{R}}, \ell_{\mathscr{P}}, \ell_{\mathscr{R}}))$-irregular triads $\hat{\mathcal{Q}} \in \hat{\mathscr{Q}}$ is at most $\delta_{\mathscr{Q}} t_{\mathscr{P}}^3 t_{\mathscr{Q}}^3 \ell_{\mathscr{P}}^3 \ell_{\mathscr{Q}}^3$. Hence, the expected number of such irregular triads contained in the random representative $\mathscr{R}(\varphi,\psi)$ is at most

$$\delta_{\mathscr{Q}} t_{\mathscr{P}}^3 \ell_{\mathscr{P}}^3 \overset{(24)}{\le} \frac{1}{3}\,.$$

Let B_2 be the event for which the random representative $\mathscr{R}(\varphi,\psi)$ contains at least one $(\delta_{\mathscr{R}}(t_{\mathscr{P}}, \ell_{\mathscr{P}}), *, r_{\mathscr{R}}(t_{\mathscr{P}}, t_{\mathscr{R}}, \ell_{\mathscr{P}}, \ell_{\mathscr{R}}))$-irregular triad. Thus again by Markov's inequality we infer $\mathbb{P}(\mathscr{R}(\varphi,\psi) \in B_2) \le 1/3$, i.e.,

$$\mathbb{P}\big(\mathscr{R}(\varphi,\psi) \text{ satisfies (iv) of Lemma 16}\big) \ge \frac{2}{3}\,. \tag{35}$$

Combining (34) and (35) implies that the probability that $\mathscr{R}(\varphi,\psi)$ satisfies (iii) and (iv) of Lemma 16 is at least 1/3. Hence, there exist representative satisfying properties (iii) and (iv), and Lemma 16 follows from (27) and (33). □

5. Concluding remarks. We close this paper with a few remarks concerning extensions of Theorem 2 to monotone properties of general k-uniform hypergraphs and hereditary properties of hypergraphs.

5.1. Monotone properties of k-uniform hypergraphs. As we mentioned earlier, the proof of Theorem 2 presented in this paper extends without any major modification from 3-uniform to k-uniform hypergraphs. This is because the two main tools used in the proof, namely, the hypergraph regularity lemma (Theorems 11 and 13) and the hypergraph counting lemma (Theorem 8), were already proved for general k-uniform hypergraphs (see [23]). While the philosophy of the regularity method for general uniform hypergraphs and its application in this context stays the same, the general case of k-uniform hypergraphs brings a more complicated and technical notation. For example, the concept of a (t, ℓ)-partition extends to a family of partitions of the vertices, pairs, triples, and so on, and $(k-1)$-tuples of vertices. Due to this more complicated structure of the partition provided by the general regularity lemma, the notion of an appropiate representative is more involved. In particular, it cannot be described through such explicit labels as, e.g., used in (ii) of Definition 14 or in (31).

We believe that the special (and more explicit) case of 3-uniform hypergraphs provides a good balance between generality and clarity and that, due to the less complex notation, the proof is more readable. Therefore we restricted ourselves to 3-uniform hypergraphs here.

5.2. Hereditary properties of hypergraphs. Another interesting generalization of Theorem 2 is the extension from monotone to hereditary properties. A hypergraph property is called hereditary if it is closed under taking induced subhypergraphs. Monotone properties are a special case of hereditary properties. Recently Alon and Shapira [4] and later Lovász and Szegedy [19] (see also [8]) proved that every hereditary graph property is testable. In particular, Alon and Shapira use a strengthened version of Szemerédi's regularity lemma from [2], which in some sense corresponds to the representative lemma, Lemma 16, from our proof (see also [17] for a similar lemma). We believe that the proof of Alon and Shapira can be adapted to k-uniform hypergraphs by using the extension of the representative lemma given in this paper. Here again the main obstacles seem to be of a technical nature. In particular, dealing with edges which are *not crossing* in the partition seems to present additional problems of a technical nature.

Inspired by the work of Lovász and Szegedy, Rödl and Schacht [22] found a way to merge some ideas from [19] with that of Alon et al. [2]. This yields a somewhat different proof, which circumvents the technical issues mentioned above.

REFERENCES

[1] N. Alon, *Testing subgraphs in large graphs*, Random Structures Algorithms, 21 (2002), pp. 359–370.

[2] N. Alon, E. Fischer, M. Krivelevich, and M. Szegedy, *Efficient testing of large graphs*, Combinatorica, 20 (2000), pp. 451–476.

[3] N. Alon and M. Krivelevich, *Testing k-colorability*, SIAM J. Discrete Math., 15 (2002), pp. 211–227.

[4] N. Alon and A. Shapira, *A characterization of the (natural) graph properties testable with one-sided error*, in Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos, CA, 2005, pp. 429–438.

[5] N. Alon and A. Shapira, *Every monotone graph property is testable*, in Proceedings of the 37th Annual ACM Symposium on Theory of Computing, Baltimore, MD, ACM, New York, 2005, pp. 128–137.

[6] N. Alon and A. Shapira, *Homomorphisms in graph property testing–a survey*, in Topics in Discrete Mathematics, Algorithms Combin. 26, M. Klazar, J. Kratochvil, M. Loebl, J. Matoušek, R. Thomas, and P. Valtr, eds., Springer, Berlin, 2006, pp. 281–313.

[7] B. Bollobás, P. Erdős, M. Simonovits, and E. Szemerédi, *Extremal graphs without large forbidden subgraphs*, in Advances in Graph Theory, Ann. Discrete Math. 3, Elsevier–North Holland, Amsterdam, 1978, pp. 29–41.

[8] C. Borgs, J. Chayes, L. Lovász, V. T. Sós, B. Szegedy, and K. Vesztergombi, *Graph limits and parameter testing*, in Proceedings of the 38th Annual ACM Symposium on Theory of Computing, Association for Computing Machinery, New York, 2006, pp. 261–270.

[9] A. Czumaj and C. Sohler, *Testing hypergraph colorability*, Theoret. Comput. Sci., 331 (2005), pp. 37–52.

[10] R. A. Duke and V. Rödl, *On graphs with small subgraphs of large chromatic number*, Graphs Combin., 1 (1985), pp. 91–96.

[11] P. Erdős, *Problems and results on graphs and hypergraphs: Similarities and differences*, in Mathematics of Ramsey Theory, Algorithms Combin. 5, J. Nešetřil and V. Rödl, eds., Springer, Berlin, 1990, pp. 12–28.

[12] P. Erdős, P. Frankl, and V. Rödl, *The asymptotic number of graphs not containing a fixed subgraph and a problem for hypergraphs having no exponent*, Graphs Combin., 2 (1986), pp. 113–121.

[13] P. Frankl and V. Rödl, *Extremal problems on set systems*, Random Structures Algorithms, 20 (2002), pp. 131–164.

[14] O. Goldreich, S. Goldwasser, and D. Ron, *Property testing and its connection to learning and approximation*, J. ACM, 45 (1998), pp. 653–750.

[15] W. T. Gowers, *Hypergraph regularity and the multidimensional Szemerédi theorem*, submitted.

[16] W. T. Gowers, *Quasirandomness, counting and regularity for 3-uniform hypergraphs*, Combin. Probab. Comput., 15 (2006), pp. 143–184.

[17] Y. Kohayakawa, B. Nagle, and V. Rödl, *Efficient testing of hypergraphs (extended abstract)*, in Automata, Languages and Programming, Lecture Notes in Comput. Sci. 2380, Springer, Berlin, 2002, pp. 1017–1028.

[18] J. Komlós, A. Shokoufandeh, M. Simonovits, and E. Szemerédi, *The regularity lemma and its applications in graph theory*, in Theoretical Aspects of Computer Science Lecture Notes in Comput. Sci. 2292, Springer, Berlin, 2002, pp. 84–112.

[19] L. Lovász and B. Szegedy, *Graph limits and testing hereditary graph properties*, Technical report TR-2005-110, Microsoft Research, Redmond, WA, 2005.

[20] B. Nagle and V. Rödl, *Regularity properties for triple systems*, Random Structures Algorithms, 23 (2003), pp. 264–332.

[21] B. Nagle, V. Rödl, and M. Schacht, *The counting lemma for regular k-uniform hypergraphs*, Random Structures Algorithms, 28 (2006), pp. 113–179.

[22] V. Rödl and M. Schacht, *Generalizations of the removal lemma*, submitted.

[23] V. Rödl and M. Schacht, *Regular partitions of hypergraphs*, Combin. Probab. Comput., to appear.

[24] V. Rödl and J. Skokan, *Regularity lemma for k-uniform hypergraphs*, Random Structures Algorithms, 25 (2004), pp. 1–42.

[25] V. Rödl and J. Skokan, *Applications of the regularity lemma for uniform hypergraphs*, Random Structures Algorithms, 28 (2006), pp. 180–194.

[26] R. Rubinfeld and M. Sudan, *Robust characterizations of polynomials with applications to program testing*, SIAM J. Comput., 25 (1996), pp. 252–271.

[27] I. Z. Ruzsa and E. Szemerédi, *Triple systems with no six points carrying three triangles*, in Combinatorics, Vol. II, Colloq. Math. Soc. János Bolyai, ed., Elsevier–North Holland, Amsterdam, 1978, pp. 939–945.

[28] E. Szemerédi, *Regular partitions of graphs*, in Problèmes combinatoires et Théorie des Graphes Colloq. Internat. CNRS 260, Centre National de la Recherche Scientifique, Paris, 1978, pp. 399–401.

[29] T. Tao, *A variant of the hypergraph removal lemma*, J. Combin. Theory Ser. A, 113 (2006), pp. 1257–1280.

SIAM J. DISCRETE MATH.
Vol. 21, No. 1, pp. 93–106

A GENERALIZATION OF KOTZIG'S THEOREM AND ITS APPLICATION*

RICHARD COLE[†], ŁUKASZ KOWALIK[‡], AND RISTE ŠKREKOVSKI[§]

Abstract. An edge of a graph is *light* when the sum of the degrees of its end-vertices is at most 13. The well-known Kotzig theorem states that every 3-connected planar graph contains a light edge. Later, Borodin [*J. Reine Angew. Math.*, 394 (1989), pp. 180–185] extended this result to the class of planar graphs of minimum degree at least 3. We deal with generalizations of these results for planar graphs of minimum degree 2. Borodin, Kostochka, and Woodall [*J. Combin. Theory Ser. B*, 71 (1997), pp. 184–204] showed that each such graph contains a light edge or a member of two infinite sets of configurations, called 2-alternating cycles and 3-alternators. This implies that planar graphs with maximum degree $\Delta \geq 12$ are Δ-edge-choosable. We prove a similar result with 2-alternating cycles and 3-alternators replaced by five fixed bounded-sized configurations called crowns. This gives another proof of Δ-edge-choosability of planar graphs with $\Delta \geq 12$. However, we show *efficient* choosability; i.e., we describe a linear-time algorithm for $\max\{\Delta, 12\}$-edge-list-coloring planar graphs. This extends the result of Chrobak and Yung [*J. Algorithms*, 10 (1989), pp. 35–51].

Key words. Kotzig's theorem, planar graph, light edge, choosability, list-coloring, algorithm

AMS subject classifications. 05C75, 05C15, 05C85, 68R10

DOI. 10.1137/050646196

1. Introduction.

One of the best-known facts concerning planar graphs states that every planar graph contains a vertex of degree at most 5. Let the *weight* of edge $e = uv$, denoted by $w(e)$, be the sum of the degrees of its end-vertices; i.e., $w(e) = \deg_G(u) + \deg_G(v)$. We say that an edge is *light* when its weight is at most 13. In 1955 Kotzig [12] showed the following theorem.

THEOREM 1.1 (Kotzig). *Every 3-connected planar graph contains a light edge.*

This result was an inspiration for dozens of papers, which form the now so-called *light graph theory* (see the surveys by Jendrol' and Voss [10, 11] and the introduction in [13]).

Kotzig's theorem was generalized in several directions; see, e.g., [2, 7, 15]. In particular, Erdős conjectured that it is valid also for planar graphs with vertices of degree at least 3, and this was proved by Borodin [1].

THEOREM 1.2 (Borodin). *Every simple planar graph with minimum degree $\delta \geq 3$ contains a light edge.*

A light edge is not always present if the graph under consideration has vertices of degree 2; for example, consider the bipartite complete graph $K_{2,k}$ for any $k \geq 12$. In this example each vertex of degree $d \geq 12$ has many 2-neighbors. However, one can guarantee the existence of a light edge by bounding the number of 2-neighbors.

*Received by the editors November 25, 2005; accepted for publication (in revised form) August 14, 2006; published electronically February 9, 2007.
http://www.siam.org/journals/sidma/21-1/64619.html

[†]Computer Science Department, New York University, 251 Mercer Street, New York, NY 10012 (cole@cs.nyu.edu). This author's work was supported by the National Science Foundation under grants CCR 0105678 and CCF 0515127.

[‡]Institute of Informatics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland, and Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany (kowalik@mimuw.edu.pl). This author's work was supported in part by KBN grant 4T11C04425.

[§]Department of Mathematics, University of Ljubljana, Jedranska 19, 1111 Ljubljana, Slovenia (skreko@fmf.uni-lj.si). This author's work was supported in part by Ministry of Science and Technology of Slovenia, Research Program P1-0297.

Proposition 1.3. *Let G be a simple planar graph with minimum degree $\delta \geq 2$ such that each d-vertex, $d \geq 12$, has at most $d - 11$ neighbors of degree 2. Then G contains a light edge.*

Proof. We may assume that every 2-vertex of G is adjacent to two vertices of degree at least 12 for otherwise there is a light edge in G. Consider the graph G' obtained from G by replacing each path uxw such that $\deg(x) = 2$ by an edge joining u and w. Additionally we replace multiple edges by single ones. Clearly G' is a simple planar graph with vertices of degree at least 3, and by Theorem 1.2, G' contains an edge of weight at most 13. Consider such an edge uw.

First assume that u has a 2-neighbor x in G. Then $\deg_G(u) \geq 12$ and in G vertex u has at least 11 neighbors of degree at least 3, which implies that $\deg_{G'}(u) \geq 11$ and hence uw has weight at least 14, a contradiction.

Hence we may assume that u has no 2-neighbor in G and that the same holds for w. It follows that uw belongs to G. Also, $\deg_G(u) = \deg_{G'}(u)$ and $\deg_G(w) = \deg_{G'}(w)$, and hence uw has in G the same weight as in G'. □

Borodin, Kostochka, and Woodall [3] proved the following result, where the number of 2-neighbors is not bounded.

Theorem 1.4 (Borodin, Kostochka, and Woodall). *Every planar graph with minimum degree $\delta \geq 2$ contains a light edge, a 2-alternating cycle, or a 3-alternator.*

In the above theorem a 2-*alternating cycle* is an even length cycle with every second vertex of degree 2, while a 3-*alternator* is a bipartite subgraph F with partite sets U, W such that, for each $u \in U$, $2 \leq \deg_F(u) = \deg_G(u) \leq 3$, and for each $w \in W$, either $\deg_F(w) \geq 3$ or w has exactly two neighbors in U, both of degree $14 - \deg_G(w)$ (the latter case is possible only if $\deg_G(w) = 11$ or 12).

In this paper, we give a similar result involving only five small fixed subgraphs, called *crowns* (see section 2 for the definition and see Figure 2.1 for an illustration), instead of 2-alternating cycles and 3-alternators.

Theorem 1.5. *Every planar graph with minimum degree $\delta \geq 2$ contains a light edge or a k-crown, for some $k \in \{1, \ldots, 5\}$.*

Unlike 2-alternating cycles and 3-alternators the five crowns have bounded size and are contained in the "neighborhood" of one vertex.

1.1. Applications. Let G be a graph. An *edge-list assignment* $L : E(G) \to \mathcal{P}(\boldsymbol{N})$ is a function that assigns to each edge e of G a set (or a *list*) $L(e)$ of *admissible* colors. A function $\lambda : E(G) \to \boldsymbol{N}$ is an *L-edge-coloring* if $\lambda(e) \in L(e)$ for every $e \in E(G)$, and $\lambda(e) \neq \lambda(f)$ for every pair of incident edges $e, f \in E(G)$. If G admits an L-edge-coloring, it is *L-edge-colorable.* For $k \in \boldsymbol{N}$, a graph G is *k-edge-choosable* if it has an L-edge-coloring for every edge-list assignment L such that $|L(e)| \geq k$ for each $e \in E(G)$.

Throughout the paper $\Delta(G)$ will denote the *maximum degree* of graph G, i.e., the largest of the vertex degrees in G. Usually it is clear which graph we refer to and then we simply write Δ.

Although it is conjectured that if a graph is k-edge-colorable, then it is also k-edge-choosable, there is no analogue of Vizing's theorem for list-coloring; i.e., it is not known whether every graph is $\Delta + \mathcal{O}(1)$-choosable. However, Borodin, Kostochka, and Woodall [3] showed the following theorem.

Theorem 1.6 (Borodin, Kostochka, and Woodall). *Every planar graph with maximum degree $\Delta \geq 12$ is Δ-edge-choosable.*

A subgraph of a planar graph is *reducible* when it cannot appear in a minimal counterexample for Theorem 1.6. In this sense, a light edge is reducible (see the paragraph with the heading "Edges of bounded weight" below). In section 3 we show that crowns are reducible. Together with our main result this gives a new proof of Theorem 1.6.

We also consider efficient algorithms for edge-list-coloring planar graphs. Then given an n-vertex graph G and an edge-list assignment L such that lists have length $\max\{\Delta, 12\}$, one has to compute an L-edge-coloring of G. Note that the size of the input is $\Theta(|E(G)|\Delta)$, which is bounded by $\mathcal{O}(n\Delta)$ when G is planar. Hence $\mathcal{O}(n\Delta)$-time algorithms are considered to be linear. Additionally, we assume that each list of admissible colors is sorted. If this assumption is not met, the lists can be bucket-sorted in $\mathcal{O}(|E(G)|\Delta + M)$ time, where M denotes the value of the largest color in the lists. Hence, equivalently one can assume that $M = \mathcal{O}(|E(G)|\Delta)$, which seems to be very natural. We will refer to it as the *small colors assumption*.

The proof of the 2-choosability criterion by Erdős, Rubin, and Taylor [8] (proved earlier by Vizing [14]) yields a linear-time algorithm for optimally edge-list-coloring graphs with $\Delta = 2$. For $\Delta = 3$ there is a linear-time algorithm for 4-edge-list-coloring general graphs due to Gabow and Skulrattanakulchai [9]. For higher values of Δ one can use simple algorithms which rely on the existence in a planar graph of an edge of low weight.

Edges of bounded weight. Assume we want to edge-list-color a planar graph G with maximum degree Δ and with lists of length at least D. When an algorithm finds in G an edge e of weight at most $D + 1$, then this edge is removed and the resulting graph is colored recursively. Since there are at most $D - 1$ edges incident with e, these edges do not use all colors from list $L(e)$, and we can color e with one of the remaining colors. Observe that this proves that light edges are reducible. Also note that when $\Delta = \mathcal{O}(1)$ this algorithm has linear-time complexity. When Δ is not bounded, but the small color assumption holds, the algorithm can also be implemented to work in linear time (see Lemma 4.1). Clearly, any graph can be edge-list-colored from lists of length $D = 2\Delta - 1$, since then any edge has weight at most $D + 1$. For $\Delta = 4, 5$ nothing better is known, even for planar graphs; just note that for these values of Δ, there are planar graphs with all edges of weight 2Δ. For example, consider the octahedron and the dodecahedron. For $\Delta = 6, \ldots, 10$ we can use the result of Borodin [1]: every planar graph of minimum degree at least 4 contains an edge of weight at most 11. Hence any planar graph contains an edge of weight at most $\max\{\Delta + 3, 11\}$ and can be edge-list-colored in linear time from lists of length $\max\{\Delta + 2, 10\}$. For $\Delta \geq 11$ we can take advantage of Theorem 1.2. As before, it immediately yields a linear-time algorithm which requires lists containing $\max\{\Delta + 1, 12\}$ colors. Table 1.1 contains the list of linear-time algorithms for list-edge-coloring planar graphs.

Ordinary edge-coloring. Chrobak and Yung [5] presented a linear-time algorithm for $\max\{\Delta, 19\}$-edge-coloring planar graphs. Although it was not mentioned explicitly, their algorithm can be easily adapted to the list version of the problem. Then its time complexity increases to $\mathcal{O}(n\Delta)$, provided that the small colors assumption holds. There is also an $\mathcal{O}(n \log n)$-time algorithm due to Chrobak and Nishizeki [4] and a very recent $\mathcal{O}(n)$-time algorithm by Cole and Kowalik [6], both for $\max\{\Delta, 9\}$-edge-coloring planar graphs. However, as far as we know neither of these two algorithms can be extended to the edge-list-coloring problem.

Our algorithm. We show an $\mathcal{O}(\Delta n)$-time algorithm for $\max\{\Delta, 12\}$-list-coloring planar graphs. The algorithm does not require a plane embedding of the input graph. This extends the algorithm of Chrobak and Yung [5].

2. The main result. In this section we present the main result of the paper, i.e., a generalization of Kotzig's theorem.

Table 1.1

Linear-time algorithms for list-edge-coloring planar graphs. For $\Delta = 4, 5, \ldots, 11$ the algorithms consist of finding a reducible edge whose existence is obvious or guaranteed by the cited paper.

Δ	Length of lists	Time	Paper
2	optimal	$\mathcal{O}(n)$	Vizing [14]; Erdős, Rubin, and Taylor [8]
3	$\Delta + 1$	$\mathcal{O}(n)$	Gabow, Skulrattanakulchai [9]
$4, 5$	$2\Delta - 1$	$\mathcal{O}(n)$	folklore
$6, 7$	10	$\mathcal{O}(n)$	Borodin [1]
$8, 9, 10$	$\Delta + 2$	$\mathcal{O}(n)$	Borodin [1]
11	$\Delta + 1$	$\mathcal{O}(n)$	Borodin [1]
≥ 12	Δ	$\mathcal{O}(\Delta n)$	this work

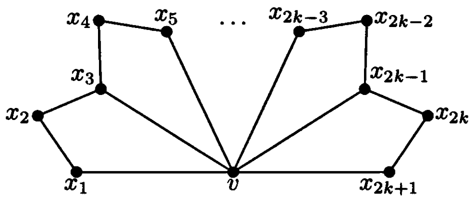


Fig. 2.1. *A k-crown.*

Definition 2.1. *Let G be a multigraph, and let S be a subgraph of G, whose vertices are $v, x_1, x_2, \ldots, x_{2k+1}$ for some $k \geq 1$. We call S a* crown of size k around v *(for short, a k-*crown *or just a* crown*; see Figure 2.1) if the following conditions are satisfied:*

(i) $E(S) = \{vx_{2i+1} : i = 0, \ldots, k\} \cup \{x_i x_{i+1} : i = 1, \ldots, 2k\}$,

(ii) $\deg_G(x_1) = \deg_G(x_{2k+1}) = 2$,

(iii) *for each $i = 1, 2, \ldots, k-1$, $\deg_G(x_{2i+1}) = 3$, and*

(iv) *vertices $v, x_1, x_2, \ldots, x_{2k+1}$ are all distinct.*

Moreover, a crown of size at most 5 will be called a small crown.

Observe that a crown S is not necessarily an induced subgraph of G. Thus, for example, G may have edges vx_2 or x_2x_4 which are not in S. We note here that every edge of a crown S in a graph G has an end-vertex of degree 2 or 3 in G. Thus, if in G one connects two vertices of degree ≥ 3 by an additional edge, then a new crown is not introduced. These remarks will be used later in some arguments. Now we are ready to prove the main result of the paper.

Proof of Theorem 1.5. Clearly, it suffices to prove the result for connected graphs. In this proof we identify planar graphs with their fixed plane embeddings. This allows us to consider faces of these graphs. The *length* of a face f, denoted by $\ell(f)$, is the length of the shortest closed walk induced by all edges incident with f. In order to make the proof easier, we will allow multiple edges and loops in our graphs (where each loop contributes 2 to the degree of its end-vertex) with the following restrictions:

(a) each face of G is of length ≥ 3;

(b) for each 2-vertex, at least one of the faces incident with it is not a triangle, and the two edges incident with it are not parallel.

Clearly, every simple planar graph, except C_3, satisfies these conditions. However, for C_3 the theorem holds trivially.

Suppose that G is a counterexample of the theorem on $|V(G)|$ vertices with the maximum possible number of edges. Let G^* be the graph obtained from G by removing all its 2-vertices.

Claim 1. *G^* is a triangulation.*

The proof is by contradiction. We will show that if G^* is not a triangulation, then G is not maximal, i.e., that one can add an edge to G so that it is still a counterexample for the theorem.

First assume that G^* is disconnected. Then there is a 2-vertex $x \in V(G)$ with neighbors u and v, each of degree ≥ 12, such that u and v belong to different components of G^*. Consider the graph $G \cup \{uv\}$ such that the added edge uv is embedded in a face of G containing u and v. Clearly it is a plane multigraph with neither light edges nor crowns. Note that u and v are not adjacent in G for otherwise they are also adjacent in G^*. Hence $G \cup \{uv\}$ satisfies conditions (a) and (b). This contradicts maximality of G and so G^* is connected. Now it remains to show that every face of G^* is of length 3.

Graph G^* does not contain a face of length 1 for otherwise G contains a 2-vertex incident with parallel edges. If G^* contains a 2-face $f = xyx$, it implies that f contains at least one 2-vertex of graph G. If f contains precisely one 2-vertex, then G violates (b), a contradiction. If f contains at least two 2-vertices, G contains two adjacent 2-vertices or a 1-crown, a contradiction again. Hence each face of G^* has length at least 3.

Suppose that f is a face of G^* of length $k \geq 4$. Since G^* is connected, f has a *facial walk*, i.e., the shortest walk consisting of edges incident with f. Let $x_0x_1, \ldots, x_{x-1}x_k$ be the vertices of this walk in clockwise order, $x_0 = x_k$.

We first prove that if f contains a 2-vertex from G, say w, and x_i, x_j denote the neighbors of w, then $i = j \pm 1 \pmod k$. Otherwise, consider the graph G' obtained from G by connecting x_i and x_j by a new edge x_ix_j. Obviously, G' is planar, because a plane embedding of G' can be obtained from a plane embedding of G by drawing edge x_ix_j in a face of G that contains the 2-walk x_iwx_j. Moreover, G' satisfies the restrictions (a) and (b). Since x_i, x_j have a 2-neighbor, each of them is of degree ≥ 12 in G, which implies that G' has no light edge. Finally observe that conditions $\deg(x_i) \geq 12$ and $\deg(x_j) \geq 12$ imply that no crown contains the new edge x_ix_j, and consequently G' contains no crown. Hence, G' contradicts the maximality of G. This establishes our auxiliary claim, that $i = j \pm 1 \pmod k$.

Since each of x_0x_1 and x_2x_3 has weight at least 14, it easily follows that $\deg(x_0) + \deg(x_2) \geq 14$ or $\deg(x_1) + \deg(x_3) \geq 14$; say the latter holds. Consider the graph $G + x_1x_3$, where x_1x_3 is inserted in f. The above auxiliary claim implies that x_1 and x_3 belong to a common face in G, and hence the resulting graph is planar. Again, one can show that this graph contradicts the maximality of G. This establishes Claim 1.

Note that the above claim implies that G has no bridges, and so the length of a face is the same as the number of (distinct) edges incident with it. Claim 1 and the fact that there are no 1-crowns in G easily imply the following claim.

Claim 2. *Every face f of G is of length $\ell(f) = 3$, 4, 5, or 6. Moreover, for $\ell(f) = 4, 5, 6$ face f is incident with $\ell(f) - 3$ vertices of degree 2.*

Initial charge. Let $F(G)$ denote the set of faces of G. We assign a charge to each vertex and face of G. For every $x \in V(G)$, we define the initial charge $c(x) = \deg(x) - 4$. Similarly, for every $f \in F(G)$, let $c(f) = \ell(f) - 4$. By Euler's formula the total sum of charge assigned to vertices and faces is

$$\begin{aligned}\sum_{x \in V(G) \cup F(G)} c(x) &= \sum_{v \in V(G)} (\deg(v) - 4) + \sum_{f \in F(G)} (\ell(f) - 4) \\ &= 2|E(G)| - 4|V(G)| + 2|E(G)| - 4|F(G)| = -8. \qquad (2.1)\end{aligned}$$

Notice that only 2-vertices, 3-vertices, and 3-faces have negative initial charge. Our goal is to redistribute charge between vertices and faces according to prescribed rules in such a way that the total sum of charge will be nonnegative, which will contradict (2.1). This contradiction will settle the theorem.

Rules. We use the following discharging rules to redistribute charge between vertices and faces.

(R1) A 2-vertex receives 1 unit from each of its two neighbors.

(R2) A 3-vertex receives 1/3 of a unit from each of its three neighbors.

(R3) A 3-face $v_1v_2v_3$ with $\deg(v_1) \leq 5$ receives 1/2 of a unit from each of v_2 and v_3.

Let f be a face and let v_1, v_2, v_3 be three consecutive vertices incident with f such that $\deg(v_2) \geq 6$.

(R4) If both v_1 and v_3 are of degree ≥ 6, then v_2 sends 1/3 of a unit to f.

(R5) If $\ell(f) \geq 4$, one of v_1, v_3 is of degree 2, and the other is of degree ≥ 6, then v_2 receives 1/6 from f.

(R6) If $\ell(f) \geq 4$ and both of v_1, v_3 are of degree 2, then v_2 receives 2/3 from f.

Since we deal with multigraphs, the multiple incidence/adjacency is considered in the application of these rules. Thus, for example, if a 3-vertex x is adjacent to a vertex v by two edges, then v sends the amount $\frac{1}{3} + \frac{1}{3}$ of a unit of charge to x by (R2).

Final charge. Here we will prove that for each $x \in V(G) \cup F(G)$, the final charge $c^*(x)$ is nonnegative; i.e., $c^*(x) \geq 0$. Let f be an arbitrary face of G. By Claim 2, $\ell(f) \in \{3, 4, 5, 6\}$. Hence we consider four cases:

$\ell(f) = 3$: If f contains a vertex of degree at most 5, then $c^*(f) = 0$ by (R3). Otherwise, all three neighbors are of degree ≥ 6, so f gets 1/3 from each of them by (R4). Hence, $c^*(f) = 0$.

$\ell(f) = 4$: In this case, by Claim 2, f contains exactly one 2-vertex. Let $f = x_1x_2x_3x_4$ with $\deg(x_4) = 2$. If $\deg(x_2) \leq 5$, then f sends no charge, and so $c(f) = c^*(f) = 0$. If $\deg(x_2) \geq 6$, f gets 1/3 from x_2 by (R4) and sends 1/6 to each of x_1 and x_3 by (R5). This yields $c^*(f) = 0$.

$\ell(f) = 5$: By Claim 2, f contains exactly two 2-vertices, and so we can assume that $f = x_1x_2x_3x_4x_5$ with $\deg(x_1) = \deg(x_3) = 2$. Then f sends 1/6 to each of x_4, x_5 by (R5) and sends 2/3 to x_2 by (R6). Hence, $c^*(f) = 0$.

$\ell(f) = 6$: By Claim 2, f has three 2-vertices alternating with three vertices of degree at least 12. Each of the latter receives 2/3 by (R6), which implies that the final charge of f is 0.

We consider now the final charge of the vertices. By rules (R1) and (R2), it is obvious that 2- and 3-vertices have nonnegative final charge and that 4- and 5-vertices do not alter their charge, which is nonnegative.

Suppose now that a vertex v is of degree $d \in \{6, 7, 8\}$. Then, it may send charge only to incident faces by rule (R4). Moreover, if some incident face is a triangle, then its two other vertices have degrees at least 6, which implies that each such triangle receives 1/3 from v. Hence,

$$c^*(v) \geq d - 4 - \frac{d}{3} \geq 0.$$

Next suppose that v is of degree $d \in \{9, 10\}$. It may send charge only to incident faces by rules (R3) and (R4), and each such face receives at most 1/2 from v. Hence,

$$c^*(v) \geq d - 4 - \frac{d}{2} \geq 0.$$

Suppose now that v is of degree 11. Notice that v is not adjacent to a 2-vertex, and so it sends charge to a neighbor only if it is a 3-vertex. Since by Claim 2, no two 3-neighbors of v are consecutive in clockwise order around v, the number of 3-neighbors is at most 5. Notice that v sends 1/2 to at most 10 faces, and to the remaining faces it sends at most 1/3. Hence,

$$c^*(v) \geq 7 - \frac{10}{2} - \frac{1}{3} - \frac{1}{3} \cdot 5 = 0.$$

Finally suppose that $d \geq 12$. Let $x_0, x_1, \ldots, x_{d-1}$ be the neighbors of v enumerated in clockwise order around v, and let f_i be the face incident with the walk $x_i v x_{i+1}$ (throughout this proof we take the indices in x_i modulo d). We consider a few cases regarding the number d_2 of 2-vertices adjacent to v.

Case 1: $d_2 = 0$. Since v sends at most 1/2 to each incident face and has at most $\lfloor \frac{d}{2} \rfloor$ adjacent 3-neighbors, its final charge is

$$c^*(v) \geq d - 4 - \frac{d}{2} - \frac{1}{3}\left\lfloor \frac{d}{2} \right\rfloor \geq \frac{d}{3} - 4 \geq 0.$$

Case 2: $d_2 = 1$. Let x_1 be the 2-neighbor of v. By Claim 1, without loss of generality we may assume that f_0 is a 3-face and f_1 is a face of length 4 or 5 (f_1 cannot be a face of length 6 since then f_1 contains two 2-neighbors of v, and so $d_2 \geq 2$). Notice that v sends 1 to x_1 and 1/2 to f_0. Next, it sends nothing to f_1 and $\leq 1/2$ to each of the $d-2$ remaining faces. Finally, it sends at most $\frac{1}{3}\left\lfloor \frac{d-1}{2} \right\rfloor$ to its adjacent 3-vertices. If $d \geq 13$, then

$$c^*(v) \geq d - 4 - 1 - \frac{1}{2} - \frac{d-2}{2} - \frac{1}{3}\left\lfloor \frac{d-1}{2} \right\rfloor \geq 0.$$

Now assume that $d = 12$. We consider two subcases regarding the degree of x_2. If $\deg(x_2) \geq 6$, then f_1 sends $\frac{1}{6}$ to v by (R5), and we conclude

$$c^*(v) \geq d - 4 - 1 - \frac{1}{2} - \frac{d-2}{2} - \frac{1}{3}\left\lfloor \frac{d-1}{2} \right\rfloor + \frac{1}{6} = 0.$$

Finally, since d is even, if $\deg(x_2) \leq 5$, then there is a face distinct from f_1 that receives at most 1/3 from v. In that case, we obtain

$$c^*(v) \geq d - 4 - 1 - \frac{1}{2} - \frac{d-3}{2} - \frac{1}{3} - \frac{1}{3}\left\lfloor \frac{d-1}{2} \right\rfloor = 0.$$

Case 3: $d_2 \geq 2$. Observe that since the rules move charge only between incident faces and vertices, while calculating the charge sent by v we can restrict ourselves only to v and its adjacent vertices and incident faces. In order to make the argument shorter, we use the following claim.

Claim 3. *We can modify the neighborhood of v so that every 2-vertex x_i is adjacent to x_{i-1} and the final charge $c^*(v)$ stays the same.*

Let $\deg_G(x_i) = 2$. Then by Claim 1, x_i is adjacent to x_{i-1} or x_{i+1}. Assume that it is adjacent to x_{i+1}. Then x_{i+2} is not a 2-vertex adjacent to x_{i+1}, since G does not contain a 1-crown. Then we remove x_i and draw it inside face f_{i+1} together with the edges to v and x_{i+1}. In the new drawing, let us rename the vertices and faces so that they are still enumerated in clockwise order. In particular, x_{i+1} is renamed as x'_i, x_i

is renamed as x'_{i+1}, and for every $j \neq i, i+1$, vertex x_j is renamed as x'_j. In the new drawing, let f'_i be the face incident with the walk $x'_i v x'_{i+1}$. Let c_j (respectively, c'_j) be the charge sent from v to f_j (respectively, f'_j) minus the charge received by v from f_j (respectively, f'_j). Obviously the charge sent/received by v to/from neighbors of v has not changed. Also, $c'_j = c_j$ for $j \neq i-1, i+1$. If $\deg_G(x_{i-1}) = 2$, then by Claim 2, f_{i-1} is of length 5 or 6, so by (R5) and (R6), $c'_{i-1} - c_{i-1} = -1/6 - (-2/3) = 1/2$. If $\deg_G(x_{i-1}) = 3, 4, 5$, then there is no 2-vertex adjacent to x_{i-1}, so f_{i-1} is a 4-face and f'_{i-1} is a 3-face; hence by (R3), $c'_{i-1} - c_{i-1} = 1/2$. Finally, when $\deg_G(x_{i-1}) \geq 6$, then f_{i-1} is of length 4 or 5, so by (R4) and (R5), $c'_{i-1} - c_{i-1} = 1/3 - (-1/6) = 1/2$. Hence $c'_{i-1} - c_{i-1} = 1/2$ in all cases. Analogously one can verify that no matter what the degree of x_{i+2} is, $c'_{i+1} - c_{i+1} = -1/2$. Hence the charge sent from v remains the same. This settles the claim.

We modify the neighborhood of v as described in Claim 3. Note that if x_i is a 2-vertex, then its neighbor x_{i-1} is of degree ≥ 12. Obviously, this redrawing in Claim 3 introduces neither a crown nor a pair of consecutive v neighbors of v of degree 3, 4, or 5. Also, G^* stays unchanged.

In what follows, we will bound the amount of charge sent by v to faces. Denote by $d_{4,5}$ the number of 4- and 5-neighbors of v. Denote by $f_{-1/6}$ and $f_{1/3}$ the number of faces which send 1/6 to v or receive 1/3 from v, respectively. Let x_i and x_j be two distinct 2-neighbors of v, such that for each $k \in \{i+1, \dots, j-1\}$, $\deg(x_k) > 2$. If there is a crown whose vertices belong to $\{v, x_{i-1}, x_i, x_{i+1}, \dots, x_j\}$, we call the (ordered) pair (x_i, x_j) *bad*; otherwise it is *good*. Let b denote the number of bad pairs. Note that there are $d_2 - b$ good pairs.

Claim 4. *For any good pair (x_i, x_j) one of the following conditions holds:*

(A) $\deg_G(x_{i+1}) \geq 6$,

(B) *for some* $k \in \{i+1, \dots, j-2\}$, $\deg(x_k) \geq 6$ *and* $\deg(x_{k+1}) \geq 6$, *or*

(C) *for some* $k \in \{i+1, \dots, j-2\}$, $\deg_G(x_k) \in \{4, 5\}$.

Assume that none of the above conditions holds. Note that by Claim 3, $j \neq i+1$. Then the following property holds: for each $k \in \{i+1, \dots, j-1\}$, $\deg_G(x_k) \geq 6$ if k has the same parity as i, and $\deg_G(x_k) = 3$ otherwise. Let H be the subgraph of G with $V(H) = \{v, x_{i-1}, x_i, x_{i+1}, \dots, x_j\}$ and $E(H) = \{vx_k : k \in \{i-1, i, \dots, j\}\} \cup \{x_{i-1}x_i, x_{i-1}x_{i+1}\} \cup \{x_k x_{k+1} : k \in \{i+1, \dots, j-1\}\}$. Then H is a crown around v, unless some pair of its vertices x_a, x_b coincide. Notice that then $\deg(x_a) = \deg(x_b) \geq 6$. As long as there is such a pair in H we remove from H all the vertices and edges inside the 2-cycle $vx_a x_b$ and we remove edge vx_b. The resulting subgraph H is a crown around v with vertices in the set $\{v, x_{i-1}, x_i, x_{i+1}, \dots, x_j\}$, which is a contradiction. This settles the claim.

Observe that in case (A) face f_i sends 1/6 to v by (R5), and in case (B) face f_k receives precisely 1/3 from v by (R4). As there are $d_2 - b$ good pairs, it follows that $f_{-1/6} + f_{1/3} + d_{4,5} \geq d_2 - b$. Thus, some $d_2 - b - f_{-1/6} - d_{4,5}$ faces receive precisely 1/3 from v. Note that for any 2-vertex x_i, the face f_i does not receive a charge from v. Thus, there are d_2 faces which do not receive any charge from v. Each of the remaining $d - d_2 - (d_2 - b - f_{-1/6} - d_{4,5})$ faces receives at most 1/2 unit from v. Now we bound the total charge sent from v to faces minus the charge received from faces. It amounts to at most

$$\frac{1}{3}\left(d_2 - b - f_{-1/6} - d_{4,5}\right) + \frac{1}{2}\left[d - d_2 - (d_2 - b - f_{-1/6} - d_{4,5})\right] - \frac{1}{6} f_{-1/6}$$

$$= \frac{d}{2} - \frac{2}{3} d_2 + \frac{b}{6} + \frac{d_{4,5}}{6}. \tag{2.2}$$

In what follows we estimate the charge v sends to neighbors. We start from bounding the number of 3-neighbors of v. Consider (cyclically) the degree sequence $S_0 = \deg(x_0), \deg(x_1), \dots, \deg(x_{d-1})$. First remove elements with value 2 from this sequence. If two consecutive elements of the resulting sequence S_1 each have value at least 6, we will call them a *big pair*. Observe that if (A) holds in Claim 4, then by Claim 3 $\deg_G(x_{i-1}) \geq 12$ and, consequently, $\deg_G(x_{i-1})$ and $\deg_G(x_{i+1})$ are a big pair. Hence by Claim 4, in S_1 there are at least $(d_2-b)-d_{4,5}$ big pairs (we consider the last element of S_1 to be consecutive with the first one). Next, as long as the sequence contains a big pair we remove one of the elements of the pair, unless the sequence consists of only two elements, each of value at least 6. In the latter case both these elements are removed. After these two steps, the resulting sequence S_2 has length $\leq d - d_2 - (d_2 - b - d_{4,5})$. By Claim 1, and because edges have weight at least 14, it follows that in G^* vertex v has no pair of consecutive neighbors both of degree 3, 4, or 5. It follows that sequence S_2 does not contain a pair of consecutive elements equal to 3, 4, or 5. Thus, S_2 contains at most $\lfloor \frac{d-d_2-(d_2-b-d_{4,5})}{2} \rfloor = \lfloor \frac{d+b+d_{4,5}}{2} \rfloor - d_2$ elements equal to 3, 4, or 5, and hence this is an upper bound for the number of 3-, 4-, and 5-neighbors of v. It follows that v has at most $\lfloor \frac{d+b+d_{4,5}}{2} \rfloor - d_2 - d_{4,5} = \lfloor \frac{d+b-d_{4,5}}{2} \rfloor - d_2$ neighbors of degree 3. Thus, the total charge sent from v to its neighbors is at most

$$d_2 + \frac{1}{3}\left(\left\lfloor \frac{d+b-d_{4,5}}{2} \right\rfloor - d_2\right). \tag{2.3}$$

Finally, by (2.2) and (2.3) we conclude that

$$\begin{aligned} c^*(v) &\geq d - 4 - d_2 - \frac{1}{3}\left(\left\lfloor \frac{d+b-d_{4,5}}{2} \right\rfloor - d_2\right) - \left(\frac{d}{2} - \frac{2}{3}d_2 + \frac{b}{6} + \frac{d_{4,5}}{6}\right) \\ &\geq \frac{d}{3} - 4 - \frac{b}{3}. \end{aligned}$$

Each k-crown contains $k-1$ vertices of degree 3, which are neighbors of v. For each bad pair (x_i, x_j) there is a crown with vertices from $\{v, x_{i-1}, \dots, x_j\}$. Since small crowns are excluded, such a crown contains at least five 3-neighbors of v. Hence v has at least $5b$ neighbors of degree 3. By Claim 1, each 3-neighbor of v is incident in G^* with two triangular faces containing v. Each of these faces also contains a neighbor of v of degree at least 11, as light edges are excluded. The edge joining v and its neighbor can belong to at most 2 of these faces. Consequently there are at least $5b$ edges joining v and its neighbors of degree at least 11. Finally, v has at least b neighbors of degree 2. It follows that $\deg_G(v) \geq 11b$ and so $b \leq \lfloor \frac{d}{11} \rfloor$.

Hence for $d \geq 14$, we get $c^*(v) \geq \frac{d}{3} - 4 - \frac{1}{3} \cdot \frac{d}{11} > 0$. For $d = 13$, we get $c^*(v) \geq \frac{d}{3} - 4 - \frac{1}{3} = 0$. Observe that Claim 1 implies that all the vertices of a crown around v, except for v, are adjacent to v. Hence a crown around v implies that at least 13 edges are incident with v, for it has size at least 6. Consequently, for $d = 12$, there are no crowns around v and $c^*(v) \geq \frac{d}{3} - 4 = 0$.

This completes the case $d \geq 12$. We infer that every vertex and face has nonnegative charge after the rules are applied, which is a contradiction. This establishes the proof. □

In Theorem 1.5 the number 5 is best possible in the sense that there is a planar graph with minimum degree 2 with no crowns of size smaller than 5 and with no light edges. To construct such a graph take a triangulation T with vertices of degree 5 and 6 such that 5-vertices are at distance at least 5 from each other; for example, the duals of some fullerens are such graphs. Then, for each 5-vertex x of T we choose one incident triangle and remove its edge not incident with x. As a result we get a graph T' with faces of length 3 and 4. Next, we put a vertex into each face of T' and join it with the vertices incident with the face. Denote the resulting triangulation by T''. Observe that every light edge in T'' joins a 3-vertex with a 10-vertex. Moreover, the 10-vertex is adjacent to a 4-vertex. For each 4-vertex $y \in V(T'')$ let its neighbors be y_0, y_1, y_2, y_3 in clockwise order. Finally, for each $i \in \{0, 1, 2, 3\}$ we add a new 2-vertex connected to y_i and y_{i+1} (indices modulo 4). Clearly, the resulting graph G has vertices of degree 2, 3, 12, and 14 only. Vertices of degree 2 and 3 are adjacent to vertices of degree 12 or 14. Hence there are no light edges in G. One may verify that G contains crowns of size 5 and 6 but no crowns of smaller size.

3. Reducibility of crowns. In this section we show that crowns are reducible. Although we use crowns of size at most 5, here we consider all crowns. In the next lemma we will use the well-known fact that every even cycle is 2-edge-choosable.

LEMMA 3.1. *Let G be a graph of maximum degree Δ and let S be a k-crown in G, $k \geq 1$. Let L be a list assignment of G such that $|L(e)| \geq \Delta$ for every edge $e \in E(G)$. Then any L-coloring of $G - E(S)$ can be extended to an L-coloring of G.*

Proof. Let λ be an arbitrary L-edge-coloring of $G - E(S)$. For every $e \in E(S)$, let $I(e)$ denote the set of edges from $E(G) - E(S)$ that are incident with e and let $L'(e) = L(e) \setminus \bigcup_{f \in I(e)} \lambda(f)$. Let us denote the vertices of S as in Figure 2.1. Recall that $\deg_G(x_1) = \deg_G(x_{2k+1}) = 2$ and for every $i = 3, 5, \ldots, 2k-1$, $\deg_G(x_i) = 3$. Note that for $i = 1, 3, \ldots, 2k+1$, $|L'(vx_i)| \geq k+1$, and for $i = 1, 2, \ldots, 2k$, $|L'(x_i x_{i+1})| \geq 2$. Without loss of generality we may assume that for $i = 1, 3, \ldots, 2k+1$, $|L'(vx_i)| = k+1$, and for $i = 1, 2, \ldots, 2k$, $|L'(x_i x_{i+1})| = 2$. Clearly in order to extend λ to an L-coloring of G it suffices to L'-color the graph S. Thus our objective will be to construct an L'-coloring of S, where L' is any list assignment with the above prescribed lengths of lists. We do so by induction on k. For $k = 1$ we must 2-list-color a 4-cycle, but even-length cycles are 2-choosable [8, 14].

Now, we consider the case $k = 2$. We may assume that $L'(vx_3) \subseteq L'(x_2x_3) \cup L'(x_3x_4)$, for otherwise we color vx_3 with a color from $L'(vx_3) \setminus [L'(x_2x_3) \cup L'(x_3x_4)]$ and then we are left with the problem of 2-list-coloring of a 6-cycle. Since $|L'(vx_3)| = 3$, it follows that $L'(x_2x_3) \neq L'(x_3x_4)$. Then we color x_2x_3 with a color not in $L'(x_3x_4)$ and we color x_1x_2 with a free color. We assume now that vx_3 has two free colors; otherwise we remove one. We may also assume that vx_3 and x_4x_5 do not have a common free color, for otherwise we color them both with such a color and then we can color vx_1, vx_5, x_3x_4, in this order, always using a free color. Since vx_5 has three free colors and both vx_3, x_4x_5 have two free colors, either vx_3 or x_4x_5 has a free color $p \notin L'(vx_5)$. In the case $p \in L'(vx_3)$ we color vx_3 with p and then we color the remaining edges in the following order: vx_1, x_3x_4, x_4x_5, vx_5. In the latter case we assign color p to x_4x_5 and color x_3x_4, vx_3, vx_1, vx_5, in this order, always using a free color. This settles the case $k = 2$.

Now assume $k \geq 3$. We consider the following two possibilities.

Case 1: $L'(x_2x_3) = L'(x_3x_4)$. Let r be a color from $L'(vx_3) \setminus L'(x_2x_3)$. We remove x_3 and identify x_2 with x_4. For each $i = 1, 3, 4, \ldots, k+1$, let $L''(vx_{2i-1}) = L'(vx_{2i-1}) \setminus \{r\}$. The resulting graph is a $(k-1)$-crown, and it is L''-colorable by the induction hypothesis. Let λ'' be such a coloring. We extend λ'' to an L'-coloring of S as follows. Let $p \in L'(x_2x_3) \setminus \{\lambda''(x_1x_2)\}$ and $q \in L'(x_3x_4) \setminus \{\lambda''(x_4x_5)\}$. Since $L'(x_2x_3) = L'(x_3x_4)$ and $\lambda''(x_1x_2) \neq \lambda''(x_4x_5)$, it follows that $p \neq q$. Hence we can color x_2x_3 with p, x_3x_4 with q, and vx_3 with r.

Case 2: $L'(x_2x_3) \neq L'(x_3x_4)$. Let $L'(x_2x_3) = \{a, b\}$ and $c \in L'(x_3x_4)$, $c \notin \{a, b\}$. Then we color vx_3 with a color distinct from a, b, and c. This is possible since $|L'(vx_3)| = k + 1 \geq 4$. Next, we color x_3x_4 with c and we color $x_4x_5, x_5x_6, \ldots, x_{2k}x_{2k+1}$, in this order, always using a free color. Now for every $i = 5, 7, \ldots, 2k - 1$, vx_i has at least $k - 2$ free colors and vx_{2k+1} has at least $k - 1$ free colors. Hence, we may color them greedily, i.e., in the order $vx_5, vx_7, \ldots, vx_{2k+1}$, always using a free color. Afterwards vx_1 has at least one free color, and both x_1x_2, x_2x_3 have two free colors, so we color them greedily as well. □

Theorem 1.5 and Lemma 3.1 imply the following corollary.

COROLLARY 3.2. *Every planar graph with maximum degree $\Delta \geq 12$ is Δ-edge-choosable.*

4. List-edge-coloring algorithm. In this section we describe a linear-time algorithm which, for a given simple planar graph G and an edge-list assignment L, computes an L-edge-coloring of G, provided that for every $e \in E(G)$, $|L(e)| = \max\{\Delta(G), 12\}$. The algorithm does not need a plane embedding of graph G. In fact, one can use the algorithm for any class of graphs which can replace planar graphs in Theorem 1.5.

We assume that the input graph G is given in the form of adjacency lists. Also the list assignment is stored as an array of lists, one list for each edge. Additionally, we assume that each list of admissible colors is sorted. Equivalently, one can assume that the largest color has value $\mathcal{O}(|E(G)|\Delta)$. Then the lists can be sorted in linear time using bucket-sort.

In the following subsection we describe some tools used by our coloring algorithm. Then we describe the main body of the algorithm and analyze its time complexity.

4.1. Efficient coloring and finding small crowns.

LEMMA 4.1. *Let G be a graph of maximum degree Δ containing an edge e of weight at most $\max\{\Delta + 1, 13\}$. Let L be an edge-list assignment of G such that $|L(e)| \geq \max\{\Delta, 12\}$ for every edge $e \in E(G)$. Then any L-edge-coloring of $G - \{e\}$ can be extended to an L-edge-coloring of G in $\mathcal{O}(\Delta)$ time.*

Proof. Let λ denote the L-edge-coloring of $G - \{e\}$, let $I(e)$ denote the set of edges incident with e, and let $L'(e) = L(e) \setminus \bigcup_{f \in I(e)} \lambda(f)$. Clearly $|L'(e)| \geq 1$. The algorithm simply colors e with any color from $L'(e)$. In order to find $L'(e)$ efficiently, each vertex x in graph G stores a sorted list $\mathsf{Used}(x)$ of colors used by the already colored incident edges. As the list $L(e)$ is also sorted, the set $L'(e)$ can be easily found in $\mathcal{O}(\Delta)$ time. Additionally, after coloring the edge $e = xy$, both lists $\mathsf{Used}(x)$ and $\mathsf{Used}(y)$ are updated in $\mathcal{O}(\Delta)$ time. □

The following lemma states that the proof of Lemma 3.1 can be transformed into an efficient algorithm when $k = \mathcal{O}(1)$.

LEMMA 4.2. *Let G be a graph of maximum degree Δ and let S be a k-crown in G, $k = \mathcal{O}(1)$. Let L be an edge-list assignment of G such that $|L(e)| \geq \Delta$ for every edge $e \in E(G)$. Then any L-edge-coloring of $G - E(S)$ can be extended to an L-edge-coloring of G in $\mathcal{O}(\Delta)$ time.*

Proof. We consider the algorithm arising from the proof of Lemma 3.1. Each of the sets $L'(e)$ from the proof of Lemma 3.1 is computed in $\mathcal{O}(\Delta)$ time, as described in the proof of Lemma 4.1. As $k = \mathcal{O}(1)$, this whole phase takes $\mathcal{O}(\Delta)$ time. Afterwards, we deal with bounded-sized graphs and bounded-sized list assignments; hence the remaining part of the coloring algorithm takes constant time. Finally, as in the proof of Lemma 4.1, relevant sets $\mathsf{Used}(\cdot)$ are updated in $\mathcal{O}(\Delta)$ time. □

Now we consider algorithm SEARCHSMALLCROWN(G, x) (see Algorithm 4.1), which will be used for searching for small crowns.

ALGORITHM 4.1. SEARCHSMALLCROWN(G, x): Searching for a small crown.

```
 1: for each $v \in N(x)$ do
 2:     $H \leftarrow (\emptyset, \emptyset)$                                  ▷ $H$ is the empty graph
 3:     for each $y \in N(v)$ do
 4:         if $\deg_G(y) \in \{2, 3\}$ then
 5:             for each $z \in N(y) \setminus \{v\}$ do
 6:                 if $\deg_G(z) \leq 3$ then
 7:                     return $\{yz\}$                          ▷ $yz$ is a light edge
 8:                 else
 9:                     $V(H) \leftarrow V(H) \cup \{y, z\}$; $E(H) \leftarrow E(H) \cup \{yz\}$
10:     Find in $H$ a vertex $\bar{y}$ such that $\deg_G(\bar{y}) = 2$ and $\mathrm{dist}_H(x, \bar{y})$ is as small as possible.
11:     if $\bar{y}$ exists then
12:         $P \leftarrow$ the shortest path in $H$ between $\bar{y}$ and another vertex of degree 2 in $G$
13:         if $P \neq \emptyset$ and $|E(P)| \leq 10$ then
14:             $C \leftarrow E(P) \cup \{vw : w \in V(P)$ and $\deg_G(w) \in \{2, 3\}\}$
15:             return $C$                                   ▷ $C$ is an $(|E(P)|/2)$-crown.
16: return $\emptyset$
```

LEMMA 4.3. *Let x and v be distinct vertices in a graph G and let $\deg_G(x) \in \{2, 3\}$. Assume that in G there is a small crown around v containing x. Then the algorithm* SEARCHSMALLCROWN(G, x) *returns a light edge or the edges of a small crown. Moreover, its time complexity is $\mathcal{O}(\Delta)$.*

Proof. First assume that the algorithm returns set C in line 15. We will show that C contains the edges of a small crown. Since a light edge was not returned in line 7, then for some vertex v, which is a neighbor of x,

$$E(H) = \{yz : y \in N(v),\ \deg_G(y) \in \{2, 3\},\ z \in N(y) - \{v\},\ \text{and}\ \deg_G(z) > 3\}. \tag{4.1}$$

Note that H is a bipartite graph with partite sets $Y = \{y \in V(H) : \deg_G(y) \in \{2, 3\}\}$ and $Z = \{z \in V(H) : \deg_G(z) > 3\}$. Hence P has even length, as both its ends have degree 2 in G. Let $y_0, z_1, y_1, z_2, y_2, \ldots, z_{|E(P)|/2}, y_{|E(P)|/2}$ be the successive vertices of P. Note that these vertices are all distinct for otherwise P is not the shortest path in H between $\bar{y}$ and another 2-vertex. By (4.1) each vertex y_i of path P is adjacent to v. Note that P contains at least two edges, since it has distinct ends. It follows that C contains edges of a crown around v of size $|E(P)|/2 \leq 5$.

Now it suffices to show that the algorithm returns a light edge in line 7 or returns set C in line 15. Assume that neither of these happens. Let S be a small crown around v containing x, $v \neq x$. (S exists by the assumptions of the lemma.) Let k denote the size of S. In lines 2 to 9 the algorithm finds the subgraph $H \subseteq G$ with edge set described in (4.1). Let x_1, x_2 be the neighbors of v in S with degree 2 in G. Observe that $E(S - v) \subseteq E(H)$. In line 10 the algorithm finds some vertex $\bar{y}$, because S contains x, x_1, and x_2 (possibly $x = x_1$ or $x = x_2$). If $\bar{y} = x_1$ (respectively, $\bar{y} = x_2$), then there is a path in H from $\bar{y}$ to another vertex of degree 2 in G, namely x_2 (respectively, x_1). Consequently, when $\bar{y} \in \{x_1, x_2\}$, the algorithm finds some path P in line 12, and $|E(P)| \leq 2k$. Also if $\bar{y} \notin \{x_1, x_2\}$, then H contains a path from $\bar{y}$ to x and a path from x to x_1; hence some path P is found. Moreover, then

$\text{dist}_H(x, \bar{y}) \leq \min\{\text{dist}_H(x, x_1), \text{dist}_H(x, x_2)\}$ and so

$$\begin{aligned} |E(P)| &\leq \text{dist}_H(\bar{y}, x) + \min\{\text{dist}_H(x, x_1), \text{dist}_H(x, x_2)\} \\ &\leq 2\min\{\text{dist}_H(x, x_1), \text{dist}_H(x, x_2)\} \\ &\leq 2k. \end{aligned}$$

It follows that $|E(P)| \leq 10$. Hence line 15 is executed, a contradiction. This proves that the algorithm returns a small crown or a light edge.

Clearly, graph H has $\mathcal{O}(\Delta)$ size and building it takes $\mathcal{O}(\Delta)$ time. The other part of the algorithm can be easily implemented using breadth first search, and then it takes time linear with respect to the size of H, i.e., $\mathcal{O}(\Delta)$ time. □

4.2. Main body of the algorithm. Now we describe algorithm EdgeListColor, which edge-list-colors an input simple planar graph G with edge color lists of length $\max\{\Delta(G), 12\}$. Our algorithm uses a queue Q which stores vertices around which one should look for light edges and small crowns. It is initialized with the set of all vertices of G. However, one vertex may appear several times in Q.

Algorithm 4.2. EdgeListColor(G): List-edge-coloring planar graph G.

```
RecursiveColor($G$)
 1: $C \leftarrow \emptyset$
 2: while $C = \emptyset$ do
 3:     $x \leftarrow$ a vertex from queue $Q$
 4:     if $\deg_G(x) = 1$ then
 5:         $y \leftarrow$ the sole neighbor of $x$; $C \leftarrow \{xy\}$
 6:     else if $x$ is incident with a light edge $xy$ then
 7:         $C \leftarrow \{xy\}$
 8:     else if $\deg_G(x) \in \{2, 3\}$ then
 9:         $C \leftarrow$ SearchSmallCrown($G, x$)
10:     if $C = \emptyset$ then $Q \leftarrow Q \setminus \{x\}$
11: $Q \leftarrow Q \cup V(C)$
12: $E(G) \leftarrow E(G) \setminus E(C)$
13: if $E(G) \neq \emptyset$ then
14:     RecursiveColor($G$)
15: $E(G) \leftarrow E(G) \cup E(C)$
16: Color edges from $E(C)$ according to Lemma 4.1 or Lemma 4.2

EdgeListColor($G$)
 1: $Q \leftarrow V(G)$
 2: RecursiveColor($G$)
```

After the initialization the algorithm calls a recursive routine RecursiveColor (see Algorithm 4.2). Let us consider one such recursive call. Consider the following assertion:

> *Q contains all 1-vertices and endpoints of light edges in G; for any small crown C around v in G, queue Q contains a vertex $x \in V(C) \setminus \{v\}$, $\deg_G(x) \in \{2, 3\}$.*

Obviously, the assertion holds after initialization. Then, each time some set of edges is removed from the graph, the endpoints of these edges are added to Q in line 11. Also, if a vertex x is removed from Q and not inserted again, then $\deg_G(x) \neq 1$, there is no light edge incident with x, and either $\deg_G(x) \notin \{2,3\}$ or there is no small crown containing it. This proves that the assertion always holds at the beginning of the RecursiveColor routine. The assertion together with Theorem 1.5 and Lemma 4.3 guarantees that in line 11 set C contains a single edge of weight $\Delta + 1$, a single light edge, or the edges of a small crown. This easily implies the following corollary.

Corollary 4.4. *Algorithm* EdgeListColor *properly colors the input planar graph.*

Proposition 4.5. *Algorithm* EdgeListColor *works in* $\mathcal{O}(|V(G)|\Delta)$ *time.*

Proof. Since in each recursive call at least one edge is removed, there are $\mathcal{O}(|E(G)|) = \mathcal{O}(|V(G)|)$ recursive calls. In each recursive call $\mathcal{O}(1)$ vertices are added to Q; hence in total $\mathcal{O}(|V(G)|)$ vertices are added to Q. A straightforward implementation of line 6 works in $\mathcal{O}(\Delta)$ time. Line 9 takes $\mathcal{O}(\Delta)$ time by Lemma 4.3. Hence the total time spent on lines 1–10 is $\mathcal{O}(|V(G)|\Delta)$.

Finally, as the number of recursive calls is $\mathcal{O}(|V(G)|$, by Lemmas 4.1 and 4.2 the total time spent on lines 11–16 is $\mathcal{O}(|V(G)|\Delta)$. This settles the proof. □

Acknowledgment. The authors thank anonymous referees for reading the paper carefully and for providing many helpful comments.

REFERENCES

[1] O. V. Borodin, *On the total coloring of planar graphs*, J. Reine Angew. Math., 394 (1989), pp. 180–185.

[2] O. V. Borodin, *Joint extension of two theorems of Kotzig on* 3*-polytopes*, Combinatorica, 13 (1993), pp. 121–125.

[3] O. V. Borodin, A. V. Kostochka, and D. R. Woodall, *List edge and list total colourings of multigraphs*, J. Combin. Theory Ser. B, 71 (1997), pp. 184–204.

[4] M. Chrobak and T. Nishizeki, *Improved edge-coloring algorithms for planar graphs*, J. Algorithms, 11 (1990), pp. 102–116.

[5] M. Chrobak and M. Yung, *Fast algorithms for edge-coloring planar graphs*, J. Algorithms, 10 (1989), pp. 35–51.

[6] R. Cole and Ł. Kowalik, *New linear-time algorithms for edge-coloring planar graphs*, Algorithmica, to appear.

[7] Z. Dvořák and R. Škrekovski, *A theorem about a contractible and light edge*, SIAM J. Discrete Math., 20 (2006), pp. 55–61.

[8] P. Erdős, A. L. Rubin, and H. Taylor, *Choosability in graphs*, Congr. Numer., 26 (1980), pp. 125–157.

[9] H. Gabow and S. Skulrattanakulchai, *Coloring algorithms on subcubic graphs*, Internat. J. Found. Comput. Sci., 15 (2004), pp. 21–40.

[10] S. Jendrol' and H.-J. Voss, *Light subgraphs of graphs embedded in the plane and in the projective plane—a survey*, Discrete Math., to appear.

[11] S. Jendrol' and H.-J. Voss, *Light subgraphs of graphs embedded in* 2*-dimensional manifolds of Euler characteristic* ≤ 0*—a survey*, in Paul Erdős and His Mathematics II, Bolyai Soc. Math. Stud. 11, G. Halász, L. Lovász, M. Simonovits, and V. T. Sós, eds., János Bolyai Math. Soc., Budapest, 2002, pp. 375–411.

[12] A. Kotzig, *Contribution to the theory of Eulerian polyhedra*, Mat.-Fyz. Časopis. Slovensk. Akad. Vied, 5 (1955), pp. 101–113.

[13] B. Mohar, R. Škrekovski, and H.-J. Voss, *Light subgraphs in planar graphs of minimum degree* 4 *and edge-degree* 9, J. Graph Theory, 44 (2003), pp. 261–295.

[14] V. G. Vizing, *Vertex colorings with given colors*, Metody Diskret. Analiz, 29 (1976), pp. 3–10 (in Russian).

[15] J. Zaks, *Extending Kotzig's theorem*, Israel J. Math., 45 (1983), pp. 281–296.

SIAM J. DISCRETE MATH.
Vol. 21, No. 1, pp. 107–129

OPERATIONS ON M-CONVEX FUNCTIONS ON JUMP SYSTEMS*

YUSUKE KOBAYASHI[†], KAZUO MUROTA[†], AND KEN'ICHIRO TANAKA[†]

Abstract. A jump system is a set of integer points with an exchange property, which is a generalization of a matroid, a delta-matroid, and a base polyhedron of an integral polymatroid (or a submodular system). Recently, the concept of M-convex functions on constant-parity jump systems was introduced by Murota as a class of discrete convex functions that admit a local criterion for global minimality. M-convex functions on constant-parity jump systems generalize valuated matroids, valuated delta-matroids, and M-convex functions on base polyhedra. This paper reveals that the class of M-convex functions on constant-parity jump systems is closed under a number of natural operations such as splitting, aggregation, convolution, composition, and transformation by networks. The present results generalize hitherto-known similar constructions for matroids, delta-matroids, valuated matroids, valuated delta-matroids, and M-convex functions on base polyhedra.

Key words. jump system, M-convex function, aggregation, convolution, transformation by networks

AMS subject classifications. 90C10, 90C25, 90C35, 90C27

DOI. 10.1137/060652841

1. Introduction.

A jump system [6] is a set of integer points with an exchange property (to be described later); see also [16], [18]. It is a generalization of a matroid [8], a delta-matroid [4], [7], [9], and a base polyhedron of an integral polymatroid (or a submodular system) [14].

Study of nonseparable nonlinear functions on matroidal structures was started with valuated matroids [10], [12], which have come to be accepted as discrete concave functions; see [20], [22]. This concept has been generalized to M-convex functions on base polyhedra [21], which play a central role in discrete convex analysis [23]. Valuated delta-matroids [11] afford another generalization of valuated matroids. As a common generalization of valuated delta-matroids and M-convex functions on base polyhedra, the concept of M-convex functions on constant-parity jump systems was introduced in [25]. To distinguish between M-convex functions on base polyhedra and those on constant-parity jump systems, we sometimes refer to the former as M^{B}-convex functions and the latter as M^{J}-convex functions. A separable convex function in the degree sequences of a graph is a typical example of M^{J}-convex functions. In all these generalizations global optimality is equivalent to local optimality defined in an appropriate manner. In addition, discrete duality theorems such as discrete separation and min-max formula hold for valuated matroids and M^{B}-convex functions, whereas they fail for valuated delta-matroids and M^{J}-convex functions.

A number of operations can be defined on matroidal structures and functions.

For example, union (or sum) can be defined for two matroids to yield another matroid. When translated in terms of incidence vectors, union can be understood

[*] Received by the editors February 24, 2006; accepted for publication (in revised form) August 25, 2006; published electronically February 9, 2007. This work was supported by the 21st Century COE Program on Information Science and Technology Strategic Core, and by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

http://www.siam.org/journals/sidma/21-1/65284/html

[†] Graduate School of Information Science and Technology, University of Tokyo, Tokyo 113-8656, Japan (Yusuke_Kobayashi@mist.i.u-tokyo.ac.jp, murota@mist.i.u-tokyo.ac.jp, kenitiro@misojiro.t.u-tokyo.ac.jp). The third author's research was supported by the Research Fellowship of the Japan Society for the Promotion of Science for Young Scientists.

Table 1.1
Sum of discrete structures.

Matroids	Rado (1942) [27] (see [28]) (explicitly by Edmonds (1968) [13])
Base polyhedra	McDiarmid (1975) [19]
Delta-matroids	Bouchet (1989) [5]
(Constant-parity) jump systems	Bouchet and Cunningham (1995) [6]

Table 1.2
Convolution of discrete functions.

Valuated matroids	Murota (1996) [21] (see also [22])
M^{B}-convex functions	Murota (1996) [21]
Valuated delta-matroids	
M^{J}-convex functions	This paper

as the Minkowski sum, followed by truncation by the vector $\mathbf{1} = (1, 1, \ldots, 1)$. Sum can also be defined for delta-matroids, base polyhedra, and (constant-parity) jump systems (see Table 1.1).

Convolution (or infimum convolution) of functions is a quantitative extension of sum, and the first result of the present paper (Theorem 12) is that M^{J}-convex functions are closed under convolution. This generalizes the known fact that valuated matroids and M^{B}-convex functions are closed under convolution (see Table 1.2).

Aggregation is another fundamental operation. For instance, it is known that any polymatroid can be obtained as an aggregation of a matroid [14] and that any jump system can be obtained as an aggregation of a delta-matroid [16]. The second result of the present paper (Theorem 11) is that M^{J}-convex functions are closed under aggregation. It is mentioned that the first result on convolution can be derived from this. A kind of converse of aggregation operation is splitting, which divides variables into several copies and generates a new function on a higher dimensional space. We show that splitting of M^{J}-convex functions is again M^{J}-convex.

Transformation (or induction) by graphs or networks is one of the most general operations. The fundamental fact in this direction is that a matroid can be transformed to another matroid through matchings in a bipartite graph. This construction also works for delta-matroids [4]. As for functions, valuated matroids are closed under transformation by bipartite graphs defined in an appropriate manner [21], [22], and M^{B}-convex functions are closed under transformation by networks [21]. The third result of the present paper (Theorem 14) is that this construction extends to M^{J}-convex functions; that is, transformation of M^{J}-convex functions by networks, to be defined precisely in section 6, preserves M^{J}-convexity. Aggregation, convolution, and splitting may be obtained as special cases of this construction, whereas our proof for the network transformation is based on the combination of aggregation, splitting, and other basic operations.

Here is a remark on the proof technique of the present paper. Our proofs consist of repeated applications of the defining exchange axiom of M^{J}-convex functions. This is particularly true of the proof given in section 7. For M^{B}-convex functions, on the other hand, an alternative “geometric” or “polyhedral” approach is possible on the basis of the convex extension of the functions. To be specific, such “polyhedral” proofs are known for convolution and network transformation of M^{B}-convex functions (see [14], [21], [24]). M^{J}-convex functions, however, seem to deny such a “polyhedral” approach, because jump systems can have “holes” within the convex hull and, accordingly, jump systems are not determined by their convex hulls. It is also noted that M^{J}-convex functions are not necessarily extensible to ordinary convex functions, although they possess a number of nice properties that justify the name of "convex functions."

2. Definitions and exchange axioms. Let V be a finite set. For $x = (x(v)), y = (y(v)) \in \mathbf{Z}^V$ define

$$x(V) = \sum_{v \in V} x(v),$$
$$||x||_1 = \sum_{v \in V} |x(v)|,$$
$$[x, y] = \{z \in \mathbf{Z}^V \mid \min(x(v), y(v)) \leq z(v) \leq \max(x(v), y(v)) \ \forall v \in V\}.$$

We denote by $\mathbf{0}$ the zero vector of an appropriate dimension. For $u \in V$ we denote by χ_u the *characteristic vector* of u, with $\chi_u(u) = 1$ and $\chi_u(v) = 0$ for $v \neq u$. A vector $s \in \mathbf{Z}^V$ is called an (x, y)-*increment* if $s = \chi_u$ or $s = -\chi_u$ for some $u \in V$ and $x + s \in [x, y]$. An (x, y)-*increment pair* will mean a pair of vectors (s, t) such that s is an (x, y)-increment and t is an $(x + s, y)$-increment.

A nonempty set $J \subseteq \mathbf{Z}^V$ is said to be a *jump system* if it satisfies an exchange axiom, called the 2-*step axiom*: for any $x, y \in J$ and for any (x, y)-increment s with $x+s \notin J$, there exists an $(x+s, y)$-increment t such that $x+s+t \in J$. A set $J \subseteq \mathbf{Z}^V$ is a *constant-sum system* if $x(V) = y(V)$ for any $x, y \in J$, and a *constant-parity system* if $x(V) - y(V)$ is even for any $x, y \in J$.

For constant-parity jump systems, Geelen [15] introduced a stronger exchange axiom:

(J-EXC) For any $x, y \in J$ and for any (x, y)-increment s, there exists an $(x + s, y)$-increment t such that $x + s + t \in J$ and $y - s - t \in J$.

This property characterizes a constant-parity jump system, a fact communicated to one of the authors by Geelen (see [25] for a proof).

THEOREM 1 (Geelen [15]). *A nonempty set J is a constant-parity jump system if and only if it satisfies* (J-EXC).

Next we turn to functions defined on integer points J. We call $f : J \to \mathbf{R}$ an *M^J-convex function* if it satisfies the following exchange axiom:

(M^{J}-EXC) For any $x, y \in J$ and for any (x, y)-increment s, there exists an $(x + s, y)$-increment t such that $x + s + t \in J$, $y - s - t \in J$, and

$$f(x) + f(y) \geq f(x + s + t) + f(y - s - t).$$

It follows from (M^{J}-EXC) that J satisfies (J-EXC) and hence is a constant-parity jump system.

We adopt the convention that $f(x) = +\infty$ for $x \notin J$. For a function $f : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ we denote the *effective domain* of f by

$$\mathrm{dom} f = \{x \in \mathbf{Z}^V \mid f(x) < +\infty\}.$$

Then it can be seen that if $f : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ satisfies (M^{J}-EXC), then its effective domain J satisfies (J-EXC).

It is known that if J satisfies (J-EXC), the exchange axiom (M^{J}-EXC) is equivalent to a local exchange axiom:

(M^{J}-$\mathrm{EXC}_{\mathrm{loc}}$) For any $x, y \in J$ with $||x - y||_1 = 4$ there exists an (x, y)-increment pair (s, t) such that $x + s + t \in J$, $y - s - t \in J$, and

$$f(x) + f(y) \geq f(x + s + t) + f(y - s - t).$$

THEOREM 2 (see [25]). *A function $f : J \to \mathbf{R}$ defined on a constant-parity jump system J satisfies* (M^{J}-EXC) *if and only if it satisfies* (M^{J}-$\mathrm{EXC}_{\mathrm{loc}}$).

In what follows, we refer to M^{J}-convexity simply as M-convexity; in particular, when we talk about an M-convex function it is presumed that its effective domain is a constant-parity jump system.

The definition of an M-convex function is consistent with the previously considered special cases where (i) J is a constant-sum jump system, and (ii) J is a constant-parity jump system contained in $\{0, 1\}^V$. Case (i) is equivalent to J being the set of integer points in the base polyhedron of an integral submodular system [14], and then the M-convex function is the same as the M^{B}-convex function investigated in [21], [23]. Case (ii) is equivalent to J being an even delta-matroid [30], [31], and then f is M-convex if and only if $-f$ is a valuated delta-matroid in the sense of [11].

For an M-convex function, it is known that global optimality (minimality) is guaranteed by local optimality in the neighborhood of ℓ_1-distance two, which generalizes the optimality criterion in [1] for separable convex function minimization over a jump system. The efficient algorithm for the minimization problem of M-convex functions follows from the optimality criterion [25], [26].

THEOREM 3 (see [25]). *Let $f : J \to \mathbf{R}$ be an M-convex function on a constant-parity jump system J, and let $x \in J$. Then $f(x) \leq f(y)$ for all $y \in J$ if and only if $f(x) \leq f(y)$ for all $y \in J$ with $||x - y||_1 \leq 2$.*

It is also known that global optimality (minimality) for constrained minimization on a hyperplane of a constant component sum is guaranteed by local optimality in the neighborhood of ℓ_1-distance four.

THEOREM 4 (see [25]). *Let $f : J \to \mathbf{R}$ be an M-convex function on a constant-parity jump system $J \subseteq \mathbf{Z}^V$, let $J_k = \{x \in J \mid x(V) = k\}$, and let $x \in J_k$. Then $f(x) \leq f(y)$ for all $y \in J_k$ if and only if $f(x) \leq f(y)$ for all $y \in J_k$ with $||x-y||_1 \leq 4$.*

This optimality criterion for M-convex functions helps us deepen our understanding of the result of Apollonio and Sebő [2], [3]. They provided a polynomial algorithm for the minconvex factor problem, which is, given an undirected graph possibly containing loops and parallel edges and a separable convex function on the degree sequences, to find a subgraph with a specified number of edges that minimizes the function. The key observation in [2], [3] is that global optimality is guaranteed by local optimality in the neighborhood of ℓ_1-distance at most four in the space of degree sequences. Since a separable convex function on the degree sequences of a graph is an M-convex function, this result can be seen as a special case of Theorem 4.

3. Basic operations.

Let $f : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ be an M-convex function. We introduce some basic operations on f that preserve M-convexity. Though too simple to be interesting in their own right, these operations are stated explicitly in view of their use in our proofs.

For a subset $U \subseteq V$ and a superset $W \supseteq V$, we define the *coordinate inversion* $f_U^- : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ of U, the *restriction* $f_U : \mathbf{Z}^U \to \mathbf{R} \cup \{+\infty\}$ to U, and the 0-*augmentation* $f^W : \mathbf{Z}^W \to \mathbf{R} \cup \{+\infty\}$ to W by

$$\begin{aligned} f_U^-(y, z) &= f(-y, z) \quad (y \in \mathbf{Z}^U, z \in \mathbf{Z}^{V\setminus U}), \\ f_U(y) &= f(y, \mathbf{0}) \quad (y \in \mathbf{Z}^U, \mathbf{0} \in \mathbf{Z}^{V\setminus U}), \\ f^W(y, z) &= \begin{cases} f(y) & \text{if } z = \mathbf{0} \\ +\infty & \text{otherwise} \end{cases} \quad (y \in \mathbf{Z}^V, z \in \mathbf{Z}^{W\setminus V}), \end{aligned}$$

respectively. For a linear function $p : \mathbf{Z}^V \to \mathbf{R}$, we define $f[p] : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ by

$$f[p](x) = f(x) + p(x).$$

It is obvious that they are M-convex.

We say that $\varphi : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ is a *separable convex function* if it is represented as

$$\varphi(x) = \sum_{u \in V} \varphi_u(x(u)),$$

where for each $u \in V$, $\varphi_u : \mathbf{Z} \to \mathbf{R} \cup \{+\infty\}$ is a convex function; that is, for any integers $\xi < \eta$

$$\varphi_u(\xi) + \varphi_u(\eta) \geq \varphi_u(\xi + 1) + \varphi_u(\eta - 1).$$

Note that this condition is equivalent to the following: for any integer ξ

$$\varphi_u(\xi - 1) + \varphi_u(\xi + 1) \geq 2\varphi_u(\xi).$$

For a separable convex function φ, we define $f + \varphi : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ by

$$(f + \varphi)(x) = f(x) + \varphi(x).$$

Theorem 5. *If f is M-convex and φ is a separable convex function, then $f + \varphi$ is M-convex.*

Proof. It suffices to show that for a one-dimensional convex function φ_u with a particular $u \in V$ the function $g(x) = f(x) + \varphi_u(x(u))$ is M-convex. Suppose that $x = (x(v)) \in \mathbf{Z}^V$, $y = (y(v)) \in \mathbf{Z}^V$, and s is an (x, y)-increment. By M-convexity of f, there exists an $(x + s, y)$-increment t such that

$$f(x) + f(y) \geq f(x + s + t) + f(y - s - t),$$

and it holds that

$$\varphi_u(x(u)) + \varphi_u(y(u)) \geq \varphi_u(x(u) + s(u) + t(u)) + \varphi_u(y(u) - s(u) - t(u))$$

by convexity of φ_u. Thus we have

$$g(x) + g(y) \geq g(x + s + t) + g(y - s - t),$$

which completes the proof. □

4. Splitting.

Splitting is an operation which generates a new function by dividing some variables. The objective of this section is to show that if a given function is M-convex, then the function obtained by splitting is also M-convex (Theorem 7). Although splitting is a simple operation, it plays an important role when we deal with transformation by networks in section 6.

First we introduce an elementary operation, called elementary splitting, which divides one variable into two variables. Elementary splitting preserves M-convexity, from which we can show that splitting preserves M-convexity.

For a function $f : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$, the *elementary splitting* of f at $v \in V$ is a function $f' : \mathbf{Z}^{V'} \to \mathbf{R} \cup \{+\infty\}$ defined by

$$f'(x_0; x(v'), x(v'')) = f(x_0; x(v') + x(v'')),$$

where $V' = (V \setminus \{v\}) \cup \{v', v''\}$ and $x_0 \in \mathbf{Z}^{V\setminus\{v\}}$.

LEMMA 6. *If f is M-convex, then its elementary splitting f' is M-convex.*

Proof. For a concise description, let $V = \{1, 2, \ldots, n\}$ and $V' = \{1, 2, \ldots, n-1, a, b\}$. We show that if f is M-convex, then its elementary splitting f' at n defined by

$$f'(x_0; x_a, x_b) = f(x_0; x_a + x_b)$$

is M-convex. For $u \in V'$ we denote by χ'_u the characteristic vector of u in V'. It suffices to show that f' satisfies (M$^{\mathrm{J}}$-EXC); that is, for any two vectors $x' = (x_0; x_a, x_b) \in \mathrm{dom} f'$, $y' = (y_0; y_a, y_b) \in \mathrm{dom} f'$, and for any (x', y')-increment s', there exists an $(x' + s', y')$-increment t' such that

$$f'(x') + f'(y') \geq f'(x' + s' + t') + f'(y' - s' - t').$$

We put $\xi = x_a + x_b$ and $\eta = y_a + y_b$. We also put $x = (x_0; \xi)$ and $y = (y_0; \eta)$.

Case 1. Suppose that $s' = \pm\chi'_k$ is an (x', y')-increment, where $1 \leq k \leq n-1$. We denote $\pm\chi_k$ by s. Since f is M-convex and s is an (x, y)-increment, there exists an $(x + s, y)$-increment t such that

$$f(x) + f(y) \geq f(x + s + t) + f(y - s - t).$$

If $t = \pm\chi_l$ with $1 \leq l \leq n-1$, then $t' = \pm\chi'_l$ is an $(x' + s', y')$-increment and

$$f'(x') + f'(y') \geq f'(x' + s' + t') + f'(y' - s' - t').$$

Otherwise we have $l = n$. Without loss of generality, we may assume that $\xi < \eta$ and $t = \chi_n$. Since $\xi < \eta$ implies that at least one of $x_a < y_a$ and $x_b < y_b$ holds, at least one of χ'_a and χ'_b, say t', is an $(x' + s', y')$-increment and it holds that

$$f'(x') + f'(y') = f(x) + f(y) \geq f(x+s+t) + f(y-s-t) = f'(x'+s'+t') + f'(y'-s'-t').$$

Case 2. Suppose that $s' = \pm\chi'_a$ or $\pm\chi'_b$ is an (x', y')-increment. In this case, without loss of generality, we may assume that $s' = \chi'_b$ and $x_b < y_b$.

If $x_a > y_a$, then $t' = -\chi'_a$ is an $(x' + s', y')$-increment and

$$f'(x') + f'(y') = f(x) + f(y) = f'(x' + s' + t') + f'(y' - s' - t').$$

Suppose that $x_a \leq y_a$. Then we have $\xi < \eta$ and χ_n is an (x, y)-increment. Since f is M-convex, by applying (M$^{\mathrm{J}}$-EXC) with $s = \chi_n$, there exists an $(x + s, y)$-increment t such that

$$f(x) + f(y) \geq f(x + s + t) + f(y - s - t).$$

If $t = \pm\chi_k$ with $1 \leq k \leq n-1$, then $t' = \pm\chi'_k$ is an $(x' + s', y')$-increment and

$$f'(x') + f'(y') = f(x) + f(y) \geq f(x+s+t) + f(y-s-t) = f'(x'+s'+t') + f'(y'-s'-t').$$

Otherwise, we have $t = \chi_n$ and $\xi + 2 \leq \eta$. Thus at least one of $x_b + 2 \leq y_b$ and $x_a + 1 \leq y_a$ holds, and hence at least one of χ'_b and χ'_a, say t', is an $(x' + s', y')$-increment. We then have

$$f'(x') + f'(y') = f(x) + f(y) \geq f(x+s+t) + f(y-s-t) = f'(x'+s'+t') + f'(y'-s'-t').$$

This shows the existence of t' in Case 2. □

Suppose that we are given a finite set $V = \{v_1, v_2 \ldots, v_n\}$ and a family of nonempty disjoint sets $\{U_v \mid v \in V\}$ indexed by $v \in V$. Let $U = \bigcup_{v \in V} U_v$. For a function $f : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$, we define the *splitting* of f to U as a function $f' : \mathbf{Z}^U \to \mathbf{R} \cup \{+\infty\}$ given by

$$f'(\tilde{x}_{v_1}, \tilde{x}_{v_2}, \ldots, \tilde{x}_{v_n}) = f(\xi_{v_1}, \xi_{v_2}, \ldots, \xi_{v_n}),$$

where $\tilde{x}_v \in \mathbf{Z}^{U_v}$ and $\xi_v = \tilde{x}_v(U_v)$ for $v \in V$. We now have the following theorem.

THEOREM 7. *If f is M-convex, then its splitting f' is M-convex.*

Proof. We can obtain splitting f' by applying elementary splittings $\sum_{v \in V}(|U_v| - 1)$ times. Hence, by Lemma 6, f' is M-convex. □

Theorem 7 implies that if dom$f$ is a constant-parity jump system, then domf' is also a constant-parity jump system.

5. Aggregation and convolution.

Minkowski sum is a fundamental operation on matroid structures, and jump systems are closed under Minkowski sum. In this section, we deal with an operation for functions called convolution, which is a quantitative extension of sum, and also a related operation called aggregation. The objective of this section is to show that M-convexity is preserved under these operations. As with splitting, aggregation plays an important role when we deal with transformations by networks in section 6.

For two jump systems $J_1 \subseteq \mathbf{Z}^V$ and $J_2 \subseteq \mathbf{Z}^V$, their *sum* $J_1 + J_2 \subseteq \mathbf{Z}^V$ is defined by

$$J_1 + J_2 = \{x_1 + x_2 \mid x_1 \in J_1, x_2 \in J_2\},$$

which is known to be a jump system.

THEOREM 8 (see [6]). *The sum of two jump systems is a jump system.*

While this theorem is shown directly in [6], Kabadi and Sridhar [16] gave an alternative proof by showing that jump systems are closed under a related elementary operation. They showed that if $J \subseteq \mathbf{Z}^V$ is a jump system, then its *elementary aggregation* $\tilde{J} \subseteq \mathbf{Z}^{\tilde{V}}$ at $v_1 \in V$ and $v_2 \in V$ defined by

$$\tilde{J} = \{(x_0, x(v_1) + x(v_2)) \mid (x_0, x(v_1), x(v_2)) \in J\}$$

is also a jump system, where $\tilde{V} = (V \setminus \{v_1, v_2\}) \cup \{v\}$ and $x_0 \in \mathbf{Z}^{V \setminus \{v_1, v_2\}}$. Theorem 8 can be derived from the following fact.

LEMMA 9 (see [16]). *An elementary aggregation of a jump system is a jump system.*

Convolution is a quantitative extension of sum. For two functions $f_1 : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ and $f_2 : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$, we define their (infimum) *convolution* as a function $f_1 \Box f_2 : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty, -\infty\}$ given by

$$(f_1 \Box f_2)(x) = \inf\{f_1(x_1) + f_2(x_2) \mid x_1 + x_2 = x,\ x_1 \in \mathbf{Z}^V,\ x_2 \in \mathbf{Z}^V\}.$$

To show that convolution preserves M-convexity (Theorem 12), we introduce a quantitative extension of elementary aggregation.

For a function $f : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$, the *elementary aggregation* of f at $v_1 \in V$ and $v_2 \in V$ is a function $\tilde{f} : \mathbf{Z}^{\tilde{V}} \to \mathbf{R} \cup \{+\infty, -\infty\}$ defined by

$$\tilde{f}(x_0; \xi) = \inf\{f(x_0; x(v_1), x(v_2)) \mid \xi = x(v_1) + x(v_2)\},$$

where $\tilde{V} = (V \setminus \{v_1, v_2\}) \cup \{v\}$ and $x_0 \in \mathbf{Z}^{V \setminus \{v_1, v_2\}}$. Then we can show that if f is M-convex, then $\tilde{f}$ is M-convex; the proof is given in section 7.

LEMMA 10. *If f is M-convex, then its elementary aggregation $\tilde{f}$ is M-convex, provided that $\tilde{f} > -\infty$.*

A general aggregation is defined as the result of repeated applications of elementary aggregations. More formally, let V be a finite set and π be its partition $V = V_1 \cup V_2 \cup \cdots \cup V_n$ into disjoint subsets. For a function $f : \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$, we define the *aggregation* of f with respect to π as a function $\tilde{f} : \mathbf{Z}^n \to \mathbf{R} \cup \{+\infty, -\infty\}$ given by

$$\tilde{f}(\xi_1, \xi_2, \dots, \xi_n) = \inf \left\{ f(x_1, x_2, \dots, x_n) \mid x_i \in \mathbf{Z}^{V_i},\ x_i(V_i) = \xi_i \right\}.$$

Then we have the following theorem.

THEOREM 11. *If f is M-convex, then its aggregation $\tilde{f}$ is M-convex, provided that $\tilde{f} > -\infty$.*

Proof. By applying elementary aggregations $|V| - n$ times, we can obtain $\tilde{f}$, which is M-convex by Lemma 10. □

We are now ready to show that convolution preserves M-convexity.

THEOREM 12. *If f_1 and f_2 are M-convex functions, then their convolution $f_1 \Box f_2$ is M-convex, provided that $f_1 \Box f_2 > -\infty$.*

Proof. First we make the direct sum $f : \mathbf{Z}^V \times \mathbf{Z}^V \to \mathbf{R} \cup \{+\infty\}$ of f_1 and f_2 defined by

$$f(x_1, x_2) = f_1(x_1) + f_2(x_2),$$

where $x_1, x_2 \in \mathbf{Z}^V$. Then f is M-convex because f_1 and f_2 are M-convex. Let π be the partition consisting of pairs of the corresponding elements. Then the aggregation of f coincides with $f_1 \Box f_2$. Hence, by Theorem 11, $f_1 \Box f_2$ is M-convex. □

Finally, we consider another operation, called composition. Let $f_1 : \mathbf{Z}^{S_1} \to \mathbf{R} \cup \{+\infty\}$ and $f_2 : \mathbf{Z}^{S_2} \to \mathbf{R} \cup \{+\infty\}$ be M-convex functions. Put $V_0 = S_1 \cap S_2$, $V_1 = S_1 \setminus V_0$, and $V_2 = S_2 \setminus V_0$. We define the *composition* of f_1 and f_2 to be a function $f : \mathbf{Z}^{V_1 \cup V_2} \to \mathbf{R} \cup \{+\infty, -\infty\}$ given by

$$f(x_1, x_2) = \inf\{f_1(x_1, y_1) + f_2(x_2, y_2) \mid y_1 = y_2 \in \mathbf{Z}^{V_0}\} \quad (x_1 \in \mathbf{Z}^{V_1}, x_2 \in \mathbf{Z}^{V_2}).$$

THEOREM 13. *The composition of two M-convex functions is M-convex, provided that it does not take the value $-\infty$.*

Proof. Consider M-convex functions $\tilde{f}_1$ and $\tilde{f}_2$ defined by

$$\begin{aligned} \tilde{f}_1(x_1, y_1, \mathbf{0}) &= f_1(x_1, y_1) \quad (x_1 \in \mathbf{Z}^{V_1}, y_1 \in \mathbf{Z}^{V_0}, \mathbf{0} \in \mathbf{Z}^{V_2}), \\ \tilde{f}_2(\mathbf{0}, (-y_2), x_2) &= f_2(x_2, y_2) \quad (\mathbf{0} \in \mathbf{Z}^{V_1}, (-y_2) \in \mathbf{Z}^{V_0}, x_2 \in \mathbf{Z}^{V_2}). \end{aligned}$$

Their convolution $\tilde{f}_1 \Box \tilde{f}_2$ is M-convex by Theorem 12, and the restriction of $\tilde{f}_1 \Box \tilde{f}_2$ to $V_1 \cup V_2$ coincides with the composition. □

Note that the composition of M-convex functions is a generalization of the *composition* of (constant-parity) jump systems. It is known that the composition of two jump systems is a jump system [6], and Theorem 13 generalizes this fact.

6. Transformation by networks. In this section, we consider the transformation of an M-convex function through a network. We show that it preserves M-convexity on the basis of splitting, aggregation, and other basic operations discussed above.

Let $G = (V, A; S, T)$ be a directed graph with vertex set V, arc set A, entrance set S, and exit set T, where S and T are disjoint subsets of V. For each $a \in A$, the cost of integer-flow in a is represented by a function $\varphi_a : \mathbf{Z} \to \mathbf{R} \cup \{+\infty\}$, which is assumed to be convex.

Given a function $f : \mathbf{Z}^S \to \mathbf{R} \cup \{+\infty\}$ associated with the entrance set S of the network, we define a function $\tilde{f} : \mathbf{Z}^T \to \mathbf{R} \cup \{+\infty, -\infty\}$ on the exit set T by

$$\tilde{f}(y) = \inf_{\xi, x} \left\{ f(x) + \sum_{a \in A} \varphi_a(\xi(a)) \mid \partial\xi = (x, -y, \mathbf{0}), \right.$$

$$\left. \xi \in \mathbf{Z}^A, (x, -y, \mathbf{0}) \in \mathbf{Z}^S \times \mathbf{Z}^T \times \mathbf{Z}^{V \setminus (S \cup T)} \right\} \quad (y \in \mathbf{Z}^T),$$

where $\partial\xi \in \mathbf{Z}^V$ is the vector given by

$$\partial\xi(v) = \sum_{a:\ a \text{ leaves } v} \xi(a) - \sum_{a:\ a \text{ enters } v} \xi(a) \quad (v \in V).$$

If such (ξ, x) does not exist, we define $\tilde{f}(y) = +\infty$. We may think of $\tilde{f}(y)$ as the minimum cost to meet a demand specification y at the exit, where the cost consists of two parts, the cost $f(x)$ of supply or production of x at the entrance and the cost $\sum_{a \in A} \varphi_a(\xi(a))$ of transportation through arcs; the sum of these is to be minimized over varying supply x and flow ξ subject to the flow conservation constraint $\partial\xi = (x, -y, \mathbf{0})$. We regard $\tilde{f}$ as a result of *transformation* (or *induction*) of f by the network.

THEOREM 14. *Assume that f is M-convex and φ_a is convex for each $a \in A$. Then the function $\tilde{f}$ induced by a network $G = (V, A; S, T)$ is M-convex, provided that $\tilde{f} > -\infty$.*

To prove this theorem, we first show that transformations by some simple bipartite networks preserve M-convexity. When $V = S \cup T$, we denote the graph G simply by $G = (S, T; A)$. It is noted that some arcs are directed from S to T and the others from T to S.

LEMMA 15. *Let $G = (S, T; A)$ be a bipartite network, where each vertex in T has exactly one incident arc (see Figure 1). If f is M-convex and $\varphi_a = 0$ for each $a \in A$, the function $\tilde{f}$ induced by G is M-convex.*

Proof. We can obtain $\tilde{f}$ from f by restriction and splitting. Hence, if f is M-convex, then $\tilde{f}$ is M-convex by Theorem 7. □

LEMMA 16. *Let $G = (S, T; A)$ be a bipartite network, where each vertex in S has exactly one incident arc (see Figure 2). If f is M-convex and $\varphi_a = 0$ for each $a \in A$, the function $\tilde{f}$ induced by G is M-convex, provided that $\tilde{f} > -\infty$.*

Proof. We can obtain $\tilde{f}$ from f by aggregation and 0-augmentation. Hence, if f is M-convex, then $\tilde{f}$ is M-convex by Theorem 11. □

LEMMA 17. *Let $G = (S, T; A)$ be a bipartite network, as in Figure 3, where $S = \{s_1, \ldots, s_n\}$, $T = \{t_1, \ldots, t_n\}$, and $A = \{a_1, \ldots, a_n\}$ with $a_i = (s_i, t_i)$ or $a_i = (t_i, s_i)$ for $i = 1, \ldots, n$. If f is M-convex and φ_a is convex for each $a \in A$, the function $\tilde{f}$ induced by G is M-convex.*

Proof. We may assume that

$$S^+ = \{s_1, \ldots, s_m\}, \qquad S^- = \{s_{m+1}, \ldots, s_n\},$$
$$T^+ = \{t_1, \ldots, t_m\}, \qquad T^- = \{t_{m+1}, \ldots, t_n\},$$

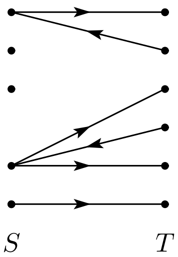


FIG. 1. *Splitting.*

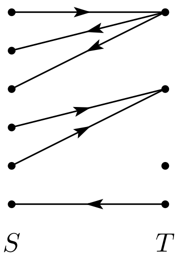


FIG. 2. *Aggregation.*

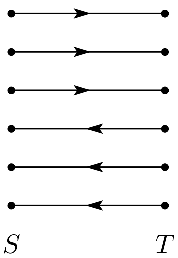


FIG. 3. *Addition.*

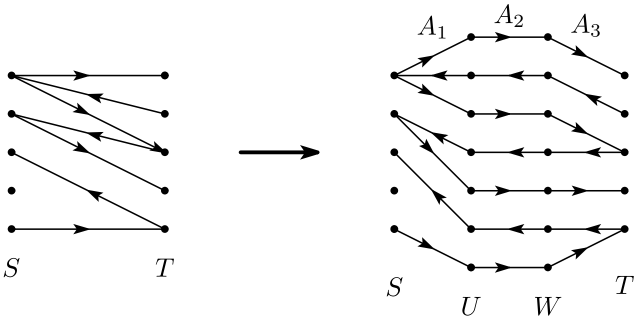


FIG. 4. *Transformation by a bipartite network.*

$$A^+ = \{(s_i, t_i) \mid i = 1, \ldots, m\}, \qquad A^- = \{(t_i, s_i) \mid i = m+1, \ldots, n\},$$

and $A = A^+ \cup A^-$. Then, for $x = (x_i) \in \mathbf{Z}^n$, $\tilde{f}$ is expressed as

$$\tilde{f}(x) = f(x) + \sum_{i=1}^{m} \varphi_a(x_i) + \sum_{i=m+1}^{n} \varphi_a(-x_i),$$

and if $\varphi_a(x)$ is convex, then $\varphi_a(-x)$ is convex for $a \in A^-$. Thus we can obtain $\tilde{f}$ by adding a separable convex function to f. Hence, if f is M-convex, then $\tilde{f}$ is M-convex by Theorem 5. □

Using the above lemmas, we see that transformation by bipartite networks preserves M-convexity.

THEOREM 18. *Assume that f is M-convex, φ_a is convex for each $a \in A$, and $G = (S, T; A)$ is a bipartite network. Then the function $\tilde{f}$ induced by G is M-convex, provided that $\tilde{f} > -\infty$.*

Proof. We construct a new network that represents the same transformation as the original network. The new network is obtained by subdividing each arc of G into three arcs, as illustrated in Figure 4. For each arc $a \in A$ we consider two new vertices u_a and w_a; if a is directed from S to T, i.e., $a = (s, t)$ with $s \in S$ and $t \in T$, we will have three arcs $a_1 = (s, u_a)$, $a_2 = (u_a, w_a)$, and $a_3 = (w_a, t)$, and if $a = (t, s)$ with $t \in T$ and $s \in S$, we will have $a_3 = (t, w_a)$, $a_2 = (w_a, u_a)$, and $a_1 = (u_a, s)$. The cost φ_a is associated with arc a_2, whereas the arcs a_1 and a_3 are given 0 as the cost. Thus the new network consists of three bipartite graphs connected in series, $G_1 = (S, U; A_1)$, $G_2 = (U, W; A_2)$, and $G_3 = (W, T; A_3)$, where $U = \{u_a \mid a \in A\}$, $W = \{w_a \mid a \in A\}$, and $A_i = \{a_i \mid a \in A\}$ $(i = 1, 2, 3)$.

The given M-convex function f on S is transformed through G_1 to a function $f_1 : \mathbf{Z}^U \to \mathbf{R} \cup \{+\infty\}$, which is M-convex by Lemma 15. Then f_1 is transformed through G_2 to a function $f_2 : \mathbf{Z}^W \to \mathbf{R} \cup \{+\infty\}$, which is M-convex by Lemma 17.

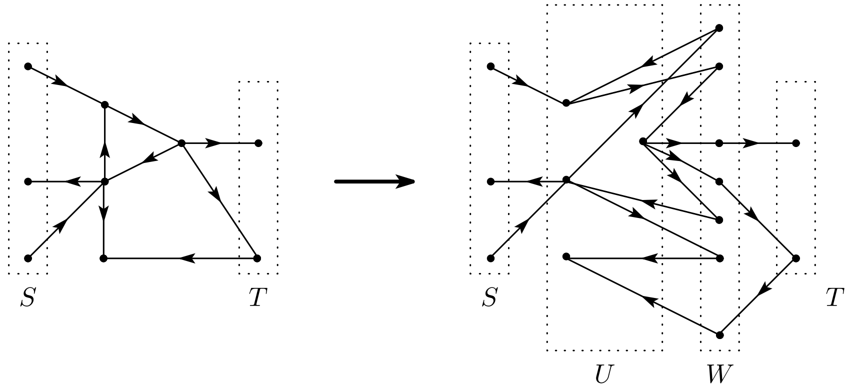


FIG. 5. *Transformation by a general network.*

Finally, f_2 is transformed through G_3 to a function $f_3 : \mathbf{Z}^T \to \mathbf{R} \cup \{+\infty, -\infty\}$, which is M-convex by Lemma 16. The resulting function f_3 coincides with the function $\tilde{f}$ induced from f by G. □

We are now ready to show Theorem 14.

Proof of Theorem 14. We construct a new network that represents the same transformation as the original network. The new network is obtained by subdividing each arc of G into some arcs, as illustrated in Figure 5. We may assume, by subdividing arcs, that no arcs exist between the two vertices in $S \cup T$. Let $U = V \setminus (S \cup T)$, let A_{UT} be the set of arcs connecting U and T, and define A_{SU} and A_{UU} similarly. For each arc $a \in A_{UT}$, we consider a new vertex w_a; if a is directed from U to T, i.e., $a = (u, t)$ with $u \in U$ and $t \in T$, we will have two arcs $a_1 = (u, w_a)$, $a_2 = (w_a, t)$, and if $a = (t, u)$ with $t \in T$ and $u \in U$, we will have $a_2 = (t, w_a)$, $a_1 = (w_a, u)$. For each arc $a = (u_1, u_2) \in A_{UU}$ with $u_1, u_2 \in U$, we consider a new vertex w_a and have two arcs $a_1 = (u_1, w_a)$, $a_2 = (w_a, u_2)$. Thus the new network consists of three bipartite graphs connected in series, $G_1 = (S, U; A_1)$, $G_2 = (U, W; A_2)$, and $G_3 = (W, T; A_3)$, where $W = \{w_a \mid a \in A_{UT} \cup A_{UU}\}$, $A_1 = A_{SU}$, $A_2 = \{a_1 \mid a \in A_{UT}\} \cup \{a_1 \mid a \in A_{UU}\} \cup \{a_2 \mid a \in A_{UU}\}$, and $A_3 = \{a_2 \mid a \in A_{UT}\}$.

By Theorem 18, transformations by the networks G_1, G_2, and G_3 preserve M-convexity. Since the transformation by G can be represented as a combination of the above three transformations, the function $\tilde{f}$ transformed from f by G is M-convex. □

As we mentioned in section 1, transformations by networks also preserve M^{B}-convexity. Two kinds of proofs for this fact are known (see [21], [22], [29]); one uses a dual variable, and the other is a complicated algorithmic proof. We can see that our proof of Theorem 14 also works for M^{B}-convex functions; that is, by proving that splitting, aggregation, and other basic operations preserve M^{B}-convexity, we can show that transformations by networks preserve M^{B}-convexity.

It is also noted that the transformation by networks can be generalized by replacing networks by linking systems, and that transformations by linking systems also preserve M-convexity [17].

7. Proof of Lemma 10 for elementary aggregation.

In this section, we give a proof of Lemma 10. For a concise description, we denote $V = \{1, 2, \dots, n-1, n\}$ and $\tilde{V} = \{1, 2, \dots, n-2, a\}$. We show that if f is M-convex, then $\tilde{f}$ defined by

$$\tilde{f}(x_0; \xi) = \inf\{f(x_0; x_{n-1}, x_n) \mid \xi = x_{n-1} + x_n\}$$

is M-convex. For $u \in \tilde{V}$, we denote by $\tilde{\chi}_u$ the characteristic vector of u in $\tilde{V}$.

We first deal with the case where the effective domain of f is bounded, whereas the general case is treated in section 7.4.

7.1. Case of bounded effective domain.

LEMMA 19. *If f is M-convex and* dom$f$ *is bounded, then its elementary aggregation $\tilde{f}$ is M-convex.*

Proof. Let J and $\tilde{J}$ be the effective domains of f and $\tilde{f}$, respectively. If f is M-convex, then J is a constant-parity jump system, which implies by Lemma 9 that $\tilde{J}$ is also a constant-parity jump system. Hence, by Theorem 2, it is enough to show that $\tilde{f}$ satisfies (M$^{\rm J}$-EXC$_{\rm loc}$); that is, for any $\tilde{x} = (x_0; \xi), \tilde{y} = (y_0; \eta) \in \tilde{J}$ with $||\tilde{x} - \tilde{y}||_1 = ||x_0 - y_0||_1 + |\xi - \eta| = 4$, there exists an $(\tilde{x}, \tilde{y})$-increment pair (s, t) such that

$$(*) \qquad \tilde{f}(\tilde{x}) + \tilde{f}(\tilde{y}) \geq \tilde{f}(\tilde{x} + s + t) + \tilde{f}(\tilde{y} - s - t).$$

Without loss of generality, we may assume that $\xi \geq \eta$. Take $x_{n-1}, x_n, y_{n-1}, y_n$ with the minimum value of $|x_{n-1} - y_{n-1}| + |x_n - y_n|$ such that

$$\tilde{f}(x_0; \xi) = f(x_0; x_{n-1}, x_n) \quad ((x_0; x_{n-1}, x_n) \in J,\ \xi = x_{n-1} + x_n),$$
$$\tilde{f}(y_0; \eta) = f(y_0; y_{n-1}, y_n) \quad ((y_0; y_{n-1}, y_n) \in J,\ \eta = y_{n-1} + y_n).$$

Note that such $x_{n-1}, x_n, y_{n-1}, y_n$ exist, because J is finite and $(x_0; \xi), (y_0; \eta) \in \tilde{J}$.

If $x_{n-1} = y_{n-1}$ or $x_n = y_n$, then it is obvious by (M$^{\rm J}$-EXC) for f that there exists an $(\tilde{x}, \tilde{y})$-increment pair (s, t) satisfying $(*)$. Since $x_{n-1} \geq y_{n-1}$ or $x_n \geq y_n$ holds by the assumption $\xi \geq \eta$, we may assume that $x_{n-1} > y_{n-1}$ and $x_n \neq y_n$.

Case 1. Suppose that $\xi \geq \eta + 2$. By (M$^{\rm J}$-EXC) for f with $s = -\chi_{n-1}$, we have

$$(1) \quad f(x_0; x_{n-1}, x_n) + f(y_0; y_{n-1}, y_n)$$
$$\geq \min \left\{ \begin{array}{c} f(x_0; x_{n-1} - 1, x_n \pm 1) + f(y_0; y_{n-1} + 1, y_n \mp 1), \\ f(x_0; x_{n-1} - 2, x_n) + f(y_0; y_{n-1} + 2, y_n), \\ f(x_0 + t_0; x_{n-1} - 1, x_n) + f(y_0 - t_0; y_{n-1} + 1, y_n) \end{array} \right\},$$

where $t_0 \in \mathbf{Z}^{n-2}$ is an (x_0, y_0)-increment. Note that the signs in (1) are determined by the relations of components, and the second term exists only if $x_{n-1} - y_{n-1} \geq 2$. If the second term or the third term achieves the minimum, then $(s, t) = (-\tilde{\chi}_a, -\tilde{\chi}_a)$ or $(-\tilde{\chi}_a, \tilde{t})$, where $\tilde{t} = (t_0, 0) \in \mathbf{Z}^{\tilde{V}}$, is an $(\tilde{x}, \tilde{y})$-increment pair satisfying $(*)$. Otherwise, we have

$$f(x_0; x_{n-1}, x_n) + f(y_0; y_{n-1}, y_n) \geq f(x_0; x_{n-1} - 1, x_n + 1) + f(y_0; y_{n-1} + 1, y_n - 1)$$

or

$$f(x_0; x_{n-1}, x_n) + f(y_0; y_{n-1}, y_n) \geq f(x_0; x_{n-1} - 1, x_n - 1) + f(y_0; y_{n-1} + 1, y_n + 1).$$

If $x_n > y_n$, then we have $f(x_0; x_{n-1}, x_n) + f(y_0; y_{n-1}, y_n) \geq f(x_0; x_{n-1} - 1, x_n - 1) + f(y_0; y_{n-1} + 1, y_n + 1)$, and so $\tilde{f}(x_0; \xi) + \tilde{f}(y_0; \eta) \geq \tilde{f}(x_0; \xi - 2) + \tilde{f}(y_0; \eta + 2)$. Thus $(s, t) = (-\tilde{\chi}_a, -\tilde{\chi}_a)$ is an $(\tilde{x}, \tilde{y})$-increment pair satisfying $(*)$.

If $x_n < y_n$, we have $f(x_0; x_{n-1}, x_n) + f(y_0; y_{n-1}, y_n) \geq f(x_0; x_{n-1} - 1, x_n + 1) + f(y_0; y_{n-1}+1, y_n-1)$. By the definition of $x_{n-1}, x_n, y_{n-1}, y_n$, we have $f(x_0; x_{n-1}, x_n) =$

$f(x_0; x_{n-1}-1, x_n+1)$ and $f(y_0; y_{n-1}, y_n) = f(y_0; y_{n-1}+1, y_n-1)$. This contradicts the minimality of $|x_{n-1}-y_{n-1}| + |x_n - y_n|$.

Case 2. Suppose that $\xi = \eta$. It suffices to show that if $\tilde{y} = \mathbf{0}$ and $\tilde{x} = (1,1,1,1;0)$, $(2,1,1;0)$, $(3,1;0)$, $(2,2;0)$, or $(4;0)$, there exists an $(\tilde{x},\tilde{y})$-increment pair (s,t) satisfying $(*)$. This is shown in section 7.2.

Case 3. Suppose that $\xi = \eta+1$. It suffices to show that if $\tilde{y} = \mathbf{0}$ and $\tilde{x} = (1,1,1;1)$, $(2,1;1)$, or $(3;1)$, there exists an $(\tilde{x},\tilde{y})$-increment pair (s,t) satisfying $(*)$. This is shown in section 7.3. □

7.2. Case 2 in the proof of Lemma 19. In this section, we deal with Case 2 in the proof of Lemma 19. First we show the essential case when $\tilde{x} = (1,1,1,1;0)$, whereas the other cases can be derived from this using the splitting technique discussed in section 4.

7.2.1. The main lemma. Let $f : \mathbf{Z}^6 \to \mathbf{R} \cup \{+\infty\}$ be an M-convex function with a bounded effective domain, and define

$$\tilde{f}(x_1,x_2,x_3,x_4;\xi) = \inf\{f(x_1,x_2,x_3,x_4;x_5,x_6) \mid x_5+x_6 = \xi\}.$$

We now show that if $\tilde{y} = \mathbf{0} \in \tilde{J}$ and $\tilde{x} = (1,1,1,1;0) \in \tilde{J}$, then there exists an $(\tilde{x},\tilde{y})$-increment pair (s,t) satisfying $(*)$. We may assume that $\tilde{f}(0,0,0,0;0) = f(0,0,0,0;0,0)$ and $\tilde{f}(1,1,1,1;0) = f(1,1,1,1;k,-k)$ with $k > 0$. We denote $\mathbf{0} = (0,0,0,0;0,0)$, $\mathbf{1}_k = (1,1,1,1;k,-k)$, and $\chi_{1234} = \chi_1+\chi_2+\chi_3+\chi_4$.

Lemma 20. *Suppose that* $\tilde{f}(0,0,0,0;0) = f(\mathbf{0})$ *and* $\tilde{f}(1,1,1,1;0) = f(\mathbf{1}_k)$ *with* $k > 0$. *Then we have*

$$f(\mathbf{0}) + f(\mathbf{1}_k) \geq \min\left\{\begin{array}{c}\tilde{f}(1,1,0,0;0) + \tilde{f}(0,0,1,1;0),\\ \tilde{f}(1,0,1,0;0) + \tilde{f}(0,1,0,1;0),\\ \tilde{f}(1,0,0,1;0) + \tilde{f}(0,1,1,0;0)\end{array}\right\}. \tag{2}$$

Proof. First, by (M^{J}-EXC) for f with $s = \chi_1$, we have

$$f(\mathbf{0}) + f(\mathbf{1}_k) \geq \min\left\{\begin{array}{c} f(1,1,0,0;0,0) + f(0,0,1,1;k,-k),\\ f(1,0,1,0;0,0) + f(0,1,0,1;k,-k),\\ f(1,0,0,1;0,0) + f(0,1,1,0;k,-k),\\ f(1,0,0,0;1,0) + f(0,1,1,1;k-1,-k),\\ f(1,0,0,0;0,-1) + f(0,1,1,1;k,-k+1)\end{array}\right\}.$$

If one of the first three terms achieves the minimum, the desired inequality holds. Otherwise, we have

$$f(\mathbf{0}) + f(\mathbf{1}_k) \geq \min\left\{\begin{array}{c} f(1,0,0,0;1,0) + f(0,1,1,1;k-1,-k),\\ f(1,0,0,0;0,-1) + f(0,1,1,1;k,-k+1)\end{array}\right\}. \tag{3}$$

We consider the following bipartite digraph $G = (U_G, V_G; A_G)$. The vertex sets U_G and V_G are defined by

$$U_G = \{u_{(p,i)} \mid 1 \leq p \leq k,\ 1 \leq i \leq 4,\ f(\chi_{1234} - \chi_i + p\chi_5 - (p-1)\chi_6) < +\infty\},$$
$$V_G = \{v_{(r,j)} \mid 1 \leq r \leq k,\ 1 \leq j \leq 4,\ f(\chi_j + r\chi_5 - (r-1)\chi_6) < +\infty\}.$$

The arc set A_G is defined as follows. For $u_{(p,i)} \in U_G$ an arc exists from $u_{(p,i)}$ to $v_{(r,j)}$ with $r \in \{1,\ldots,k\}$ and $j \in \{1,2,3,4\} \setminus \{i\}$ if there exists q such that $0 \leq q \leq k$ and

$$\begin{aligned} &f(\mathbf{0}) + f(\chi_{1234} - \chi_i + p\chi_5 - (p-1)\chi_6)\\ &\qquad\geq f(\chi_j + r\chi_5 - (r-1)\chi_6) + f(\chi_{1234} - \chi_i - \chi_j + q\chi_5 - q\chi_6).\end{aligned}$$

Note that this inequality guarantees $v_{(r,j)} \in V_G$. Similarly, for $v_{(r,j)} \in V_G$ an arc exists from $v_{(r,j)}$ to $u_{(p,i)}$ with $p \in \{1, \dots, k\}$ and $i \in \{1,2,3,4\} \setminus \{j\}$ if there exists q such that $0 \leq q \leq k$ and

$$\begin{aligned} &f(\mathbf{1}_k) + f(\chi_j + r\chi_5 - (r-1)\chi_6) \\ &\qquad\qquad \geq f(\chi_{1234} - \chi_i + p\chi_5 - (p-1)\chi_6) + f(\chi_i + \chi_j + q\chi_5 - q\chi_6). \end{aligned}$$

Note that this inequality guarantees $u_{(p,i)} \in U_G$.

Then the following lemma holds, which we prove in section 7.2.2.

LEMMA 21. *The out-degree of each vertex in G is at least one.*

We mention here that $U_G \neq \emptyset$ and $V_G \neq \emptyset$. For it follows from the inequality (3) that $u_{(k,1)} \in U_G$ or $v_{(1,1)} \in V_G$. Then Lemma 21 implies that $U_G \neq \emptyset$ and $V_G \neq \emptyset$.

By Lemma 21, G has a directed cycle

$$C = (u_{(p_1,i_1)}, v_{(p_2,i_2)}, u_{(p_3,i_3)}, v_{(p_4,i_4)}, \dots, u_{(p_{2m-1},i_{2m-1})}, v_{(p_{2m},i_{2m})}).$$

This means, by the definition of A_G, that there exist $q_1, q_2, \dots, q_{2m}$ such that

$$\begin{aligned} &f(\mathbf{0}) + f(\chi_{1234} - \chi_{i_1} + p_1\chi_5 - (p_1-1)\chi_6) \\ &\qquad \geq f(\chi_{i_2} + p_2\chi_5 - (p_2-1)\chi_6) + f(\chi_{1234} - \chi_{i_1} - \chi_{i_2} + q_1\chi_5 - q_1\chi_6), \\ &f(\mathbf{1}_k) + f(\chi_{i_2} + p_2\chi_5 - (p_2-1)\chi_6) \\ &\qquad \geq f(\chi_{1234} - \chi_{i_3} + p_3\chi_5 - (p_3-1)\chi_6) + f(\chi_{i_3} + \chi_{i_2} + q_2\chi_5 - q_2\chi_6), \\ &f(\mathbf{0}) + f(\chi_{1234} - \chi_{i_3} + p_3\chi_5 - (p_3-1)\chi_6) \\ &\qquad \geq f(\chi_{i_4} + p_4\chi_5 - (p_4-1)\chi_6) + f(\chi_{1234} - \chi_{i_3} - \chi_{i_4} + q_3\chi_5 - q_3\chi_6), \\ &f(\mathbf{1}_k) + f(\chi_{i_4} + p_4\chi_5 - (p_4-1)\chi_6) \\ &\qquad \geq f(\chi_{1234} - \chi_{i_5} + p_5\chi_5 - (p_5-1)\chi_6) + f(\chi_{i_5} + \chi_{i_4} + q_4\chi_5 - q_4\chi_6), \\ &\cdots \\ &f(\mathbf{0}) + f(\chi_{1234} - \chi_{i_{2m-1}} + p_{2m-1}\chi_5 - (p_{2m-1}-1)\chi_6) \\ &\qquad \geq f(\chi_{i_{2m}} + p_{2m}\chi_5 - (p_{2m}-1)\chi_6) \\ &\qquad\quad + f(\chi_{1234} - \chi_{i_{2m-1}} - \chi_{i_{2m}} + q_{2m-1}\chi_5 - q_{2m-1}\chi_6), \\ &f(\mathbf{1}_k) + f(\chi_{i_{2m}} + p_{2m}\chi_5 - (p_{2m}-1)\chi_6) \\ &\qquad \geq f(\chi_{1234} - \chi_{i_1} + p_1\chi_5 - (p_1-1)\chi_6) + f(\chi_{i_1} + \chi_{i_{2m}} + q_{2m}\chi_5 - q_{2m}\chi_6). \end{aligned}$$

By adding these inequalities, we obtain

$$\begin{aligned} m(f(\mathbf{0}) + f(\mathbf{1}_k)) \geq\ & f(\chi_{1234} - \chi_{i_1} - \chi_{i_2} + q_1\chi_5 - q_1\chi_6) \\ &+ f(\chi_{i_3} + \chi_{i_2} + q_2\chi_5 - q_2\chi_6) \\ &+ f(\chi_{1234} - \chi_{i_3} - \chi_{i_4} + q_3\chi_5 - q_3\chi_6) \\ &+ f(\chi_{i_5} + \chi_{i_4} + q_4\chi_5 - q_4\chi_6) \\ &+ \cdots \\ &+ f(\chi_{1234} - \chi_{i_{2m-1}} - \chi_{i_{2m}} + q_{2m-1}\chi_5 - q_{2m-1}\chi_6) \\ &+ f(\chi_{i_1} + \chi_{i_{2m}} + q_{2m}\chi_5 - q_{2m}\chi_6). \end{aligned}$$

Then we have

$$\begin{aligned} m(f(\mathbf{0})+f(\mathbf{1}_k)) \geq\ & \tilde{f}(\tilde{\chi}_{1234}-\tilde{\chi}_{i_1}-\tilde{\chi}_{i_2})+\tilde{f}(\tilde{\chi}_{i_3}+\tilde{\chi}_{i_2}) \\ & + \tilde{f}(\tilde{\chi}_{1234}-\tilde{\chi}_{i_3}-\tilde{\chi}_{i_4})+\tilde{f}(\tilde{\chi}_{i_5}+\tilde{\chi}_{i_4}) \\ & + \cdots \\ & + \tilde{f}(\tilde{\chi}_{1234}-\tilde{\chi}_{i_{2m-1}}-\tilde{\chi}_{i_{2m}})+\tilde{f}(\tilde{\chi}_{i_1}+\tilde{\chi}_{i_{2m}}), \end{aligned}$$

where $\tilde{\chi}_{1234}=\tilde{\chi}_1+\tilde{\chi}_2+\tilde{\chi}_3+\tilde{\chi}_4$.

Here we note that

$$\begin{aligned} m\tilde{\chi}_{1234} =\ & (\tilde{\chi}_{1234}-\tilde{\chi}_{i_1}-\tilde{\chi}_{i_2})+(\tilde{\chi}_{i_3}+\tilde{\chi}_{i_2}) \\ & + (\tilde{\chi}_{1234}-\tilde{\chi}_{i_3}-\tilde{\chi}_{i_4})+(\tilde{\chi}_{i_5}+\tilde{\chi}_{i_4}) \\ & + \cdots \\ & + (\tilde{\chi}_{1234}-\tilde{\chi}_{i_{2m-1}}-\tilde{\chi}_{i_{2m}})+(\tilde{\chi}_{i_1}+\tilde{\chi}_{i_{2m}}). \end{aligned}$$

Then the desired inequality (2) follows from Lemma 22 below. □

LEMMA 22. *If*

$$m\tilde{\chi}_{1234} = \sum_{1\leq i<j\leq 4} m_{ij}(\tilde{\chi}_i+\tilde{\chi}_j) \tag{4}$$

and

$$m(f(\mathbf{0})+f(\mathbf{1}_k)) \geq \sum_{1\leq i<j\leq 4} m_{ij}\tilde{f}(\tilde{\chi}_i+\tilde{\chi}_j)$$

for some nonnegative integers m_{ij} and m, then

$$f(\mathbf{0})+f(\mathbf{1}_k) \geq \min\left\{ \begin{matrix} \tilde{f}(\tilde{\chi}_1+\tilde{\chi}_2)+\tilde{f}(\tilde{\chi}_3+\tilde{\chi}_4), \\ \tilde{f}(\tilde{\chi}_1+\tilde{\chi}_3)+\tilde{f}(\tilde{\chi}_2+\tilde{\chi}_4), \\ \tilde{f}(\tilde{\chi}_1+\tilde{\chi}_4)+\tilde{f}(\tilde{\chi}_2+\tilde{\chi}_3) \end{matrix} \right\}.$$

Proof. On the right-hand side of (4), the sum of the coefficients of $\tilde{\chi}_1$ and $\tilde{\chi}_2$ is $2m_{12}+m_{13}+m_{14}+m_{23}+m_{24}$. Meanwhile, that of $\tilde{\chi}_3$ and $\tilde{\chi}_4$ is $2m_{34}+m_{13}+m_{14}+m_{23}+m_{24}$. Hence $m_{12}=m_{34}$. We can show $m_{13}=m_{24}$, $m_{14}=m_{23}$ in the same way. Thus we have

$$\begin{aligned} m(f(\mathbf{0})+f(\mathbf{1}_k)) \geq\ & m_{12}(\tilde{f}(\tilde{\chi}_1+\tilde{\chi}_2)+\tilde{f}(\tilde{\chi}_3+\tilde{\chi}_4)) \\ & + m_{13}(\tilde{f}(\tilde{\chi}_1+\tilde{\chi}_3)+\tilde{f}(\tilde{\chi}_2+\tilde{\chi}_4)) \\ & + m_{14}(\tilde{f}(\tilde{\chi}_1+\tilde{\chi}_4)+\tilde{f}(\tilde{\chi}_2+\tilde{\chi}_3)) \end{aligned}$$

and

$$m_{12}+m_{13}+m_{14}=m,$$

which imply the desired inequality. □

7.2.2. Proof of Lemma 21. The out-degree of vertex $u_{(p,i)}$ is nonzero by Lemma 24 below, which relies on the following lemma.

LEMMA 23. *For any integers $p \leq q$ and for any $i \in \{1,2,3,4\}$,* (A) *or* (B) *holds.*

(A) *There exists an integer r such that $p \leq r \leq q+1$ and*

$$
\begin{aligned}
&f((p+2)\chi_5 - p\chi_6) + f(\chi_{1234} - \chi_i + q\chi_5 - (q+1)\chi_6) \\
&\qquad \geq \min \left\{ \begin{array}{l}
f(\chi_{j_1} + (p+q+2-r)\chi_5 - (p+q+1-r)\chi_6) \\
\qquad\qquad + f(\chi_{j_2} + \chi_{j_3} + r\chi_5 - r\chi_6), \\
f(\chi_{j_2} + (p+q+2-r)\chi_5 - (p+q+1-r)\chi_6) \\
\qquad\qquad + f(\chi_{j_3} + \chi_{j_1} + r\chi_5 - r\chi_6), \\
f(\chi_{j_3} + (p+q+2-r)\chi_5 - (p+q+1-r)\chi_6) \\
\qquad\qquad + f(\chi_{j_1} + \chi_{j_2} + r\chi_5 - r\chi_6)
\end{array} \right\},
\end{aligned}
$$

where $\{j_1, j_2, j_3\} = \{1,2,3,4\} \setminus \{i\}$.

(B) *There exists an integer r such that $p+1 \leq r \leq q+1$ and*

$$
\begin{aligned}
&f((p+2)\chi_5 - p\chi_6) + f(\chi_{1234} - \chi_i + q\chi_5 - (q+1)\chi_6) \\
&\geq f(r\chi_5 - r\chi_6) + f(\chi_{1234} - \chi_i + (p+q+2-r)\chi_5 - (p+q+1-r)\chi_6).
\end{aligned}
$$

Proof. We show the proof by induction on $q-p$.

If $q-p=0$, then, by (M$^{\mathrm{J}}$-EXC) with $s = -\chi_6$, we have

$$
\begin{aligned}
&f((p+2)\chi_5 - p\chi_6) + f(\chi_{1234} - \chi_i + q\chi_5 - (q+1)\chi_6) \\
&\quad \geq \min \left\{ \begin{array}{c}
f(\chi_{j_1} + (p+2)\chi_5 - (p+1)\chi_6) + f(\chi_{1234} - \chi_{j_1} - \chi_i + q\chi_5 - q\chi_6), \\
f(\chi_{j_2} + (p+2)\chi_5 - (p+1)\chi_6) + f(\chi_{1234} - \chi_{j_2} - \chi_i + q\chi_5 - q\chi_6), \\
f(\chi_{j_3} + (p+2)\chi_5 - (p+1)\chi_6) + f(\chi_{1234} - \chi_{j_3} - \chi_i + q\chi_5 - q\chi_6), \\
f((p+1)\chi_5 - (p+1)\chi_6) + f(\chi_{1234} - \chi_i + (q+1)\chi_5 - q\chi_6)
\end{array} \right\},
\end{aligned}
$$

where $\{j_1, j_2, j_3\} = \{1,2,3,4\} \setminus \{i\}$. If one of the first three terms achieves the minimum, then (A) holds with $r = q$; otherwise (B) holds with $r = p+1$.

If $q-p=1$, then, by (M$^{\mathrm{J}}$-EXC) with $s = -\chi_5$, we have

$$
\begin{aligned}
&f((p+2)\chi_5 - p\chi_6) + f(\chi_{1234} - \chi_i + q\chi_5 - (q+1)\chi_6) \\
&\geq \min \left\{ \begin{array}{c}
f(\chi_{j_1} + (p+1)\chi_5 - p\chi_6) + f(\chi_{1234} - \chi_{j_1} - \chi_i + (q+1)\chi_5 - (q+1)\chi_6), \\
f(\chi_{j_2} + (p+1)\chi_5 - p\chi_6) + f(\chi_{1234} - \chi_{j_2} - \chi_i + (q+1)\chi_5 - (q+1)\chi_6), \\
f(\chi_{j_3} + (p+1)\chi_5 - p\chi_6) + f(\chi_{1234} - \chi_{j_3} - \chi_i + (q+1)\chi_5 - (q+1)\chi_6), \\
f((p+1)\chi_5 - (p+1)\chi_6) + f(\chi_{1234} - \chi_i + (q+1)\chi_5 - q\chi_6)
\end{array} \right\},
\end{aligned}
$$

where $\{j_1, j_2, j_3\} = \{1,2,3,4\} \setminus \{i\}$. If one of the first three terms achieves the minimum, then (A) holds with $r = q+1$; otherwise (B) holds with $r = p+1$.

Suppose that $q-p \geq 2$. By (M$^{\mathrm{J}}$-EXC) with $s = -\chi_6$, we have

$$
\begin{aligned}
&f((p+2)\chi_5 - p\chi_6) + f(\chi_{1234} - \chi_i + q\chi_5 - (q+1)\chi_6) \\
&\quad \geq \min \left\{ \begin{array}{c}
f(\chi_{j_1} + (p+2)\chi_5 - (p+1)\chi_6) + f(\chi_{1234} - \chi_{j_1} - \chi_i + q\chi_5 - q\chi_6), \\
f(\chi_{j_2} + (p+2)\chi_5 - (p+1)\chi_6) + f(\chi_{1234} - \chi_{j_2} - \chi_i + q\chi_5 - q\chi_6), \\
f(\chi_{j_3} + (p+2)\chi_5 - (p+1)\chi_6) + f(\chi_{1234} - \chi_{j_3} - \chi_i + q\chi_5 - q\chi_6), \\
f((p+3)\chi_5 - (p+1)\chi_6) + f(\chi_{1234} - \chi_i + (q-1)\chi_5 - q\chi_6), \\
f((p+2)\chi_5 - (p+2)\chi_6) + f(\chi_{1234} - \chi_i + q\chi_5 - (q-1)\chi_6)
\end{array} \right\},
\end{aligned}
$$

where $\{j_1, j_2, j_3\} = \{1,2,3,4\} \setminus \{i\}$. Note that the fourth term exists only if $q-p \geq 3$. If one of the first three terms achieves the minimum, then (A) holds with $r = q$, and if the last term achieves the minimum, then (B) holds with $r = p+2 \leq q$. To the fourth term, the induction applies and yields (A) with $p+1 \le r \le q$ or (B) with $p+2 \le r \le q$. □

LEMMA 24. *For any integer $1 \le p \le k$ and for any $i \in \{1,2,3,4\}$, there exist integers q and r such that $0 \le q \le k-1$, $1 \le r \le k$, and*

$$\begin{aligned} &f(\mathbf{0}) + f(\chi_{1234} - \chi_i + p\chi_5 - (p-1)\chi_6) \\ &\qquad \ge \min \left\{ \begin{matrix} f(\chi_{j_1} + r\chi_5 - (r-1)\chi_6) + f(\chi_{j_2} + \chi_{j_3} + q\chi_5 - q\chi_6), \\ f(\chi_{j_2} + r\chi_5 - (r-1)\chi_6) + f(\chi_{j_3} + \chi_{j_1} + q\chi_5 - q\chi_6), \\ f(\chi_{j_3} + r\chi_5 - (r-1)\chi_6) + f(\chi_{j_1} + \chi_{j_2} + q\chi_5 - q\chi_6) \end{matrix} \right\}, \end{aligned}$$

where $\{j_1, j_2, j_3\} = \{1,2,3,4\} \setminus \{i\}$.

Proof. It suffices to consider p which minimizes $f(\chi_{1234} - \chi_i + p\chi_5 - (p-1)\chi_6)$. Let p be the minimum minimizer. By (M$^{\mathrm{J}}$-EXC) with $s = \chi_5$, we have

$$\begin{aligned} &f(\mathbf{0}) + f(\chi_{1234} - \chi_i + p\chi_5 - (p-1)\chi_6) \\ &\qquad \ge \min \left\{ \begin{matrix} f(\chi_{j_1} + \chi_5) + f(\chi_{j_2} + \chi_{j_3} + (p-1)\chi_5 - (p-1)\chi_6), \\ f(\chi_{j_2} + \chi_5) + f(\chi_{j_3} + \chi_{j_1} + (p-1)\chi_5 - (p-1)\chi_6), \\ f(\chi_{j_3} + \chi_5) + f(\chi_{j_1} + \chi_{j_2} + (p-1)\chi_5 - (p-1)\chi_6), \\ f(\chi_5 - \chi_6) + f(\chi_{1234} - \chi_i + (p-1)\chi_5 - (p-2)\chi_6), \\ f(2\chi_5) + f(\chi_{1234} - \chi_i + (p-2)\chi_5 - (p-1)\chi_6) \end{matrix} \right\}. \end{aligned}$$

Note that the last two terms exist only if $p \ge 2$.

If one of the first three terms achieves the minimum, the claim holds with $q = p-1$ and $r = 1$.

To consider the fourth term, suppose that $p \ge 2$ and

$$\begin{aligned} &f(\mathbf{0}) + f(\chi_{1234} - \chi_i + p\chi_5 - (p-1)\chi_6) \\ &\qquad \ge f(\chi_5 - \chi_6) + f(\chi_{1234} - \chi_i + (p-1)\chi_5 - (p-2)\chi_6). \end{aligned}$$

Then, since $f(\mathbf{0}) = \tilde{f}(0,0,0,0;0) \le f(\chi_5 - \chi_6)$, we have

$$f(\chi_{1234} - \chi_i + p\chi_5 - (p-1)\chi_6) \ge f(\chi_{1234} - \chi_i + (p-1)\chi_5 - (p-2)\chi_6),$$

which contradicts the definition of p.

To consider the fifth term, suppose that $p \ge 2$ and

$$\begin{aligned} &f(\mathbf{0}) + f(\chi_{1234} - \chi_i + p\chi_5 - (p-1)\chi_6) \\ &\qquad \ge f(2\chi_5) + f(\chi_{1234} - \chi_i + (p-2)\chi_5 - (p-1)\chi_6). \end{aligned}$$

By Lemma 23, at least one of (A) and (B) holds.

(A) There exists an integer r' such that $0 \le r' \le p-1$ and

$$\begin{aligned} &f(2\chi_5) + f(\chi_{1234} - \chi_i + (p-2)\chi_5 - (p-1)\chi_6) \\ &\qquad \ge \min \left\{ \begin{matrix} f(\chi_{j_1} + (p-r')\chi_5 - (p-r'-1)\chi_6) \\ \qquad + f(\chi_{j_2} + \chi_{j_3} + r'\chi_5 - r'\chi_6), \\ f(\chi_{j_2} + (p-r')\chi_5 - (p-r'-1)\chi_6) \\ \qquad + f(\chi_{j_3} + \chi_{j_1} + r'\chi_5 - r'\chi_6), \\ f(\chi_{j_3} + (p-r')\chi_5 - (p-r'-1)\chi_6) \\ \qquad + f(\chi_{j_1} + \chi_{j_2} + r'\chi_5 - r'\chi_6) \end{matrix} \right\}. \end{aligned}$$

(B) There exists an integer r' such that $1 \le r' \le p-1$ and

$$\begin{aligned} &f(2\chi_5) + f(\chi_{1234} - \chi_i + (p-2)\chi_5 - (p-1)\chi_6) \\ &\qquad \ge f(r'\chi_5 - r'\chi_6) + f(\chi_{1234} - \chi_i + (p-r')\chi_5 - (p-r'-1)\chi_6). \end{aligned}$$

In case of (A) we have

$$\begin{aligned} &f(\mathbf{0}) + f(\chi_{1234} - \chi_i + p\chi_5 - (p-1)\chi_6) \\ &\qquad \ge \min \left\{ \begin{matrix} f(\chi_{j_1} + (p-r')\chi_5 - (p-r'-1)\chi_6) + f(\chi_{j_2} + \chi_{j_3} + r'\chi_5 - r'\chi_6), \\ f(\chi_{j_2} + (p-r')\chi_5 - (p-r'-1)\chi_6) + f(\chi_{j_3} + \chi_{j_1} + r'\chi_5 - r'\chi_6), \\ f(\chi_{j_3} + (p-r')\chi_5 - (p-r'-1)\chi_6) + f(\chi_{j_1} + \chi_{j_2} + r'\chi_5 - r'\chi_6) \end{matrix} \right\}, \end{aligned}$$

which implies the desired claim with $q = r'$ and $r = p - r'$.

In case of (B) we have $1 \le r' \le p-1$ and

$$\begin{aligned} &f(\mathbf{0}) + f(\chi_{1234} - \chi_i + p\chi_5 - (p-1)\chi_6) \\ &\qquad \ge f(r'\chi_5 - r'\chi_6) + f(\chi_{1234} - \chi_i + (p-r')\chi_5 - (p-r'-1)\chi_6). \end{aligned}$$

Since $f(\mathbf{0}) = \tilde{f}(0,0,0,0;0) \le f(r'\chi_5 - r'\chi_6)$, we have

$$f(\chi_{1234} - \chi_i + p\chi_5 - (p-1)\chi_6) \ge f(\chi_{1234} - \chi_i + (p-r')\chi_5 - (p-r'-1)\chi_6),$$

which contradicts the definition of p. □

In the same way as Lemma 24, we have the following lemma, which means that the out-degree of vertex $v_{(r,j)}$ is nonzero.

LEMMA 25. *For any integer $1 \le r \le k$ and for any $j \in \{1,2,3,4\}$, there exist integers p and q such that $1 \le p \le k$, $1 \le q \le k$, and*

$$\begin{aligned} &f(\mathbf{1}_k) + f(\chi_j + r\chi_5 - (r-1)\chi_6) \\ &\qquad \ge \min \left\{ \begin{matrix} f(\chi_{1234} - \chi_{i_1} + p\chi_5 - (p-1)\chi_6) + f(\chi_{i_1} + \chi_j + q\chi_5 - q\chi_6), \\ f(\chi_{1234} - \chi_{i_2} + p\chi_5 - (p-1)\chi_6) + f(\chi_{i_2} + \chi_j + q\chi_5 - q\chi_6), \\ f(\chi_{1234} - \chi_{i_3} + p\chi_5 - (p-1)\chi_6) + f(\chi_{i_3} + \chi_j + q\chi_5 - q\chi_6) \end{matrix} \right\}, \end{aligned}$$

where $\{i_1, i_2, i_3\} = \{1,2,3,4\} \setminus \{j\}$.

Proof. We consider the coordinate transformation from $(x_1, x_2, x_3, x_4; x_5, x_6)$ to $(1-x_1, 1-x_2, 1-x_3, 1-x_4; k+x_6, -k+x_5)$. Applying Lemma 24 in the new coordinate system, we see the following fact:

> For any integer $1 \le p' \le k$ and for any $j \in \{1,2,3,4\}$, there exist integers q' and r' such that $0 \le q' \le k-1$, $1 \le r' \le k$, and
>
> $$\begin{aligned} &f(\mathbf{1}_k) + f(\chi_j + (k-p'+1)\chi_5 - (k-p')\chi_6) \\ &\qquad \ge \min \left\{ \begin{matrix} f(\chi_{1234} - \chi_{i_1} + (k-r'+1)\chi_5 - (k-r')\chi_6) \\ \qquad + f(\chi_{i_1} + \chi_j + (k-q')\chi_5 - (k-q')\chi_6), \\ f(\chi_{1234} - \chi_{i_2} + (k-r'+1)\chi_5 - (k-r')\chi_6) \\ \qquad + f(\chi_{i_2} + \chi_j + (k-q')\chi_5 - (k-q')\chi_6), \\ f(\chi_{1234} - \chi_{i_3} + (k-r'+1)\chi_5 - (k-r')\chi_6) \\ \qquad + f(\chi_{i_3} + \chi_j + (k-q')\chi_5 - (k-q')\chi_6) \end{matrix} \right\}, \end{aligned}$$
>
> where $\{i_1, i_2, i_3\} = \{1,2,3,4\} \setminus \{j\}$.

By setting $p = k - r' + 1$, $r = k - p' + 1$, and $q = k - q'$, we obtain the claim. □

Lemma 21 immediately follows from Lemmas 24 and 25.

7.2.3. Other cases in Case 2. The other cases in Case 2 are treated here with the aid of the splitting technique.

LEMMA 26. *If $\tilde{y} = \mathbf{0}$ and $\tilde{x} = (1,1,1,1;0)$, $(2,1,1;0)$, $(3,1;0)$, $(2,2;0)$, or $(4;0)$, then there exists an $(\tilde{x},\tilde{y})$-increment pair (s,t) satisfying $\tilde{f}(\tilde{x})+\tilde{f}(\tilde{y}) \geq \tilde{f}(\tilde{x}+s+t)+\tilde{f}(\tilde{y}-s-t)$.*

Proof. If $\tilde{x} = (1,1,1,1;0)$, then the claim follows from Lemma 20.

Suppose that $\tilde{x} = (2,1,1;0)$. In this case, we may assume $\tilde{f}(0,0,0;0) = f(0,0,0;0,0)$ and $\tilde{f}(2,1,1;0) = f(2,1,1;k,-k)$ with $k > 0$. We define f' as $f'(x_1,x_2,x_3,x_4;x_5,x_6) = f(x_1+x_2,x_3,x_4;x_5,x_6)$, and $\tilde{f}'$ as

$$\tilde{f}'(x_1,x_2,x_3,x_4;\xi) = \inf\{f'(x_1,x_2,x_3,x_4;x_5,x_6) \mid x_5+x_6 = \xi\}.$$

Then $\tilde{f}'(x_1,x_2,x_3,x_4;0) = \tilde{f}(x_1+x_2,x_3,x_4;0)$. Since f' is a splitting of f, it is M-convex by Theorem 7. By Lemma 20, we have

$$\begin{aligned} f(0,0,0;0,0)+f(2,1,1;k,-k) &= f'(0,0,0,0;0,0)+f'(1,1,1,1;k,-k) \\ &\geq \min\left\{\begin{array}{l} \tilde{f}'(1,1,0,0;0)+\tilde{f}'(0,0,1,1;0), \\ \tilde{f}'(1,0,1,0;0)+\tilde{f}'(0,1,0,1;0), \\ \tilde{f}'(1,0,0,1;0)+\tilde{f}'(0,1,1,0;0) \end{array}\right\} \\ &= \min\left\{\begin{array}{l} \tilde{f}(2,0,0;0)+\tilde{f}(0,1,1;0), \\ \tilde{f}(1,1,0;0)+\tilde{f}(1,0,1;0), \\ \tilde{f}(1,0,1;0)+\tilde{f}(1,1,0;0) \end{array}\right\}, \end{aligned}$$

which means that there exists an $(\tilde{x},\tilde{y})$-increment pair (s,t) satisfying $\tilde{f}(\tilde{x})+\tilde{f}(\tilde{y}) \geq \tilde{f}(\tilde{x}+s+t)+\tilde{f}(\tilde{y}-s-t)$.

Suppose that $\tilde{x} = (3,1;0)$. In this case, we may assume $\tilde{f}(0,0;0) = f(0,0;0,0)$ and $\tilde{f}(3,1;0) = f(3,1;k,-k)$ with $k > 0$. We define f' as $f'(x_1,x_2,x_3,x_4;x_5,x_6) = f(x_1+x_2+x_3,x_4;x_5,x_6)$, and $\tilde{f}'$ as

$$\tilde{f}'(x_1,x_2,x_3,x_4;\xi) = \inf\{f'(x_1,x_2,x_3,x_4;x_5,x_6) \mid x_5+x_6 = \xi\}.$$

Then $\tilde{f}'(x_1,x_2,x_3,x_4;0) = \tilde{f}(x_1+x_2+x_3,x_4;0)$. Since f' is a splitting of f, it is M-convex by Theorem 7. By Lemma 20, we have

$$\begin{aligned} f(0,0;0,0)+f(3,1;k,-k) &= f'(0,0,0,0;0,0)+f'(1,1,1,1;k,-k) \\ &\geq \min\left\{\begin{array}{l} \tilde{f}'(1,1,0,0;0)+\tilde{f}'(0,0,1,1;0), \\ \tilde{f}'(1,0,1,0;0)+\tilde{f}'(0,1,0,1;0), \\ \tilde{f}'(1,0,0,1;0)+\tilde{f}'(0,1,1,0;0) \end{array}\right\} \\ &= \min\left\{\begin{array}{l} \tilde{f}(2,0;0)+\tilde{f}(1,1;0), \\ \tilde{f}(2,0;0)+\tilde{f}(1,1;0), \\ \tilde{f}(1,1;0)+\tilde{f}(2,0;0) \end{array}\right\} \\ &= \tilde{f}(2,0;0)+\tilde{f}(1,1;0), \end{aligned}$$

which means that there exists an $(\tilde{x},\tilde{y})$-increment pair (s,t) satisfying $\tilde{f}(\tilde{x})+\tilde{f}(\tilde{y}) \geq \tilde{f}(\tilde{x}+s+t)+\tilde{f}(\tilde{y}-s-t)$.

Suppose that $\tilde{x} = (2,2;0)$. In this case, we may assume $\tilde{f}(0,0;0) = f(0,0;0,0)$ and $\tilde{f}(2,2;0) = f(2,2;k,-k)$ with $k > 0$. We define f' as $f'(x_1,x_2,x_3,x_4;x_5,x_6) = f(x_1+x_2,x_3+x_4;x_5,x_6)$, and $\tilde{f}'$ as

$$\tilde{f}'(x_1,x_2,x_3,x_4;\xi) = \inf\{f'(x_1,x_2,x_3,x_4;x_5,x_6) \mid x_5+x_6 = \xi\}.$$

Then $\tilde{f}'(x_1, x_2, x_3, x_4; 0) = \tilde{f}(x_1 + x_2, x_3 + x_4; 0)$. Since f' is a splitting of f, it is M-convex by Theorem 7. By Lemma 20, we have

$$\begin{aligned} f(0,0;0,0) + f(2,2;k,-k) &= f'(0,0,0,0;0,0) + f'(1,1,1,1;k,-k) \\ &\geq \min \left\{ \begin{array}{l} \tilde{f}'(1,1,0,0;0) + \tilde{f}'(0,0,1,1;0), \\ \tilde{f}'(1,0,1,0;0) + \tilde{f}'(0,1,0,1;0), \\ \tilde{f}'(1,0,0,1;0) + \tilde{f}'(0,1,1,0;0) \end{array} \right\} \\ &= \min \left\{ \begin{array}{l} \tilde{f}(2,0;0) + \tilde{f}(0,2;0), \\ \tilde{f}(1,1;0) + \tilde{f}(1,1;0), \\ \tilde{f}(1,1;0) + \tilde{f}(1,1;0) \end{array} \right\}, \end{aligned}$$

which means that there exists an $(\tilde{x}, \tilde{y})$-increment pair (s,t) satisfying $\tilde{f}(\tilde{x}) + \tilde{f}(\tilde{y}) \geq \tilde{f}(\tilde{x} + s + t) + \tilde{f}(\tilde{y} - s - t)$.

Suppose that $\tilde{x} = (4;0)$. In this case, we may assume $\tilde{f}(0;0) = f(0;0,0)$ and $\tilde{f}(4;0) = f(4;k,-k)$ with $k > 0$. We define f' as $f'(x_1, x_2, x_3, x_4; x_5, x_6) = f(x_1 + x_2 + x_3 + x_4; x_5, x_6)$, and $\tilde{f}'$ as

$$\tilde{f}'(x_1, x_2, x_3, x_4; \xi) = \inf \{ f'(x_1, x_2, x_3, x_4; x_5, x_6) \mid x_5 + x_6 = \xi \}.$$

Then $\tilde{f}'(x_1, x_2, x_3, x_4; 0) = \tilde{f}(x_1 + x_2 + x_3 + x_4; 0)$. Since f' is a splitting of f, it is M-convex by Theorem 7. By Lemma 20, we have

$$\begin{aligned} f(0;0,0) + f(4;k,-k) &= f'(0,0,0,0;0,0) + f'(1,1,1,1;k,-k) \\ &\geq \min \left\{ \begin{array}{l} \tilde{f}'(1,1,0,0;0) + \tilde{f}'(0,0,1,1;0), \\ \tilde{f}'(1,0,1,0;0) + \tilde{f}'(0,1,0,1;0), \\ \tilde{f}'(1,0,0,1;0) + \tilde{f}'(0,1,1,0;0) \end{array} \right\} \\ &= \tilde{f}(2;0) + \tilde{f}(2;0), \end{aligned}$$

which means that there exists an $(\tilde{x}, \tilde{y})$-increment pair (s,t) satisfying $\tilde{f}(\tilde{x}) + \tilde{f}(\tilde{y}) \geq \tilde{f}(\tilde{x} + s + t) + \tilde{f}(\tilde{y} - s - t)$. □

7.3. Case 3 in the proof of Lemma 19. In this section, we deal with Case 3 in the proof of Lemma 19. First we focus on the case of $\tilde{x} = (1,1,1;1)$, whereas the other cases are treated later using the splitting technique discussed in section 4.

Let $f : \mathbf{Z}^5 \to \mathbf{R} \cup \{+\infty\}$ be an M-convex function, and put

$$\tilde{f}(x_1, x_2, x_3; \xi) = \inf \{ f(x_1, x_2, x_3; x_4, x_5) \mid x_4 + x_5 = \xi \}.$$

We now show that if $\tilde{y} = \mathbf{0} \in \tilde{J}$ and $\tilde{x} = (1,1,1;1) \in \tilde{J}$, then there exists an $(\tilde{x}, \tilde{y})$-increment pair (s,t) satisfying $(*)$. We may assume $\tilde{f}(0,0,0;0) = f(0,0,0;0,0)$ and $\tilde{f}(1,1,1;1) = f(1,1,1;k+1,-k)$ with $k > 0$.

Lemma 27. *Suppose that* $\tilde{f}(0,0,0;0) = f(0,0,0;0,0)$ *and* $\tilde{f}(1,1,1;1) = f(1,1,1;k+1,-k)$ *with* $k > 0$. *Then we have*

$$f(0,0,0;0,0) + f(1,1,1;k+1,-k) \geq \min \left\{ \begin{array}{l} \tilde{f}(1,1,0;0) + \tilde{f}(0,0,1;1), \\ \tilde{f}(1,0,1;0) + \tilde{f}(0,1,0;1), \\ \tilde{f}(0,1,1;0) + \tilde{f}(1,0,0;1) \end{array} \right\}.$$

Proof. We define $f' : \mathbf{Z}^6 \to \mathbf{R} \cup \{+\infty\}$ as

$$f'(x_1, x_2, x_3, x_4; x_5, x_6) = \begin{cases} f(x_1, x_2, x_3; x_5, x_6) & \text{if } x_4 = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Since f is M-convex, f' is also M-convex. By Lemma 24 applied to f' with $i = 4$, we have the following fact:

For any integer $1 \le p \le k'$, there exist integers q and r such that $0 \le q \le k' - 1$, $1 \le r \le k'$, and

$$f'(0,0,0,0;0,0) + f'(1,1,1,0;p,-(p-1)) \ge \min\left\{\begin{matrix} f'(1,0,0,0;r,-(r-1)) + f'(0,1,1,0;q,-q), \\ f'(0,1,0,0;r,-(r-1)) + f'(1,0,1,0;q,-q), \\ f'(0,0,1,0;r,-(r-1)) + f'(1,1,0,0;q,-q) \end{matrix}\right\}.$$

By taking $k' \ge k+1$ and $p = k+1$ in the above, we have

$$\begin{aligned} &f(0,0,0;0,0) + f(1,1,1;k+1,-k) \\ &\quad = f'(0,0,0,0;0,0) + f'(1,1,1,0;k+1,-k) \\ &\quad \ge \min\left\{\begin{matrix} f'(1,0,0,0;r,-(r-1)) + f'(0,1,1,0;q,-q), \\ f'(0,1,0,0;r,-(r-1)) + f'(1,0,1,0;q,-q), \\ f'(0,0,1,0;r,-(r-1)) + f'(1,1,0,0;q,-q) \end{matrix}\right\} \\ &\quad = \min\left\{\begin{matrix} f(1,0,0;r,-(r-1)) + f(0,1,1;q,-q), \\ f(0,1,0;r,-(r-1)) + f(1,0,1;q,-q), \\ f(0,0,1;r,-(r-1)) + f(1,1,0;q,-q) \end{matrix}\right\} \\ &\quad \ge \min\left\{\begin{matrix} \tilde{f}(1,0,0;1) + \tilde{f}(0,1,1;0), \\ \tilde{f}(0,1,0;1) + \tilde{f}(1,0,1;0), \\ \tilde{f}(0,0,1;1) + \tilde{f}(1,1,0;0) \end{matrix}\right\}, \end{aligned}$$

which implies the lemma. □

LEMMA 28. *If $\tilde{y} = \mathbf{0}$ and $\tilde{x} = (1,1,1;1)$, $(2,1;1)$, or $(3;1)$, then there exists an $(\tilde{x},\tilde{y})$-increment pair (s,t) satisfying $\tilde{f}(\tilde{x}) + \tilde{f}(\tilde{y}) \ge \tilde{f}(\tilde{x}+s+t) + \tilde{f}(\tilde{y}-s-t)$.*

Proof. If $\tilde{x} = (1,1,1;1)$, then the claim follows from Lemma 27.

Suppose that $\tilde{x} = (2,1;1)$. In this case, we may assume $\tilde{f}(0,0;0) = f(0,0;0,0)$ and $\tilde{f}(2,1;1) = f(2,1;k+1,-k)$ with $k > 0$. We define f' as $f'(x_1,x_2,x_3;x_4,x_5) = f(x_1+x_2,x_3;x_4,x_5)$, and $\tilde{f}'$ as

$$\tilde{f}'(x_1,x_2,x_3;\xi) = \inf\{f'(x_1,x_2,x_3;x_4,x_5) \mid x_4 + x_5 = \xi\}.$$

Then $\tilde{f}'(x_1,x_2,x_3;\xi) = \tilde{f}(x_1+x_2,x_3;\xi)$. Since f' is a splitting of f, it is M-convex by Theorem 7. By Lemma 27, we have

$$\begin{aligned} f(0,0;0,0) + f(2,1;k+1,-k) &= f'(0,0,0;0,0) + f'(1,1,1;k+1,-k) \\ &\ge \min\left\{\begin{matrix} \tilde{f}'(1,0,0;1) + \tilde{f}'(0,1,1;0), \\ \tilde{f}'(0,1,0;1) + \tilde{f}'(1,0,1;0), \\ \tilde{f}'(0,0,1;1) + \tilde{f}'(1,1,0;0) \end{matrix}\right\} \\ &= \min\left\{\begin{matrix} \tilde{f}(1,0;1) + \tilde{f}(1,1;0), \\ \tilde{f}(1,0;1) + \tilde{f}(1,1;0), \\ \tilde{f}(0,1;1) + \tilde{f}(2,0;0) \end{matrix}\right\}, \end{aligned}$$

which means that there exists an $(\tilde{x},\tilde{y})$-increment pair (s,t) satisfying $\tilde{f}(\tilde{x}) + \tilde{f}(\tilde{y}) \ge \tilde{f}(\tilde{x}+s+t) + \tilde{f}(\tilde{y}-s-t)$.

Suppose that $\tilde{x} = (3;1)$. In this case, we may assume $\tilde{f}(0;0) = f(0;0,0)$ and $\tilde{f}(3;1) = f(3;k+1,-k)$. We define f' as $f'(x_1,x_2,x_3;x_4,x_5) = f(x_1+x_2+x_3;x_4,x_5)$, and $\tilde{f}'$ as

$$\tilde{f}'(x_1,x_2,x_3;\xi) = \inf\{f'(x_1,x_2,x_3;x_4,x_5) \mid x_4+x_5=\xi\}.$$

Then $\tilde{f}'(x_1,x_2,x_3;\xi) = \tilde{f}(x_1+x_2+x_3;\xi)$. Since f' is a splitting of f, it is M-convex by Theorem 7. By Lemma 27, we have

$$\begin{aligned} f(0;0,0) + f(3;k+1,-k) &= f'(0,0,0;0,0) + f'(1,1,1;k+1,-k) \\ &\geq \min\left\{\begin{matrix} \tilde{f}'(1,0,0;1) + \tilde{f}'(0,1,1;0), \\ \tilde{f}'(0,1,0;1) + \tilde{f}'(1,0,1;0), \\ \tilde{f}'(0,0,1;1) + \tilde{f}'(1,1,0;0) \end{matrix}\right\} \\ &= \tilde{f}(1;1) + \tilde{f}(2;0), \end{aligned}$$

which means that there exists an $(\tilde{x},\tilde{y})$-increment pair (s,t) satisfying $\tilde{f}(\tilde{x}) + \tilde{f}(\tilde{y}) \geq \tilde{f}(\tilde{x}+s+t) + \tilde{f}(\tilde{y}-s-t)$. □

7.4. Case of unbounded effective domain. We now deal with the general case of Lemma 10 without assuming boundedness on the effective domain.

Proof of Lemma 10. For $R = 1,2,\ldots$, we define $f^{(R)} : \mathbf{Z}^n \to \mathbf{R} \cup \{+\infty\}$ by

$$f^{(R)}(x) = \begin{cases} f(x) & \text{if } \max_{v\in V} |x(v)| \leq R \\ +\infty & \text{otherwise} \end{cases} \qquad (x \in \mathbf{Z}^n),$$

which is an M-convex function with a bounded effective domain, provided that R is large enough for $\mathrm{dom} f^{(R)} \neq \emptyset$. For each R an elementary aggregation $\tilde{f}^{(R)}$ of $f^{(R)}$ is M-convex by Lemma 19. Take $x, y \in \mathrm{dom}\tilde{f}$. There exists $R_0 = R_0(x,y)$, depending on x and y, such that $x, y \in \mathrm{dom}\tilde{f}^{(R)}$ for every $R \geq R_0$. Since $\tilde{f}^{(R)}$ is M-convex, there exists an (x,y)-increment pair (s_R, t_R) such that

$$\tilde{f}^{(R)}(x) + \tilde{f}^{(R)}(y) \geq \tilde{f}^{(R)}(x+s_R+t_R) + \tilde{f}^{(R)}(y-s_R-t_R).$$

Since the set of all (x,y)-increment pairs is finite, at least one (x,y)-increment pair appears infinitely many times in the sequence $(s_{R_0}, t_{R_0}), (s_{R_0+1}, t_{R_0+1}), \ldots$. More precisely, there exist an (x,y)-increment pair (s,t) and an increasing subsequence $R_1 < R_2 < \cdots$ such that $(s_{R_i}, t_{R_i}) = (s,t)$ for $i = 1,2,\ldots$. By letting $R \to \infty$ along this subsequence in the above inequality, we obtain

$$\tilde{f}(x) + \tilde{f}(y) \geq \tilde{f}(x+s+t) + \tilde{f}(y-s-t).$$

Thus $\tilde{f}$ satisfies (M^{J}-$\mathrm{EXC}_{\mathrm{loc}}$). This completes the proof of Lemma 10. □

REFERENCES

[1] K. Ando, S. Fujishige, and T. Naitoh, *A greedy algorithm for minimizing a separable convex function over a finite jump system*, J. Oper. Res. Soc. Japan, 38 (1995), pp. 362–375.

[2] N. Apollonio and A. Sebő, *Minsquare factors and maxfix covers of graphs*, in Integer Programming and Combinatorial Optimization, Lecture Notes in Comput. Sci. 3064, D. Bienstock and G. Nemhauser, eds., Springer-Verlag, Berlin, 2004, pp. 388–400.

[3] N. Apollonio and A. Sebő, *Minconvex Factors of Prescribed Size in Graphs*, Technical report 145, Les Cahiers du Laboratoire Leibniz, Grenoble, France, 2006.

[4] A. Bouchet, *Greedy algorithm and symmetric matroids*, Math. Programming, 38 (1987), pp. 147–159.

[5] A. Bouchet, *Matchings and Δ-matroids*, Discrete Appl. Math., 24 (1989), pp. 55–62.

[6] A. Bouchet and W. H. Cunningham, *Delta-matroids, jump systems, and bisubmodular polyhedra*, SIAM J. Discrete Math., 8 (1995), pp. 17–32.

[7] R. Chandrasekaran and S. N. Kabadi, *Pseudomatroids*, Discrete Math., 71 (1988), pp. 205–217.

[8] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, and A. Schrijver, *Combinatorial Optimization*, John Wiley and Sons, New York, 1998.

[9] A. W. M. Dress and T. Havel, *Some combinatorial properties of discriminants in metric vector spaces*, Adv. in Math., 62 (1986), pp. 285–312.

[10] A. W. M. Dress and W. Wenzel, *Valuated matroid: A new look at the greedy algorithm*, Appl. Math. Lett., 3 (1990), pp. 33–35.

[11] A. W. M. Dress and W. Wenzel, *A greedy-algorithm characterization of valuated Δ-matroids*, Appl. Math. Lett., 4 (1991), pp. 55–58.

[12] A. W. M. Dress and W. Wenzel, *Valuated matroids*, Adv. Math., 93 (1992), pp. 214–250.

[13] J. Edmonds, *Matroid partition*, in Mathematics of the Decision Sciences, Part I, G. B. Danzig and A. F. Veinott, Jr., eds., AMS, Providence, RI, 1968, pp. 335–345.

[14] S. Fujishige, *Submodular Functions and Optimization*, 2nd ed., Ann. Discrete Math. 58, Elsevier, Amsterdam, 2005.

[15] J. F. Geelen, *private communication*, 1996.

[16] S. N. Kabadi and R. Sridhar, *Δ-matroid and jump system*, J. Appl. Math. Decis. Sci., 9 (2005), pp. 95–106.

[17] Y. Kobayashi and K. Murota, *Induction of M-Convex Functions by Linking Systems*, METR 2006-43, Department of Mathematical Informatics, University of Tokyo, Tokyo, Japan, 2006.

[18] L. Lovász, *The membership problem in jump systems*, J. Combin. Theory Ser. B, 70 (1997), pp. 45–66.

[19] C. J. H. McDiarmid, *Rado's theorem for polymatroids*, Math. Proc. Cambridge Philos. Soc., 78 (1975), pp. 263–281.

[20] K. Murota, *Valuated matroid intersection* I: *Optimality criteria*, SIAM J. Discrete Math., 9 (1996), pp. 545–561.

[21] K. Murota, *Convexity and Steinitz's exchange property*, Adv. Math., 124 (1996), pp. 272–311.

[22] K. Murota, *Matrices and Matroids for Systems Analysis*, Springer-Verlag, Berlin, 2000.

[23] K. Murota, *Discrete Convex Analysis*, SIAM Monogr. Discrete Math. Appl. 10, SIAM, Philadelphia, 2003.

[24] K. Murota, *On infimal convolution of M-convex functions*, RIMS Kokyuroku, 1371 (2004), pp. 20–26.

[25] K. Murota, *M-convex functions on jump systems: A general framework for minsquare graph factor problem*, SIAM J. Discrete Math., 20 (2006), pp. 213–226.

[26] K. Murota and K. Tanaka, *A steepest descent algorithm for M-convex functions on jump systems*, IEICE Trans. Fund. Elec., Commun., Comput. Sci., E89-A (2006), pp. 1160–1165.

[27] R. Rado, *A theorem on independence relations*, Quart. J. Math. Oxford Ser., 13 (1942), pp. 83–89.

[28] A. Schrijver, *Combinatorial Optimization*, Springer-Verlag, Heidelberg, 2003.

[29] A. Shioura, *A constructive proof for the induction of M-convex functions through networks*, Discrete Appl. Math., 82 (1998), pp. 271–278.

[30] W. Wenzel, *Pfaffian forms and Δ-matroids*, Discrete Math., 115 (1993), pp. 253–266.

[31] W. Wenzel, *Δ-matroids with the strong exchange conditions*, Appl. Math. Lett., 6 (1993), pp. 67–70.

SIAM J. Discrete Math.
Vol. 21, No. 1, pp. 130–140

A NATURAL FAMILY OF FLAG MATROIDS[*]

ANNA DE MIER[†]

Abstract. A flag matroid can be viewed as a chain of matroids linked by quotients. Flag matroids, of which relatively few interesting families have previously been known, are a particular class of Coxeter matroids. In this paper we give a family of flag matroids arising from an enumeration problem that is a generalization of the tennis ball problem. These flag matroids can also be defined in terms of lattice paths, and they provide a generalization of the lattice path matroids of [J. Bonin, A. de Mier, and M. Noy, *J. Combin. Theory Ser. A*, 104 (2003), pp. 63–94].

Key words. matroid, flag matroid, lattice path matroid, tennis ball problem

AMS subject classifications. 05B35, 05A15

DOI. 10.1137/050627873

1. Introduction and preliminaries.

Flag matroids are a subclass of Coxeter matroids, but they can also be described in pure matroid-theoretical terms. Roughly speaking, a flag matroid is a collection of matroids on the same ground set that form a chain in the strong order (i.e., they are quotients of each other). Flag matroids play an important role in the theory of Coxeter matroids and also shed light on ordinary matroid theory. Nevertheless, not many classes of flag matroids have been studied up to now. The goal of this paper is to introduce a new family of flag matroids based on an enumeration problem and show how these flag matroids can be interpreted in terms of lattice paths. We refer to [4], especially to Chapter 1, for an introduction to flag matroids and the ideas behind them.

We assume the reader is familiar with the basic concepts of matroid theory; we follow the notation of Oxley's book [8]. We recall here only the notion of quotient. Given two matroids M and N on the same ground set, M is a *quotient* of N if every flat of M is a flat of N (one can also say that M is a strong map image of N). In this case, the rank of M is at most the rank of N, with equality holding if and only if M and N are equal.

We also need to say a few words about lattice path matroids. We do not need lattice path matroids in general as defined in [2], but only the subclass of nested matroids. These matroids have independently arisen several times in the literature since at least 1965, and have been given a variety of names; see [1, 2] and the references therein for definitions and results (in these papers, nested matroids are called "generalized Catalan matroids").

Let P be a lattice path from $(0,0)$ to (m,r) with steps $E=(1,0)$ and $N=(0,1)$. Let $\mathcal{P}$ be the set of paths from $(0,0)$ to (m,r) with steps E and N and that do not go above P. For each path $Q \in \mathcal{P}$, let $Q_N = \{i : \text{step } i \text{ in } Q \text{ is } N\}$. We denote by $[n]$ the set $\{1,2,\ldots,n\}$.

Theorem 1.1. *The set $\{Q_N : Q \in \mathcal{P}\}$ is the collection of bases of a matroid $M[P]$ on the ground set $[m+r]$.*

A matroid is *nested* if it is isomorphic to $M[P]$ for some path P. Hence, the

[*] Received by the editors March 28, 2005; accepted for publication (in revised form) September 5, 2006; published electronically February 15, 2007.
http://www.siam.org/journals/sidma/21-1/62787.html

[†] Mathematical Institute, University of Oxford, 24–29 St Giles, Oxford OX1 3LB, UK (ademier@gmail.com).

bases of a nested matroid are in bijection with the lattice paths that do not go above a certain fixed path P (see the left side of Figure 1 for an example of a nested matroid on the set [15]; the path highlighted corresponds to the basis $\{2, 5, 8, 11, 12, 15\}$). The name *nested* comes from the fact that a nested matroid can also be defined as a transversal matroid whose presentation consists of nested sets, and also because of the following characterization of nested matroids in terms of cyclic flats (recall that a flat is *cyclic* if it is a union of circuits).

THEOREM 1.2. *A matroid is nested if and only if its cyclic flats form a chain under inclusion. Furthermore, the proper nontrivial cyclic flats of the matroid* $M[P]$ *are the initial segments* $[t]$ *of* $[m + r]$*, where* t *is such that step* t *of* P *is* E *and step* $t + 1$ *is* N*.*

Our view on flag matroids is slightly different from that of [4], but it is easy to see that the two perspectives are equivalent. The definition in [4] is in terms of flags of sets, whereas ours relies on what we call ordered partitions of a set. For the reader already familiar with the theory of flag matroids, changing from one definition to the other should be straightforward.

DEFINITION 1.3. *An* ordered k-partition *of a set* S *is a* k*-tuple* $(A_1, \dots, A_k)$ *of nonempty sets with* $A_1 \cup A_2 \cup \cdots \cup A_k = S$ *and* $A_i \cap A_j = \emptyset$ *whenever* $i \neq j$*. For positive integers* $r_1, \dots, r_k$ *such that* $r_1 + \cdots + r_k = |S|$*, an* $(r_1, r_2, \dots, r_k)$-partition *of* S *is an ordered* k*-partition* $(A_1, \dots, A_k)$ *of* S *such that* $|A_i| = r_i$ *for all* i *with* $1 \leq i \leq k$*.*

The bases of a matroid M on a set S trivially determine a collection of ordered 2-partitions of S: take all pairs of the form $(B, S - B)$, where B is a basis of M. The first axiom for flag matroids generalizes this idea; the other two axioms arise from the definition of a flag matroid in terms of Coxeter groups (see [4]). Given an ordered k-partition B, we denote by B_i the ith set in the k-tuple B.

DEFINITION 1.4. *A* flag matroid F *is a pair* $(S, \mathcal{F})$ *such that* $\mathcal{F}$ *is a collection of ordered* k*-partitions of the set* S *satisfying the following properties:*

(F1) *For* $1 \leq i \leq k$*, the set* $\mathcal{B}_i = \{\cup_{1 \leq j \leq i} B_j : B \in \mathcal{F}\}$ *is the set of bases of a matroid* M_i*;*

(F2) *for* $1 \leq i \leq k - 1$*,* M_i *is a quotient of* M_{i+1}*;*

(F3) *if* $(A_1, \dots, A_k)$ *is an ordered* k*-partition of* S *such that, for all* i *with* $1 \leq i \leq k$*, the set* $A_1 \cup \cdots \cup A_i$ *is a basis of the matroid* M_i*, then* $(A_1, \dots, A_k)$ *is in* $\mathcal{F}$*.*

Because of the similarity with matroids, we call the elements of $\mathcal{F}$ the *flag bases* of F. Note that it follows from property (F1) that there exist integers $r_1, \dots, r_k$ adding up to $|S|$ such that all ordered partitions in $\mathcal{F}$ are in fact $(r_1, \dots, r_k)$-partitions. The k-tuple $(r_1, \dots, r_k)$ will be called the *flag rank* of F. The matroids $M_1, \dots, M_k$ above are called the *constitutents* of the flag matroid F. Notice that M_k is the free matroid on S and that M_i has rank $r_1 + \cdots + r_i$.

A trivial example of a flag matroid is the *uniform* flag matroid, having as flag bases all possible $(r_1, \dots, r_k)$-partitions of a set S. Other examples come from chains of subspaces of a vector space, giving rise to *representable* flag matroids. Also, given a matroid M, the *underlying flag matroid* has as constituents the matroids $M_i = T^i(M)$, the truncations of M to ranks 1 to $r(M)$. A flag matroid with flag rank $(1, 1, \dots, 1)$ can also be viewed as a Gaussian greedoid [5].

Flag matroids, as is true of Coxeter matroids in general, are usually viewed in terms of their polytopes. For instance, the polytope of the uniform flag matroid of flag rank $(1, 1, \dots, 1)$ is the permutahedron; the polytope of the underlying flag matroid is studied in [3].

2. The tennis ball problem. The tennis ball problem is a problem in enumeration that can be phrased in terms of balls-and-bins and in terms of lattice paths. We need both approaches here. We first define the original problem and show that its solution amounts to counting bases of a certain type of nested matroid. Then we generalize the problem and show that it gives rise to a family of flag matroids.

DEFINITION 2.1. *Let l_1 and l_2 be positive integers. Suppose we have infinitely many balls numbered $1, 2, \dots$ and two bins labeled A and B. In the first turn, balls $1, 2, \dots, l_1 + l_2$ go into bin A, and then l_2 of those are moved to bin B. In the second turn, balls $l_1 + l_2 + 1, \dots, 2(l_1 + l_2)$ go into bin A, and of the $2l_1 + l_2$ balls there, l_2 are moved to bin B. At each turn, the next $l_1 + l_2$ balls go into bin A, and of the balls in A, l_2 are moved to bin B. An n-*configuration *is an ordered 2-partition of $[(l_1 + l_2)n]$ giving a possible distribution of balls in the bins after n turns. The (l_1, l_2)–*tennis ball problem *asks for the number of n-configurations.*

This problem was solved in [7] using the following relationship with nested matroids.

THEOREM 2.2. *The number of n-configurations of the (l_1, l_2)–tennis ball problem is the number of bases of the nested matroid $M[(N^{l_1}E^{l_2})^n]$.*

The proof is straightforward by the bijection that sends a basis $\{n_1, \dots, n_r\}$ of $M[(N^{l_1}E^{l_2})^n]$ to the configuration having the balls $\{n_1, \dots, n_r\}$ in bin A (see Figure 1). For nonnegative integers a, b, in what follows we call the matroid $M[(N^aE^b)^n]$ the *nth (a, b)-tbp matroid.*

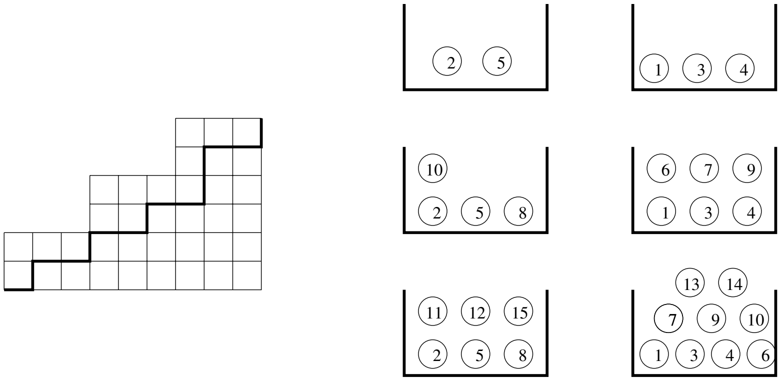


FIG. 1. *A diagram representing the matroid $M[(N^2E^3)^3]$. The path highlighted corresponds to the 3-configuration shown on the right.*

Theorem 2.2 can be rephrased by saying that the tennis ball problem with two bins gives the bases of a matroid. The main result of this section is that the tennis ball problem with k bins, which we next define, gives the flag bases of a flag matroid.

DEFINITION 2.3. *Let $(l_1, l_2, \dots, l_k)$ be a k-tuple of positive integers; let $L = l_1 + l_2 + \cdots + l_k$. Suppose we have infinitely many balls numbered $1, 2, \dots$ and k bins labeled $\Gamma_1, \Gamma_2, \dots, \Gamma_k$. In the first turn, balls $1, 2, \dots, L$ go into bin Γ_1; of those, $L - l_1$ are moved to bin Γ_2; of those, $L - l_1 - l_2$ are moved to bin Γ_3, and so on until l_k balls are moved to bin Γ_k. In the second turn, balls $L + 1, \dots, 2L$ go into bin Γ_1, and of the $l_1 + L$ balls there, $L - l_1$ are moved to bin Γ_2; of the balls now in bin Γ_2, $L - l_1 - l_2$ are moved to bin Γ_3, and so on. At each turn, the next L balls go into bin Γ_1, and of the balls in Γ_1, $L - l_1$ are moved to bin Γ_2, etc. An n-*configuration *is an ordered k-partition of $[Ln]$ corresponding to a possible distribution of the balls* *in the bins after n turns. The* $(l_1, \dots, l_k)$–tennis ball problem *asks for the number of n-configurations, that is, the number of $(nl_1, nl_2, \dots, nl_k)$-partitions of the set $[nL]$ that we can obtain after n turns.*

Note that we are interested only in n-configurations, not in the movements of the balls that lead to them; an n-configuration can typically be obtained by several different movements of the balls. Let $F_n^{(l_1,\dots,l_k)}$ be the collection of $(nl_1, \dots, nl_k)$-partitions of $[nL]$ that we get as n-configurations. These ordered partitions are the flag bases of a flag matroid whose constituent matroids are nested matroids.

Theorem 2.4. *The set $F_n^{(l_1,\dots,l_k)}$ is the collection of flag bases of a flag matroid on the set $[nL]$. Moreover, for i with $1 \leq i \leq k$, the ith constituent of the flag matroid is the nth $(l_1 + \cdots + l_i,\, l_{i+1} + \cdots + l_k)$-tbp matroid.*

Proof. We need to check that axioms (F1)–(F3) hold for $\mathcal{F} = F_n^{(l_1,\dots,l_k)}$. It is easy to see that the set $\mathcal{B}_i = \{\cup_{1 \leq j \leq i} B_j : B \in \mathcal{F}\}$ is the set of bases of the nth $(l_1+\cdots+l_i,$ $l_{i+1} + \cdots + l_k)$-tbp matroid, so (F1) holds.

To show that (F2) holds it is enough to prove that if $a + b = a' + b'$ and $a < a'$, then the nth (a, b)-tbp matroid M is a quotient of the nth (a', b')-tbp matroid M'. We show that each flat F of M is a flat of M'. If F is a cyclic flat, this follows from the characterization of cyclic flats of nested matroids in Theorem 1.2. Otherwise, F is $F' \cup I$, where F' is a cyclic flat of M and I is the set of isthmuses of F. So F' is an initial segment of $[(a + b)n]$ whose length is a multiple of $a + b$. Since $a + b = a' + b'$, by Theorem 1.2 again we have that F' is a cyclic flat of M'. For $F' \cup I$ to be a flat of M, the set I has to be such that $|I \cap [t(a + b)]| < a$ for all t. Since $a + b = a' + b'$ and $a < a'$, we also have that $|I \cap [t(a' + b')]| < a'$ for all t; hence $F' \cup I$ is a flat of M'.

To show that (F3) holds, let $(A_1, \dots, A_k)$ be an $(nl_1, \dots, nl_k)$-partition of $[nL]$ such that for all i, $\cup_{1 \leq j \leq i} A_j$ is a basis of the nth $(l_1 + \cdots + l_i,\, l_{i+1} + \cdots + l_k)$-tbp matroid. We show that $(A_1, \dots, A_k)$ is in the collection $\mathcal{F}$ by showing how to get the n-configuration $(A_1, \dots, A_k)$ by suitably moving the balls. Let C_i be $\cup_{j=1}^{i} A_j$. To avoid wordiness, we identify balls with the integers of their labels. We start with all bins empty and explain how to perform n turns with the condition that at the end of each turn, the set of balls in bin Γ_i is a subset of $[nL] - C_{i-1}$ for all i with $2 \leq i \leq k$. Suppose we have performed $t - 1$ such turns, for t with $0 < t \leq n$, and let us describe turn t. The assumption that $C_{k-1}, C_{k-2}, \dots, C_1$ are bases of their respective tbp matroids gives the following facts:

(1) At least tl_k of the elements of $[tL]$ are in $[tL] - C_{k-1}$;

(2) at least $t(l_k + l_{k-1})$ of the elements of $[tL]$ are in $[tL] - C_{k-2}$;

$\vdots$

(a) at least $t(l_k + \cdots + l_3)$ of the elements of $[tL]$ are in $[tL] - C_2$;

(b) at least $t(l_k + \cdots + l_2)$ of the elements of $[tL]$ are in $[tL] - C_1$.

Since after the first $t - 1$ turns there are $(t - 1)l_j$ elements in bin Γ_j, we can deduce from facts (b)–(1) that at this point, for all i with $1 \leq i \leq k-1$, at least $l_{i+1} + \cdots + l_k$ integers from $[tL]$ are in $[tL] - C_i$ but not in $\Gamma_{i+1} \cup \cdots \cup \Gamma_k$.

Now move $(t-1)n+1, (t-1)n+2, \dots, tn$ to bin Γ_1. From all the integers in Γ_1, choose $L - l_1$ to be moved to bin Γ_2, starting with as many integers as possible from among those in $[tL] - C_{k-1}$ that are still in Γ_1, then take as many integers as possible from among those in $[tL] - C_{k-2}$, and so on until $L - l_1$ integers are obtained. The remarks in the previous paragraph show that it is possible to choose integers in this way. We move them to bin Γ_2. Since $[tL] - C_{k-1} \subset [tL] - C_{k-2} \subset \cdots \subset [tL] - C_1$, the integers now in Γ_2 are a subset of $[tL] - C_1$, as required. Moreover, by the way the balls are chosen, we have that among the balls that are at this point in Γ_2, at least $l_3 + \cdots + l_k$ are in $[tL] - C_2$.

We describe generally how to move $l_i + \cdots + l_k$ balls to bin Γ_i from bin Γ_{i-1} in a way such that the balls in Γ_i are a subset of $[tL] - C_{i-1}$, and, moreover, at least $l_{i+1} + \cdots + l_k$ of them are in $[tL] - C_i$. From the balls in bin Γ_{i-1}, pick $l_i + \cdots + l_k$ starting with as many as possible from $[tL] - C_{k-1}$; if there are not still $l_i + \cdots + l_k$, then take as many as possible from $[tL] - C_{k-2}$, and so on, until taking as many as possible from $[tL] - C_{i-1}$. By the same reason as above, such integers exist; the integers now in Γ_i are a subset of $[tL] - C_{i-1}$ and at least $l_{i+1} + \cdots + l_k$ of them are in $[tL] - C_i$.

At the end of n turns, we have nl_i balls in bin Γ_i, and these are a subset of $[nL] - C_{i-1}$ for all i. This implies that bin Γ_k contains exactly the balls in A_k, and hence bin Γ_{k-1} contains the balls in A_{k-1}, and so on. Therefore the ordered partition $(A_1, \ldots, A_k)$ is an n-configuration, thus it belongs to the collection $F_n^{(l_1,\ldots,l_k)}$, and (F3) follows. □

3. Interpretation in terms of lattice paths. The tennis ball problem with two bins has a simple interpretation in terms of lattice paths: we associate bin A with steps N and bin B with steps E, and then each n-configuration corresponds to a path that does not go above $(N^{l_1}E^{l_2})^n$. For the tennis ball problem with k bins, we can associate with each bin a direction in $\mathbb{N}^k$. Then the flag bases of $F_n^{(l_1,\ldots,l_k)}$ are in bijection with certain paths in $\mathbb{N}^k$. We characterize those paths combinatorially, and for $k = 3$ we describe them as the set of lattice paths that do not cross a certain border.

Let $e_1, \ldots, e_k$ be the unit coordinate vectors in $\mathbb{R}^k$. With each n-configuration of the $(l_1, \ldots, l_k)$–tennis ball problem, we associate a path $s_1 s_2 \cdots s_{nL}$ from $(0, \ldots, 0)$ to $(nl_1, \ldots, nl_k)$ with steps defined as $s_i = e_j$ if ball i is in bin Γ_j. We call this path an *n-configuration path.* Hence, an n-configuration path can be seen as a sequence of elements from $\{e_1, \ldots, e_k\}$. It is easy to characterize which such sequences give configuration paths.

LEMMA 3.1. *A path from $(0, \ldots, 0)$ to $(nl_1, \ldots, nl_k)$ is an n-configuration path for the $(l_1, \ldots, l_k)$–tennis ball problem if and only if, for all t with $1 \le t \le n$ and all i with $1 \le i \le k-1$, among the first tL steps there are at most $t(l_1 + \cdots + l_i)$ whose type belongs to $\{e_1, \ldots, e_i\}$.*

Proof. After t turns, for $1 \le t \le n$, there are exactly $t(l_1 + \cdots + l_i)$ balls of $[tL]$ in the first i bins. Since balls can move only to bins with a higher index, at the end of n turns there are at most $t(l_1 + \cdots + l_i)$ balls of $[tL]$ in $\Gamma_1 \cup \cdots \cup \Gamma_i$. Hence, among the first tL steps of a configuration path there are at most $t(l_1 + \cdots + l_i)$ steps whose type is in $\{e_1, \ldots, e_i\}$.

For the converse, assume we have a path π that satisfies the condition. Let A_i be the set of integers s such that step s in π is of type e_i. Consider the $(nl_1, \ldots, nl_k)$-partition $(A_1, \ldots, A_k)$ of $[nL]$ obtained in this way. The condition on the path implies that the set $\cup_{j=1}^{i} A_j$ is a basis of the $(l_1 + \cdots + l_i,\ l_{i+1} + \cdots + l_k)$-tbp matroid, for all i with $1 \le i \le k-1$. Therefore $(A_1, \ldots, A_k)$ is a flag basis of the flag matroid $F_n^{(l_1,\ldots,l_k)}$, and hence π is an n-configuration path, as required. □

The following is an immediate corollary.

COROLLARY 3.2. *Given an n-configuration path π, the path obtained by switching a pair of steps $s_i = e_l$ and $s_j = e_m$ is a configuration path if $i < j$ and $l \le m$. Moreover, let π' be an initial segment of π with t'_j steps of type e_j for all $1 \le j \le k$.*

Let n' be the minimum integer such that $t'_j \leq n'l_j$ for all j. For $n'' \geq 0$, consider the path obtained from π' followed by $(n'+n'')l_k - t'_k$ steps e_k, then $(n'+n'')l_{k-1} - t'_{k-1}$ steps e_{k-1}, and so on, until finishing with $(n'+n'')l_1 - t'_1$ steps e_1. Then this path is an $(n'+n'')$-configuration path.

The *n-diagram* for the $(l_1, \dots, l_k)$–tennis ball problem is the set of points in $\mathbb{N}^k$ that are contained in some n-configuration path. Our goal is to study what these diagrams look like for $k = 3$. The following is another corollary of Lemma 3.1.

COROLLARY 3.3. *If (x, y, z) is in the n-diagram, then (x', y, z') is in the n-diagram for all $z \leq z' \leq nl_3$ and all $x' \leq x$.*

A direct consequence of this corollary is that to describe the n-diagram it is enough to give, for each (x, y), the minimum value of z such that (x, y, z) is a point of the n-diagram; this minimum is denoted $m_n(x, y)$ (trivially the maximum value of z is nl_3). If no such z exists, because no point of the form $(x, y, *)$ is in the n-diagram, we set $m_n(x, y) = *$. So the n-diagram is described by an $(nl_1+1) \times (nl_2+1)$ matrix $\mathcal{M}_n$ with entries in the set $\{0, 1, \dots, nl_3\} \cup \{*\}$ and such that in row x and column y we have $m_n(x-1, y-1)$. If $n = 1$, then trivially $\mathcal{M}_1$ is the zero matrix; the corresponding 1-diagram is represented in Figure 2. In all figures below, the direction of the third coordinate has been reversed for a better view of the picture, and, as pointed out above, all points under a point that is shown are in the diagram as well.

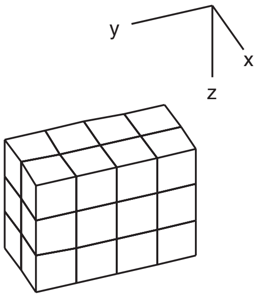


FIG. 2. *The* 1*-diagram for the* $(2, 4, 3)$*–tennis ball problem.*

The 2-diagram for the $(2, 4, 3)$–tennis ball problem is shown in Figure 3. We first give the matrices $\mathcal{M}_n$ and then prove that they give the right diagrams.

The matrix $\mathcal{M}_2$ is made up of four blocks,

$$\mathcal{M}_2 = \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right),$$

where A is the $(l_1+1) \times (l_2+1)$ matrix, all of whose entries are zero (hence, it is $\mathcal{M}_1$); D is an $l_1 \times l_2$ matrix, all of whose entries are l_3; B is an $(l_1+1) \times l_2$ matrix with

$$b_{i,j} = \begin{cases} 0 & \text{if } i \leq l_1 + 1 - j, \\ l_3 & \text{otherwise;} \end{cases}$$

and C is the $l_1 \times (l_2+1)$ matrix with $\min\{l_2, l_3\}+1$ non-$*$ columns, with the elements in the last column being l_3 and every other non-$*$ column being obtained by adding $+1$ to the next, that is,

$$\begin{pmatrix} * & \cdots & * & 2l_3 & 2l_3-1 & 2l_3-2 & \cdots & l_3+1 & l_3 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ * & \cdots & * & 2l_3 & 2l_3-1 & 2l_3-2 & \cdots & l_3+1 & l_3 \end{pmatrix}.$$

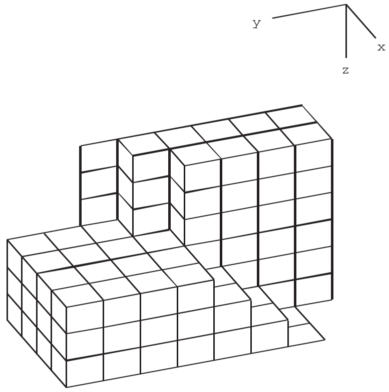


FIG. 3. *The* 2*-diagram for the* $(2,4,3)$*–tennis ball problem.*

We now define recursively the $(nl_1+1)\times(nl_2+1)$ matrix $\mathcal{M}_n$ that gives the n-diagram. The matrix $\mathcal{M}_n$ also decomposes into four blocks,

$$\mathcal{M}_n = \left(\begin{array}{c|c} A_n & B_n \\ \hline C_n & D_n \end{array}\right),$$

where A_n has $(n-1)l_1+1$ rows and $(n-1)l_2+1$ columns. The matrix A_n is $\mathcal{M}_{n-1}$. The entry in row i and column j of the matrix B_n, for $1 \le i \le (n-1)l_1+1$ and $1 \le j \le l_2$, is given by $(n-s)l_3$, where s is the only integer for which $(n-s)(l_1+l_2) < i-1+j+(n-1)l_2 \le (n-s+1)(l_1+l_2)$. Roughly speaking, B_n consists of diagonal stripes of width l_1+l_2; see the examples below. The matrix C_n has l_1 rows and $(n-1)l_2+1$ columns; all entries in the last column are $(n-1)l_3$ and each column is obtained by adding one to the next, until we reach nl_3; hence the matrix is given by

$$\begin{pmatrix} * & \cdots & * & nl_3 & nl_3-1 & nl_3-2 & \cdots & (n-1)l_3+1 & (n-1)l_3 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ * & \cdots & * & nl_3 & nl_3-1 & nl_3-2 & \cdots & (n-1)l_3+1 & (n-1)l_3 \end{pmatrix}.$$

Finally, D_n is the $l_1 \times l_2$ matrix all of whose entries are $(n-1)l_3$. The matrix $\mathcal{M}_3$ is shown below for the $(2,4,3)$– and $(3,2,2)$–tennis ball problems, and the corresponding 3-diagrams are shown in Figures 4 and 5.

$$\left(\begin{array}{ccccc|cccc|cccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 3 & 3 & 3 & 3 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 & 3 & 3 & 3 & 3 & 3 & 6 \\ 0 & 0 & 0 & 0 & 0 & 3 & 3 & 3 & 3 & 3 & 3 & 6 & 6 \\ \cline{1-9} * & 6 & 5 & 4 & 3 & 3 & 3 & 3 & 3 & 3 & 6 & 6 & 6 \\ * & 6 & 5 & 4 & 3 & 3 & 3 & 3 & 3 & 6 & 6 & 6 & 6 \\ \hline * & * & * & * & \multicolumn{1}{c}{*} & 9 & 8 & 7 & 6 & 6 & 6 & 6 & 6 \\ * & * & * & * & \multicolumn{1}{c}{*} & 9 & 8 & 7 & 6 & 6 & 6 & 6 & 6 \end{array}\right) \qquad \left(\begin{array}{ccc|cc|cc} 0 & 0 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 2 & 2 & 2 \\ 0 & 0 & 0 & 2 & 2 & 2 & 2 \\ \cline{1-5} 4 & 3 & 2 & 2 & 2 & 2 & 2 \\ 4 & 3 & 2 & 2 & 2 & 2 & 4 \\ 4 & 3 & 2 & 2 & 2 & 4 & 4 \\ \hline * & * & \multicolumn{1}{c}{6} & 5 & 4 & 4 & 4 \\ * & * & \multicolumn{1}{c}{6} & 5 & 4 & 4 & 4 \\ * & * & \multicolumn{1}{c}{6} & 5 & 4 & 4 & 4 \end{array}\right)$$

We now show that the matrix $\mathcal{M}_n$ gives the n-diagram and, moreover, that all paths contained in the n-diagram are n-configuration paths.

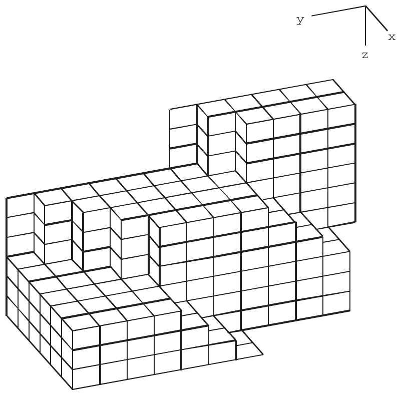


Fig. 4. *The* 3*-diagram for the* $(2, 4, 3)$*–tennis ball problem.*

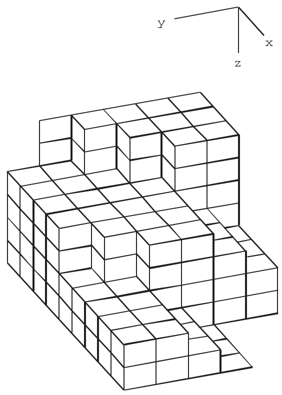


Fig. 5. *The* 3*-diagram for the* $(3, 2, 2)$*–tennis ball problem.*

Theorem 3.4. *The* n*-diagram for the* (l_1, l_2, l_3)*–tennis ball problem is given by the matrix* $\mathcal{M}_n$*. Furthermore, the* n*-configuration paths are exactly those contained in the* n*-diagram.*

Proof. The proof is by induction on n. As seen above, the case $n = 1$ is trivial. Assume $\mathcal{M}_{n-1}$ is the matrix of the $(n-1)$-diagram for the (l_1, l_2, l_3)–tennis ball problem. Let $\mathcal{N}$ be the matrix of the n-diagram; we prove that $\mathcal{N} = \mathcal{M}_n$. The matrix $\mathcal{N}$ has dimensions $(nl_1 + 1) \times (nl_2 + 1)$.

Recall that the entry in row $x + 1$ and column $y + 1$ of $\mathcal{N}$ is $m_n(x, y)$. Given x and y with $0 \le x \le nl_1$ and $0 \le y \le nl_2$, let z be $m_n(x, y)$. We show that z is the entry in row $x + 1$ and column $y + 1$ of the matrix $\mathcal{M}_n$. The proof has three cases.

Case 1. $x \le (n-1)l_1$ and $y \le (n-1)l_2$.

In this case we need to show that $z = m_{n-1}(x, y)$. Let π be an $(n-1)$-configuration path that contains the point $(x, y, m_{n-1}(x, y))$. By Corollary 3.2, π can be extended to an n-configuration path; hence $z \le m_{n-1}(x, y)$. To show equality, assume $z < m_{n-1}(x, y) \le (n-1)l_3$ and let ρ be an n-configuration path that contains the point (x, y, z). Let ρ' be the initial segment corresponding to the first $x + y + z$ steps. By Corollary 3.2 again, ρ' can be extended to an $(n-1)$-configuration path contradicting the induction hypotheses.

Case 2. $x > (n-1)l_1$.

We want to show in this case that z is given by the entries of the matrices C_n and D_n, depending on the value of y. Observe first that Lemma 3.1 implies that $x+y+z > (n-1)L$, and hence also that $y+z \geq (n-1)(l_2+l_3)$.

We discuss now the subcase $y \geq (n-1)l_2$, showing that $z = (n-1)l_3$. Let π be any $(n-1)$-configuration path. Extend π by adding $y-(n-1)l_2$ steps e_2 followed by $x-(n-1)l_1$ steps e_1 and then l_3 steps e_3, and add the remaining steps in any way. This path clearly satisfies the condition on Lemma 3.1; hence $z \leq (n-1)l_3$. To complete the proof of the claim, suppose that $z < (n-1)l_3$. Then by Lemma 3.1, $x+y+z < (n-1)L$. This contradicts the first conclusion of the previous paragraph.

The other subcase left is $y < (n-1)l_2$. We show that in this case $z = (n-1)(l_2+l_3)-y$. Since we already know that z is at least $(n-1)(l_2+l_3)-y$ it is enough to show that there is an n-configuration path containing the point $(x, y, (n-1)(l_2+l_3)-y)$. Consider the path

$$\sigma = (e_3)^{(n-1)(l_2+l_3)-y}(e_2)^y(e_1)^x(e_3)^{nl_3-(n-1)(l_2+l_3)+y}(e_2)^{nl_2-y}(e_1)^{nl_1-x}.$$

It is easy to check that this path satisfies the condition of Lemma 3.1, and hence it is an n-configuration path containing the point $(x, y, (n-1)(l_2+l_3)-y)$, as required.

Case 3. $x < (n-1)l_1$ and $y > (n-1)l_2$.

In this case the value of z has to be the one given by the matrix B_n. We have to show that if s is such that $(n-s)(l_1+l_2) \leq x+y < (n-s+1)(l_1+l_2)$, then $z = (n-s)l_3$.

Since from Case 2 we have that the point $(nl_1, y, (n-1)l_3)$ is in the n-diagram, by Corollary 3.3 it follows that the point $(x, y, (n-1)l_3)$ is in the n-diagram as well; hence $z \leq (n-1)l_3$. If $x+y > (n-1)(l_1+l_2)$, then by Lemma 3.1 $x+y+z > (n-1)L$, and hence $z \geq (n-1)l_3$, so in this case $z = (n-1)l_3$.

Now suppose $(n-s)(l_1+l_2) < x+y \leq (n-s+1)(l_1+l_2)$. As in the previous paragraph, Lemma 3.1 implies that $z \geq (n-s)l_3$. To show that we have equality, we give an n-configuration path containing the point $(x, y, (n-s)l_3)$. Consider the path

$$\sigma = (e_3)^{(n-s)l_3}(e_2)^y(e_1)^x(e_3)^{sl_3}(e_2)^{nl_2-y}(e_1)^{nl_1-x}.$$

By using that $y > (n-1)l_2$ and that $x+y \leq (n-s+1)(l_1+l_2)$ it is easy to show that σ satisfies the condition of Lemma 3.1, and hence it is an n-configuration path.

To finish the proof we have to show that any path contained in the n-diagram is an n-configuration path. Let π be such a path; we check that π satisfies the condition in Lemma 3.1. Let (X, Y, Z) be a point in the path with $X+Y+Z = tL$ for some t with $1 \leq t \leq n-1$; our goal is to show that $X \leq tl_1$ and $X+Y \leq t(l_1+l_2)$. Consider the point $p = (X, Y, m_n(X, Y))$. The proof above shows that there is an n-configuration path π' that goes through p; since $Z \geq m_n(X, Y)$, we can apply Corollary 3.2 to obtain from π' an n-configuration path containing the point (X, Y, Z). Since all n-configuration paths satisfy the condition in Lemma 3.1, we have that $X \leq tl_1$ and $X+Y \leq t(l_1+l_2)$. □

4. Concluding remarks. The results of the previous section show that some sets of lattice paths in three dimensions can be interpreted in terms of flag matroids; hence, the flag matroids one obtains from the tennis ball problem naturally generalize lattice path matroids. This might lead to the suspicion that any set of paths in $\mathbb{N}^k$ with a "reasonable" border also gives rise to flag matroids. Unfortunately, it is very easy to produce counterexamples to this. For instance, consider the diagram in Figure 6. If the paths contained in that diagram were in correspondence with the

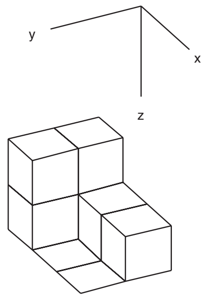


FIG. 6. *The set of paths contained in the diagram does not give rise to a flag matroid.*

flag bases of a flag matroid F, we would have that $B_1 = \{2, 6\}$ and $B_2 = \{4, 5\}$ are cobases of the second constituent of F. Hence, it should be possible to replace 2 in B_1 with either 4 or 5. But no path contained in the diagram has $\{4, 6\}$ or $\{5, 6\}$ as its set of steps in the direction e_3.

A question that remains open is to solve the $(l_1, \ldots, l_k)$–tennis ball problem, or even the $(1, 1, 1)$–tennis ball problem. The approaches used previously to solve the case $k = 2$ do not seem to generalize easily. In particular, the strategy from [7] would suggest the use of a Tutte polynomial-like invariant to count flag bases. There are some generalizations of the Tutte polynomial to pairs of matroids and to chains of matroids related by strong maps [6, 11], but unfortunately they do not seem to include the number of flag bases as a specialization. Following the Tutte polynomial approach for flag matroids would require defining first the suitable generalization.

We finish with some easy bounds. A trivial upper bound for the number of n-configurations is given by the total number of $(nl_1, \ldots, nl_k)$-partitions of $[nL]$, which is the multinomial coefficient

$$\binom{nL}{nl_1, \ldots, nl_k} = \frac{(nL)!}{(nl_1)! \cdots (nl_k)!}.$$

The following connection with Young tableaux gives a lower bound on the number of n-configurations of the $(l_1, \ldots, l_k)$–tennis ball problem when $l_k \geq l_{k-1} \geq \cdots \geq l_1$. Consider sequences of length $n(l_1 + \cdots + l_k)$ over the alphabet $\{e_1, \ldots, e_k\}$ containing nl_i copies of e_i and such that in any initial subsequence the number of symbols e_i is greater than or equal to the number of symbols e_{i-1} for all i with $2 \leq i \leq k$. Since all these sequences trivially satisfy the condition in Lemma 3.1, they give n-configuration paths. The number of such sequences equals the number of standard Young tableaux of shape $(nl_k, \ldots, nl_1)$, and this is given by the hook-length formula (see [9, Chapter 7]). In the case $l_1 = \cdots = l_k = l$, this is

$$\frac{(nlk)!}{(nl)!^k \prod_{i=1}^{k-1} (\frac{nl}{i} + 1)^{k-i}}.$$

The general case gets more involved and we omit it since not much insight is gained from it.

For the $(1, 1, 1)$–tennis ball problem, the first order approximation of the lower and upper bounds are $C27^n n^{-7/2}$ and $C' 27^n n^{-1/2}$, respectively, for some constants C and C'. Computational evidence seems to suggest that the right number lies closer to the lower bound, and that the exponent in the term on n is -3 [10].

A general lower bound for the $(l_1,\dots,l_k)$–tennis ball problem can be obtained as follows. Let $t(a,b,n)$ be the number of n-configurations of the (a,b)–tennis ball problem. Then the number of n-configurations of the $(l_1,\dots,l_k)$–tennis ball problem is at least

$$t(l_1, l_2+\cdots+l_k, n)t(l_2, l_3+\cdots+l_k, n)\cdots t(l_{k-1}, l_k, n),$$

since we can think of the $(l_1,\dots,l_k)$–tennis ball problem as n turns of the $(l_1, l_2+\cdots+l_k)$–tennis ball problem, followed by n turns of the $(l_2, l_3+\cdots+l_k)$–tennis ball problem on the result of the first, and so on. The bound is strict since each $t(a,b,n)$ counts the number of n-configurations of the (a,b)–tennis ball problem, but each of these can usually be reached by several movements of the balls, and that is relevant for the version with k bins.

Acknowledgments. The author thanks Dominic Welsh for asking a question that suggested this work, and Joseph Bonin and Sergi Elizalde for helpful discussions.

REFERENCES

[1] J. Bonin and A. de Mier, *Lattice path matroids: Structural properties*, European J. Combin., 27 (2006), pp. 701–738.

[2] J. Bonin, A. de Mier, and M. Noy, *Lattice path matroids: Enumerative aspects and Tutte polynomials*, J. Combin. Theory Ser. A, 104 (2003), pp. 63–94.

[3] A. V. Borovik, I. M. Gelfand, A. Vince, and N. White, *The lattice of flats and its underlying flag matroid polytope*, Ann. Comb., 1 (1997), pp. 17–26.

[4] A. V. Borovik, I. M. Gelfand, and N. White, *Coxeter Matroids*, Progr. Math. 216, Birkhäuser Boston, Boston, MA, 2003.

[5] V. Bryant and I. Sharpe, *Gaussian, strong and transversal greedoids*, European J. Combin., 20 (1999), pp. 259–262.

[6] M. Las Vergnas, *On the Tutte polynomial of a morphism of matroids*, Ann. Discrete Math., 8 (1980), pp. 7–20.

[7] A. de Mier and M. Noy, *A solution to the tennis ball problem*, Theoret. Comput. Sci., 346 (2005), pp. 254–264.

[8] J. G. Oxley, *Matroid Theory*, Oxford University Press, Oxford, UK, 1992.

[9] R. P. Stanley, *Enumerative Combinatorics*, Vol. 2, Cambridge University Press, Cambridge, UK, 1999.

[10] D. van der Zypen, *private communication*, 2005.

[11] D. Welsh and K. Kayibi, *A linking polynomial of two matroids*, Adv. in Appl. Math., 32 (2004), pp. 391–419.

SIAM J. Discrete Math.
Vol. 21, No. 1, pp. 141–157

ALGORITHMS FOR FAULT-TOLERANT ROUTING IN CIRCUIT-SWITCHED NETWORKS[*]

AMITABHA BAGCHI[†], AMITABH CHAUDHARY[‡], CHRISTIAN SCHEIDELER[§], AND PETR KOLMAN[¶]

Abstract. In this paper we consider the *k edge-disjoint paths problem* (k-EDP), a generalization of the well-known edge-disjoint paths problem. Given a graph $G = (V, E)$ and a set of terminal pairs (or requests) T, the problem is to find a maximum subset of the pairs in T for which it is possible to select paths such that each pair is connected by k edge-disjoint paths and the paths for different pairs are mutually disjoint. To the best of our knowledge, no nontrivial result is known for this problem for $k > 1$. To measure the performance of our algorithms we use the recently introduced flow number F of a graph. This parameter is known to fulfill $F = O(\Delta\alpha^{-1} \log n)$, where Δ is the maximum degree, α is the edge expansion of G, and n is the number of vertices in G. We show that a simple greedy online algorithm achieves a competitive ratio of $O(k^3F)$ which naturally extends the best known bound of $O(F)$ for $k = 1$ to higher k. To achieve this competitive ratio, we introduce a new method of converting a system of k disjoint paths into a system of k length-bounded disjoint paths. We also show that any deterministic online algorithm has a competitive ratio of $\Omega(kF)$. In addition, we study the *k disjoint flows problem* (k-DFP), which is a generalization of the previously studied unsplittable flow problem. The difference between the k-DFP and the k-EDP is that now we consider a graph with edge capacities and our requests are allowed to have arbitrary demands d_i. The aim is to find a subset of requests of maximum total demand for which it is possible to select flow paths such that all the capacity constraints are maintained and each selected request with demand d_i is connected by k disjoint paths, each of flow value d_i/k. The k-EDP and k-DFP problems have important applications in fault-tolerant (virtual) circuit switching, which plays a key role in optical networks.

Key words. edge-disjoint paths, fault-tolerant routing, flow number, greedy algorithms, multicommodity flow

AMS subject classifications. 68W40, 68R10

DOI. 10.1137/S0895480102419743

1. Introduction.

This paper was motivated by a talk given by Rakesh Sinha from Ciena Inc. The speaker pointed out in his talk that standard problems such as the edge-disjoint paths problem (EDP) and the unsplittable flow problem (UFP) are insufficient for practical purposes: They do not allow a rapid adaptation to edge faults or heavy load conditions. Instead of having just one path for each request, it would be much more desirable to determine a collection of alternative independent paths for each accepted request that can flexibly be used to ensure rapid adaptability. The paths, however, should be chosen so that not too much bandwidth is wasted under normal conditions. Keeping this in mind, we introduce two optimization problems

[*] Received by the editors December 12, 2002; accepted for publication (in revised form) November 15, 2006; published electronically February 15, 2007. A preliminary version of this work appeared in the Proceedings of the ACM Symposium on Parallel Algorithms and Architectures, 2002.

http://www.siam.org/journals/sidma/21-1/41974.html

[†] Department of Computer Science and Engineering, Indian Institute of Technology Delhi, India (bagchi@cse.iitd.ernet.in).

[‡] Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556 (chaudhary.1@nd.edu).

[§] Department of Informatics, Technische Universität München, Germany (scheideler@in.tum.de).

[¶] Department of Applied Mathematics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic (kolman@kam.mff.cuni.cz). The work of this author was supported in part by the European Commission within the 6th Framework Programme under contract 001907 “Dynamically Evolving, Large Scale Information Systems” (DELIS) and by project 1M0021620808 (ITI) of the Ministry of Education of the Czech Republic.

which have not been studied before, to the best of our knowledge: the k edge-disjoint paths problem (k-EDP) and the k disjoint flows problem (k-DFP).

In the k-EDP we are given an undirected graph $G = (V, E)$ and a set of terminal pairs (or requests) T. The problem is to find a maximum subset of the pairs in T such that each chosen pair can be connected by k disjoint paths and, moreover, the paths for different pairs are mutually disjoint.

Similarly, in the k-DFP we are given an undirected network $G = (V, E)$ with edge capacities and a set of terminal pairs T with demands d_i, $1 \leq i \leq |T|$. The problem is to find a subset of the pairs of maximum total demand such that each chosen pair can be connected by k disjoint paths, each path is carrying d_i/k units of flow and no capacity constraint is violated.

In order to demonstrate that the k-DFP can be used to achieve fault tolerance together with a high utilization of the network resources and rapid adaptability, consider a network G in which new edge faults may occur continuously, but the total number of faulty edges at the same time is at most f. In this case, given a request with demand d, the strategy is to reserve $k + f$ disjoint flow paths for it for some $k \geq 1$ with total demand $(1 + f/k)d$. As long as at most f edge faults appear at the same time, it will still be possible to ship a demand of d along the remaining paths. Furthermore, under fault-free conditions, only a fraction f/k of the reserved bandwidth is wasted, which can be made as small as required by setting k sufficiently large, within the constraints placed by the properties of the network.

Regarding the connectivity properties of the networks we consider, note that we do not require the network to be k-edge connected.

1.1. Previous results.

Since we are not aware of previous results for the k-EDP and the k-DFP for $k > 1$, we will just survey the heavily studied case of $k = 1$, i.e., the EDP and the more general UFP. We denote by m the number of edges and by n the number of vertices in the graph G.

Several results are known about the approximation ratio and competitive ratio achievable for the UFP under the assumption that the maximum demand of a commodity, $d_{\max}$, does not exceed the minimum edge capacity, $c_{\min}$, often referred to as the *no-bottleneck assumption* [1, 15, 5, 8, 13, 18, 17]. Baveja and Srinivasan [5] present a polynomial time algorithm with an approximation ratio $O(\sqrt{m})$. A recent paper by Chekuri, Khanna, and Shepherd [10] presents an $O(\sqrt{n})$ approximation algorithm. On the lower bound side, it was shown by Guruswami et al. [13] that on directed networks the UFP is NP-hard to approximate within a factor of $n^{1/2-\epsilon}$ for any $\epsilon > 0$. Using a new parameter called the *flow number* F of a network, Kolman and Scheideler [17] show that a simple online algorithm has a competitive ratio of $O(F)$; they also prove that $F = O(\Delta\alpha^{-1} \log n)$, where, for the EDP, Δ is the maximal degree of the network, α is the edge expansion, and n is the number of nodes and, for the UFP, Δ has to be defined as the maximal node capacity of the network and α as the expansion with respect to the the edge capacities. Combining the approach of Kolman and Scheideler [17] with the randomized rounding technique, Chakrabarti et al. [8] recently proved a randomized approximation ratio of $O(\Delta_G \alpha_G^{-1} \log^2 n)$ for the more general UFP with profits, where Δ_G and α_G stand for the maximum degree and the expansion of the given network when ignoring the capacities.

We also consider two related problems, the *integral splittable flow problem* (ISF) [13] and the *k-splittable flow problem* (k-SFP). In both cases, the input and the objective (i.e., to maximize the sum of accepted demands) are the same as in the UFP. The difference is that in the ISF all demands are integral and a flow satisfying a demand can be split into several paths, each carrying an integral amount of flow. In the k-SFP a flow of a single commodity may be split into at most k flow paths (of not necessarily integral values). Under the no-bottleneck assumption Guruswami et al. [13] give an $O(\sqrt{md_{\max}} \log^2 m)$ approximation for the ISF. The k-SFP was independently introduced by Baier, Köhler, and Skutella [4]. Since that time several approximation and hardness results about the k-SFP appeared [20, 21, 16]. The techniques of Kolman and Scheideler [17] allow us to achieve an $O(F)$ randomized competitive ratio and an $O(F)$ deterministic approximation ratio for both of these problems on unit-capacity networks. Although the ISF and the k-SFP on one side and the k-DFP on the other seem very similar at first glance, there is a serious difference between the two. Whereas the ISF and the k-SFP are *relaxations* of the UFP (they allow the use of more than one path for a single request, and the paths are *not* required to be disjoint), the k-DFP is actually a *more complex* version of the UFP since it requires several *disjoint* paths for a single request.

1.2. New results. This paper's main results are

- a deterministic online algorithm for the k-EDP with competitive ratio $O(k^3F)$ (subsequent work yields an improved competitive ratio $O(k^2F)$,
- a deterministic offline algorithm for the k-DFP on unit-capacity networks with an approximation ratio of $O(k^3F\log(kF))$,
- a lower bound $\Omega(kF)$ for the competitive ratio of any deterministic online algorithm for the k-EDP (and thus, obviously, for the k-DFP).

Thus, for constant k, we have matching upper and lower bounds for the k-EDP.

Furthermore, we demonstrate that disjointness of the k paths for every single request seems to be the crucial condition that makes these problems harder than other problems such as the ISF or the k-SFP.

We also show, using previously known techniques, how to transform the online algorithm for the k-EDP into an offline algorithm for the k-EDP with profits and how to convert the offline algorithm for the k-DFP into a randomized online algorithm for the k-DFP with an expected competitive ratio of $O(k^3F\log(kF))$.

Our algorithms for the k-EDP and k-DFP are based on a simple concept, a natural extension of the *bounded greedy algorithm* (BGA) that has already been studied in several papers [15, 18, 17]: For a given request if we can find k disjoint flow paths of total length at most L, without violating capacity constraints given the connections we have already made, select any such system of k paths for this request. The core of this paper is in the analysis of this simple algorithm. The problem is to show that this strategy works even if the optimal offline algorithm connects many requests via k disjoint paths of total length more than L. In order to solve this problem we use a new technique, based on Menger's theorem and the Lovász local lemma that converts large systems of k disjoint paths into small systems of k disjoint paths. Previously, shortening strategies were known only for $k = 1$ [18, 17].

1.3. Basic notation and techniques. Many of the previous techniques for the EDP and related problems do not allow us to prove strong upper bounds on approximation or competitive ratios due to the use of inappropriate parameters. If n is the only parameter used, the upper bound of $O(\sqrt{n})$ is essentially the best possible for the case of directed networks [13, 10]. Much better ratios can be shown if the expansion or the routing number [22] of a network are used. These measures give very good bounds for low-degree networks with uniform edge capacities but are usually very poor when applied to networks of high-degree or highly nonuniform degrees or edge capacities. To get more precise bounds for the approximation and competitive ratios of algorithms, Kolman and Scheideler [17] introduced a new network measure, the *flow number* F. Not only does the flow number lead to more precise results, it also has the major advantage that, in contrast to the expansion or the routing number, it can be computed exactly in polynomial time. Hence we use the flow number in this paper as well.

Before we introduce the flow number, we need some notation. In a *concurrent multicommodity flow problem* there are k commodities, each with two terminal nodes s_i and t_i and a demand d_i. A *feasible* solution is a set of flow paths for the commodities that obey capacity constraints but need not meet the specified demands. An important difference between this problem and the UFP is that the commodity between s_i and t_i can be routed along multiple paths. The *(relative) flow value of a feasible solution* is the maximum f such that at least $f \cdot d_i$ units of commodity i are simultaneously routed for each i. The *max-flow* for a concurrent multicommodity flow problem is defined as the maximum flow value over all feasible solutions. For a path p in a solution, the *flow value of* p is the amount of flow routed along it. A special class of the concurrent multicommodity flow problems is the *product multicommodity flow problem* (PMFP). In a PMFP, a nonnegative weight $\pi(u)$ is associated with each node $u \in V$, and there is a commodity with demand $\pi(u) \cdot \pi(v)$ for every pair of nodes (u, v).

Suppose we have a network $G = (V, E)$ with arbitrary nonnegative edge capacities. For every node v, let the *capacity* of v be defined as $c(v) = \sum_{w:\{v,w\}\in E} c(v, w)$ and the capacity of G be defined as $\Gamma = \sum_v c(v)$. Given a concurrent multicommodity flow problem with feasible solution $\mathcal{S}$, let the *dilation* $D(\mathcal{S})$ of $\mathcal{S}$ be defined as the length of the longest flow path in $\mathcal{S}$ and the *congestion* $C(\mathcal{S})$ of $\mathcal{S}$ be defined as the inverse of its flow value (i.e., the congestion tells us how many times the edge capacities would have to be increased in order to fully satisfy all the original demands along the paths of $\mathcal{S}$). Let I_0 be the PMFP in which $\pi(v) = c(v)/\sqrt{\Gamma}$ for every node v; i.e., each pair of nodes (v, w) has a commodity with demand $c(v) \cdot c(w)/\Gamma$. The flow number $F(G)$ of a network G is the minimum of $\max\{C(\mathcal{S}), D(\mathcal{S})\}$ over all feasible solutions $\mathcal{S}$ of I_0. When there is no risk of confusion, we simply write F instead of $F(G)$. Note that the flow number of a network is invariant to scaling of capacities.

The smaller the flow number, the better are the communication properties of the network. For example, $F(\text{line}) = \Theta(n)$, $F(\text{mesh}) = \Theta(\sqrt{n})$, $F(\text{hypercube}) = \Theta(\log n)$, $F(\text{butterfly}) = \Theta(\log n)$, and $F(\text{expander}) = \Theta(\log n)$, where the expanders we refer to are constant degree graphs of constant edge expansion.

The *shortening lemma* [17] will be a useful tool for the analysis of our algorithms.

Lemma 1.1 (shortening lemma). *For any network with flow number F, the following holds: For any $\epsilon \in (0, 1]$ and any feasible solution $\mathcal{S}$ to an instance of the concurrent multicommodity flow problem with a flow value of f, there exists a feasible solution with flow value $f/(1 + \epsilon)$ that uses paths of length at most $2 \cdot F(1 + 1/\epsilon)$. Moreover, the flow through any edge e not used by $\mathcal{S}$ is at most $\epsilon \cdot c(e)/(1 + \epsilon)$.*

Another useful class of concurrent multicommodity flow problems is the *balanced multicommodity flow problem* (BMFP). A BMFP is a multicommodity flow problem in which the sum of the demands of the commodities originating and the commodities terminating in a node v is at most $c(v)$ for every $v \in V$. We make use of the following property of the problem [17].

Lemma 1.2. *For any network G with flow number F and any instance I of a BMFP for G, there is a feasible solution for I with congestion and dilation at most $2F$.*

Apart from the flow number we also need Chernoff bounds [14], the symmetric form of the Lovász local lemma [12], and Menger's theorem [7, p. 75].

LEMMA 1.3 (Chernoff bound). *Consider any set of* n *independent binary random variables* $X_1, \ldots, X_n$. *Let* $X = \sum_{i=1}^n X_i$ *and* μ *be chosen so that* $\mu \geq \mathrm{E}[X]$. *Then it holds for all* $\delta \geq 0$ *that*

$$\Pr[X \geq (1+\delta)\mu] \leq \mathrm{e}^{-\min[\delta^2,\, \delta]\cdot\mu/3} .$$

LEMMA 1.4 (Lovász local lemma). *Let* $A_1, \ldots, A_n$ *be "bad" events in an arbitrary probability space. Suppose that each event is mutually independent of all other events but is at most* b, *and that* $\Pr[A_i] \leq p$ *for all* i. *If* $\mathrm{e}p(b+1) \leq 1$, *the probability of no bad event occurring is greater than* 0.

LEMMA 1.5 (Menger's theorem). *Let* s *and* t *be distinct vertices of* G. *The minimal number of edges separating* s *from* t *is equal to the maximal number of edge-disjoint* s-t *paths.*

In the following, a k-*system* is a set of k edge-disjoint paths connecting the same pair of vertices. A k-system is *small* if it uses at most L edges for some fixed parameter L depending on network properties. The *flow value* of a k-system is the total amount of flow routed along the k paths in it. We require the flow to be the same along all the k paths. For a set M of k-systems, let $||M||$ denote the total amount of flow sent along all of them, i.e., the sum of the flow values. For a path p let $|p|$ denote the number of edges of p, i.e., its *length*.

1.4. Organization of this paper.

In section 2 we present our upper and lower bounds for the k-EDP and some related problems, and in section 3 we present our upper bounds for the k-DFP. The paper ends with a conclusion and open problems.

2. Algorithms for the k-EDP.

Consider the following extension of the BGA: Let L be a suitably chosen parameter. Given a request, if it is possible to find a small k-system for it that is disjoint with all previously selected k-systems, then accept the request, and select any such k-system for it. Otherwise, reject the request. Let us call this algorithm the k bounded greedy algorithm (k-BGA).

Note that the problem of finding k edge-disjoint paths of total length at most L between the same pair of nodes, i.e., the problem of finding a small k-system, can be reduced to the classical min-cost (integral) flow problem, which can be solved by standard methods in polynomial time [11, Chapter 4]. The k-BGA can therefore also be used offline as an approximation algorithm. It is worth mentioning that if there were a bound of L/k on the length of every path (instead of the bound L/k on the average path length), the problem would not be tractable (cf. [6]).

2.1. The upper bound.

THEOREM 2.1. *Given a network* G *of flow number* F, *the competitive ratio of the* k-*BGA with parameter* $L = 24k^3F$ *is* $O(k^3F)$.

Proof. Let $\mathcal{B}$ be the solution obtained by the k-BGA and $\mathcal{O}$ be the optimal solution. For notational simplicity we allow a certain ambiguity. Sometimes $\mathcal{B}$ and $\mathcal{O}$ refer to the subsets of T of the satisfied requests, and sometimes $\mathcal{B}$ and $\mathcal{O}$ refer to the actual k-systems that realize the satisfied requests. We say that a k-system $q \in \mathcal{B}$ is a *witness* for a k-system p if p and q share an edge. Obviously, a request with a small k-system in the optimal solution that was rejected by the k-BGA must have a witness in $\mathcal{B}$.

Let $\mathcal{O}' \subseteq \mathcal{O}$ denote the set of all k-systems in $\mathcal{O}$ that are larger than L and that correspond to requests *not* accepted by the k-BGA and that do *not* have a witness in

$\mathcal{B}$. Then each k-system in $\mathcal{O}-\mathcal{O}'$ either has a witness or was accepted by the k-BGA. Since the k-systems in $\mathcal{O}-\mathcal{O}'$ are edge-disjoint, each request accepted by the k-BGA can be a witness to at most L requests in $\mathcal{O}-\mathcal{O}'$. Hence, $|\mathcal{O}-\mathcal{O}'| \leq (1+L)|\mathcal{B}|$.

It remains to prove an upper bound on $|\mathcal{O}'|$. To achieve this, we transform the k-systems in $\mathcal{O}'$ into a set $\mathcal{P}$ of possibly overlapping but *small* k-systems. Since these small k-systems would have been candidates for the k-BGA but were not picked, each of them has at least one witness in $\mathcal{B}$. Then we show that the small k-systems in $\mathcal{P}$ do not overlap much, and thus many k-systems from $\mathcal{B}$ are needed in order to provide a witness for every k-system in $\mathcal{P}$.

Note that the set $\mathcal{O}'$ of k-systems can be viewed as a feasible solution of relative flow value 1 to the set of requests $\mathcal{O}'$ of the concurrent multicommodity flow problem where each request has demand k. The shortening lemma with parameter $\epsilon = 1/(2k)$ immediately implies the following fact.

Fact 2.2. The k-systems in $\mathcal{O}'$ can be transformed into a set $\mathcal{R}$ of flow systems transporting the same amount of flow such that every flow path has a length of at most $5kF$. Furthermore, $\mathcal{R}$ has the property that the flow at every edge that is used by some k-system in $\mathcal{O}'$ is at most $1+1/(2k)$ and the flow at every other edge is at most $1/(2k)$.

This does not immediately provide us with small k-systems for the requests in $\mathcal{O}'$. However, it is possible to extract small k-systems from the flow system $\mathcal{R}$.

LEMMA 2.3. *For every request in $\mathcal{O}'$, a set of small k-systems can be extracted out of its flow system in $\mathcal{R}$ with a total flow value of at least* $1/4$.

Proof. Let (s_i, t_i) be a fixed request from $\mathcal{O}'$, and let E_i be the set of all edges that are traversed by the flow system for (s_i, t_i) in $\mathcal{R}$. Consider any set of $k-1$ edges in E_i. Since the flow through any edge in $\mathcal{R}$ is at most $1+1/(2k)$, the total amount of flow in the flow system for (s_i, t_i) in $\mathcal{R}$ that traverses the $k-1$ edges is at most $(k-1)(1+1/(2k)) < k-1/2$. Thus, the minimal s_i-t_i-cut in the graph (V, E_i) consists of at least k edges. Hence, Menger's theorem [7] implies that there are k edge-disjoint paths between s_i and t_i in E_i. We take any such k paths and denote them as the k-system σ_1. We associate a *weight* (i.e., total flow) of $k\cdot\epsilon_1$ with σ_1, where ϵ_1 is the minimum flow from s_i to t_i through an edge in E_i belonging to the k-system σ_1.

Assume now that we have already found ℓ k-systems $\sigma_1, \sigma_2, \ldots, \sigma_\ell$ for some $\ell \geq 1$. If $\sum_{j=1}^{\ell} k\cdot\epsilon_j \geq \frac{1}{2}$ we stop the process of defining σ_j. Otherwise, the minimal s_i-t_i-cut in (V, E_i) must still be at least k, because the total flow along any $k-1$ edges in E_i is still less than the total remaining flow from s_i to t_i. Thus, we can apply Menger's theorem again. This allows us to find another k-system $\sigma_{\ell+1}$ between s_i and t_i, and in the same way as above we associate with it a weight $\epsilon_{\ell+1}$. Let $\hat{\ell}$ be the number of k-systems at the end of the process.

So far there is no guarantee that any of the k-systems defined above will be small nor that they will transport enough flow between the terminal pair s_i and t_i. However, after a simple procedure they will satisfy our needs.

According to Fact 2.2, all flow paths in $\mathcal{R}$ have a length of at most $5kF$. Hence, the total amount of edge capacity consumed by a flow system in $\mathcal{R}$ representing a request in $\mathcal{O}'$ is at most $5k^2F$. If there were k-systems in $\sigma_1, \ldots, \sigma_{\hat{\ell}}$ of total weight at least $1/4$ that use more than $20k^3F$ edges each, then they would not fit into the available edge capacity, because $20k^3F \cdot 1/(4k) = 5k^2F$. Thus, there exists a subset of the k-systems $\sigma_1, \ldots, \sigma_{\hat{\ell}}$ with total weight at least $1/4$ such that each of them is small, i.e., each of them uses at most $20k^3F$ edges. □

We are ready to bound $|\mathcal{O}'|$, the number of k-systems in $\mathcal{O}'$, in terms of $|\mathcal{B}|$. Let $\mathcal{S}_i$ denote the set of small k-systems for request $(s_i, t_i) \in \mathcal{O}'$, given by Lemma 2.3, and let $\mathcal{S}$ be the set of all $\mathcal{S}_i$. By the definition of $\mathcal{S}$,

$$||\mathcal{O}'|| \leq 4k \cdot ||\mathcal{S}|| \,. \tag{1}$$

Since the k-systems in $\mathcal{S}$ connect requests from $\mathcal{O}'$ and they are small, each of them must have a witness in $\mathcal{B}$. Let $E_{\mathcal{S}}$ denote the set of all edges on which a k-system from $\mathcal{S}$ has a witness. According to the definition of $\mathcal{O}'$, no edge in $E_{\mathcal{S}}$ can be part of a k-system in $\mathcal{O}'$. It follows from Fact 2.2 that the flow belonging to k-systems in $\mathcal{S}$ on any one of the edges in $E_{\mathcal{S}}$ is at most $1/(2k)$. Thus, it holds for the total flow along k-systems in $\mathcal{S}$ that

$$||\mathcal{S}|| \leq k \cdot \left(\tfrac{1}{2k} \cdot |E_{\mathcal{S}}|\right) . \tag{2}$$

Let $E_{\mathcal{B}}$ be the set of all edges used by $\mathcal{B}$. Then

$$|E_{\mathcal{S}}| \leq |E_{\mathcal{B}}| \leq L \cdot |\mathcal{B}| \,. \tag{3}$$

Since $|\mathcal{O}'| = \frac{1}{k} \cdot ||\mathcal{O}'||$, combining inequalities (1)–(3) gives $|\mathcal{O}'| \leq 2L \cdot |\mathcal{B}|$ and completes the proof. □

The above upper bound on the competitive ratio for the k-BGA with parameter $L = 20k^3F$ is the best possible, since a k-system of size $\Theta(k^3F)$ may prevent $\Theta(k^3F)$ other k-systems from being selected.

Stronger bound. In a subsequent work [3], the ideas from the previous proof were exploited in the proof of the following lemma (a kind of shortening lemma for flow along k-systems).

LEMMA 2.4. *Given a unit network with flow number F, a set T of pairs of vertices, and a feasible flow $\mathcal{F}$ such that there are k units of flow between each pair from T, there exists a k-flow $\bar{\mathcal{F}}$ such that*

- *there are k units of flow between each pair from T,*
- *the flow through every edge is at most 4,*
- *each k-system used in $\bar{\mathcal{F}}$ has size at most $20 \cdot k^2F$.*

Moreover, if the k-flow $\mathcal{F}$ is integral (i.e., each pair from T is connected by a unit k-system), stronger bounds hold:

- *The flow in $\bar{\mathcal{F}}$ through every edge is at most 2, and*
- *each k-system used in $\bar{\mathcal{F}}$ has size at most $8 \cdot k^2F$.*

The lemma provides an alternative way to analyze the k-BGA algorithm. Given the optimal solution $\mathcal{O}$, we apply Lemma 2.4 and transform it into a fractional solution $\mathcal{O}'$ with flow at most 2 on every edge that consists of small k-systems only. Then we use the witnessing argument again and obtain the stronger bound.

THEOREM 2.5. *Given a network G of flow number F, the competitive ratio of the k-BGA with parameter $L = 8k^2F$ is $O(k^2F)$.*

2.2. General online lower bound. We show there is a lower bound on the competitive ratio of any deterministic online algorithm for the k-EDP problem which is not far away from the performance of the k-BGA.

THEOREM 2.6. *For any n, k, and $F \geq \log_k n$ with $n \geq k^2 \cdot F$, there is a graph G of size $\Theta(n)$ with maximum degree $O(k)$ and flow number $\Theta(F)$ such that the competitive ratio of any deterministic online algorithm on G is $\Omega(k \cdot F)$.*

Proof. A basic building block of our construction is the following simple graph. Let D_k (*diamond*) denote the graph consisting of two bipartite graphs $K_{1,k}$ and $K_{k,1}$

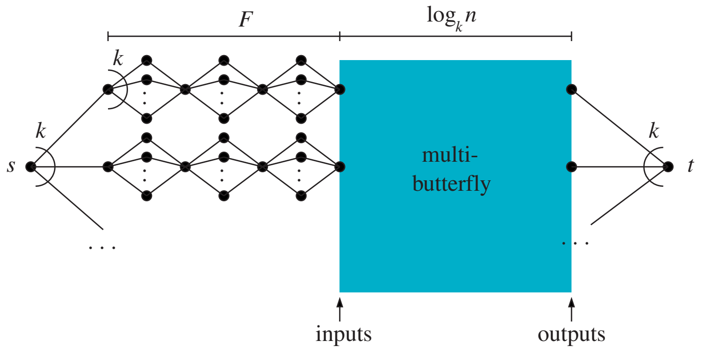


FIG. 2.1. *The graph for the lower bound.*

glued naturally together at the larger sides. The two k-degree nodes in D_k are its *endpoints*. Let C (*chaplet*) denote the graph consisting of F diamond graphs attached one to the other at the endpoints, like in an open chaplet.

The core of the graph G consists of $m = n/(k \cdot F) \geq k$ disjoint copies of the chaplet graph C attached to the inputs of a k-ary multibutterfly (Figure 2.1). In addition, a node s is connected to the first k chaplet graphs and a node t is connected to the first k output nodes of the multibutterfly. Let $s_{i,j}$ denote the first endpoint of a diamond j in a chaplet i, and let $t_{i,j} (= s_{i,j+1})$ denote the other endpoint. We use the fact that a k-ary multibutterfly with n' inputs and outputs (which is a network of degree $O(k)$) can route any r-relation from the inputs to the outputs with edge congestion and dilation at most $O(\max[r/k, \log_k n'])$ [22].

First, we show that our graph G has a flow number of $\Theta(F)$. Since the diameter of G is $\Omega(F)$, it is sufficient to prove that a PMFP with $\pi(u) = c(u)/\Gamma$ for the given graph can be solved with congestion and dilation $O(F)$. Consider each node v of degree δ_v to consist of δ_v copies of itself, and let V' be the set of all of these copies. Then the PMFP reduces to the problem of sending a packet of size $1/N$ for any pair of nodes in V', where $N = |V'|$. Such a routing problem can be split into N permutations σ_i with $\sigma_i(v) = (v + i) \bmod N$ for all $i \in \{0, \dots, N-1\}$ and $v \in V'$. Each such permutation represents a routing problem ρ in the original network where each node is the starting point and endpoint of a number of packets that is equal to its degree. We want to bound the congestion and dilation for routing such a problem.

In order to route ρ, we first move all packets to the inputs of the k-ary multibutterfly in such a way that every input node of the multibutterfly will have $O(kF)$ packets. This can clearly be done with edge congestion $O(F)$ and dilation $O(F)$. Next, we use the multibutterfly to send the packets to the rows of their destinations. Since every input has $O(k \cdot F)$ packets, this can also be done with congestion and dilation $O(F)$. Finally, all packets are sent to their correct destinations. This too causes a congestion and dilation of at most $O(F)$. Hence, routing ρ requires only a total congestion and dilation of $O(F)$.

Combining the fact that all packets are of size $1/N$ with the fact that we have N permutations σ_i, it follows that the congestion and dilation of routing the PMFP in the given graph is $O(F)$. Hence, its flow number is $\Theta(F)$.

Now consider the following two sequences of requests:

1. (s,t), and
2. $(s,t), (s_{1,1}, t_{1,1}), (s_{1,2}, t_{1,2}), \dots, (s_{1,F}, t_{1,F}), (s_{2,1}, t_{2,1}),$
   $\dots, (s_{k,F}, t_{k,F})$

Obviously, every deterministic online algorithm has to accept (s,t) to ensure a finite competitive ratio for the sequence 1. However, in this case none of the other requests in 2 can be satisfied. But the optimal solution for 2 is to reject (s,t) and to accept all other requests. Hence, the competitive ratio of any deterministic online algorithm is $\Omega(k \cdot F)$. □

2.3. Requests with profits. In the *k edge-disjoint paths with profits problem* (k-EDPP) we are given an undirected graph $G=(V,E)$ and a set of requests T. Each request $r_i=(s_i,t_i)$ has a positive profit $b(r_i)$. The problem is to find a subset S of the pairs in T of maximum profit for which it is possible to select disjoint paths such that each pair is connected by k disjoint paths.

It turns out that a simple offline variant of the k-BGA gives the same approximation ratio for the k-EDPP as we have for the k-EDP. The algorithm involves sorting the requests in decreasing order of their profits and running the k-BGA on this sorted sequence. We call this algorithm the *sorted* k-BGA.

THEOREM 2.7. *Given a network G of flow number F, the approximation ratio of the sorted k-BGA with parameter $L=20k^3F$ is $O(k^3F)$ for the k-EDPP.*

Proof. The proof is almost identical to the proof of Theorem 2.1. The only additional observation is that, since the sorted k-BGA proceeds through the requests from the most profitable, every small k-system in $\mathcal{O}-\mathcal{O}'$ and in the modified set $\mathcal{P}$ has a witness in $\mathcal{B}$ with larger or equal profit. □

Again, via Lemma 2.4 we can get a better bound $O(k^2F)$ for k-BGA with parameter $L=8k^2F$.

2.4. The multiple EDP. Another variant of the k-EDP our techniques can be applied to is the *multiple EDP* (multi-EDP) which is defined as follows: Given a graph G and a set of terminal pairs with integral demands d_i, $1 \le d_i \le \Delta$, find a maximum subset of the pairs for which it is possible to select disjoint paths so that every selected pair i has d_i disjoint paths. Let $d_{\max}$ denote the maximal demand over all requests.

A variant of the k-BGA, the *multiple BGA* (multi-BGA), can be used here as well: Given a request with demand d_i, reject it if it is not possible to find d_i edge-disjoint paths between the terminal pairs of total length at most $20d_id_{\max}^2F$. Otherwise, select any such d_i paths for it.

THEOREM 2.8. *Given a network G of flow number F, the competitive ratio of the multi-BGA is $O(d_{\max}^3F)$.*

Proof. The proof goes along the same lines as the proof of Theorem 2.1: First, the shortening lemma with parameter $\epsilon=1/(2d_{\max})$ is applied, and afterwards, the extraction procedure is used. The difference is that now we extract only d_i-systems for a request with demand d_i, not $d_{\max}$-systems. □

As in previous sections, a better bound is possible using Lemma 2.4.

2.5. All-or-nothing multicommodity flow problem. Chekuri, Khanna, and Shepherd [9] introduced the *all-or-nothing* multicommodity flow problem, a relaxation of the EDP. A version of the problem that is relevant for flows along k-systems is as follows: Given a graph G and a set of terminal pairs T, find a maximum subset U of T such that there exist k-systems of total flow value k for each pair in U and the cumulative flow of all commodities through every edge is at most one.

Since this problem is only a relaxation of the k-EDP problem and since in our analysis of the k-BGA algorithm we did not use the integrality of the optimal solution (apart from the constants in Lemma 2.4), our results for k-EDP apply also for the above version of the all-or-nothing multicommodity flow problem.

3. Algorithms for the k-DFP. We now turn to capacitated networks and consider requests with arbitrary demands. Throughout this section we will assume that the maximal demand is at most k times larger than the minimal edge capacity, which is analogous to assumptions made in almost all papers about the UFP. We call this the *weak bottleneck assumption.* Moreover, we assume that all edge capacities are the same. Since F is invariant to scaling, we simply set all edge capacities to 1. The minimal demand of a request will be denoted by $d_{\min}$. We first give an offline algorithm for the k-DFP and prove that it has a good approximation ratio, and then we mention how to convert it into a competitive online algorithm.

To solve the offline k-DFP, we first sort the requests in decreasing order of their demands. On this sorted sequence of requests we use an algorithm that is very similar to the k-BGA: Let L be a suitably chosen parameter. Given a request with a demand of d, accept it if it is possible to find for it a small k-system with flow value d that fits into the network without violating the capacity constraints. Otherwise, reject it. This extension of the k-BGA will be called *k-flow BGA.*

The next theorem demonstrates that the performance of the k-flow BGA for the k-DFP is comparable to the performance of the k-BGA for the k-EDP. It is slightly worse due to a technical reason: It is much harder to use our technique for extracting small k-systems for the k-DFP than for the k-EDP.

THEOREM 3.1. *Given a unit-capacity network G with flow number F, the approximation ratio of the k-flow BGA for the k-DFP with parameter $L = \gamma \cdot k^3 F \log(kF)$ for an appropriately large constant γ, when run on requests sorted in nonincreasing order, is $O(k^3 F \log(kF))$.*

Proof. As usual, let $\mathcal{B}$ denote the set of k-systems for the requests accepted by the BGA and $\mathcal{O}$ be the set of k-systems in the optimal solution. Each k-system consists of k disjoint flow paths which we call *streams.* For notational simplicity we will sometimes think about $\mathcal{B}$ and $\mathcal{O}$ as a set of streams (instead of k-systems).

For each stream $q \in \mathcal{B}$ or $q \in \mathcal{O}$, let $f(q)$ denote the flow along that stream. If q belongs to the request (s_i, t_i) with demand d_i, then $f(q) = d_i/k$. For a set $\mathcal{Q}$ of streams, let $||\mathcal{Q}|| = \sum_{q \in \mathcal{Q}} f(q)$. Also, for an edge $e \in E$ and a stream q, let $F(e, q)$ denote the sum of flow values of all streams in $\mathcal{B}$ passing through e whose flow is at least as large as the flow of q, i.e., $F(e,q) = ||\{p \mid p \in \mathcal{B},\ e \in p,\ f(p) \geq f(q)\}||$. A stream $p \in \mathcal{B}$ is a witness for a stream q if $f(p) \geq f(q)$ and p and q intersect in an edge e with $F(e,q) + f(q) > 1$. For each edge e let $\mathcal{W}(e, \mathcal{B})$ denote the set of streams in $\mathcal{B}$ that serve as witnesses on e. Similarly, for each edge e let $\mathcal{V}(e, \mathcal{Q})$ denote the set of streams in $\mathcal{Q}$ that have witnesses on e. We also say that a k-system has a witness on an edge e if any of its k streams has a witness on e. We start with a simple observation.

CLAIM 3.2. *For any stream $q \in \mathcal{O}$ and edge e, if q has a witness on e, then $\|\mathcal{W}(e, \mathcal{B})\| \geq 1/2$.*

Proof. Let p be a witness of q on e. Assume, by contradiction, that $F(e,q) < 1/2$. It easily follows that $f(p) < 1/2$. Since $f(q) \leq f(p)$ and $F(e,q) + f(q) > 1$ by the definition of a witness, we have a contradiction. □

Let $\mathcal{O}' \subset \mathcal{O}$ be the set of k-systems that are larger than L, that correspond to requests *not* accepted by the k-flow BGA, and that do *not* have a witness in $\mathcal{B}$. The next two bounds on $||\mathcal{O} \setminus \mathcal{O}'||$ and $||\mathcal{O}'||$ complete the proof.

LEMMA 3.3. $||\mathcal{O} \setminus \mathcal{O}'|| \leq (1 + 2L) \cdot ||\mathcal{B}||$.

Proof. We partition $\mathcal{O} \setminus \mathcal{O}'$ into two sets. Let $\mathcal{O}_1 \subseteq \mathcal{O} \setminus \mathcal{O}'$ consist of all the k-systems corresponding to requests accepted by the BGA, and let $\mathcal{O}_2 = (\mathcal{O} \setminus \mathcal{O}') \setminus \mathcal{O}_1$. Obviously, $||\mathcal{O}_1|| \leq ||\mathcal{B}||$. Note that each k-system in $\mathcal{O}_2$ must have a witness in $\mathcal{B}$.

Let $E' \subseteq E$ denote the set of all edges on which some k-system from $\mathcal{O}_2$ has a witness. We then have

$$\|\mathcal{O}_2\| \le \sum_{e\in E'} k \cdot \|\mathcal{V}(e,\mathcal{O}_2)\| \le \sum_{e\in E'} k \le \sum_{e\in E'} k \cdot 2\|\mathcal{W}(e,\mathcal{B})\|.$$

The first inequality follows from the definition of $\mathcal{V}(e,\mathcal{Q})$ and the above observation that each $q \in \mathcal{O}_2$ has a witness in $\mathcal{B}$. The second inequality holds due to the unit capacities, and the last one follows from Claim 3.2.

Since all k-systems in $\mathcal{B}$ are of length at most L, we have

$$\begin{aligned}\sum_{e\in E'} \|\mathcal{W}(e,\mathcal{B})\| &\le \sum_{\text{streams } p\in\mathcal{B}} |p| \cdot f(p)\\ &\le \sum_{k-\text{systems } s\in\mathcal{B}} L \cdot d(s)/k \le L \cdot \|\mathcal{B}\|/k\ .\end{aligned}$$

This completes the proof of Lemma 3.3. □

In the next lemma we bound $\|\mathcal{O}'\|$ by first transforming the large k-systems in $\mathcal{O}'$ into a set $\mathcal{S}$ of small k-systems and then bounding $\|\mathcal{S}\|$ in terms of $\|\mathcal{B}\|$.

LEMMA 3.4. $\|\mathcal{O}'\| = O(L \cdot \|\mathcal{B}\|)$.

Proof. In order to prove the lemma, we will transform the k-systems in $\mathcal{O}'$ into a set of k-systems $\mathcal{S}$ in which each k-system has a length at most L and therefore must have a witness in $\mathcal{B}$. To achieve this, we perform a sequence of transformations:

1. First, we scale the demands and edge capacities so that each edge in G has a capacity of $C = \lceil 3k/d_{\min} \rceil$ and all requests have demands that are integral multiples of k. More precisely, the demand of each request of original demand d is set to $d' = k \cdot \lceil C \cdot d/k \rceil$. Since $d'/C \in [d, (1+1/3)d]$, this slightly increases the demands, and therefore it also increases the flows along the streams so that the total flow along an edge is now at most $(1 + 1/3)C$. Note that slightly increasing the demands increases only $\|\mathcal{O}'\|$ and therefore makes only the bound on the relationship between $\|\mathcal{O}'\|$ and $\|\mathcal{B}\|$ more pessimistic.
2. Next, we replace each request (s_i, t_i) in $\mathcal{O}'$ by d'_i/k *elementary* requests of demand k each, shipped along the same k-system as for (s_i, t_i). For every k-system of such a request, we keep only the first $8c \cdot kF$ and the last $8c \cdot kF$ nodes along each of its k streams for some $c = O(\log(kF))$. The resulting set of (possibly disconnected) streams of a k-system will be called a *k-core.* As shown in Claim 3.5, it is possible to distribute the elementary requests into C/c sets $S_1, \dots, S_{C/c}$ so that the congestion caused by the k-cores within each set is at most $2c$ at each edge.
3. Afterwards, we consider each S_i separately. We will reconnect disconnected streams in each k-core in S_i with flow systems derived from the flow number. The reconnected k-cores will not yet consists of k disjoint streams. We will show in Claim 3.6 how to extract k-systems of length at most L from each reconnected k-core.
4. Once we have found the small k-systems, we will be able to compare $\|\mathcal{O}'\|$ with $\|\mathcal{B}\|$ with the help of witnesses.

Next we present two vital claims. The proof of the first claim requires the use of the Lovász local lemma, and the proof of the second claim is similar to the proof of Theorem 2.1.

CLAIM 3.5. *The elementary requests can be distributed into C/c sets $S_1, \dots, S_{C/c}$ for some $c = O(\log(kF))$ so that for each set S_i the edge congestion caused by its k-cores is at most $2c$.*

Proof. We first prove the claim for $c = O(\log(kCF))$ and then demonstrate how to get to $c = O(\log(kF))$.

Consider the random experiment of assigning a number $i \in \{1, \ldots, C/c\}$ to each elementary request uniformly and independently at random, and let S_i be the set of all requests that got number i. For every edge e let the random variable $X_{e,i}$ denote the number of streams assigned to S_i that traverse e. Since the maximal edge congestion is at most $4C/3$, we have $\mathrm{E}[X_{e,i}] \leq 4c/3$ for every edge e. Every edge e can be used by at most one stream of any k-core. Hence, a k-core can contribute a value of at most 1 to $X_{e,i}$, and the contributions of different k-cores are independent. We can use Chernoff bounds to derive

$$\Pr[X_{e,i} \geq (1 + 1/3) \cdot 4c/3] \leq \mathrm{e}^{-(1/3)^2 \cdot (4c/3)/3} = \mathrm{e}^{-4c/3^4} .$$

For every edge e and every $i \in \{1, \ldots, C/c\}$, let $A_{e,i}$ be the event that $X_{e,i} > 2c$. Since $(4/3)^2 \leq 2$, the above probability estimate bounds the probability that the event $A_{v,i}$ appears. Our aim is to show, with the help of the Lovász local lemma, that it is possible in the random experiment to assign numbers to the requests so that none of these events appears, which would yield our claim. To apply the Lovász local lemma we have to bound the dependencies among the events $A_{e,i}$.

Each edge e can be used by at most $4C/3 < 2C$ k-cores, and these are the only k-cores that affect the values $X_{e,i}$, $i \in \{1, \ldots, C/c\}$. Realizing that each of the k-cores contains at most $2k(8c \cdot kF)$ edges and that the k-cores choose their sets S_i independently at random, we conclude that the event $A_{e,i}$ depends on at most $32ck^2CF$ other events $A_{f,j}$.

To be able to use the Lovász local lemma, we have only to choose the value c so that

$$\mathrm{e} \cdot \mathrm{e}^{-4c/3^4}(32ck^2CF + 1) \leq 1 .$$

This can certainly be achieved by setting $c = \Theta(\ln(kCF))$ large enough.

The above procedure is sufficient for proving the lemma only if $C = (kF)^{O(1)}$. If $C = (kF)^{\Omega(1)}$ a more involved technique will be used. The k-cores will be distributed into the sets S_i not in a single step but in a sequence of refinements (an approach first used by Leighton, Maggs, and Rao [19] and subsequently by Scheideler [22]). In the first refinement, our aim is to show that for $c_1 = O(\ln^3 C)$ the k-cores can be distributed into the sets $S_1, \ldots, S_{C/c_1}$ so that the edge congestion in each S_i is at most $(1 + O(1/\ln C))4c_1/3$. For this we use the same random experiment as for c above. It follows that $\mathrm{E}[X_{e,i}] = 4c_1/3$ and that

$$\begin{aligned}\Pr[X_{e,i} \geq (1 + 1/\sqrt[3]{c_1}) \cdot 4c_1/3] &\leq \mathrm{e}^{-(1/\sqrt[3]{c_1})^2 \cdot (4c_1/3)/3} \\ &= \mathrm{e}^{-4\sqrt[3]{c_1}/9} .\end{aligned}$$

Hence, to be able to use the Lovász local lemma, we have to choose the value c_1 so that

$$\mathrm{e} \cdot \mathrm{e}^{-4\sqrt[3]{c_1}/9}(32c_1k^2CF + 1) \leq 1 .$$

This can certainly be achieved by setting $c_1 = \Theta(\ln^3 C)$ large enough, which completes the first refinement step.

In the second refinement step, each S_i is refined separately. Consider some fixed S_i. Our aim is to show that for $c_2 = O(\ln^3 c_1)$ the k-cores in S_i can be distributed into the sets $S_{i,1},\ldots,S_{i,c_1/c_2}$ so that the edge congestion in each $S_{i,j}$ is at most $(1+1/\sqrt[3]{c_2})(1+1/\sqrt[3]{c_1})4c_2/3$. The proof for this follows exactly the same lines as for c_1. Thus, overall C/c_2 sets $S_{i,j}$ are produced in the second step, with the corresponding congestion bound.

In general, in the $(\ell+1)$st refinement step, each set S established in refinement ℓ is refined separately, using $c_{\ell+1}=O(\ln^3 c_\ell)$, until $c_{\ell+1}=O(\ln(kF))$ for the first time. Note that in this case, $c_\ell=\omega(\ln(kF))$ and $c_\ell=(kF)^{O(1)}$. At this point we use the method presented at the beginning of the proof for the parameter c to obtain C/c' sets $S_1,\ldots,S_{C/c'}$ for some $c'=O(\ln(kF))$ with a congestion of at most

$$\left(\prod_{j=1}^{\ell}(1+1/\sqrt[3]{c_j})\right)\cdot(4/3)^2\cdot c',$$

where l is the total number of refinement steps. Using the facts that $1+x\le \mathrm{e}^x$ for all $x\ge 0$ and $\mathrm{e}^x\le 1+2x$ for all $0\le x\le 1/2$, it holds for the product that

$$\prod_{j=1}^{\ell}(1+1/\sqrt[3]{c_j})\le \mathrm{e}^{\sum_{j=0}^{\ell}1/\sqrt[3]{c_j}}\le \mathrm{e}^{\epsilon}\le 1+2\epsilon$$

for a constant $0<\epsilon\le 1/2$ that can be made arbitrarily small by making sure that c_ℓ is above a certain constant value depending on ϵ. Hence, it is possible to select the values $c_1,\ldots,c_\ell,c'$ so that the congestion in each S_i at the end is at most $2c'$. □

CLAIM 3.6. *For every set S_i, every elementary request in S_i can be given k-systems of total flow value at least $1/4$ such that each of them consists of at most L edges. Furthermore, the congestion of every edge used by an original k-system in S_i is at most $2c+1/(2k)$, and the congestion of every other edge is at most $1/(2k)$.*

Proof. For an elementary request r let $p_1^r,\ldots,p_{\ell_r}^r$ be all the disconnected streams in its k-core, $1\le \ell_r\le k$. Let the first $8c\cdot kF$ nodes in p_i^r be denoted by $a_{i,1}^r,\ldots,a_{i,8c\cdot kF}^r$ and the last $8c\cdot kF$ nodes in p_i^r be denoted by $b_{i,1}^r,\ldots,b_{i,8c\cdot kF}^r$. Consider the set of pairs

$$\mathcal{L}=\bigcup_{r\in S_1}\bigcup_{i=1}^{\ell_r}\bigcup_{j=1}^{8c\cdot kF}\{(a_{i,j}^r,b_{i,j}^r)\}\ .$$

Due to the congestion bound in Claim 3.5, a node v of degree δ can be a starting point or an endpoint of at most $2c\delta$ pairs in $\mathcal{L}$. From Lemma 1.2 we know that for any network G with flow number F and any instance I of the BMFP on G there is a feasible solution for I with congestion and dilation at most $2F$. Hence, it is possible to connect all of the pairs in $\mathcal{L}$ by flow systems of length at most $2F$ and flow value $f(p_i^r)$ so that the edge congestion is at most $2c\cdot 2F$. Let the flow system between $a_{i,j}^r$ and $b_{i,j}^r$ be denoted by $f_{i,j}^r$. For each elementary request $r=(s,t)$, each $1\le i\le \ell_r$, and each $1\le j\le 8c\cdot kF$, we define a flow system $g_{i,j}^r$: First, it moves from s to $a_{i,j}^r$ along p_i^r, then from $a_{i,j}^r$ to $b_{i,j}^r$ along $f_{i,j}^r$, and finally from $b_{i,j}^r$ to t along p_i^r, and we assign to it a flow value of $f(p_i^r)/(8c\cdot kF)$. This ensures that a total flow of $f(p_i^r)$ is still being shipped for each p_i^r. Furthermore, this allows us to reduce the flow along $f_{i,j}^r$ by a factor of $1/(8c\cdot kF)$. Hence, the edge congestion caused by the $f_{i,j}^r$ for all r,i,j reduces to at most $4c\cdot F/(8c\cdot kF)=1/(2k)$. Therefore, the additional congestion at any edge is at most $1/(2k)$, which proves the congestion bounds in the claim.

Now consider any given elementary request $r=(s,t)$. For any set of $k-1$ edges, the congestion caused by the flow systems for r is at most $(k-1)(1+1/(2k))\le k-1/2$.

Hence, according to Menger's theorem there are k edge-disjoint flows in the system from s to t. Continuing with the same arguments as in Theorem 2.1, we obtain a set of k-systems for r with properties as stated in the claim. $\square$

Now that we have small k-systems for every elementary request, we combine them back into the original requests. For a request with demand d this results in a set of k-systems of size at most L each and total flow value at least $d/(4k)$ (Claim 3.6). Let the set of all these k-systems for all requests be denoted by $\mathcal{S}$. Since every k-system has a size at most L, it could have been a candidate for the BGA. Thus, each of these k-systems must have a witness. Crucially, every edge that has witnesses for these k-systems must be an edge that is not used by *any* of the original k-systems in $\mathcal{O}'$. (This follows directly from the definition of $\mathcal{O}'$.) According to Claim 3.6, the amount of flow from $\mathcal{S}$ traversing any of these edges is at most $1/(2k)$. Let E' be the set of all witness edges.

For each request we now choose one of its k-systems independently at random, with probability proportional to the flow values of the k-systems. This will result in a set of k-systems $\mathcal{P}$ in which each request has exactly one k-system and in which the expected amount of flow traversing any edge in E' is at most $1/(2k)$. Next, we assign the original demand of the request to each of these k-systems. This causes the expected amount of flow that traverses any edge in E' to increase from at most $1/(2k)$ to at most $4k \cdot 1/(2k) = 2$.

We are now ready to bound $\|\mathcal{P}\|$ in terms of $\|\mathcal{B}\|$. For every k-system $h \in \mathcal{S}$, let the indicator variable X_h take the value 1 if and only if h is chosen to be in $\mathcal{P}$. We shall look upon $\|\mathcal{P}\|$ as a random variable (though it always has the same value) and bound its value by bounding its expected value $E[\|\mathcal{P}\|]$. In the following we assume that $f(h)$ is the flow along a stream of the k-system h and $d(h)$ is the demand of the request corresponding to h. Also, recall that the total flow value of k-systems in $\mathcal{S}$ belonging to a request with demand d is at least $d/(4k)$.

$$\begin{aligned}
\mathrm{E}[\|\mathcal{P}\|] &\leq \mathrm{E}\left[\sum_{e\in E'} k \cdot \|\mathcal{V}(e,\mathcal{P})\|\right] \\
&\leq \sum_{e\in E'} k \cdot \mathrm{E}\left[\sum_{p\in\mathcal{S}\colon\, e\in p} X_p \cdot \frac{d(p)}{k}\right] \\
&\leq \sum_{e\in E'} k \cdot \sum_{p\in\mathcal{S}\colon\, e\in p} \frac{k\cdot f(p)}{d(p)/(4k)} \cdot \frac{d(p)}{k} \\
&\leq \sum_{e\in E'} k \cdot 4k \sum_{p\in\mathcal{S}\colon\, e\in p} f(p) \\
&\leq \sum_{e\in E'} 4k^2 \cdot \frac{1}{2k} \leq \sum_{e\in E'} 2k \\
&\leq 4k \sum_{e\in E'} \|\mathcal{W}(e,\mathcal{B})\| \leq \cdots \leq 4L \cdot \|\mathcal{B}\|,
\end{aligned}$$

where the last calculations are done in the same way as in the proof of Lemma 3.3. $\square$

Combining the two lemmas proves the theorem. $\square$

We note that if the minimum demand of a request, $d_{\min}$, fulfills $d_{\min} \geq k/\log(kF)$, then one would not need Claim 3.5. In particular, if $d_{\min}$ were known in advance, then the k-flow BGA could choose $L = O(k^3F/(d_{\min}/k))$ to achieve an approximation ratio of $O(k^3F/(d_{\min}/k))$. This would allow a smooth transition from the bounds for the k-EDP (where $d_{\min} = k$) to the k-DFP.

3.1. An online algorithm for the *k*-DFP. In this section we present a randomized online algorithm for the k-DFP. This algorithm, which we shall call the *randomized* k-flow BGA, is an extension of the k-flow BGA algorithm for the offline k-DFP. The technique we present for making offline algorithms online has been used before [2, 17].

Consider, first, the set $\mathcal{O}$ of k-systems for requests accepted by the optimal algorithm. Let $\mathcal{O}_1 \subseteq \mathcal{O}$ consist of k-systems each with demand at least $k/2$, and let $\mathcal{O}_2 = \mathcal{O} \setminus \mathcal{O}_1$. Either $||\mathcal{O}_1|| \geq 1/2 \cdot ||\mathcal{O}||$, or $||\mathcal{O}_2|| > 1/2 \cdot ||\mathcal{O}||$.

The randomized k-flow BGA begins by *guessing* which of these two events will happen. If it guesses the former, it ignores all requests with demand less than $k/2$ and runs the regular k-flow BGA on the rest of the requests. If it guesses the latter, it ignores all requests with demand at least $k/2$ and runs the k-flow BGA on the rest.

THEOREM 3.7. *Given a unit-capacity network G with flow number F, the expected competitive ratio of the randomized k-flow BGA for the online k-DFP is $O(k^3F\log(kF))$ when run with parameter $L = \gamma \cdot k^3F\log(kF))$ for an appropriately large constant γ.*

Proof. The proof runs along exactly the same lines as the proof for Theorem 3.1, but we have to prove Claim 3.2 for the changed situation. Note that the original proof for Claim 3.2 relies on the fact that requests are sorted in a nondecreasing order before being considered. That need not be true here. Let $\mathcal{B}$ denote, as usual, the k-systems for requests accepted by the randomized k-flow BGA.

Consider the case when the algorithm guesses that $||\mathcal{O}_1|| \geq 1/2 \cdot ||\mathcal{O}||$. We claim that, for any stream $q \in \mathcal{O}_1$ and edge e, if q has a witness on e, then $\|\mathcal{W}(e, \mathcal{B})\| \geq 1/2$. Let $p \in \mathcal{B}$ be the stream witnessing q on e. Since the algorithm considers only requests with demand at least $k/2$, $f(p) \geq 1/2$. The claim follows since $\|\mathcal{W}(e, \mathcal{B})\| \geq f(p)$. Following the rest of the proof for Theorem 3.1, substituting $\mathcal{O}_1$ for $\mathcal{O}$, shows that in this case the randomized k-flow BGA will have a competitive ratio of $O(k^3F\log(kF))$.

Now consider the case when the algorithm guesses $||\mathcal{O}_2|| \geq 1/2 \cdot ||\mathcal{O}||$. We claim that even in this case for any stream $q \in \mathcal{O}_2$ and edge e, if q has a witness on e, then $\|\mathcal{W}(e, \mathcal{B})\| \geq 1/2$. From the definition of witnessing, we have $F(e, q) + f(q) > 1$. Next, from the definition of $\mathcal{O}_2$, $f(q) < 1/2$. The claim follows as $\|\mathcal{W}(e, \mathcal{B})\| \geq F(e, q)$. As in the previous case, the rest of the proof for Theorem 3.1 applies here too; substitute $\mathcal{O}_2$ for $\mathcal{O}$.

The competitive ratio in both cases is $O(k^3F\log(kF))$. Note that an incorrect guess just reduces the expected competitive ratio by a factor of 2. □

3.2. Comparison with other flow problems. In this section we demonstrate that the k-DFP is harder to approximate than other related flow problems because of the requirement that the k paths for every request must be disjoint.

The k-SFP and the ISF have been defined in the introduction. As already mentioned there, previous proof techniques [17] imply the following result under the no-bottleneck assumption (i.e., the maximal demand is at most equal to the minimal edge capacity).

THEOREM 3.8. *For a unit-capacity network G with flow number F, the approximation ratio of the 1-BGA with parameter $L = 4F$ for the k-SFP and for the ISF, when run on requests ordered according to their demands starting from the largest, is $O(F)$.*

Proof. The crucial point is that in the analysis of the BGA algorithm for the UFP problem in the previous work [17] the solution of the BGA is compared with an optimal solution of a relaxed problem, namely, the *fractional* maximum multicommodity flow problem, and this problem is also a relaxation for both the ISF and the k-SFP. It follows that the approximation guarantee $O(F)$ of the BGA proved for the UFP problem holds for the k-SFP and the ISF problems as well. □

Using the standard techniques mentioned earlier, the algorithm can be converted into a randomized online algorithm with the same expected competitive ratio. If there is a guarantee that the ratio between the maximal and the minimal demand is at most 2 (or some other constant) or that the maximal demand is at most 1/2 (or some other constant smaller than 1, the edge capacity), the online algorithm can be made even deterministic with the same competitive ratio (cf. [18]). Taking into account the online lower bound of Theorem 2.6, this shows that the k-SFP and the ISF are indeed simpler problems than the k-DFP.

The techniques of the current paper imply results for the ISF even when the no-bottleneck assumption does not hold and only the weak bottleneck assumption is guaranteed (i.e., the maximal demand is at most k times larger than the minimal edge capacity). Under this assumption, on unit-capacity networks the ISF resembles the multi-EDP problem from section 2.4, and it is possible to use the multi-BGA algorithm for it and get the same guarantee as in Theorem 2.8.

COROLLARY 3.9. *Given a unit-capacity network G with flow number F, the competitive ratio of the multi-BGA for the ISF under the weak bottleneck assumption is $O(d_{\max}^3 F)$.*

4. Conclusion.

In this paper we introduced the k-EDP and the k-DFP problems and presented upper and lower bounds for them as well as for other related problems. Many questions remain open. For example, what is the best competitive ratio a deterministic algorithm can achieve for the k-EDP? We suspect that it is $O(kF)$, but it seems very hard to prove. Is it possible to simplify the proof for the k-DFP and improve the upper bound? We suspect that it should be possible to prove an $O(kF)$ upper bound here as well. Even an improvement of the $O(k^3F\log(kF))$ bound k-DFP to $O(k^3F)$ would be interesting.

Acknowledgments. We would like to thank Rakesh Sinha for bringing these problems to our attention and Alan Frieze for helpful insights. We are also grateful to the journal's anonymous referees for their useful comments.

REFERENCES

[1] B. Awerbuch, Y. Azar, and S. Plotkin, *Throughput-competitive on-line routing*, in Proceedings of the 34th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 1993, pp. 32–40.

[2] Y. Azar and O. Regev, *Combinatorial algorithms for the unsplittable flow problem*, Algorithmica, 44 (2006), pp. 49–66.

[3] A. Bagchi, A. Chaudhary, and P. Kolman, *Short length Menger's theorem and reliable optical routing*, Theoret. Comput. Sci., 339 (2005), pp. 315–332.

[4] G. Baier, E. Köhler, and M. Skutella, *On the k-splittable flow problem*, in Proceedings of the 10th Annual European Symposium on Algorithms (ESA), Lecture Notes in Comput. Sci. 2461, Springer, Berlin, 2002, pp. 101–113.

[5] A. Baveja and A. Srinivasan, *Approximation algorithms for disjoint paths and related routing and packing problems*, Math. Oper. Res., 25 (2000), pp. 255–280.

[6] A. Bley, *On the complexity of vertex-disjoint length-restricted path problems*, Comput. Complexity, 12 (2003), pp. 131–149.

[7] B. Bollobás, *Modern Graph Theory*, Springer, New York, 1998.

[8] A. Chakrabarti, C. Chekuri, A. Gupta, and A. Kumar, *Approximation algorithms for the unsplittable flow problem*, in Proceedings of the 5th International Workshop on Approxi-

mation Algorithms for Combinatorial Optimization (APPROX), Lecture Notes in Comput. Sci. 2462, Springer, Berlin, 2002.

[9] C. Chekuri, S. Khanna, and F. B. Shepherd, *The all-or-nothing multicommodity flow problem*, in Proceedings of the 36th annual ACM Symposium on Theory of Computing (STOC), 2004, pp. 156–165.

[10] C. Chekuri, S. Khanna, and F. B. Shepherd, *An $o(\sqrt{n})$ approximation and integrality gap for disjoint paths and unsplittable flow*, Theory of Comput., 2 (2006), pp. 137–146.

[11] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, and A. Schrijver, *Combinatorial Optimization*, John Wiley, New York, 1997.

[12] P. Erdős and L. Lovász, *Problems and results on* 3*-chromatic hypergraphs and some related questions*, in Infinite and Finite Sets (to Paul Erdős on his 60th birthday), A. Hajnal, R. Rado, and V. Sós, eds., North–Holland, Amsterdam, 1975, pp. 609–627.

[13] V. Guruswami, S. Khanna, R. Rajaraman, B. Shepherd, and M. Yannakakis, *Near-optimal hardness results and approximation algorithms for edge-disjoint paths and related problems*, J. Comput. Syst. Sci., 67 (2003), pp. 473–496.

[14] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc., 58 (1963), pp. 13–30.

[15] J. Kleinberg, *Approximation Algorithms for Disjoint Paths Problems*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1996.

[16] R. Koch, M. Skutella, and I. Spenke, *Approximation and complexity of k-splittable flows*, in Proceedings of the 3rd Workshop on Approximation and Online Algorithms (WAOA), Lecture Notes in Comput. Sci. 3879, Springer, Berlin, 2005, pp. 244–257.

[17] P. Kolman and C. Scheideler, *Improved bounds for the unsplittable flow problem*, J. Algorithms, 61 (2006), pp. 20–44.

[18] P. Kolman and C. Scheideler, *Simple on-line algorithms for the maximum disjoint paths problem*, Algorithmica, 39 (2004), pp. 209–233.

[19] F. T. Leighton, B. M. Maggs, and S. B. Rao, *Packet routing and job-shop scheduling in O(congestion + dilation) steps*, Combinatorica, 14 (1994), pp. 167–186.

[20] M. Martens and M. Skutella, *Flows on few paths: Algorithms and lower bounds*, Networks, 48 (2006), pp. 68–76.

[21] M. Martens and M. Skutella, *Length-bounded and dynamic k-splittable flows*, in Operations Research Proceedings 2005, Springer, Berlin, 2006, pp. 297–302.

[22] C. Scheideler, *Universal Routing Strategies for Interconnection Networks*, Lecture Notes in Comput. Sci. 1390, Springer, Berlin, 1998.

SIAM J. DISCRETE MATH.
Vol. 21, No. 1, pp. 158–169

THE COMPLEXITY OF COMBINATORIAL OPTIMIZATION PROBLEMS ON d-DIMENSIONAL BOXES*

MIROSLAV CHLEBÍK† AND JANKA CHLEBÍKOVÁ‡

Abstract. The MAXIMUM INDEPENDENT SET problem in d-box graphs, i.e., in intersection graphs of axis-parallel rectangles in $\mathbb{R}^d$, is known to be NP-hard for any fixed $d \geq 2$. A challenging open problem is that of how closely the solution can be approximated by a polynomial time algorithm. For the restricted case of d-boxes with bounded aspect ratio a PTAS exists [T. Erlebach, K. Jansen, and E. Seidel, *SIAM J. Comput.*, 34 (2005), pp. 1302–1323]. In the general case no polynomial time algorithm with approximation ratio $o(\log^{d-1} n)$ for a set of n d-boxes is known. In this paper we prove APX-hardness of the MAXIMUM INDEPENDENT SET problem in d-box graphs for any fixed $d \geq 3$. We give an explicit lower bound $\frac{245}{244}$ on efficient approximability for this problem unless P = NP. Additionally, we provide a generic method how to prove APX-hardness for other graph optimization problems in d-box graphs for any fixed $d \geq 3$.

Key words. independent set, geometric intersection graphs, rectangle graphs

AMS subject classifications. 68Q25, 68R10

DOI. 10.1137/050629276

1. Introduction.

Many optimization problems like MAXIMUM CLIQUE, MAXIMUM INDEPENDENT SET, and MINIMUM (VERTEX) COLORING are NP-hard in general graphs but solvable in polynomial time in interval graphs [14]. However, many of the problems, e.g., MAXIMUM INDEPENDENT SET [13], [16] and MINIMUM COLORING [21], are known to be NP-hard already in 2-dimensional models of geometric intersection graphs as in unit disk graphs or in intersection graphs of axis-parallel rectangles in $\mathbb{R}^d$ for any fixed $d \geq 2$ (for short, d-box intersection graphs or d-box graphs). Among basic NP-hard graph optimization problems, only MAXIMUM CLIQUE is known to be solvable in polynomial time in d-box graphs [4], [20], [25]. In most cases geometric restrictions on input instances allow one to obtain better approximation algorithms for problems that are extremely hard to approximate in general graphs. On the other hand, geometric restrictions make the task of achieving hardness results more difficult.

The most studied problem in d-box intersection graphs, MAXIMUM INDEPENDENT SET (MAX-IS), can be formulated as follows: for a given set $\mathcal{R}$ of n axis-parallel d-dimensional boxes (for short, d-boxes) find a maximum cardinality subset $\mathcal{R}^* \subseteq \mathcal{R}$ of pairwise disjoint boxes. The problem has attracted the attention of many researchers (e.g., [1], [5], [6], [12], [15], [17], [24]) due to its applications in map labeling, data mining, VLSI design, image processing, and point location in d-dimensional Euclidean space. As the problem is NP-hard for any fixed $d \geq 2$ [13], [16], attention is focused on efficient approximation algorithms. Let us briefly describe known approximability results for it; a more detailed overview of them can be found in [6]. The earliest result

*Received by the editors April 15, 2005; accepted for publication (in revised form) April 18, 2006; published electronically February 26, 2007. An extended abstract of this work appeared in the Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2005, pp. 267–276.

http://www.siam.org/journals/sidma/21-1/62927.html

†Max Planck Institute for Mathematics in the Sciences, Inselstraße 22-26, D-04103 Leipzig, Germany (chlebik@mis.mpg.de).

‡Faculty of Mathematics, Physics, and Informatics, Comenius University, Mlynská dolina, 842 48 Bratislava, Slovakia (chlebikova@fmph.uniba.sk). The work of this author was supported by VEGA grant 1/3106/06.

was a shifting grid method based PTAS by Hochbaum and Maass [15] in the case of unit d-cubes. This method works for any collection of *fat* objects in $\mathbb{R}^d$ of roughly the same size, and it requires $n^{O(k^{d-1})}$ time to guarantee an approximation factor of $(1+\frac{1}{k})$. Moreover, this approach can be generalized to objects not necessarily fat but whose projections to the last $(d-1)$ coordinates are fat and of roughly the same size. This was essentially established by Agarwal, van Kreveld, and Suri [1] in their work on unit-height rectangles in $\mathbb{R}^2$. Generalizing in another direction, Erlebach, Jansen, and Seidel [12] and Chan [6] obtained a PTAS for fat objects of possibly varying sizes, such as arbitrary d-cubes or bounded aspect ratio d-boxes. For *arbitrary* d-boxes, even for $d = 2$, the existence of a PTAS or a constant factor approximation is an open problem. As has been observed in several papers [1], [17], a logarithmic approximation factor is possible in this case. For example, the results of Agarwal, van Kreveld, and Suri [1] imply a $O(n \log_2^{d-1} n)$-time algorithm with factor at most $\lceil \log_2 n \rceil^{d-1}$. Nielsen [24] independently described an algorithm with optimum-sensitive approximation factor $(1 + \log_2(is(\mathcal{R})))^{d-1}$, where $is(\mathcal{R})$ is the maximum number of independent boxes of $\mathcal{R}$. Currently, no polynomial time algorithm is known with $o(\log^{d-1} n)$-approximation factor, although Berman et al. [5] have observed that a $\log_2^{d-1} n$ bound can be reduced by arbitrary multiplicative constant. However, in spite of many efforts, understanding the limits on the approximability of the Maximum Independent Set problem in intersection graphs of d-boxes remains an open problem.

1.1. Our results. In this paper we present the proof of APX-hardness for the Maximum Independent Set problem in axis-parallel d-dimensional boxes for any fixed $d \geq 3$. It follows, in particular, that for any fixed $d \geq 3$ the existence of a PTAS for the problem restricted to d-boxes with bounded aspect ratio [12] cannot be generalized to arbitrary axis-parallel d-boxes, unless P = NP.

The idea of our proof is based on the following two results:

(i) In section 3 we observe that Maximum Independent Set, Minimum Vertex Cover, and some other graph optimization problems are APX-hard even in certain subdivisions of graphs with low maximum degree. For example, for any fixed integer $k \geq 0$ the Maximum Independent Set problem is APX-hard in graphs obtained from 3-regular graphs by $2k$ subdivision of each edge.

(ii) In section 2 we prove that each graph obtained from another one by at least 2-subdivision of each edge is an intersection graph of axis-parallel d-boxes for any fixed $d \geq 3$. Moreover, a d-box intersection representation of such graphs can be provided in polynomial time.

Both results (i) and (ii) are very general and can be of independent interest. Using them we provide a method how to achieve approximation hardness results in d-box graphs for other graph optimization problems, e.g., for covering and domination problems. The method used allows us to provide also explicit lower bounds on efficient approximability. This is demonstrated on the problems Maximum Independent Set and Minimum Vertex Cover in d-box graphs (for any fixed $d \geq 3$) proving NP-hardness to achieve an approximation factor of $1 + \frac{1}{244}$ and $1 + \frac{1}{249}$, respectively. One can notice that the best known approximation algorithms for graph optimization problems in d-box graphs assume that an intersection representation of an input graph by d-boxes is given. Therefore it should be emphasized that our hardness results apply to this setting as well. Moreover, they hold for instances in which no point of $\mathbb{R}^d$ is simultaneously covered by more than two d-boxes and each d-box intersects at most three others.

1.2. Definitions and notations. Recall that a *d-dimensional box* (for short, d-box) is a subset of $\mathbb{R}^d$ that is a Cartesian product of d intervals in $\mathbb{R}$. For convenience, terms an *interval* and a *rectangle* are used for a 1-box and a 2-box, respectively.

Definition 1. *The intersection graph of a family of sets S_v, $v \in V$, is a graph with vertex set V such that for any $u, v \in V$ a vertex u is adjacent to a vertex v if and only if $S_u \cap S_v \neq \emptyset$. The family $\{S_v, v \in V\}$ is an intersection representation of the intersection graph. The intersection graphs of families of axis-parallel d-dimensional boxes are called* d-box intersection graphs *or simply* d-box graphs.

Definition 2. *Let G be a simple graph with vertex set V and edge set E. If G contains a cycle, then the* girth *of G is the length of its shortest cycle. A vertex $v \in V$ is said to* cover *itself, all edges incident with v, and all vertices adjacent to v. An edge $\{u, v\} \in E$ is said to* cover *itself, vertices u and v, and all edges incident with u or v. Two elements of $V \cup E$ are* independent *if neither covers the other.*

For a graph G, a vertex cover *is a subset of V that covers all edges E, a* dominating set *is a subset of V that covers all vertices V, and an* edge dominating set *is a subset of E that covers all edges E.*

The goal of the Maximum Independent Set problem is to find an independent set of maximum cardinality in a graph G; let $is(G)$ denote its cardinality. The Minimum Vertex Cover problem (Min-VC) asks us to find a vertex cover of minimum cardinality in G; let $vc(G)$ denote its optimum value. The problems Minimum Dominating Set (Min-DS), Minimum Independent Dominating Set (Min-IDS), and Minimum Edge Dominating Set (Min-EDS) ask for a dominating set, an independent dominating set, and an edge dominating set of minimum size in G, respectively. Let $ds(G)$, $ids(G)$, and $eds(G)$ stand, respectively, for the corresponding minima.

Definition 3. *Let $G = (V, E)$ be a given graph. For an integer $k \geq 0$, a* k-subdivision *of an edge $e = \{u, v\} \in E$ in G is defined as a replacement of e by a path with endvertices u and v and with k new internal vertices. A k-subdivision of G, denoted by $\operatorname{div}_k(G)$, is a graph obtained from G by a k-subdivision of each edge e from E. (All added paths are pairwise disjoint.)*

We will consider also subdivisions of $G = (V, E)$ that are not uniform but are edge dependent. In such case an edge function $s := s_G$ from E to nonnegative integers will be given and the resulting graph will be obtained by $s(e)$-subdivision of each edge $e \in E$.

For the basic optimization terminology we refer the reader to Ausiello et al. [3]. For any NPO optimization problem Q, I_Q is the set of instances of Q, $\mathrm{sol}_Q(x)$ is the set of feasible solutions for $x \in I_Q$, and for each pair (x, y) such that $x \in I_Q$ and $y \in \mathrm{sol}_Q(x)$, $m_Q(x, y)$ is the value of a feasible solution y. The optimal value for an instance $x \in I_Q$ is denoted by $\mathrm{OPT}_Q(x)$.

Definition 4. *Let Q and Q' be two* NPO *problems and f be a polynomial time computable function that maps instances of Q to instances of Q'. Then f is said to be an* L-reduction *from Q to Q', if there are constants α, $\beta \in (0, \infty)$ and a polynomial time computable function g such that for every $x \in I_Q$ (i) $\mathrm{OPT}_{Q'}(f(x)) \leq \alpha \mathrm{OPT}_Q(x)$, (ii) for every $y' \in \mathrm{sol}_{Q'}(f(x))$, $g(x, y') \in \mathrm{sol}_Q(x)$ so that $|\mathrm{OPT}_Q(x) - m_Q(x, g(x, y'))| \leq \beta |OPT_{Q'}(f(x)) - m_{Q'}(f(x), y')|$.*

To show APX-completeness of a problem $Q \in$ APX it is enough to show that there is an L-reduction from some APX-complete problem to Q.

Remark 1. Let us recall that all problems Maximum Independent Set, Minimum Vertex Cover, Minimum Dominating Set, Minimum Edge Dominating Set, and Minimum Independent Dominating Set are APX-complete in bounded degree graphs. Their inclusion in APX follows from easy counting arguments, when restricted to graphs of degree at most B, $B \geq 3$, $is(G) \geq ids(G) \geq ds(G) \geq \frac{|V|}{B+1}$, $vc(G) \geq \frac{|V|}{B+1}$, and $eds(G) \geq \frac{|V|}{2B}$. (For some of these inequalities it is necessary to confine ourselves to graphs without isolated vertices.) Hence for any of the above *minimization* problems in bounded degree graphs any feasible solution approximates the optimal one within a constant. For Maximum Independent Set, the lower bounds given above apply to any inclusionwise maximal independent set. This provides a constant factor approximation in all cases. In most cases the proof of APX-hardness even in 3-regular graphs is known (see [2], [28], [23], and references therein).

2. Intersection graphs of axis-parallel boxes. Roberts [26] proved that each graph can be realized as an intersection graph of axis-parallel d-dimensional boxes for some d depending on the graph. For any fixed $d \geq 2$, the recognition of d-box graphs is NP-hard [18], [27], and hence the reconstruction of their representation by d-boxes is NP-hard as well. In this section we prove that highly nontrivial subclasses of general graphs are d-box graphs for any $d \geq 3$. Namely, each graph obtained from another one by at least 2-subdivision of each edge is an intersection graph of d-boxes for any fixed $d \geq 3$ and its intersection representation can be found in polynomial time.

Theorem 1. *Let $G = (V, E)$ be a graph, and let an integer $s(e) \geq 2$ be given for each edge $e \in E$. Denote by G' a graph obtained from G by a $s(e)$-subdivision of each edge e. Then for any fixed integer $d \geq 3$, the graph G' can be realized as an intersection graph of a set of axis-parallel d-dimensional boxes. Moreover, such realization can be done in time polynomial in $|V| + \sum_e s(e)$.*

Proof. Let $G = (V, E)$, $s : E \to \{2, 3, \dots\}$, and G' be given as above. First, we describe the realization of G' as an intersection graph of a set $\{R_1, R_2, \dots, R_N\}$ of axis-parallel boxes in $\mathbb{R}^3$, where $N = |V| + \sum_e s(e)$.

We can assume that $V = \{1, 2, \dots, |V|\}$ and assign each edge $e \in E$ a number n_e using a bijection $e \in E \mapsto n_e \in \{1, 2, \dots, |E|\}$ between E and $\{1, 2, \dots, |E|\}$. Each vertex $i \in \{1, 2, \dots, |V|\}$ will be represented by a 3-box $R_i = [2i-1, 2i] \times [2i-1, 2i] \times [1, 2|E|]\}$ (see Figure 1).

The graph G' is obtained from G replacing each edge $e = \{i, j\} \in E$ (assume $i < j$) by a path with vertices i, A_e^1, A_e^2, ... , $A_e^{s(e)}$, j. Now we define the boxes R_e^1, ... , $R_e^{s(e)}$ representing vertices A_e^1, A_e^2, ... , $A_e^{s(e)}$, respectively. The projection on the third coordinate axis is chosen to be $[2n_e - 1, 2n_e]$ to ensure that no two boxes R_e^i and $R_{e'}^j$, which correspond to distinct edges e and e', intersect. More precisely, define $R_e^1 := [2i-1, 2j] \times [2i-1, 2i] \times [2n_e - 1, 2n_e]$, and further put $R_e' := [2j-1, 2j] \times [2i-1, 2j] \times [2n_e - 1, 2n_e]$ (see Figure 1). If $s(e) = 2$, one can simply put $R_e^2 := R_e'$. If $s(e) \geq 3$, then boxes R_e^2, R_e^3, ... , $R_e^{s(e)}$ will be taken as subboxes of R_e' of the form $R_e^l := [2j-1, 2j] \times [c_l, d_l] \times [2n_e - 1, 2n_e]$ for $l = 2, 3, \dots, s(e)$, where c_l, d_l are rationals such that $c_2 = 2i-1$, $d_{s(e)} = 2j$, $2i < c_3 < d_2$, $c_{s(e)} < d_{s(e)-1} < 2j-1$ and, if $s(e) \geq 4$, $c_{l+1} < d_l < c_{l+2}$ whenever $2 \leq l \leq s(e) - 2$ (see Figure 1). One can easily check that the intersection graph of the set $\{R_1, R_2, \dots, R_{|V|}\} \cup \bigcup_{e \in E} \{R_e^1, R_e^2, \dots, R_e^{s(e)}\}$ of axis-parallel boxes in $\mathbb{R}^3$ is (isomorphic to) G'. Moreover, the time complexity of this construction is polynomial in $|V| + \sum_e s(e)$.

To obtain the corresponding realization in $\mathbb{R}^d$ for $d > 3$, one can take the set $\{R_i \times [0, 1]^{d-3} : i = 1, 2, \dots, N\}$ of d-boxes. □

Remark 2. The graph G' from Theorem 1 is of girth at least 9. In any realization of G' by axis-parallel d-dimensional boxes no point of $\mathbb{R}^d$ is simultaneously covered by more than two boxes. For the 2-dimensional case a $4K$-approximation algorithm is

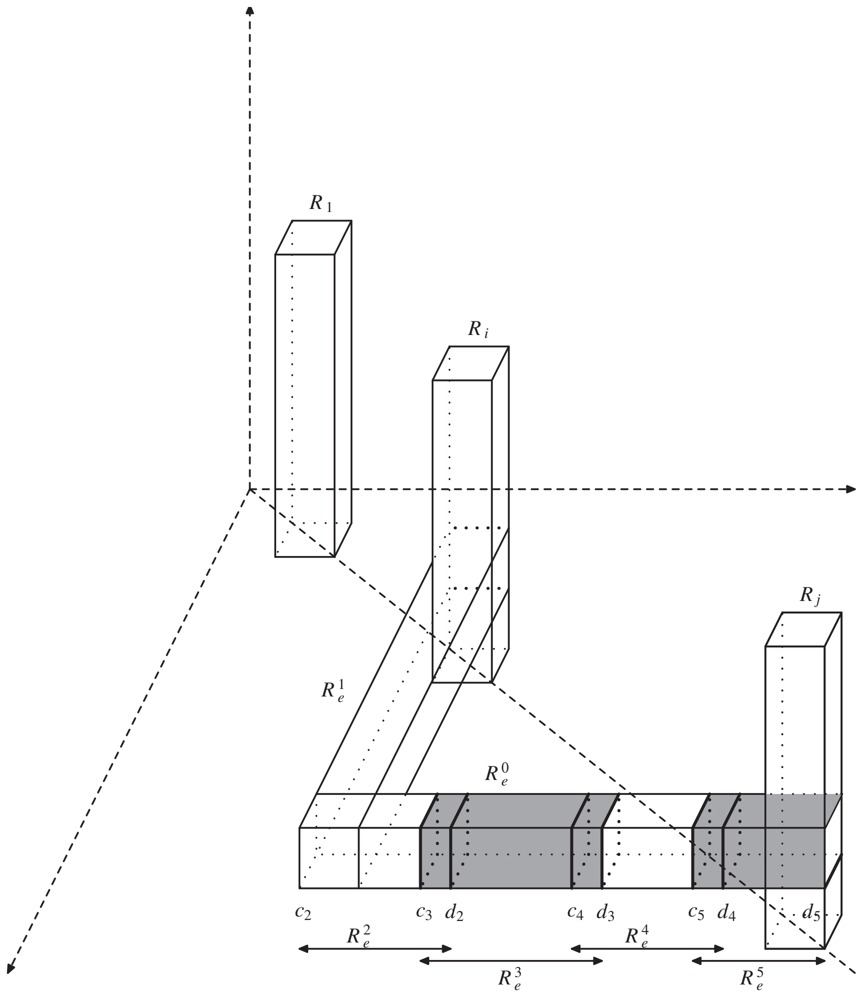


Fig. 1.

known for finding a maximum weighted independent set in a given set $\mathcal{R}$ of weighted axis-parallel rectangles, where K is the maximum number of rectangles in $\mathcal{R}$ that simultaneously cover a point in $\mathbb{R}^2$ [22].

Theorem 1 shows that for every fixed $d \geq 3$ the intersection graphs of sets of d-boxes are from a topological point of view as complex as general graphs. It is far from clear whether 2-box graphs have much simpler topological structure. However, the complexity of intersection graphs of axis-parallel lines significantly differs in dimensions 2 and 3. In the following theorem we show that, similar to the case of axis-parallel boxes, highly nontrivial subclasses of general graphs are already intersection graphs of sets of axis-parallel lines in $\mathbb{R}^d$ for any $d \geq 3$. On the other hand, for the 2-dimensional case intersection graphs of axis-parallel lines are exactly complete bipartite graphs, for which classical optimization problems are easily solvable.

Theorem 2. *Let $G = (V, E)$ be a given graph. Suppose that for each edge $e \in E$ an integer $s(e)$ with $s(e) \in \{2, 3\} \cup \{k : k \geq 5\}$ is given; denote by G' a graph obtained from G by a $s(e)$-subdivision of each edge e. Then the graph G' can be realized as an*

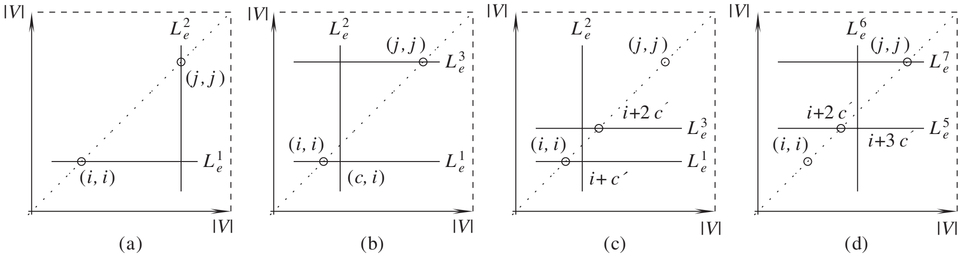


FIG. 2. *The realization of lines belonging to an edge $e = \{i, j\}$ for different values of $s(e)$. The lines parallel to z-axes are displayed as a circle in the corresponding vertex.* (a) *$s(e) = 2$, the cross-section with the plane $z = n_e$.* (b) *$s(3) = 3$, the cross-section with the plane $z = n_e$.* (c) *and* (d) *$s(3) = 7$, the cross-section with the planes $z = n_e$ and $z = n_e + c'$, respectively.*

intersection graph of a set of axis-parallel lines in $\mathbb{R}^3$. Moreover, such realization can be done in time polynomial in $|V| + \sum_e s(e)$.

Proof. Let a graph $G = (V, E)$, $s : E \to \{2, 3, 5, 6, \dots\}$, and G' be given as above. In what follows we describe the realization of G' as an intersection graph of a set of N axis-parallel lines in $\mathbb{R}^3$, where $N = |V| + \sum_e s(e)$. Assume that $V = \{1, 2, \dots, |V|\}$, and number the edges from E by a bijection $e \mapsto n_e$ between E and $\{1, 2, \dots, |E|\}$. Each vertex $i \in \{1, 2, \dots, |V|\}$ will be represented as the line $L_i = (i, i, \cdot)$ parallel to the z-axis.

The graph G' is obtained from G, replacing each edge $e = \{i, j\} \in E$ (assume $i < j$) by a path with vertices i, A_e^1, A_e^2, $\dots$, $A_e^{s(e)}$, j. Keeping one such e fixed, we define lines L_e^1, L_e^2, $\dots$, $L_e^{s(e)}$ representing vertices A_e^1, A_e^2, $\dots$, $A_e^{s(e)}$, respectively.

(a) Assume first that $s(e) \in \{2, 3, 7\}$. In all these three cases, we take as L_e^1 the line $(\cdot, i, n_e)$. If $s(e) = 2$, we put $L_e^2 := (j, \cdot, n_e)$ (see Figure 2(a)). If $s(e) = 3$, then we take $L_e^2 = (c, \cdot, n_e)$ and $L_e^3 = (\cdot, j, n_e)$ for some $i < c < i + 1$ (see Figure 2(b)). In case $s(e) = 7$, let $L_e^2 = (i + c', \cdot, n_e)$, $L_e^3 = (\cdot, i + 2c', n_e)$, $L_e^4 = (i + 2c', i + 2c', \cdot)$, $L_e^5 = (\cdot, i+2c', n_e+c')$, $L_e^6 = (i+3c', \cdot, n_e+c')$, $L_e^7 = (\cdot, j, n_e+c')$ for some $0 < c' < \frac{1}{4}$ (see Figure 2(c) and (d)).

(b) Assume now that $s(e) = a + 3m$ for some $a \in \{2, 3, 7\}$ and $m \geq 1$. We will proceed in two steps. In the first one we realize $3m$ subdivision of e, which reduces the task to the above case of a-subdivision, where $a \in \{2, 3, 7\}$. Choose $i^{(1)} < i^{(2)} < \cdots < i^{(m)}$ from $(i, i+1)$ (we can ensure that for distinct edges e, e' these sets are disjoint), and $n_e < n^{(1)} < n^{(2)} < \cdots < n^{(m)} < n_e + 1$. Take $L_e^1 := (\cdot, i, n_e)$, $L_e^2 := (i^{(1)}, \cdot, n_e)$, $L_e^3 := (i^{(1)}, i^{(1)}, \cdot)$, $L_e^4 := (\cdot, i^{(1)}, n^{(1)})$, $L_e^5 := (i^{(2)}, \cdot, n^{(1)})$, $L_e^6 := (i^{(2)}, i^{(2)}, \cdot)$, $\dots$, $L_e^{3m-2} := (\cdot, i^{(m-1)}, n^{(m-1)})$, $L_e^{3m-1} := (i^{(m)}, \cdot, n^{(m-1)})$, $L_e^{3m} := (i^{(m)}, i^{(m)}, \cdot)$. Now, in the second step, it suffices to insert a lines, where $a \in \{2, 3, 7\}$. The construction is the same as in (i), but the role of $(i, i, \cdot)$ and n_e is now played by $(i^{(m)}, i^{(m)}, \cdot)$ and $n^{(m)}$, respectively. It is also easy to see that parameters can be chosen in such a way that the intersection graph of set $\{L_1, L_2, \dots, L_{|V|}\} \cup \bigcup_{e \in E} \{L_e^1, L_e^2, \dots, L_e^{s(e)}\}$ is (isomorphic to) G'. Moreover, time complexity of the construction is polynomial in $|V| + \sum_e s(e)$. □

3. Approximation hardness results in subdivisions of graphs. Let $\mathcal{C}$ denote the collection of the following problems: MAXIMUM INDEPENDENT SET, MINIMUM VERTEX COVER, MINIMUM DOMINATING SET, MINIMUM EDGE DOMINATING SET, and MINIMUM INDEPENDENT DOMINATING SET. Each problem from $\mathcal{C}$ is well known to be APX-complete when restricted to graphs of degree at most 3 or even to

3-regular graphs (see Remark 1). Moreover, explicit NP-hard gap-type results and explicit lower bounds on their efficient approximability are known for several of them [10], [7], [8]. In this section we show APX-completeness for each problem from $\mathcal{C}$ even when restricted to certain subdivisions of low-degree graphs.

First we prove that for Maximum Independent Set and Minimum Vertex Cover the optimum value for a graph and for its certain subdivisions are in a simple relation.

Lemma 1. *Let $G = (V, E)$ be a graph, and let $e \in E$ be a given edge. Denote by G' a graph obtained from G by a 2-subdivision of the edge e. Then $vc(G') = vc(G) + 1$ and $is(G') = is(G) + 1$.*

Proof. Suppose that the edge $e = \{u, v\}$ is replaced by a path u, u', v', and v with new vertices u' and v'. For every vertex cover C in G either $C \cup \{u'\}$ or $C \cup \{v'\}$ is the vertex cover in G'; hence $vc(G') \leq vc(G) + 1$.

Now we prove the opposite inequality $vc(G) \leq vc(G') - 1$. Let C' be a vertex cover in G'. We can modify it to a vertex cover C in G with $|C| \leq |C'| - 1$ as follows. If $C' \cap \{u, v\} \neq \emptyset$, we take $C := C' \setminus \{u', v'\}$. If $C' \cap \{u, v\} = \emptyset$, then clearly both $u', v' \in C'$ and we take $C := \{u\} \cup C' \setminus \{u', v'\}$.

The claim for independent sets follows in a straightforward way. □

It is easy to see that the proof of Lemma 1 is constructive and that the corresponding algorithm applies to all feasible solutions and not only to optimal ones. Applying iteratively its steps we can obtain the following theorem.

Theorem 3. *Let $G = (V, E)$ be a graph, and for each edge $e \in E$ let an integer $s(e) \geq 0$ be given. Denote by G' a graph obtained from G by a $2s(e)$-subdivision of each edge $e \in E$. Let Q be either the problem* Minimum Vertex Cover *or* Maximum Independent Set. *Then*

(A) $\mathrm{OPT}_Q(G') = \mathrm{OPT}_Q(G) + \sum_e s(e)$;

(B) *every $y \in \mathrm{sol}_Q(G)$ can be transformed in polynomial time (in size of G and $\sum_e s(e)$) to $y' \in \mathrm{sol}_Q(G')$ such that $|y'| = |y| + \sum_e s(e)$;*

(C) *every $y' \in \mathrm{sol}_Q(G')$ can be transformed in polynomial time to $y \in \mathrm{sol}_Q(G)$ such that $|y'| - \sum_e s(e) \leq |y|$ if Q is a maximization problem (respectively, $|y| \leq |y'| - \sum_e s(e)$ if Q is minimization problem).*

Proof. We can assume that Q is the Minimum Vertex Cover problem (and for Maximum Independent Set we can argue analogously).

Let $K := \sum_e s(e)$ and assume that $K > 0$. We can find a sequence of graphs $G_0 := G$, G_1, ... , $G' := G_K$ such that for each $i = 1, 2, \ldots, K$ the graph G_i is created from G_{i-1} as in Lemma 1 (by a 2-subdivision of one of its edge). To prove the property (B), consider a vertex cover C in G. Put $C_0 := C$ and as in the proof of Lemma 1 find, for each $i = 1, 2, \ldots, K$, a vertex cover C_i in G_i with $|C_i| = |C_{i-1}| + 1$. Then $C' := C_K$ is a vertex cover in G' with $|C'| = |C| + K$. This also shows that $vc(G') \leq vc(G) + K$. To prove the property (C), consider a vertex cover C' in G'. Now as in the proof of Lemma 1 find, for each $i = K, K - 1, \ldots, 2, 1$, a vertex cover C_{i-1} in G_{i-1} with $|C_{i-1}| \leq |C_i| - 1$. Then $C := C_0$ is a vertex cover in G with $|C| \leq |C'| - K$, and hence $vc(G) \leq vc(G') - K$. Consequently, $vc(G') = vc(G) + K$, and the property (A) is proved as well. □

Also the optimum of several other graph optimization problems behaves well under certain subdivision operations, similarly as for Maximum Independent Set and Minimum Vertex Cover. We will demonstrate that for Minimum Dominating Set and Minimum Edge Dominating Set.

Lemma 2. *Let $G = (V, E)$ be a given graph. Denote by G' a graph obtained* *from G by a 3-subdivision of an edge $e \in E$. Then* (i) $ds(G') = ds(G) + 1$ *and* (ii) $eds(G') = eds(G) + 1$.

Proof. Let G' be a graph obtained from G by a 3-subdivision of the edge $e = \{u, v\}$, i.e., replacing e by a path u, u', w, v', v with new vertices u', w, and v'.

(i) To prove $ds(G') \leq ds(G) + 1$, consider a dominating set D in G. Adding one of vertices u', w, v' to D we can obtain a dominating set D' in G' with $|D'| = |D| + 1$ as follows: (I) If $(u \in D \;\&\; v \in D)$ or $(u \notin D \;\&\; v \notin D)$ we take $D' := D \cup \{w\}$. (II) If $(u \in D \;\&\; v \notin D)$ we take $D' := D \cup \{v'\}$. (III) If $(v \in D \;\&\; u \notin D)$ we take $D' := D \cup \{u'\}$.

Notice that $D \subset D'$ and that the restriction of D' to the path u, u', w, v', v is an independent set. This observation will be used later in the proof of Theorem 4.

To prove $ds(G') \geq ds(G) + 1$, consider a dominating set D' in G'. We can modify it to a dominating set D in G with $|D| \leq |D'| - 1$ as follows. If $D' \cap V$ is a dominating set in G, we take $D := D' \cap V$. If $D' \cap V$ is not a dominating set in G then clearly $u, v \notin D'$, $|D' \cap \{u', w, v'\}| \geq 2$, and we take $D := \{u\} \cup D' \cap V$.

(ii) To prove $eds(G') \leq eds(G) + 1$, consider an edge dominating set M in G and denote $V(M)$ the set of end vertices of edges in M. We modify M to an edge dominating set M' in G' with $|M'| = |M| + 1$ as follows: (I) If $u \notin V(M)$ we take $M' := M \cup \{\{u', w\}\}$. (II) If $v \notin V(M)$ and $u \in V(M)$ we take $M' := M \cup \{\{v', w\}\}$. (III) If $u, v \in V(M)$ and $e \notin M$ we take $M' := M \cup \{\{u', w\}\}$. (IV) If $e \in M$ we take $M' := M \setminus \{e\} \cup \{\{u, u'\}, \{v, v'\}\}$.

To prove $eds(G') \geq eds(G) + 1$, consider an edge dominating set M' in G' and put $M_0 := M' \cap \{\{u, u'\}, \{u', w\}, \{v', w\}, \{v, v'\}\}$. Clearly $M_0 \neq \emptyset$ and if $|M_0| = 1$ then either $\{u', w\} \in M'$ or $\{v', w\} \in M'$. We can modify M' to an edge dominating set M in G with $|M| \leq |M'| - 1$ as follows. If $|M_0| \geq 2$ we take $M := M' \setminus M_0 \cup \{e\}$. If $|M_0| = 1$ we take $M := M' \setminus M_0$. □

Using steps of the proof of the previous lemma we can obtain the following theorem.

THEOREM 4. *Let $G = (V, E)$ be a graph, and for each edge $e \in E$ let an integer $s(e) \geq 0$ be given. Denote by G' a graph obtained from G by a $3s(e)$-subdivision of each edge $e \in E$. Then the properties* (A)–(C) *from Theorem* 3 *are fulfilled for both problems* MINIMUM DOMINATING SET *and* MINIMUM EDGE DOMINATING SET.

Moreover, if $s(e) > 0$ for each $e \in E$, then $ids(G') = ds(G')$ and every dominating set D in G can be transformed in polynomial time to an independent dominating set D' in G' with $|D'| = |D| + \sum_e s(e)$.

Proof. We provide the proof for the MINIMUM DOMINATING SET problem; the proof for the second problem is analogous using the corresponding part of the proof of Lemma 2. Let G' be a graph obtained from G by a $3s(e)$-subdivision of each edge e, i.e., replacing the edge $e = \{u, v\}$ by a path with endvertices u and v, and $3s(e)$ new vertices (the paths are pairwise disjoint). Let $K := \sum_e s(e)$. We can assume that $K > 0$, and find $G_0 := G$, G_1, … , $G_K := G'$ as in the proof of Theorem 3.

To prove the property (B), consider a dominating set D in G. Put $D_0 := D$ and as in the proof of Lemma 2 find, for each $i = 1, 2, \ldots, K$, a dominating set D_i in G_i such that $|D_i| = |D_{i-1}| + 1$, $D_{i-1} \subset D_i$, and the restriction of D_i to the path used to create G_i from G_{i-1} is an independent set. Then $D' := D_K$ is a dominating set in G' with $|D'| = |D| + K$. This also shows that $ds(G') = ds(G) + K$. Moreover, if $s(e) > 0$ for every $e \in E$, then the set D' is an independent dominating set in G', and $ds(G') = ids(G') \leq ds(G) + K$ in this case. To prove the property (C), consider a dominating set D' in G' and put $D_K := D'$. As in the proof of Lemma 2 find, for each $i = K, K-1, \dots, 2, 1$, a dominating set D_{i-1} in G_{i-1} with $|D_{i-1}| \le |D_i| - 1$. Then $D := D_0$ is a dominating set in G with $|D| \le |D'| - K$. This also shows that $ds(G) \le ds(G') - K$. Consequently, $ds(G') = ds(G) + K$. If $s(e) > 0$ for every $e \in E$, then as it follows from the proof, there is a minimum dominating set in G', which is also independent; hence $ids(G') = ds(G')$. □

Remark 3. In Theorem 4, if $s(e)$ is an *odd* integer for each edge e then the graph G' is bipartite.

Now using Theorems 3 and 4 we can easily prove APX-completeness of each of the basic optimization problems Maximum Independent Set, Minimum Vertex Cover, Minimum Dominating Set, Minimum Edge Dominating Set, and Minimum Independent Dominating Set even when restricted to certain subdivisions of graphs of degree at most 3.

Theorem 5. (i) *The problems* Maximum Independent Set *and* Minimum Vertex Cover *are* APX*-complete when restricted to* $2k$*-subdivisions of* 3*-regular graphs for any fixed integer* $k \ge 0$.

(ii) *The problems* Minimum Dominating Set*,* Minimum Edge Dominating Set*, and* Minimum Independent Dominating Set *are* APX*-complete when restricted to* $3k$*-subdivisions of degree at most* 3 *graphs for any fixed integer* $k \ge 0$.

Proof. Let $k \ge 0$ be a fixed integer. Without loss of generality we can consider only graphs without isolated vertices. As was mentioned in Remark 1, all considered problems are in APX when restricted to graphs of degree at most 3. Hence, to prove APX-completeness of each of the problems Maximum Independent Set, Minimum Vertex Cover, Minimum Dominating Set, Minimum Edge Dominating Set, and Minimum Independent Dominating Set restricted to certain subdivisions of low-degree graphs, it is enough to show that such subdivision operations are in bounded degree graphs in fact L-reductions to the same problems.

(i) Let us start with Max-IS and a $2k$-subdivision operation. To verify the first condition of an L-reduction we have to check that there is a constant c such that $is(\mathrm{div}_{2k}(G)) \le c \cdot is(G)$ for every graph G of maximum degree B, $B \ge 3$. As follows from Theorem 3,

$$is(\mathrm{div}_{2k}(G)) = is(G) + |E|k. \tag{3.1}$$

Recall that for a graph $G = (V, E)$ of maximum degree B the following inequalities hold: $|E| \le \frac{|V|}{2}B$ and $is(G) \ge \frac{|V|}{B+1}$. Now one can see that the choice $\alpha := 1 + \frac{B(B+1)k}{2}$ will do. The second condition from the definition of an L-reduction is satisfied with $\beta = 1$ by Theorem 3. Hence the operation that transforms a graph to its $2k$-subdivision is an L-reduction that self-reduces Max-IS restricted to graphs of maximum degree B.

We can argue similarly for Min-VC using Theorem 3 and simple lower bound $vc(G) \ge \frac{|V|}{B+1}$.

(ii) The same approach as in (i) can be used also for problems Min-DS, and Min-EDS, to prove that a $3k$-subdivision is an L-reduction for them, when restricted to graphs of maximum degree B. It is enough to consider Theorem 4 together with lower bounds $is(G) \ge ids(G) \ge ds(G) \ge \frac{|V|}{B+1}$, and $eds(G) \ge \frac{|V|}{2B}$. Moreover, a $3k$-subdivision for $k > 0$ reduces Min-DS to Min-IDS and it is again an L-reduction when restricted to graphs of maximum degree B. □

Remark 4. Notice that the theorem above shows hardness results for graphs with low maximum degree and large girth. The part (ii) for k odd claims APX-completeness results in bipartite graphs of maximum degree 3 and of girth at least $9k + 3$.

For the later applications, we formulate also the explicit NP-hard gap-type results for Maximum Independent Set and Minimum Vertex Cover restricted to certain subdivisions of low-degree graphs.

Theorem 6. *It is* NP*-hard to approximate*

(i) Maximum Independent Set *in* 2*-subdivisions of* 3*-regular graphs within* $1 + \frac{1}{387}$*, and in* 2*-subdivisions of* 4*-regular graphs within* $1 + \frac{1}{244}$*;*

(ii) Minimum Vertex Cover *in* 2*-subdivisions of* 3*-regular graphs within* $1 + \frac{1}{390}$*, and in* 2*-subdivisions of* 4*-regular graphs within* $1 + \frac{1}{249}$*.*

Proof. (i) We will use the corresponding NP-hard gap results from [10] for Maximum Independent Set in B-regular graphs, $B \geq 3$. For any $\varepsilon > 0$ it is NP-hard to decide in B-regular graphs $G = (V, E)$ whether $is(G) < \frac{|V|}{2}(1 - 3\delta_B + \varepsilon)$ or $is(G) > \frac{|V|}{2}(1 - 2\delta_B - \varepsilon)$, where δ_B is a constant for B-regular graphs, $\delta_3 \approx 0.0103305$, and $\delta_4 \approx 0.020242915$. Using the formula (3.1) we see that this translates to the following NP-hardness result for 2-subdivisions of B-regular graphs: for any $\varepsilon > 0$ it is NP-hard to decide whether $is(\mathrm{div}_2(G)) < \frac{|V|}{2}(1 + B - 3\delta_B + \varepsilon)$ or $is(\mathrm{div}_2(G)) > \frac{|V|}{2}(1 + B - 2\delta_B - \varepsilon)$. Consequently, the approximation within any constant smaller than $1 + \frac{\delta_B}{1+B-3\delta_B}$ is NP-hard.

(ii) We can argue similarly for Minimum Vertex Cover using NP-hard gap results for it in B-regular graphs, $B \geq 3$ [10]. For any $\varepsilon > 0$ it is NP-hard to decide in B-regular graphs $G = (V, E)$ whether $vc(G) < \frac{|V|}{2}(1 + 2\delta_B + \varepsilon)$ or $vc(G) > \frac{|V|}{2}(1 + 3\delta_B - \varepsilon)$, where δ_3 and δ_4 are as above. □

4. Approximation hardness results in d-box graphs. Theorem 1 shows that any graph obtained from another one by at least 2-subdivision of each edge is a d-box graph for any $d \geq 3$. This immediately implies that many optimization problems in intersection graphs of d-boxes are as hard to approximate as in general graphs. It is rather easy to reach this conclusion for such problems as Minimum Steiner Tree or Minimum Traveling Salesman. For these problems, replacing edges by pairwise disjoint paths (and splitting edge weights properly) cannot make the problem easier to approximate. But for some optimization problems the algorithms with better approximation ratios have been designed in d-box graphs rather than in general graphs.

In this section we prove APX-hardness and hence nonexistence of a PTAS (unless P = NP) for some basic graph optimization problems in d-box graphs for any $d \geq 3$. Moreover, all our hardness results apply as well to the setting when a representation by d-boxes is given as an input, not merely its intersection graph. This makes hardness results stronger, as the problem to find a d-box intersection representation of a graph is known to be NP-hard.

Theorem 7. *Let $d \geq 3$ be a fixed integer. Each of the problems* Maximum Independent Set*,* Minimum Vertex Cover*,* Minimum Dominating Set*,* Minimum Edge Dominating Set*, and* Minimum Independent Dominating Set *is* APX*-hard when restricted to intersection graphs of sets of axis-parallel d-dimensional boxes and hence does not admit* PTAS *unless* P = NP*. These hardness results apply also to instances whose intersection graph is simultaneously of maximum degree* 3 *and of girth at least k (for any prescribed constant k) and, except* Maximum Independent Set *and* Minimum Vertex Cover*, is bipartite as well.*

Proof. The proof is straightforward using Theorems 1 and 5. □

These results could be stated as explicit NP-hard gap type results and provide explicit lower bounds on its approximability. This is demonstrated on Maximum

Independent Set and Minimum Vertex Cover to show how large explicit values can be obtained with the current methods.

Theorem 8. *For any fixed $d \geq 3$ it is* NP-*hard to approximate the* Maximum Independent Set *problem within* $1+\frac{1}{244}$ *and the* Minimum Vertex Cover *problem within* $1+\frac{1}{249}$ *in sets of axis-parallel d-dimensional boxes.*

Proof. We provide the proof for Maximum Independent Set; the proof for Minimum Vertex Cover is analogous. Let $d \geq 3$ be a fixed integer. Assume that $G' = (V', E')$ is a 2-subdivision of a 4-regular graph $G = (V, E)$. As follows from Theorem 1, G' is an intersection graph of a set $\mathcal{R}$ of d-boxes and an intersection realization of $\mathcal{R}$ can be found in polynomial time. Due to Theorem 6 it is NP-hard to decide whether the maximum number of pairwise disjoint d-boxes of $\mathcal{R}$ is less than $0.49392715|V'|$ or greater than $0.495951417|V'|$ (under the premise that one of these two cases occurs). Consequently, it is NP-hard to approximate Maximum Independent Set within $1+\frac{1}{244}$ in d-boxes for $d \geq 3$. □

Remark 5. The results of Theorems 7 and 8 hold also for intersection graphs of sets of axis-parallel lines for any fixed $d \geq 3$. The proofs are the same, only Theorem 2 is used instead of Theorem 1.

The method of this paper is rather general and can provide inapproximability results also for other combinatorial optimization problems on sets of d-boxes for any $d \geq 3$ (see [9] for more details). The question of approximation hardness of these problems in the 2-dimensional case is open. However, as shown in [9], using subdivisions of planar graphs provides a generic method of proving NP-hardness of all these problems on sets of axis-parallel rectangles (even unit squares) in the plane. Similar methods how to prove NP-hardness for problems in geometric intersection graphs of planar objects have been already used in [11] for unit disk graphs and in [19] for intersection graphs of line segments.

REFERENCES

[1] P. K. Agarwal, M. van Kreveld, and S. Suri, *Label placement by maximum independent set in rectangles*, Comput. Geom., 11 (1998), pp. 209–218.

[2] P. Alimonti and V. Kann, *Some APX-completeness results for cubic graphs*, Theoret. Comput. Sci., 237 (2000), pp. 123–134.

[3] G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi, *Complexity and Approximation*, Springer, New York, 1999.

[4] E. Balas and C.-S. Yu, *On graphs with polynomially solvable maximum-weight clique problem*, Networks, 19 (1989), pp. 247–253.

[5] P. Berman, B. DasGupta, S. Muthukrishnana, and S. Ramaswami, *Efficient approximation algorithms for tiling and packing problems with rectangles*, J. Algorithms, 41 (2001), pp. 443–470.

[6] T. M. Chan, *Polynomial-time approximation schemes for packing and piercing fat objects*, J. Algorithms, 46 (2003), pp. 178–189.

[7] M. Chlebík and J. Chlebíková, *Inapproximability results for edge dominating set problems in graphs*, J. Comb. Optim., 11 (2006), pp. 279–290.

[8] M. Chlebík and J. Chlebíková, *Approximation hardness of dominating set problems*, in Proceedings of the 12th Annual European Symposium on Algorithms, ESA, Lecture Notes in Comput. Sci. 3221, Springer, New York, 2004, pp. 192–203.

[9] M. Chlebík and J. Chlebíková, *Approximation hardness of optimization problems in intersection graphs of d-dimensional boxes*, in Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 2005, pp. 267–276.

[10] M. Chlebík and J. Chlebíková, *Complexity of approximating bounded variants of optimization problems*, Theoret. Comput. Sci., 354 (2006), pp. 320–338.

[11] B. N. Clark, C. J. Colbourn, and D. S. Johnson, *Unit disk graphs*, Discrete Math., 86 (1990), pp. 165–177.

[12] T. Erlebach, K. Jansen, and E. Seidel, *Polynomial-time approximation schemes for geometric intersection graphs*, SIAM J. Comput., 34 (2005), pp. 1302–1323.

[13] R. J. Fowler, M. S. Paterson, and S. L. Tanimoto, *Optimal packing and covering in the plane are NP-complete*, Inform. Process. Lett., 12 (1981), pp. 133–137.

[14] U. I. Gupta, D. T. Lee, and J. Y.-T. Leung, *Efficient algorithms for interval graphs and circular-arc graphs*, Networks, 12 (1982), pp. 459–467.

[15] D. S. Hochbaum and W. Maass, *Approximating schemes for covering and packing problems in image processing and VLSI*, J. ACM, 32 (1985), pp. 130–136.

[16] H. Imai and T. Asano, *Finding the connected components and a maximum clique of an intersection graph of rectangles in the plane*, J. Algorithms, 4 (1983), pp. 310–323.

[17] S. Khanna, S. Muthukrishnan, and M. Paterson, *On approximate rectangle tiling and packing*, in Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 1998, pp. 384–393.

[18] J. Kratochvíl, *A special planar satisfiability problem and a consequence of its NP-completeness*, Discrete Appl. Math. 52 (1994), pp. 233–252.

[19] J. Kratochvíl and J. Nešetřil, *Independent set and clique problems in intersection-defined classes of graphs*, Comment. Math. Univ. Carolin., 31 (1990), pp. 85–93.

[20] D. T. Lee, *Maximum clique problem of rectangle graphs*, Advances in Computing Research, 1 (1983), pp. 91–107.

[21] D. T. Lee and J. Y.-T. Leung, *On the 2-dimensional channel assignment problem*, IEEE Trans. Comput., 33 (1984), pp. 2–6.

[22] L. Lewin-Eytan, J. Naor, and A. Orda, *Routing and admission control in networks with advance reservations*, in Proceedings of the 5th International Workshop on Approximation Algorithms for Combinatorial Optimization, Lecture Notes in Comput. Sci. 2462, Springer, New York, 2002, pp. 215–228.

[23] D. F. Manlove, *On the algorithmic complexity of twelte covering and independence parameters of graphs*, Discrete Appl. Math., 91 (1999), pp. 155–175.

[24] F. Nielsen, *Fast stabbing of boxes in high dimensions*, Theoret. Comput. Sci., 246 (2000), pp. 53–72.

[25] E. Prisner, *Graphs with few cliques*, in Graph Theory, Combinatorics, and Applications, Proceedings of 7th Conference on the Theory and Applications of Graphs, Y. Alavi and A. Schwenk, eds., John Wiley and Sons, New York, 1995, pp. 945–956.

[26] F. S. Roberts, *On boxicity and cubicity of a graph*, in Recent Progress in Combinatorics, W. T. Tutte, ed., Academic Press, New York, 1969, pp. 301–310.

[27] M. Yannakakis, *The complexity of the partial order dimension problem*, SIAM J. Alg. Discrete Meth., 3 (1982), pp. 351–358.

[28] M. Yannakakis and F. Gavril, *Edge dominating sets in graphs*, SIAM J. Appl. Math., 38 (1980), pp. 364–372.

SIAM J. Discrete Math.
Vol. 21, No. 1, pp. 170–177

VERTEX-MAGIC TOTAL LABELINGS OF REGULAR GRAPHS*

IAN D. GRAY†

Abstract. In this paper it is shown that if a graph G possesses a spanning subgraph H with a strong vertex magic total labeling (VMTL) and $G - E(H)$ is even-regular, then G also has a strong VMTL. Among other things, this is used to conclude that all Hamiltonian regular graphs of odd order possess strong VMTLs. A relationship is then demonstrated between regular graphs of even degree and sparse magic squares. We next consider cubic graphs of order $2n$ consisting of two 2-factors of order n, connected by a 1-factor (*quasi-prisms*). Based on McQuillan's construction of VMTLs of such 3-regular graphs, VMTLs are derived for similar regular graphs of any odd degree. Finally, a construction is given for VMTLs of quartic graphs of order $4n+2$ consisting of two cycles of odd order n connected by a 2-factor (*simple quasi-anti-prisms*), and based on this construction VMTLs are derived for similar regular graphs of any even degree.

Key words. vertex-magic, regular graph, quasi-prism, quasi-anti-prism, sparse magic square

AMS subject classification. 05C78

DOI. 10.1137/050639594

1. VMTLs of regular graphs.

All graphs in this paper are finite, simple, and undirected. The graph G has vertex-set $V(G)$ and edge-set $E(G)$.

A *vertex magic total labeling (VMTL)* of a graph G is a mapping of the integers $1, \ldots, |V(G)| + |E(G)|$ onto the vertices and edges of G in such a way that the sum of the integers assigned to any vertex and its incident edges is the same constant k regardless of the vertex chosen. The constant k is referred to as the *magic constant.* A *strong VMTL* is a VMTL where the largest integer labels are assigned to the vertices.

Vertex magic total labelings have received a great deal of attention is recent years. Conditions have been established which rule out VMTLs for several infinite families of graphs including wheels of order greater than 11 [9] and trees with a high proportion of leaves [5].

However, it has also been proven that cycles C_n [15], complete graphs K_n [6, 8], complete bipartite graphs $K_{n,n}$ and $K_{n,n+1}$ [8], generalized Petersen graphs [1], and a variety of other graphs [5, 9] all possess such labelings. A major conceptual step forward was McQuillan's proof [11] that a large class of cubic graphs, referred to in the present paper as *quasi-prisms*, could be shown to possess VMTLs without the fine structure of the graph needing to be known.

MacDougall [10] has conjectured that all *regular* graphs other than K_2 and $2K_3$ possess VMTLs, and to date, while constructions have been derived for some families of regular graphs including those previously mentioned, no counterexamples to MacDougall's conjecture have been found. This paper adds significant further support to the conjecture by demonstrating that "almost all" regular graphs of odd order possess VMTLs, as well as many graphs of even order.

[*] Received by the editors September 5, 2005; accepted for publication (in revised form) May 1, 2006; published electronically March 2, 2007.
http://www.siam.org/journals/sidma/21-1/63959.html

[†] School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW 2308 Australia (idgray@bigpond.com).

2. Constructing VMTLs of regular graphs of odd order.

THEOREM 2.1. *If G is a graph of order n with a spanning subgraph H which possesses a strong VMTL and $G - E(H)$ is even-regular, then G also possesses a strong VMTL.*

Proof. Let $G-E(H)$ be $2r$-regular. Let $h = |E(H)|$ and assign the labels $1, \ldots, h$ to the edges of H according to its VMTL. The vertices of G will be referred to as $v_1, \ldots, v_n$. We proceed to assign labels to sets of n edges at a time in such a way that after each set of labels is assigned, the sums of the edge-labels at the vertices form a sequence of consecutive integers. After all edges are labeled, we then assign a set of consecutive integers to the vertices to complete the VMTL. We do this in the following way.

We first decompose $G - E(H)$ into r 2-factors, $A_1, \ldots , A_r$. (This will always be possible since every regular graph of even degree has a 2-factor [16, p. 125].) We label the edges of $G - E(H)$ in r iterations by labeling the 2-factors, as follows: For the jth iteration, we direct the edges of A_j so that each vertex has one incoming and one outgoing edge. At the end of the jth iteration, we let $W_j(v_i)$ be the sum of the labels of all *labeled* edges of G incident to the vertex v_i. $W_0(v_i)$ will be the sum of the labels of edges of H incident to the vertex v_i.

Let $M_j = \max_i\{W_j(v_i))\}$ be the maximum weight on any vertex at the end of the jth iteration. We assign the integer

$$\lambda_j(v_i) = M_{j-1} + h + 1 + (j-1)n - W_{j-1}(v_i)$$

to the outgoing edge of the vertex v_i. Clearly, $\lambda_j(v_i) + W_{j-1}(v_i)$ will be the same constant for all vertices. Let $\alpha_j(v_i)$ be the label on the incoming edge. Since an incoming edge for one vertex is an outgoing edge for another,

$$\begin{aligned}\{\alpha_j(v_i) : i = 1, \ldots, n\} &= \{\lambda_j(v_i) : i = 1, \ldots, n\}\\ &= \{h + i + (j-1)n : i = 1, \ldots, n\}.\end{aligned}$$

Hence, after any iteration the sums $W_j(v_i) = \alpha_j(v_i) + \lambda_j(v_i) + W_{j-1}(v_i)$ of the labeled edges incident to each vertex will form a set of consecutive integers. Thus when all of the edges have been labeled, the $W_r(v_i)$ will also be a set of consecutive integers.

The edges have now been assigned the labels $1, \ldots, h + rn$, and it remains to assign the consecutive labels $h + rn + 1, \ldots, h + rn + n$ to the vertices. In order to obtain the VMTL, to each vertex v_i we assign the complementary integer: $\lambda(v_i) = M_r + h + 1 + rn - W_r(v_i)$, and the final labeling is strong since the largest labels are on the vertices. □

COROLLARY 2.2. *Every Hamiltonian regular graph of odd order has a strong VMTL.*

Proof. By definition, a Hamiltonian graph has a spanning cycle C_n, and C_n has a strong VMTL [15]. The specifics of this construction are given in section 5. □

To illustrate the application of this theorem, let Figure 2.1(b) be the graph G and Figure 2.1(a) be H, a spanning cycle of G with the given VMTL, with $G - E(H)$ directed as shown. If we consider the vertex with label 13 in Figure 2.1(a), the initial weight at that vertex $W_0 = 6$. Reassigning the vertex labels, the outgoing edge is labeled with the vertex label 13 and the incoming edge with 11, so $W_1 = 30$. The vertex is now relabeled with 18 to give a magic constant $k = 48$. A similar process is carried out with each of the other vertices to give the final VMTL.

In [2], it is shown that every 2-connected k-regular graph of order $v \leq 3k + 1$ is Hamiltonian, so clearly this corollary applies to all such graphs. In [13], it is shown

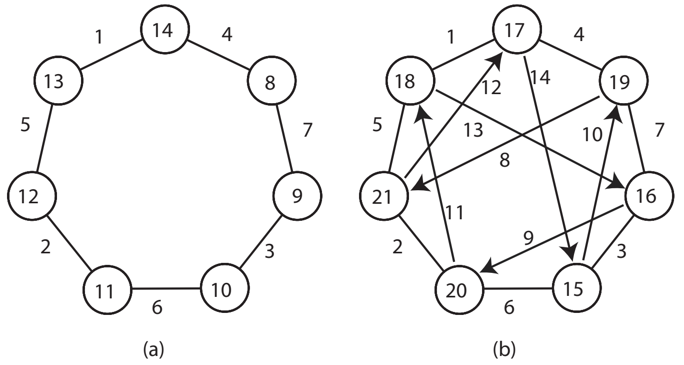
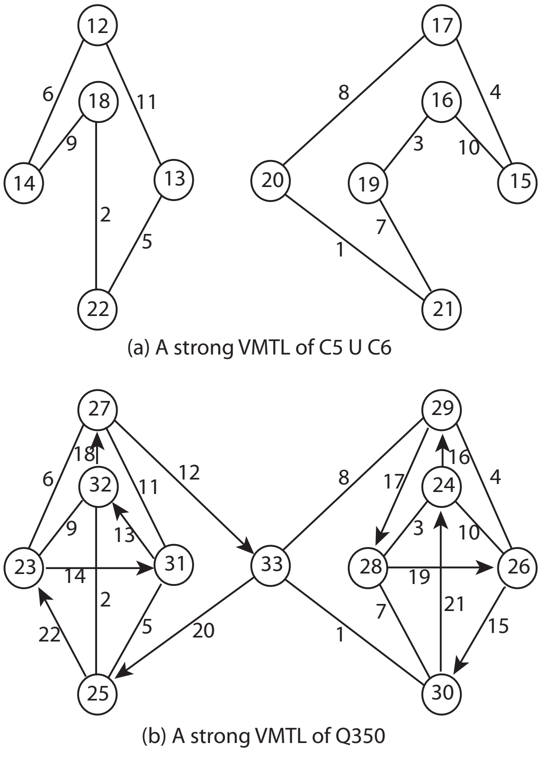


FIG. 2.1. *An example of the construction of Theorem* 2.1.

that almost all regular graphs are Hamiltonian. However, this is an asymptotic result, of course, so it is relatively easy to find examples of non-Hamiltonian graphs of odd order; in particular, any graph with a cut vertex is not Hamiltonian, the quartic of order 11, Q350 in [12], being an example. However, this graph possesses the 2-factor $C_5 \cup C_6$ which has a strong VMTL as shown in Figure 2.2(a) and we can use this labeling and apply Theorem 2.1 to construct a VMTL as shown in Figure 2.2(b).



(a) A strong VMTL of C5 U C6

(b) A strong VMTL of Q350

FIG. 2.2. *Q*350 *from* $C_5 \cup C_6$.

One way of establishing that every regular graph of odd order has a VMTL would be to show that every regular graph of odd order greater than 7 possesses a 2-factor with a strong VMTL. We don't know whether this is true, and while we do not pursue this question in the present paper, the following partial result is already known.

Corollary 2.3. *Every regular graph of odd order with a spanning subgraph consisting of isomorphic cycles has a strong VMTL.*

Proof. If a and b are both odd, then the graph aC_b has been shown in [14] to have a strong VMTL. If $n = ab$, then we let $H = aC_b$ in the theorem, and the result follows. □

However, the graph $C_3 \cup C_4$ does not have a strong VMTL, and so we must have the restriction on order mentioned above.

In [3], constructions are given for sparse antimagic squares, which are shown there to be equivalent to strong VMTLs of certain families of incomplete bipartite graphs of order $0 \bmod 4$. Clearly, there are large classes of irregular graphs which have these bipartite graphs as spanning subgraphs to which Theorem 2.1 can also be applied.

3. Strong VMTLs of even-regular graphs and sparse magic squares. A *sparse semimagic square* $S_n(d,r)$ is a square array of order n containing the integers $1,\dots,nd-r$ once each and $n^2-(nd-r)$ 0's arranged so that the row and column sums are all the same. The parameter d is referred to as the *density* of the square, and the parameter r is the *deficiency.*

These squares were first introduced in [4], which provides constructions for *regular* sparse semimagic squares for any order $n \geq 3$, i.e., where the number of entries in each row and column is the same, as well as for some classes of sparse *magic* squares.

The inspiration to investigate semimagic squares came from [8], where semimagic squares are used to construct a labeling for $K_{n,n}$. Curiously, the converse relationship also holds, as seen in the following theorem.

Theorem 3.1. *Every even-regular graph with a strong VMTL gives rise to a family of sparse semimagic squares.*

Proof. Without loss of generality, consider a $2r$-regular graph G of order n. Since G is even-regular, it possesses a (nonunique) factorization into 2-factors [7, p. 127]. We first obtain an orientation of G by orienting each 2-factor and let A be the adjacency matrix of the resulting digraph. Clearly, each vertex will have an in-degree of r and an out-degree of r; hence A will contain r 1's in each row and in each column, and since G is simple, $A(i,j)+A(j,i) \leq 1$.

Let $M_n(2r)$ be the matrix representing the VMTL of the graph. Since the VMTL is strong, the edge labels will be $1,\dots,nr$, and the vertex labels will be $nr+1,\dots,nr+n$. The edge label in cell (i,j) will equal the edge label in (j,i), so each edge label will occur once above and once below the principal diagonal. Let $M = M_n(2r) + n(r+1)A$. We claim that this is a sparse semimagic square. The addition of $n(r+1)A$ adds $nr(r+1)$ to each row and column; hence the rows and column sums of M are constant. It also adds $n(r+1)$ to one of each nonzero pair of symmetric entries. Since these entries are labeled by $1,\dots,nr$, this gives us a set of labels $\{1,\dots,nr,n(r+1)+1,\dots,n(r+1)+nr\}$, which with the diagonal entries gives us the set of integers $1,\dots,n(2r+1)$. Hence M is a regular semimagic square $\mathrm{S}_n(2r+1,0)$ of order n and density $2r+1$. Since the factorization is not unique, M is just one member of a family of such squares. The result follows. □

4. Constructing VMTLs of odd-regular graphs of even order. Let us define a *quasi-prism* as a cubic graph of order $2n$ which can be partitioned into two 2-factors, each of order n with a 1-factor between them. We will call the 2-factors *cycle-sets* and the edges of the 1-factor *struts.* Particular examples of quasi-prisms are prisms and generalized Petersen graphs.

In [11], McQuillan provided a construction for vertex magic total labelings of quasi-prisms. The following is a simple explanation of this construction. We will refer to one cycle-set as the *outer-set* and the other as the *inner-set*.

First, we orient the two cycle-sets and direct the struts from the outer to the inner set. We then denote the edges and vertices as follows, for $i = 1, \dots, n$:

e_i will be an edge of the outer-set.

v_i will be the vertex of the outer-set which is the tail of e_i.

s_i will be the strut which has the head of e_i as its tail.

f_i will be the edge of the inner-set which has the head of s_i as its tail.

u_i will be the vertex of the inner-set which is the head of f_i.

We then label the edges and vertices as follows. Let $\sigma(i)$ be any permutation of the integers $1, \dots, n$. Then

$\lambda(e_i) = \sigma(i)$,
$\lambda(s_i) = 4n + 1 - \lambda(e_i)$,
$\lambda(f_i) = \lambda(e_i) + n$,
$\lambda(v_i) = 5n + 1 - \lambda(e_i)$,
$\lambda(u_i) = 3n + 1 - \lambda(e_i)$.

Each vertex-sum will equal $9n + 2$, so the labeling is a VMTL.

In this labeling, the vertices receive distinct labels from the set $\{2n + i, 4n + i : i = 1, \dots, n\}$, and the struts receive distinct labels from the set $\{3n+i : i = 1, \dots, n\}$. This fact is important in a later construction.

We now proceed to prove that any $(2t + 3)$-regular graph having a quasi-prism as a spanning subgraph has a VMTL. The proof is constructive and relies on beginning with a labeling on the spanning cubic subgraph which has properties similar to that above.

Lemma 4.1. *Let G be a $(2r + 1)$-regular graph of order $2m$ with a VMTL such that either*

(i) *its vertices are assigned distinct labels from $\{(2r - 2)m + i, 2rm + i : i = 1, \dots, m\}$ and it has a 1-factor whose labels are distinct members of $\{(2r - 1)m + i : i = 1, \dots, m\}$ or*

(ii) *its vertices are assigned distinct labels from $\{(2r-3)m+i, (2r-1)m+i : i = 1, \dots, m\}$ and it has a 1-factor whose labels are distinct members of $\{2rm + i : i = 1, \dots, m\}$*

If a 2-factor is added to the graph, the resulting $(2r + 3)$-regular graph has a VMTL.

Proof. Consider first case (i). We add the 2-factor and orient it. For each vertex we remove the label from that vertex and reassign it instead to the outgoing edge of that vertex. For each vertex, the new vertex sum will be a constant plus the label on its incoming edge. We now add $3m$ to each label on the 1-factor, increasing each vertex sum by $3m$. Thus the set of vertex-sums is $\{a+(2r+1)m+i, a+(2r+3)m+i : i = 1, \dots, m\}$, where a is the magic constant of the original graph. We now have the labels $\{(2r - 1)m + i, (2r + 1)m + i : i = 1, \dots, m\}$ available to assign to the vertices. We can assign the label λ_i to the vertex with sum $a + (2r + 1)2m + m + 1 - \lambda_i$ to obtain a constant vertex sum. The result follows. If we let $r^* = r + 1$, then the vertex-labels become $\{(2r^* - 3)m + i, (2r^* - 1)m + i : i = 1, \dots, m\}$ and the labels on the 1-factor become $\{2r^*m + i : i = 1, \dots, m\}$, and the result is a VMTL (which satisfies the conditions of case (ii)).

For case (ii), we carry out the same procedure as for case (i), but instead of adding $3m$ to the labels of the 1-factor, we only add m to the labels of the 1-factor. The resulting set of vertex-sums will be $\{a + (2r - 2)m + i, a + (2r)m + i : i = 1, \dots, m\}$.

We now have the labels $\{2rm+i, (2r+2)m+i : i = 1, \ldots, m\}$ available to assign to the vertices. We can assign the label λ_i to the vertex with sum $a+(2r)2m+m+1-\lambda_i$ to obtain a constant vertex sum. The result follows. If we let $r^* = r+1$, then the vertex-labels become $\{(2r^*-2)m+i, 2r^*m+i : i = 1, \ldots, m\}$, the labels on the 1-factor become $\{(2r^*-1)m+i : i = 1, \ldots, m\}$, and the result is a VMTL (which satisfies the conditions of case (i)). □

THEOREM 4.2. *Every* $(2t+3)$*-regular graph which has a quasi-prism as a spanning subgraph has a VMTL.*

Proof. In [11], McQuillan provides a construction with the properties of case (i) for any quasi-prism, as previously described. Let G be the $(2t+3)$-regular graph and Q be the quasi-prism. Then $G - E(Q)$ is $2t$-regular and hence has a 2-factorization [7]. We can arbitrarily order these 2-factors and alternate between the procedures for cases (i) and (ii) in Lemma 4.1 until no 2-factors remain to be labeled. The result follows. □

Since every complete graph of even order has a quasi-prism as a spanning subgraph, this theorem permits an alternative construction for a VMTL of K_{2n} to that found in [6].

We also have the following result.

THEOREM 4.3. $(C_n)^C$ *has a VMTL for all* $n \geq 5$.

Proof. If n is odd, then $(C_n)^C$ is Hamiltonian for $n \geq 5$ and hence by Corollary 2.2 has a VMTL. For even $n \geq 6$, let $n = 2t+6$ and let Q be any quasi-prism of order $2t+6$. Clearly, $Q \subset K_{2t+6}$. The degree for each vertex of $K_{2t+6} - E(Q)$ is $2t+2$. If $2t+2 \geq t+3$, then $K_{2t+6} - E(Q)$ is Hamiltonian [16]. Hence, if $t \geq 1$, then $C_{2t+6} \subseteq K_{2t+6} - E(Q)$. Hence $C_{2t+6} \cap E(Q) = \varnothing$ and so $Q \subseteq (C_{2t+6})^C$. $(C_{2t+6})^C$ is $(2t+3)$-regular and hence by the previous construction possesses a VMTL which can be built from Q. Thus the result holds for $t \geq 1$. It can be easily determined by inspection that $(C_6)^C$ is a quasi-prism; hence the result also holds for $t = 0$. The result follows. □

5. VMTLs of even-regular graphs of even order.

The general question of whether all even-regular graphs of even order possess VMTLs seems to be much harder. As mentioned earlier, constructions are known for C_{2m}, K_{2m}, and $K_{2m,2m}$ but for few other families of graphs. In this section, we show how to construct families of $2r$-regular graphs of even order $2m$ with a VMTL for all $3 \leq r \leq m-1$.

We define a *quasi-anti-prism* as a 4-regular graph of order $2n$ which can be decomposed into two 2-factors A_1 and A_2, each of order n plus a 2-factor of order $2n$ in which each edge is incident to one vertex of A_1 and one vertex of A_2. A *simple* quasi-anti-prism will be a quasi-anti-prism in which the two 2-factors of order n are cycles. We will construct labelings of simple quasi-anti-prisms of twice-odd order (i.e., which contain $2C_n$ as a spanning subgraph).

A simple way of constructing VMTLs for cycles of odd order is as follows. Consider $C_n = C_{2t+1}$. We direct the cycle so that each vertex has one incoming and one outgoing edge. We assign the integer 1 to any edge, and for $i = 1, \ldots, n$ if an edge has a label i, then we assign the label $\mu(i)$ to the edge to which it points, where

$$\mu(i) = \begin{cases} i + \frac{n+1}{2} & \text{if } i < \frac{n+1}{2}, \\ i - \frac{n-1}{2} & \text{if } i \geq \frac{n+1}{2}. \end{cases} \tag{5.1}$$

This gives us the vertex weights,

$$wt(v_i) = i + \mu(i) = \begin{cases} 2i + \frac{n+1}{2} & \text{if } i < \frac{n+1}{2}, \\ 2i - \frac{n-1}{2} & \text{if } i \geq \frac{n+1}{2}. \end{cases} \tag{5.2}$$

If v_i is the vertex to which the edge labeled i points, then we can assign v_i the label $\lambda(v_i) = \frac{1}{2}(5n+3) - (i + \mu(i))$. The edges will be labeled with $1, \ldots, n$, the vertices will be labeled with $n+1, \ldots, 2n$, and the labeling is a strong VMTL.

Theorem 5.1. *If a quartic graph G of twice-odd order $2n$ has $2C_n$ as a spanning subgraph, then it has a VMTL.*

Proof. The proof is by construction. We begin with the above labeling for a cycle and construct a labeling of $2C_n$ as follows: We label both copies of $2C_n$ as above and then add $3n$ to the vertex labels of one copy. We then add n to both the edge labels and vertex labels of the other copy. Clearly, both copies have the same vertex-sums, which we will call k. However, the result is *not* a VMTL since we have used the integers $1, \ldots, 2n$ for the edges and $2n+1, \ldots, 3n, 4n+1, \ldots, 5n$ for the vertices, which leaves the integers $3n+1, \ldots, 4n$ unused. $G - E(2C_n)$ is a 2-factor, and we now orient this 2-factor so that each vertex has one incoming and one outgoing edge. For each vertex v_i we re-assign $\lambda(v_i)$ to its outgoing edge. If we call the label on the incoming edge $\alpha(v_i)$, then the weight on each vertex will be $W(v_i) = k + \alpha(v_i)$, and these weights constitute the set of integers $\{k+2n+1, \ldots, k+3n, k+4n+1, \ldots, k+5n\}$. We can then assign the complementary label $8n+1-\alpha(v_i)$ to v_i to give the required VMTL. □

Corollary 5.2. *If a $2r$-regular graph G of twice-odd order $2n$ has a simple quasi-anti-prism as a spanning subgraph, then it has a VMTL.*

Proof. In the case of a simple quasi-anti-prism, the labeling given by the construction above satisfies the requirements of case (i) of Lemma 4.1. □

Corollary 5.3. *$K_{4m+2} - f$ has a VMTL, where f is a 1-factor.*

Proof. Begin with any simple quasi-anti-prism of order $4m+2$ and construct its VMTL. Apply the methods of Lemma 4.1 alternately and iteratively to obtain the required result. □

It should be noted that the method of Theorem 5.1 can be applied more generally to the case where instead of $2C_n$, we have any two 2-factors of order n, each possessing a strong VMTL. For example, if a quasi-anti-prism has $C_9 \cup 3C_3$ or $3C_3 \cup 3C_3$ as its order 9 2-factors, then it possesses a VMTL since both C_9 and $3C_3$ possess strong VMTLs. Similarly, by the same argument as Corollary 5.2, a $2r$-regular graph which contains such a quasi-anti-prism as a spanning subgraph will also possess a VMTL. These results can be summarized as follows.

Theorem 5.4. *If the two 2-factors of odd order n of a quasi-anti-prism G each possess a strong VMTL, then G has a VMTL.*

Corollary 5.5. *If a $2r$-regular graph G of twice-odd order $2n$ has such a quasi-anti-prism as a spanning subgraph, then it has a VMTL.*

These final results highlight the importance of determining whether every 2-factor of odd order greater than 7 possesses a strong VMTL. On the one hand, such a result would establish that every regular graph of odd order possessed a VMTL, as previously mentioned in section 2, while on the other hand it would establish that a large class of even regular graphs of twice-odd order also possessed VMTLs. While the author has found constructions for $C_3 \cup C_{2t-2}$ and $C_4 \cup C_{2t-3}$, a general construction for strong VMTLs of 2-factors of odd order has proven to be elusive.

REFERENCES

[1] M. Bača, M. Miller, and Slamin, *Vertex-magic total labelings of generalised Petersen graphs*, Int. J. Comput. Math., 79 (2002), pp. 1259–1263.

[2] J. Bondy and M. Kouider, *Hamilton cycles in regular* 2*-connected graphs*, J. Combin. Theory Ser. B, 44 (1988), pp. 177–186.

[3] I. Gray and J. MacDougall, *Sparse anti-magic squares and vertex-magic labelings of bipartite graphs*, Discrete Math., to appear.

[4] I. Gray and J. MacDougall, *Sparse semi-magic squares and vertex-magic labelings*, Ars Combin., 80 (2006), pp. 225–242.

[5] I. Gray, J. MacDougall, J. McSorley, and W. Wallis, *Vertex-magic total labeling of complete bipartite graphs*, Ars Combin., 69 (2003), pp. 117–127.

[6] I. Gray, J. MacDougall, and W. Wallis, *On vertex-magic labeling of complete graphs*, Bull. Inst. Combin. Appl., 38 (2003), pp. 42–44.

[7] D. Horton and J. Sheehan, *The Petersen Graph*, Austral. Math. Soc. Lect. Ser. 7, Cambridge University Press, Cambridge, UK, 1993.

[8] J. MacDougall, M. Miller, Slamin, and W. Wallis, *Vertex-magic total labelings of graphs*, Util. Math., 61 (2002), pp. 3–21.

[9] J. MacDougall, M. Miller, and W. Wallis, *Vertex-magic total labelings of wheels and related graphs*, Util. Math., 62 (2002), pp. 175–183.

[10] J. A. MacDougall, *Vertex-Magic Labeling of Regular Graphs*, Lecture, DIMACS Connect Institute, July 18, 2002.

[11] D. McQuillan, *Vertex-magic cubic graphs*, J. Combin. Math. Combin. Comput., 48 (2004), pp. 103–106.

[12] R. C. Read and R. J. Wilson, *An Atlas of Graphs*, Clarendon Press, Oxford, UK, 1998.

[13] R. Robinson and N. Wormald, *Almost all regular graphs are hamiltonian*, Random Structures Algorithms, 5 (1994), pp. 363–374.

[14] V. Swaminathan and P. Jeyanthi, *Super vertex magic labeling*, Indian J. Pure Appl. Math., 34 (2003), pp. 935–939.

[15] W. Wallis, *Magic Graphs*, Birkhäuser, Boston, 2001.

[16] D. B. West, *Introduction to Graph Theory*, Prentice-Hall, Upper Saddle River, NJ, 1996.

SIAM J. Discrete Math.
Vol. 21, No. 1, pp. 178–190

TWO NEW BOUNDS FOR THE RANDOM-EDGE SIMPLEX-ALGORITHM*

BERND GÄRTNER† AND VOLKER KAIBEL‡

Abstract. We prove that the Random-Edge simplex-algorithm requires an expected number of at most $13n/\sqrt{d}$ pivot steps on any simple d-polytope with n vertices. This is the first nontrivial upper bound for general polytopes. We also describe a refined analysis that potentially yields much better bounds for specific classes of polytopes. As one application, we show that for combinatorial d-cubes the trivial upper bound of 2^d on the performance of Random-Edge can asymptotically be improved by the factor $1/d^{(1-\varepsilon)\log d}$ for every $\varepsilon > 0$.

Key words. simplex-algorithm, Random-Edge, pivot rule, cube

AMS subject classifications. 90C05, 65K05

DOI. 10.1137/05062370X

1. Introduction.

Dantzig's *simplex method* [8] is a widely used tool for solving linear programs (LP). The feasible region of an LP is a polyhedron; any algorithm implementing the simplex method traverses a sequence of vertices such that (i) consecutive vertices are equal (the *degenerate* case) or connected by a polyhedron edge and (ii) the objective function strictly improves along any traversed edge. In both theory and practice, we may assume that some initial vertex is available and that the optimal solution to the LP is attained at a vertex, if there is an optimum at all. It follows that if the algorithm does not cycle, it will eventually find an optimal solution or discover that the problem is unbounded. For a comprehensive introduction to the simplex method, see, e.g., Chvátal's book [7].

For most (complexity-)theoretic investigations, one can safely assume that the LPs that are considered are bounded as well as being both primally and dually nondegenerate [23]. Thus, we will deal only with *simple polytopes*, i.e., bounded d-dimensional polyhedra, where at each vertex exactly d facets meet, and with objective functions that are nonconstant along any edge of the polytope.

The distinguishing feature of each simplex-algorithm is the *pivot rule* according to which the next vertex in the sequence is selected in case there is a choice. Many popular pivot rules are efficient in practice, meaning that they induce a short vertex sequence in typical applications. The situation in theory is in sharp contrast to this: for most of the *deterministic* pivot rules proposed in the literature, the simplex-algorithm in the worst case is forced to traverse a number of vertices that is exponential in the number of variables and constraints of the LP. This includes most of the rules that are used in practice. It is open whether there is a pivot rule that always induces a sequence of polynomial length.

To explain simplex's excellent behavior in practice, the tools of *average case analysis* [5] and *smoothed analysis* [24] have been devised. Recently, Kelner and Spielman

*Received by the editors February 1, 2005; accepted for publication (in revised form) September 27, 2006; published electronically March 15, 2007.
http://www.siam.org/journals/sidma/21-1/62370.html

†Institut für Theoretische Informatik, ETH Zürich, CH-8092 Zürich, Switzerland (gaertner@inf.ethz.ch). This author received support from the Swiss Foundation (SNF), project 200021-100316/1.

‡Otto-von-Guericke Universität Magdeburg, Fakultät für Mathematik, Universitätsplatz 2, D-39106 Magdeburg, Germany (volker.kaibel@mathematik.uni-magdeburg.de). This author was supported by the DFG Research Center Matheon and by the DFG Research Group Algorithms, Structure, Randomness.

developed an algorithm that uses the *shadow vertex* simplex algorithm as the main subroutine. Similar to ellipsoid- and interior-point algorithms [17, 16], its running time is polynomial *in the bit size* of the input [18]. While this is a major step forward, it does not yield a bound in the sense we are interested in here: in view of the unsolved question for a *strongly* polynomial time algorithm for linear programing, we are looking for bounds that involve *only* the dimension d and the number of facets n. Moreover, in our context *simplex-algorithm* means an algorithm that proceeds along (improving) edges of a single polyhedron.

To make progress on such bounds, research has turned to *randomized* pivot rules. Indeed, Kalai [14, 15] as well as Matoušek, Sharir, and Welzl [21] could prove that the expected number of steps taken by the Random-Facet pivot rule is only *subexponential* in the worst case. These results hold under our above assumption that the feasible region of the LP is a simple and full-dimensional polytope.

Much less is known about the perhaps most natural randomized pivot rule: choose the next vertex in the sequence uniformly at random among the neighbors of the current vertex with better objective function value. This rule is called Random-Edge, and unlike Random-Facet, it has no recursive structure to peg an analysis to. Nontrivial upper bounds on its expected number of pivot steps on general polytopes do not exist. Results are known for 3-polytopes [6, 13], d-polytopes with $d+2$ facets [9], and for linear assignment problems [25]. Only recently, Balogh and Pemantle solved the long standing problem of finding a tight bound for the expected performance of Random-Edge on the d-dimensional *Klee–Minty* cube [3]. This polytope is the "mother" of many worst-case inputs for deterministic pivot rules [19, 2].

None of the existing results exclude the possibility of both Random-Facet and Random-Edge being the desired polynomial-time pivot rules. In the more general and well-studied setting of *abstract objective functions* on polytopes [1, 26, 27, 15], superpolynomial lower bounds are known for both rules, where the construction for Random-Edge [22] is very recent and much more involved than the one for Random-Facet [20]. Both approaches inherently use objective functions on cubes that are not linearly induced.

In this paper, we derive the first nontrivial upper bound for the expected performance of Random-Edge on simple polytopes, with edge orientations induced by abstract objective functions. Even when we restrict to linear objective functions on combinatorial cubes, the result is new. The general bound itself is rather weak and also achieved, e.g., by the deterministic Greatest-Decrease rule. The emphasis here is on the fact that we are able to make progress at all, given that Random-Edge has turned out to be very difficult to attack in the past. Also, our new bound separates Random-Edge from many deterministic rules (e.g., *Dantzig's rule*, *Bland's rule*, or the *shadow vertex rule*) that may visit all vertices in the worst case [2].

In the second part of the paper, we refine the analysis, with the goal of obtaining better bounds for specific classes of polytopes. Roughly speaking, these are polytopes with large and regular local neighborhoods. Our prime example is the class of combinatorial cubes, for which we improve the general upper bound by the factor $1/d^{(1-\varepsilon)\log d}$ for every $\varepsilon > 0$.

As before, this also works for abstract objective functions and thus complements the recent lower bound of Matoušek and Szabó [22] with a first nontrivial upper bound.

Our results can also be interpreted in the general framework of random walks on graphs. While results concerning stationary distributions and mixing times are known for many classes of undirected graphs, or more generally, strongly connected digraphs [11], random walks on acyclic digraphs cannot usually be analyzed by standard methods. Our paper provides some new techniques for polytope graphs induced by abstract objective functions.

2. A bound for general polytopes. Throughout this section, P is a d-dimensional simple polytope with a set V of n vertices. A directed graph $D = (V, A)$ is called an *acyclic unique sink orientation* (AUSO) of P if

(i) its underlying undirected graph is the vertex-edge graph of P,
(ii) D is acyclic, and
(iii) any subgraph of D induced by the vertices of a nonempty face of P has a unique sink.

Any linear function $\varphi : V \to \mathbb{R}$ that is *generic* (nonconstant on edges of P) induces an AUSO in a natural way: there is a directed edge $v \to w$ between adjacent vertices if and only if $\varphi(v) > \varphi(w)$. The global sink of the AUSO is the unique vertex that minimizes φ over P. If φ is any generic (not necessarily linear) function inducing an AUSO that way, φ is called an *abstract objective function.* For a given AUSO D of P, any function φ that maps vertices to their ranks w.r.t. a fixed topological sorting of D is an abstract objective function that induces D. In general, D need not be induced by a linear function, e.g., if D fails to satisfy the necessary *Holt–Klee* condition for linear realizability [12]. For the remainder of this section, we fix an AUSO D of P, an abstract objective function φ that induces D and some vertex $s \in V$.

Let π be the random variable defined as the directed path in D, starting at s and ending at the sink v_{opt} of D, induced by the Random-Edge pivot rule. From each visited vertex $v \neq v_{\text{opt}}$, π proceeds to a neighbor w of v along an outgoing edge chosen uniformly at random from all outgoing edges.

For each $v \in V$, denote by

$$\text{out}(v) \ := \ \{w \in V \ : \ (v, w) \in A\}$$

the set of all smaller (w.r.t. φ) neighbors of v. If $|\,\text{out}(v)| = k$, then v is called a *k-vertex.* We denote by V_k the set of all k-vertices.

For every vertex $v \neq v_{\text{opt}}$ on the path π, let v' be its successor on π. For such a pair (v, v') we say that π *skips* the vertices

$$S(v) \ := \ \{w \in \text{out}(v) \ : \ \varphi(v') < \varphi(w)\}$$

at v. Note that nodes that are skipped by π do not lie on π. For every $0 \leq k \leq d$ let

$$\eta_k(\pi) \ := \ \left|\{v \in \pi \cap V_k \ : \ |S(v)| \geq \lfloor \tfrac{|\,\text{out}(v)|}{2} \rfloor\}\right|$$

be the number of k-vertices on π at which π skips at least $\lfloor \frac{k}{2} \rfloor$ neighbors. Here, as in the following, we write, depending on the context, π for the set of vertices on the path π.

If we denote by $n_k(\pi)$ the total number of k-vertices on the path π, then we obtain

$$\mathbb{E}[\eta_k(\pi)] \ \geq \ \tfrac{1}{2}\mathbb{E}[n_k(\pi)] \ . \tag{1}$$

Indeed, we have

$$\mathbb{E}[\eta_k(\pi)] \ = \ \sum_{v \in V_k} \mathbb{P}[v \in \pi \text{ and } |S(v)| \geq \lfloor \tfrac{k}{2} \rfloor]$$

and

$$\mathbb{E}[n_k(\pi)] \ = \ \sum_{v \in V_k} \mathbb{P}[v \in \pi] \ .$$

The claim then follows from

$$\mathbb{P}\left[|S(v)| \geq \lfloor\tfrac{\lfloor \mathrm{out}(v)\rfloor}{2}\rfloor \mid v \in \pi\right] \;\geq\; \tfrac{1}{2}\ .$$

Due to $\varphi(v) > \varphi(w) > \varphi(v')$ for all $w \in S(v)$, the sets $S(v)$ are pairwise disjoint. Thus, exploiting the linearity of expectation, we obtain for the number length(π) of vertices on π

$$\mathbb{E}[\mathrm{length}(\pi)] \;\leq\; n - \sum_{k=0}^{d} \mathbb{E}[\eta_k(\pi)]\lfloor\tfrac{k}{2}\rfloor \;\leq\; n - \sum_{k=0}^{d} \tfrac{1}{2}\lfloor\tfrac{k}{2}\rfloor\mathbb{E}[n_k(\pi)],$$

where we used (1) for the second inequality. Clearly, we have $\mathbb{E}[\mathrm{length}(\pi)] = \sum_{k=0}^{d} \mathbb{E}[n_k(\pi)]$. Therefore, we obtain (note $\frac{1}{2}\lfloor\frac{k}{2}\rfloor \geq \frac{k-1}{4}$)

$$\mathbb{E}[\mathrm{length}(\pi)] \;\leq\; \min\left\{\sum_{k=0}^{d} \mathbb{E}[n_k(\pi)], n - \sum_{k=0}^{d} \tfrac{k-1}{4}\mathbb{E}[n_k(\pi)]\right\}. \tag{2}$$

If h_k denotes the total number of k-vertices in V, then we clearly have $0 \leq \mathbb{E}[n_k(\pi)] \leq h_k$. Thus, (2) yields

$$\mathbb{E}[\mathrm{length}(\pi)] \;\leq\; \max\left\{\min\left\{\sum_{k=0}^{d} x_k, n - \sum_{k=0}^{d} \tfrac{k-1}{4}x_k\right\} \,:\, 0 \leq x_k \leq h_k \text{ for all } k\right\}. \tag{3}$$

In (3), the maximum must be attained by some $x \in \mathbb{R}^{d+1}$ for which the minimum is attained by both $\sum x_k$ and $n - \sum \frac{k-1}{4}x_k$. Indeed, if $\sum x_k < n - \sum \frac{k-1}{4}x_k$, then due to $n = \sum h_k$, not all x_k can be at their respective upper bounds h_k. Thus one of them can be slightly increased in order to increase the minimum. If $\sum x_k > n - \sum \frac{k-1}{4}x_k$, then not all x_k can be zero, since this would yield $0 > n$. Therefore, one of them can be decreased in order to increase the minimum. Thus we conclude

$$\mathbb{E}[\mathrm{length}(\pi)] \;\leq\; \max\left\{\sum_{k=0}^{d} x_k \,:\, \sum_{k=0}^{d} \tfrac{k+3}{4}x_k = n,\ 0 \leq x_k \leq h_k \text{ for all } k\right\}. \tag{4}$$

By weak linear programming duality and exploiting $n = \sum_{k=0}^{d} h_k$ once more, we can derive from (4) the estimate

$$\mathbb{E}[\mathrm{length}(\pi)] \;\leq\; \sum_{k=0}^{d} h_k \cdot \max\{y, 1 - \tfrac{k-1}{4}y\} \tag{5}$$

for every $y \in \mathbb{R}$.

In the following discussion, we need two important results from the theory of convex polytopes. The parameters h_k are independent of the actual AUSO of the polytope. The *h-vector* formed by them is a linear transformation of the *f-vector* of the polytope, storing for each i the number of i-dimensional faces of the polytope.

The first classical result we need is the *Dehn–Sommerville equations*

$$h_k \;=\; h_{d-k} \qquad \text{for all } 0 \leq k \leq d \tag{6}$$

(see [28, sect. 8.3]). The second one is the *unimodality of the h-vector*:

$$h_0 \;\leq\; h_1 \;\leq\; \cdots \;\leq\; h_{\lfloor d/2\rfloor}. \tag{7}$$

The latter is equivalent to the *nonnegativity of the g-vector*, which is one of the hard parts of the *g-theorem for simplicial polytopes*; see [28, sect. 8.6].

From (6) and (7) we can derive

$$n \;=\; \sum_{k=0}^{d} h_k \;\geq\; \bigl(d - 8\sqrt{d}\bigr) h_{\lfloor 4\sqrt{d} \rfloor} \;,$$

which yields (for $d > 64$)

$$h_{\lfloor 4\sqrt{d} \rfloor} \;\leq\; \frac{n}{d - 8\sqrt{d}} \;. \tag{8}$$

Now we choose $y := 1/\sqrt{d}$ in (5). We have

$$\frac{1}{\sqrt{d}} \geq 1 - \frac{k-1}{4\sqrt{d}} \;\Leftrightarrow\; k \geq 4\sqrt{d} - 3 \;.$$

Thus, (5) gives

$$\mathbb{E}[\mathrm{length}(\pi)] \;\leq\; \sum_{k=0}^{\lfloor 4\sqrt{d}-3 \rfloor} h_k \left(1 - \frac{k-1}{4\sqrt{d}}\right) \;+\; \sum_{k=\lfloor 4\sqrt{d}-3 \rfloor + 1}^{d} \frac{h_k}{\sqrt{d}} \;. \tag{9}$$

By the unimodality of the h-vector and (8), the first sum in (9) can be estimated by

$$4\sqrt{d} \cdot h_{\lfloor 4\sqrt{d} \rfloor} \;\leq\; \frac{4n}{\sqrt{d} - 8} \leq \frac{12n}{\sqrt{d}}, \quad d \geq 144 \;.$$

Clearly, the second sum in (9) is bounded by $n/\sqrt{d}$. The resulting total bound of $13n/\sqrt{d}$ also holds for $d < 144$, because n is a trivial upper bound. Thus we have proved the following result.

THEOREM 1. *The expected number of vertices visited by the Random-Edge simplex-algorithm on a d-dimensional simple polytope with $n \geq d+1$ vertices, equipped with an abstract (in particular, a linear) objective function is bounded by*

$$13 \cdot \frac{n}{\sqrt{d}} \;.$$

A similar analysis reveals that the number of vertices that are visited when using the GREATEST-DECREASE-rule is bounded by

$$C \cdot \frac{n}{\sqrt{d}}$$

for some constant C. In each step, this rule selects the neighboring vertex with *smallest* φ-value, thus skipping all other neighbors of the current vertex v.

For general simple polytopes, our analysis of the bound for RANDOM-EDGE stated in (3) is essentially the best possible. This can be seen through the examples of duals of stacked simplicial polytopes (see, e.g., [4]), which are simple d-polytopes with n-vertices, $h_0 = h_d = 1$, and $h_k = \frac{n-2}{d-1}$ for all $1 \leq k \leq d-1$.

3. A bound for cubes. The core argument of the analysis presented in section 2 is the following: for every vertex on the Random-Edge path π with out-degree k, we know that π in expectation skips $k/2$ vertices *in the single step from v to its successor.* We then exploited the Dehn–Sommervile equations as well as the unimodality of the h-vector in order to argue that many vertices on π must have large out-degree—unless π is "short" anyway.

For the d-dimensional cube, we have much more information on the h-vector: $h_k = \binom{d}{k}$ for every k. Thus, "most" vertices have out-degree roughly $d/2$ in the case of cubes. We will exploit this stronger knowledge in a sharper analysis for cubes, which relies on studying larger structures around vertices than just their out-neighbors. We actually do the analysis for general polytopes and obtain a bound on the expected path length in terms of two specific quantities. Later we bound these quantities for the case of cubes.

3.1. The general approach. Within this subsection, let P be a d-dimensional simple polytope with a set V of n-vertices, $D = (V, A)$, an AUSO of P, $\varphi : V \to \mathbb{R}$ an abstract objective function inducing D, and $s \in V$ a fixed vertex. We denote by $\operatorname{dist}^{\rightarrow}(v, w)$ the length (number of arcs) of a shortest directed path from v to w. If there is no such path, then $\operatorname{dist}^{\rightarrow}(v, w)$ is defined to be ∞.

Definition 1 (t-reach and k-good). *Let $t, k \in \mathbb{N}$ and $v \in V$.*

(1) *We call*

$$\mathrm{R}_t(v) \;:=\; \{w \in V \,:\, \operatorname{dist}^{\rightarrow}(v, w) \le t\}$$

the t-reach *of v. The* boundary *of $\mathrm{R}_t(v)$, denoted by $\partial\,\mathrm{R}_t(v)$, is the set of all $w \in \mathrm{R}_t(v)$ for which there is a directed (not necessarily shortest) path of length precisely t from v to w.*

(2) *The t-reach $\mathrm{R}_t(v)$ is* k-good *if*

$$|\operatorname{out}(w)| \;\ge\; k$$

holds for all $w \in \mathrm{R}_t(v)$ with $\operatorname{dist}^{\rightarrow}(v, w) \le t - 1$.

(3) *A vertex v is* (t, k)-good *if its t-reach is k-good. The set of all (t, k)-good vertices is denoted by $\mathrm{G}(t, k)$.*

If v is (t, k)-good for $k > 0$, the optimal vertex v_{opt} may occur in the boundary of $\mathrm{R}_t(v)$ but not in its interior. For $t, k \in \mathbb{N}$, we define

$$g(t, k) \;:=\; \min\{|\partial\,\mathrm{R}_t(v)| \,:\, v \in \mathrm{G}(t, k)\}\;.$$

For every vertex $v \in V$ and some $t \in \mathbb{N}$, denote by the random variable $w_t(v)$ the vertex that is reached by the Random-Edge simplex-algorithm, started at v, after t steps. Let $w_t(v) := v_{\text{opt}}$ in case the sink is reached before step t. Generalizing the notion from section 2, we denote by

$$\tilde{S}_t(v) \;:=\; \{u \in \mathrm{R}_t(v) \,:\, \varphi(w_t(v)) < \varphi(u)\}$$

the set of vertices in $\mathrm{R}_t(v)$ left behind while walking from v to $w_t(v)$.

Lemma 1. *For every $t, k \in \mathbb{N}$ with $t \ge 1$ and $v \in \mathrm{G}(t, k)$, we have*

$$\mathbb{P}\big[|\tilde{S}_t(v)| \ge \tfrac{g(t,k)}{2}\big] \;\ge\; \frac{g(t, k)}{2d^t}\;.$$

Proof. Let $\partial\,\mathrm{R}_t(v) = \{u_1, \ldots, u_q\}$ with $\varphi(u_1) > \cdots > \varphi(u_q)$. Let $i^\star$ be the random variable for the index of $w_t(v)$ in $\partial\,\mathrm{R}_t(v)$, i.e., $w_t(v) = u_{i^\star}$. Note that $w_t(v) \in \partial\,\mathrm{R}_t(v)$ indeed (and so $w_t(v) \ne v_{opt}$) since v is (t, k)-good.

Since the out-degree at every vertex is at most d, we have for every $1 \leq i \leq q$

$$\mathbb{P}[i^\star = i] \;\geq\; \frac{1}{d^t}.$$

Therefore,

$$\mathbb{P}[i^\star > q/2] \;\geq\; \frac{q}{2d^t}$$

holds. Now $i^\star > q/2 \geq g(t,k)/2$ implies that at least $\lfloor g(t,k)/2 \rfloor$ vertices from $\partial\,\mathrm{R}_t(v)$ are left behind, and since v is left behind as well (we have $t \geq 1$), we get $|\tilde{S}_t(v)| \geq g(t,k)/2$. The claim follows. □

Now let us consider the path π followed by the Random-Edge simplex-algorithm started at s and ending in v_{opt}. For $t, k \in \mathbb{N}$, $t \geq 1$, we subdivide π into subpaths with the property that every subpath has either length one and starts at a non-(t,k)-good vertex or length t and starts at a (t,k)-good vertex (a *long subpath*).

Let $n_{t,k}(\pi)$ be the number of long subpaths in our subdivision. We denote the pairs of start and end vertices of these long paths by $(x_1, y_1), \ldots, (x_{n_{t,k}(\pi)}, y_{n_{t,k}(\pi)})$. Let

$$S_t(x_i) \;:=\; \{u \in \mathrm{R}_t(x_i) \,:\, \varphi(y_i) < \varphi(u)\},$$

and define

$$\eta_{t,k}(\pi) \;:=\; \left|\left\{i \in \{1, \ldots, n_{t,k}(\pi)\} \,:\, |S_t(x_i)| \geq \tfrac{g(t,k)}{2}\right\}\right|$$

to be the number of those long subpaths which leave behind at least $\frac{g(t,k)}{2}$ vertices from $\mathrm{R}_t(x_i)$.

The distribution of $S_t(x_i)$ conditioned on the event that x_i is the start vertex of a long subpath in the partitioning of π equals the distribution of $\tilde{S}_t(x_i)$. Thus, using Lemma 1, we can deduce the following similarly to our derivation of (1):

$$\mathbb{E}[\eta_{t,k}(\pi)] \;\geq\; \frac{g(t,k)}{2d^t}\mathbb{E}[n_{t,k}(\pi)]. \tag{10}$$

Also here, the sets $S_t(x_i)$ (for $1 \leq i \leq n_{t,k}(\pi)$) are pairwise disjoint. Thus, for each long subpath (consisting of t arcs) starting at some x_i with $|S_t(x_i)| \geq g(t,k)/2$, we can count at least $g(t,k)/2 - t$ vertices that are not visited by π. Therefore, we can conclude

$$\mathbb{E}[\mathrm{length}(\pi)] \;\leq\; n - \left(\tfrac{g(t,k)}{2} - t\right)\mathbb{E}[\eta_{t,k}(\pi)]\ .$$

Using (10) and defining

$$\tilde{g}(t,k) \;:=\; \left(\tfrac{g(t,k)}{2} - t\right)\tfrac{g(t,k)}{2d^t}\ ,$$

this yields

$$\mathbb{E}[\mathrm{length}(\pi)] \;\leq\; n - \tilde{g}(t,k)\mathbb{E}[n_{t,k}(\pi)]\ . \tag{11}$$

We assume that $\tilde{g}(t,k) > 0$—this will be satisfied by the values of t and k we use in our application to cubes below.

On the other hand, denote by

$$f(t,k) \ := \ |V \setminus \mathrm{G}(t,k)| \tag{12}$$

the total number of non-(t,k)-good vertices. From the definition of our path subdivision, we immediately obtain

$$\mathbb{E}[\mathrm{length}(\pi)] \ \le \ f(t,k) + t \cdot \mathbb{E}[n_{t,k}(\pi)] \ . \tag{13}$$

Adding up nonnegative multiples of (11) and (13) in such a way that $\mathbb{E}[n_{t,k}(\pi)]$ cancels out, one obtains the following bound:

$$\mathbb{E}[\mathrm{length}(\pi)] \ \le \ \frac{tn + \tilde{g}(t,k)(f(t,k))}{\tilde{g}(t,k) + t} \ \le \ \frac{t}{\tilde{g}(t,k)} n + f(t,k).$$

Using the definition of $\tilde{g}$, this yields the following estimation.

Lemma 2. *For $t,k \in \mathbb{N}$ with $t \ge 1$, we have*

$$\mathbb{E}[\mathrm{length}(\pi)] \ \le \ \frac{4td^t}{g(t,k)(g(t,k)-2t)} n + f(t,k) \ .$$

A general way to bound the function $f(t,k)$ is as follows.

Lemma 3. *For $t,k \in \mathbb{N}$, we have*

$$f(t,k) \ \le \ \frac{d^t - 1}{d-1} h_{<k} \ ,$$

where $h_{<k} := \sum_{j=0}^{k-1} h_j$ is the number of vertices with out-degree less than k.

Proof. If $v \in V \setminus \mathrm{G}(t,k)$, then there is some $w \in \mathrm{R}_{t-1}(v)$ with $|\operatorname{out}(w)| < k$. On the other hand, each such w is contained in at most $\sum_{i=0}^{t-1} d^i = \frac{d^t-1}{d-1}$ $(t-1)$-reaches since the undirected graph is d-regular. The claim follows. ☐

The following describes a way of bounding the function $g(t,k)$ by studying the undirected graph of the polytope.

Definition 2 ((t,k)-neighborhood, $\gamma(t,k)$). *Let $t,k \in \mathbb{N}$.*

(1) *A subset $N \subset V$ is called a* (t,k)-neighborhood *of $v \in V$ if $N = \{v\}$ in case of $t = 0$ or, if $t \ge 1$, there are k neighbors $w_1, \dots, w_k$ of v in the graph of P together with $(t-1,k)$-neighborhoods $N_1, \dots, N_k$ of $w_1, \dots, w_k$, respectively, such that $N = \bigcup_{i=1}^k N_i$.*

(2) *We define $\gamma(t,k)$ as the minimum cardinality of $\{w \in N \ : \ \mathrm{dist}(v,w) = t\}$, taken over all $v \in V$ and all (t,k)-neighborhoods of v. Here, $\mathrm{dist}(v,w)$ denotes the graph-theoretical distance between v and w in the undirected graph of P.*

If v is (t,k)-good, then it follows right from the definitions that the boundary $\partial\,\mathrm{R}_t(v)$ of its t-reach contains a (t,k)-neighborhood N of v.

Lemma 4. *For $t,k \in \mathbb{N}$, we have*

$$g(t,k) \ \ge \ \gamma(t,k) \ .$$

3.2. Specialization to cubes. In order to obtain from Lemma 2 an explicit bound for the expected number of vertices visited by the Random-Edge simplex-algorithm on the d-cube, we will derive estimates on the functions $f(\cdot,\cdot)$ and $\gamma(\cdot,\cdot)$ for the case of cubes.

LEMMA 5. *For* $0 < \beta < \frac{1}{2}$, $k(d) = \lceil \beta d \rceil$, *and* $t(d) = \mathrm{o}(\frac{d}{\log d})$, *there is some* $0 < \alpha < 1$ *such that*

$$f(t(d), k(d)) \ \leq \ 2^{\alpha d + \mathrm{o}(d)}$$

holds (where f *is the function defined in* (12) *for the case of the* d*-cube).*

Proof. For the d-cube, we have

$$h_{<k(d)} \ = \ \sum_{i=0}^{\lceil \beta d \rceil - 1} \binom{d}{i} \ = \ 2^{h(\beta) d + \mathrm{o}(d)} \ ,$$

where $h(x) = x \log \frac{1}{x} + (1-x) \log \frac{1}{1-x}$ is the binary entropy function (see, e.g., [10, Chap. 9, Ex. 42]). By Lemma 3, this implies (with $\alpha := h(\beta) < 1$ due to $\beta < \frac{1}{2}$)

$$f(t(d), k(d)) \ \leq \ d^{t(d)} 2^{\alpha d + \mathrm{o}(d)} \ ,$$

which proves the claim because of $t(d) = \mathrm{o}(\frac{d}{\log d})$. □

The final building block of our bound for the special case of cubes is the following. Here, we denote by $a^{\underline{b}}$ the product $a(a-1) \dots (a-b+1)$ for $a, b \in \mathbb{N}$.

LEMMA 6. *If the polytope* P *considered in section* 3.1 *is a* d*-cube, then the following is true:*

(1) *For each* $t, k \in \mathbb{N}$ *with* $1 \leq t, k \leq d$, *we have*

$$\gamma(t,k) \ \geq \ \frac{k^t}{t!} - \sum_{i=1}^{t-1} \frac{k^i}{t^{\underline{i}}} \binom{d-1}{t-i-1} \ . \tag{14}$$

(2) *For all* $\delta > \delta' > 0$ *there is some* $0 < \beta < \frac{1}{2}$ *such that*

$$\gamma(t(d), k(d)) \ = \Omega(2^{(1-\delta) \log^2 d})$$

holds for $t(d) = \lceil (1 - \delta') \log d \rceil$ *and* $k(d) = \lceil \beta d \rceil$.

Proof. Let us prove (1) for each fixed k, by induction on t, where the case $t = 1$ holds due to $\gamma(1,k) = k$. Thus, let us consider the case $t \geq 2$ and look at a (t,k)-neighborhood of size $\gamma(t,k)$.

We may assume that the vertex v and its neighbors $w_1, \dots, w_k$ are $v = \mathbb{0}$ and $w_i = \mathbb{e}_i$ $(1 \leq i \leq k)$. For each i, the $(t-1,k)$-neighborhood N_i of $\mathbb{e}_i$ has at least $\gamma(t-1,k)$ vertices w with $\mathrm{dist}(\mathbb{e}_i, w) = t-1$, by definition. All of them have distance $t-2$ or t from $\mathbb{0}$ (note that in the cube case, dist is simply the Hamming distance). More precisely, for all $w \in N_i$, we have

$$\begin{array}{rclcl} \mathrm{dist}(\mathbb{0}, w) & = & t-2 & \Leftrightarrow & w_i = 0, \\ \mathrm{dist}(\mathbb{0}, w) & = & t & \Leftrightarrow & w_i = 1. \end{array}$$

In total, there are only $\binom{d-1}{t-2}$ vertices with $\mathrm{dist}(\mathbb{0}, w) = t-2$ and $w_i = 0$, meaning that

$$\left| \{ w \in N_i \, : \, \mathrm{dist}(\mathbb{0}, w) = t \} \right| \ \geq \ \gamma(t-1,k) - \binom{d-1}{t-2} \ .$$

On the other hand, every vertex w with $\mathrm{dist}(\mathbb{0}, w) = t$ has $w_i = 1$ if $w \in N_i$. Therefore, a vertex $w \in \bigcup_{i=1}^{k} N_i$ counted for $\gamma(t,k)$ is contained in at most t of the neighborhoods $N_1, \dots, N_k$. Thus, we conclude

$$\gamma(t,k) \ \geq \ \frac{k\big(\gamma(t-1,k) - \binom{d-1}{t-2}\big)}{t} \ ,$$

and thus,

$$\gamma(t,k) \;\geq\; \frac{k}{t}\gamma(t-1,k) - \frac{k\binom{d-1}{t-2}}{t}\;. \tag{15}$$

Using the induction hypothesis and (15) we derive

$$\begin{aligned}
\gamma(t,k) \;&\geq\; \frac{k}{t}\left(\frac{k^{t-1}}{(t-1)!} - \sum_{i=1}^{t-2}\frac{k^i}{(t-1)^{\underline{i}}}\binom{d-1}{t-i-2}\right) - \frac{k}{t}\binom{d-1}{t-2} \\
&=\; \frac{k^t}{t!} - \sum_{i=1}^{t-2}\frac{k^{i+1}}{t^{\underline{i+1}}}\binom{d-1}{t-i-2} - \frac{k^{0+1}}{t^{\underline{0+1}}}\binom{d-1}{t-0-2} \\
&=\; \frac{k^t}{t!} - \sum_{i=0}^{t-2}\frac{k^{i+1}}{t^{\underline{i+1}}}\binom{d-1}{t-i-2}\;,
\end{aligned}$$

which, after an index shift in the sum, yields the claim.

In order to establish part (2), denote the ith summand of the sum in (14) by s_i. We first show that (for $t = t(d)$ and $k = k(d)$ as specified in the statement of part (2))

$$s_1 \geq 2s_2 \geq \cdots \geq 2^{t-2}s_{t-1} \tag{16}$$

holds for large enough d. Indeed, this follows from

$$\begin{aligned}
\frac{s_i}{s_{i+1}} \;&=\; \frac{k^i \cdot t^{\underline{i+1}} \cdot (d-1)^{\underline{t-i-1}} \cdot (t-i-2)!}{t^{\underline{i}} \cdot k^{i+1} \cdot (t-i-1)! \cdot (d-1)^{\underline{t-i-2}}} \\
&= \frac{t-i}{t-i-1} \cdot \frac{d-t+i+1}{k} \\
&\geq 1 \cdot 2\;,
\end{aligned}$$

where the last inequality holds for large enough d and for our special choices of t and k.

We now choose some $0 < \delta'' < 1$ with $\delta'' < \delta'$ and define $\beta := 2^{\frac{1}{\delta''-1}}$. In particular, we have $0 \leq \beta < \frac{1}{2}$. Due to (16), s_1 asymptotically dominates the sum of the s_i, and it suffices to show

$$s_1 \leq \mathrm{o}\left(\frac{k(d)^{t(d)}}{t(d)!}\right) \tag{17}$$

and

$$\frac{k(d)^{t(d)}}{t(d)!} \;\geq\; 2^{(1-\delta)\log^2 d} \tag{18}$$

for large enough d.

We prove (17) by the following estimation:

$$\begin{aligned}
\frac{s_1}{\frac{k(d)^{t(d)}}{t(d)!}} &= \frac{k(d)\cdot(d-1)^{\underline{t(d)-2}}\cdot t(d)!}{t(d)\cdot(t(d)-2)!\cdot k(d)^{t(d)}}\\
&\le \left(\frac{d}{k(d)}\right)^{t(d)-2}\cdot\frac{t(d)}{k(d)}\\
&\le \beta^{-(1-\delta')\log d}\cdot \mathrm{O}\left(\frac{\log d}{d}\right)\\
&= d^{\frac{1-\delta'}{1-\delta''}}\cdot \mathrm{O}\left(\frac{\log d}{d}\right)\\
&= \mathrm{o}(1)\ ,
\end{aligned}$$

where the last equation follows from $\frac{1-\delta'}{1-\delta''}<1$.

We note that this estimate is the best possible in the following sense: if we set $t(d)=\log^{1+\theta} d$ for any $\theta>0$, (17) does not hold anymore.

Finally, (18) holds due to the following chain of inequalities that is valid for large enough d:

$$\begin{aligned}
\frac{k(d)^{t(d)}}{t(d)!} &\ge \left(\frac{k(d)}{t(d)}\right)^{t(d)}\\
&\ge \left(\frac{\beta d}{\log d}\right)^{(1-\delta')\log d}\\
&\ge 2^{(1-\delta')\log^2 d+(\log\beta-\log\log d)(1-\delta')\log d}\\
&\ge 2^{(1-\delta)\log^2 d},
\end{aligned}$$

where the last inequality is due to $\delta'<\delta$. □

Now we can prove our main result. It shows that in comparison to algorithms that may need to visit essentially all vertices (like, e.g., the simplex-algorithm with Dantzig's or Bland's or the shadow vertex rule), Random-Edge achieves a super-polynomial speed up (by a factor $d^{(1-\varepsilon)\log d}$) on cubes.

Theorem 2. *For every $\varepsilon>0$, the expected number of vertices visited by the simplex-algorithm using the* Random-Edge *pivot rule on a d-dimensional cube, equipped with an abstract (in particular, a linear) objective function, is bounded by*

$$\mathrm{O}(2^{d-(1-\varepsilon)\log^2 d})\ .$$

Proof. Let π be the (random) path for some arbitrary start vertex defined by the Random-Edge simplex-algorithm on a d-cube equipped with an AUSO. Define $\delta:=\frac{\varepsilon}{2}$, choose some $\delta'>0$ with $\delta'<\delta$, and let $0<\beta<\frac{1}{2}$ as in Lemma 6(2). Let $t(d):=\lceil(1-\delta')\log d\rceil$ and $k(d):=\lceil\beta d\rceil$.

By Lemma 2, we have

$$\mathbb{E}[\text{length}(\pi)] \le \frac{4t(d)d^{t(d)}}{g(t(d),k(d))(g(t(d),k(d))-2t(d))}2^d+f(t(d),k(d))\ . \tag{19}$$

Lemmas 4 and 6(2) imply

$$g(t(d),k(d)) \ge \Omega(2^{(1-\delta)\log^2 d})\ , \tag{20}$$

which in particular yields $t(d) = \mathrm{o}(g(t(d), k(d))$, and thus (by (19))

$$\mathbb{E}[\text{length}(\pi)] \; = \mathrm{O}\left(t(d) \cdot \frac{d^{t(d)}}{g(t(d), k(d))^2} \cdot 2^d\right) + f(t(d), k(d)) \; . \tag{21}$$

From Lemma 5, we know that there is some constant $0 < \alpha < 1$ with

$$f(t(d), k(d)) \;\leq\; 2^{\alpha d + \mathrm{o}(d)} \; . \tag{22}$$

Combining (21), (20), and (22), we obtain

$$\begin{aligned} \mathbb{E}[\text{length}(\pi)] \; &= \mathrm{O}(\log d \cdot 2^{(-1-\delta'+2\delta)\log^2 d} \cdot 2^d) + 2^{\alpha d + \mathrm{o}(d)} \\ &= \mathrm{O}(2^{d-(1-\varepsilon+\delta')\log^2 d + \log\log d)}) + 2^{\alpha d + \mathrm{o}(d)} \\ &= \mathrm{O}(2^{d-(1-\varepsilon)\log^2 d}) \; , \end{aligned}$$

where we used $\delta = \frac{\varepsilon}{2}$ in the second to last and $\delta' > 0$ in the last inequality. This proves the claim. □

4. Conclusion. Probably one can extend the methods we have used for analyzing Random-Edge on cubes to other classes of polytopes (e.g., general products of simplices). However, it seems to us that it would be more interesting to find a way of sharpening our bounds by enhancing our approach with some new ideas. As mentioned at the end of section 2, the analysis of our approach is sharp in the general setting. We suspect that one cannot prove a subexponential bound for Random-Edge on cubes with our methods. Therefore, it would be most interesting to find a way of combining our kind of analysis with some other ideas.

REFERENCES

[1] I. Adler and R. Saigal, *Long monotone paths in abstract polytopes*, Math. Oper. Res., 1 (1976), pp. 89–95.

[2] N. Amenta and G. Ziegler, *Deformed products and maximal shadows of polytopes*, in Advances in Discrete and Computational Geometry, Contemp. Math. 223, J. Chazelle, J. E. Goodman, and R. Pollack, eds., Amer. Math. Soc., Providence, RI, 1999, pp. 57–90.

[3] J. Balogh and R. Pemantle, *The Klee-Minty Random Edge Chain Moves With linear Speed*, manuscript, 2004.

[4] M. M. Bayer and C. W. Lee, *Combinatorial Aspects of Convex Polytopes*, in Handbook of Convex Geom. A, B, North–Holland, Amsterdam, 1993, pp. 485–534.

[5] K. H. Borgwardt, *The Simplex Method: A Probabilistic Analysis*, in Algorithms Combin., Springer-Verlag, New York, 1987.

[6] A. Z. Broder, M. E. Dyer, A. M. Frieze, P. Raghavan, and E. Upfal, *The worst-case running time of the random simplex algorithm is exponential in the height*, Inform. Process. Lett., 56 (1995), pp. 79–81.

[7] V. Chvátal, *Linear Programming*, W. H. Freeman, New York, 1983.

[8] G. B. Dantzig, *Linear Programming and Extensions*, Princeton University Press, Princeton, New Jersey, 1963.

[9] B. Gärtner, J. Solymosi, F. Tschirschnitz, P. Valtr, and E. Welzl, *One line and n points*, Random Structures Algorithms, 23 (2003), pp. 453–471.

[10] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics*, 2nd ed., Addison-Wesley, Reading, MA, 1994.

[11] G. Grimmet and D. Stirzaker, *Probability and Random Processes*, Clarendon Press, London, 1982.

[12] F. B. Holt and V. Klee, *A proof of the strict monotone* 4*-step conjecture*, in Advances in Discrete and Computational Geometry, Contemp. Math. 233, J. Chazelle, J. E. Goodman, and R. Pollack, eds., Amer. Math. Soc., Providence, RI, 1998, pp. 201–216.

[13] V. Kaibel, R. Mechtel, M. Sharir, and G. M. Ziegler, *The simplex algorithm in dimension three*, SIAM J. Comput., 34 (2005), pp. 475–497.

[14] G. Kalai, *A subexponential randomized simplex algorithm*, Random Structures Algorithms, to appear.

[15] G. Kalai, *Linear programming, the simplex algorithm and simple polytopes*, Math. Program., 79 (1997), pp. 217–233.

[16] N. K. Karmarkar, *A new polynomial-time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.

[17] L. G. Khachiuan, *A polynomial algorithm in linear programming*, Sov. Math. Doklady, 20 (1979), pp. 1093–1096.

[18] J. A. Kelner and D. A. Spielman, *A randomized polynomial-time simplex algorithm for linear programming*, in Proceedings of the 38th Annual ACM Symposium on Theory of Computing, 2006, pp. 51–60.

[19] V. Klee and G. J. Minty, *How Good is the Simplex Algorithm?*, in Inequalities III, O. Shisha, ed., Academic Press, New York, 1972, pp. 159–175.

[20] J. Matoušek, *Lower bounds for a subexponential optimization algorithm*, Random Structures Algorithms, 5 (1994), pp. 591–607.

[21] J. Matoušek, M. Sharir, and E. Welzl, *A subexponential bound for linear programming*, Algorithmica, 16 (1996) pp. 498–516.

[22] J. Matoušek and T. Szabó, *RANDOM EDGE can be exponential on abstract cubes*, Adv. Math., 204 (2006), pp. 262–277.

[23] A. Schrijver, *Theory of Linear and Integer Programming*, Wiley-Interscience, New York, 1986.

[24] D. Spielman and S.-H. Teng, *Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time*, J. ACM, 51 (2004), pp. 385–463.

[25] C. A. Tovey, *Low order polynomial bounds on the expected performance of local improvement algorithms*, Math. Program., 35 (1986), pp. 193–224.

[26] D. Wiedemann, *Unimodal set-functions*, Congr. Numer., 50 (1985), pp. 165–169.

[27] K. Williamson Hoke, *Completely unimodal numberings of a simple polytope*, Discrete Appl. Math., 20 (1988), pp. 69–81.

[28] G. M. Ziegler, *Lectures on Polytopes*, Grad. Texts in Math. 152, Springer-Verlag, Heidelberg, 1994.

SIAM J. DISCRETE MATH.
Vol. 21, No. 1, pp. 191–198

QUADRATICALLY MANY COLORFUL SIMPLICES*

IMRE BÁRÁNY[†] AND JIŘÍ MATOUŠEK[‡]

Abstract. The colorful Carathéodory theorem asserts that if $X_1, X_2, \dots, X_{d+1}$ are sets in $\mathbf{R}^d$, each containing the origin 0 in its convex hull, then there exists a set $S \subseteq X_1 \cup \cdots \cup X_{d+1}$ with $|S \cap X_i| = 1$ for all $i = 1, 2, \dots, d+1$ and $0 \in \operatorname{conv}(S)$ (we call $\operatorname{conv}(S)$ a *colorful covering simplex*). Deza et al. [*Discrete Comput. Geom.*, 35 (2006), pp. 597–615] proved that if the X_i are in general position with respect to 0 (consequently, each X_i has at least $d+1$ points), then there are at least $2d$ colorful covering simplices, and they constructed an example with no more than d^2+1 such simplices. Under the same assumption, we show that there are at least $\frac{1}{5}d(d+1)$ colorful covering simplices, thus determining the order of magnitude. A similar result was proved independently by Stephen and Thomas [http://www.arxiv.org/abs/math.CO/0512400 (2005)]. We also obtain a lower bound of $3d$ for $d \geq 3$, which is better for small d and, in particular, together with a parity argument it settles the case $d = 3$, where the minimum possible number of colorful covering simplices is 10.

Key words. colorful simplicial depth, colorful Carathéodory theorem, convex geometry

AMS subject classifications. 52A35, 52A20, 68Q05

DOI. 10.1137/050643039

1. Introduction.

The following theorem, proved by the first author [1], has found numerous applications (see [2], [3], and [5]).

THEOREM 1.1 (colorful Carathéodory theorem). *Let $X_1, X_2, \dots, X_{d+1}$ be finite sets in $\mathbf{R}^d$ such that $0 \in \operatorname{conv}(X_i)$ for all $i = 1, 2, \dots, d+1$. Then there exists a $(d+1)$-point set $S \subseteq X_1 \cup \cdots \cup X_{d+1}$ with $|X_i \cap S| = 1$ for each i and such that $0 \in \operatorname{conv}(S)$.*

If we imagine that the points of X_i have color i, then the theorem asserts the existence of a *colorful* set S with $0 \in \operatorname{conv}(S)$, where "colorful" means "containing all colors." We call the convex hull of such an S a *colorful covering simplex*.

We will assume throughout this paper that the sets X_i as in the colorful Carathéodory theorem are *in general position with respect to* 0, meaning that $X_i \cap X_j = \emptyset$ for $i \neq j$ and no $k+1$ points of $X = X_1 \cup \cdots \cup X_{d+1}$ lie in a common k-dimensional linear subspace of $\mathbf{R}^d$ for all $k = 0, 1, \dots, d-1$. In this situation $0 \in \operatorname{conv}(X_i)$ implies $|X_i| \geq d+1$.

It was shown in [1] that if the X_i are as in the colorful Carathéodory theorem and in general position with respect to 0, then there are actually at least $d+1$ colorful covering simplices. The minimum possible number of colorful covering simplices was investigated by Deza et al. [4], who improved the lower bound to $2d$; on the other hand, they exhibited a configuration with only d^2+1 colorful covering simplices. They conjectured that this is actually the minimum possible number.

We prove that this is at least the correct order of magnitude.

[*] Received by the editors October 19, 2005; accepted for publication (in revised form) October 2, 2006; published electronically March 15, 2007.
http://www.siam.org/journals/sidma/21-1/64303.html

[†] Rényi Institute of Mathematics, Hungarian Academy of Sciences, P.O. Box 127, 1364 Budapest, Hungary, and Department of Mathematics, University College London, Gower Street, London WC1E 6BT, UK (barany@math-inst.hu). The work of this author was supported by Hungarian National Foundation grants T 046246 and T 037846.

[‡] Department of Applied Mathematics and Institute of Theoretical Computer Science (ITI), Charles University, Malostranské nám. 25, 118 00 Praha 1, Czech Republic (matousek@kam.mff.cuni.cz).

Theorem 1.2. *Let $X_1, \dots, X_{d+1}$ be sets in $\mathbf{R}^d$ in general position with respect to* 0*, each containing* 0 *in its convex hull. Then there are at least $\frac{1}{5}d(d+1)$ colorful covering simplices.*

We could get a constant little better than $\frac{1}{5}$, but since we have no reason to believe that an optimal constant could be obtained by our approach, we prefer simplicity of the numbers appearing in the proof.

Deza et al. [4] show that for $d = 2$ the smallest possible number of colorful simplices is 5, and for $d = 3$ this number is either 8 or 10. The following theorem shows that the number is 10.

Theorem 1.3. *Under the assumptions of Theorem* 1.2*, the number of colorful covering simplices is at least* $3d$ *if* $d \geq 3$*. For* $d = 3$*, the smallest possible number of colorful covering simplices equals* 10*.*

After submitting this paper for publication, we learned that Tamon Stephen and Hugh Thomas [6] independently established a result similar to Theorem 1.2, and actually slightly stronger, with at least $\lfloor (d+2)^2/4 \rfloor$ colorful covering simplices. Their proof is considerably simpler than ours.

2. Preparations.

From now on, we assume that $X_1, \dots, X_{d+1} \subset \mathbf{R}^d$ are $(d+1)$-point sets in general position with respect to 0 and with $0 \in \operatorname{conv}(X_i)$ for all i. We may also assume that all points of X lie on the unit sphere S^{d-1} (if not, we replace X by its central projection on S^{d-1}, which affects neither the assumptions nor the conclusions of our theorems).

Every d-point subset $A \subset X$ generates the convex cone

$$\operatorname{pos}(A) = \left\{ \sum_{a \in A} t_a a : t_a \geq 0 \text{ for all } a \in A \right\}.$$

We let $\sigma(A) = \operatorname{pos}(A) \cap S^{d-1}$ be the corresponding spherical simplex spanned by A. By the general position assumption, each such spherical simplex is contained in an open hemisphere.

The set X_{d+1}, the points of the last color, will play a special role in our arguments. We let $Y = X \setminus X_{d+1}$ be the subset made of the first d colors, and we let $P = -X_{d+1}$ be the points antipodal to the last color class.

A *transversal* is any subset $T \subset Y$ with $|T \cap X_i| = 1$ for all $i = 1, 2, \dots, d$, and a *partial transversal* is any subset of a transversal. Let $\mathcal{T}^d(Y)$ denote the system of all transversals of Y, and for $Y' \subseteq Y$, we let $\mathcal{T}^d(Y') = \{T \in \mathcal{T}^d(Y) : T \subseteq Y'\}$.

If $a \in S^{d-1}$ is a point and $T \in \mathcal{T}^d(Y)$, we say that T *covers* a if $a \in \sigma(T)$. Similarly, if $\mathcal{F} \subseteq \mathcal{T}^d(Y)$ is a system of transversals, we say that $\mathcal{F}$ *covers* a if at least one $S \in \mathcal{F}$ covers A.

Colorful covering simplices, the objects of interest in Theorem 1.2, are in one-to-one correspondence with ordered pairs (p, T), where $p \in P$, $T \in \mathcal{T}^d(Y)$, and T covers p. Indeed, for any such (p, T), it is easily seen that $T \cup \{-p\}$ defines a colorful covering simplex (and it is equally easy to see that the correspondence is bijective, but we won't actually need that). So we aim at bounding the number of such pairs (p, T) from below.

We will use the following stronger version of the colorful Carathéodory theorem [1].

Theorem 2.1. *Let $X_1, X_2, \dots, X_d$ be finite sets in $\mathbf{R}^d$ such that $0 \in \operatorname{conv}(X_i)$ for all $i = 1, 2, \dots, d$ and let $x \in \mathbf{R}^d$ be arbitrary. Then there exists a d-point set $S \subseteq X_1 \cup \dots \cup X_d$ with $|X_i \cap S| = 1$ for each i and such that $0 \in \operatorname{conv}(S \cup \{x\})$.*

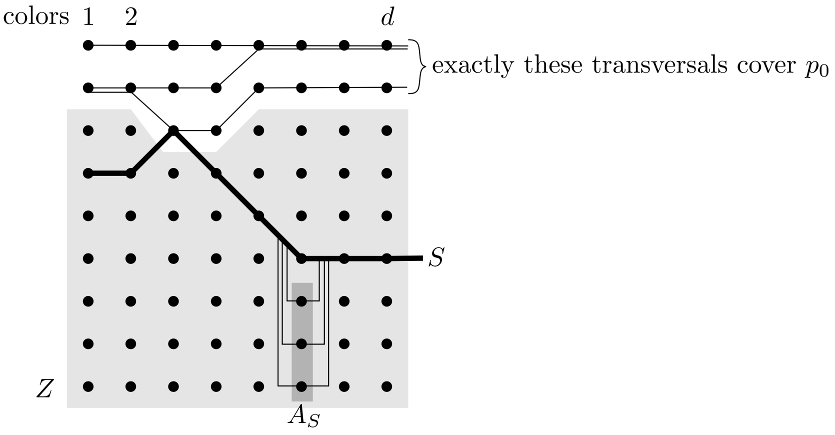


FIG. 1. *Illustration of Lemma* 3.1.

This theorem clearly implies that the set of transversals $\mathcal{T}^d(Y)$ covers every point of the unit sphere, and, in particular, it shows that the number of colorful simplices is at least $d+1$. We will actually apply the following consequence.

COROLLARY 2.2. *For every point $y \in Y$ there is $p \in P$ and a transversal $T \in \mathcal{T}^d(Y)$ that contains y and covers p.*

Proof. If y is in X_i, say, then apply Theorem 2.1 to the sets X_j, $j \neq i$, and to the point y. Then $0 \in \operatorname{conv}(S \cup \{y\})$ for a suitable S. Setting $x = S \cap X_{d+1}$, $T = S \setminus \{x\} \cup \{y\}$ is a transversal in $\mathcal{T}^d(Y)$. It is easy to see that T covers $-x$, which is a point in P. □

We will also use the following lemma, with an easy topological proof.

LEMMA 2.3 (octahedron lemma; Deza et al. [4]). *Let S, T be two disjoint transversals, and let x be a point covered by S. If $\mathcal{T}^d(S \cup T)$ doesn't cover all of S^{d-1}, then there exists $T' \in \mathcal{T}^d(S \cup T)$, $T' \neq S$, that also covers x.*

3. Proof of Theorem 1.2.

Let $Y = X_1 \cup \cdots \cup X_d$ and $P = -X_{d+1}$ be as in the previous section. For every $p \in P$, let $k(p)$ be the number of transversals $T \in \mathcal{T}^d(Y)$ that cover p. We thus want to bound $K := \sum_{p \in P} k(p)$ from below.

Let $k_{\min} = \min_{p \in P} k(p)$. If $k_{\min} \geq \frac{1}{5}(d+1)$, then $K \geq |P| \cdot \frac{1}{5}(d+1) > \frac{1}{5}d(d+1)$, and the conclusion of Theorem 1.2 holds. So from now on, we assume $k_{\min} < \frac{1}{5}(d+1)$.

We let $p_0 \in P$ be one of the points covered exactly $k_{\min}$ times by $\mathcal{T}^d(Y)$. Let $\mathcal{T}_0 \subseteq \mathcal{T}^d(Y)$ consist of the $k_{\min}$ transversals covering p_0, and let $Z = Y \setminus \bigcup \mathcal{T}_0$ be the points of Y not contained in any transversals of $\mathcal{T}_0$ (here we mean points that are *elements* of the transversals, considered as finite sets, not points covered by the transversals). Let $Z_i = X_i \cap Z$. Since $|\mathcal{T}_0| \leq \frac{1}{5}(d+1)$, we have $|Z_i| \geq \frac{4}{5}(d+1)$ for all i.

A key to producing many transversals that cover points of P is the following lemma (also see Figure 1 for an illustration).

LEMMA 3.1 (many associated transversals). *Suppose that $p \in P$ is a point covered by fewer than $\frac{1}{5}d(d+1)$ transversals of X, and let $S \in \mathcal{T}^d(X) \setminus \mathcal{T}_0$ be a transversal that covers p but doesn't cover p_0. Let us denote by s_i the point of S of color i. Then there is a color $i \in \{1, 2, \ldots, d\}$ and a subset $A_S \subseteq Z_i \cup \{s_i\}$ of at least $\frac{1}{3}(d+1)$ points such that for every $a \in A_S$, the transversal $S^a = (S \setminus \{s_i\}) \cup \{a\}$ also covers p.*

Proof. Let us set $W = Z \setminus S$. For every transversal $T \in \mathcal{T}^d(W)$, we can apply the octahedron lemma (Lemma 2.3) to S and T with $x = p$. Indeed, no $T' \in \mathcal{T}^d(S \cup T)$ can cover p_0, since $S \notin \mathcal{T}_0$ and T is disjoint from all transversals in $\mathcal{T}_0$. Hence we get that there is $T' \in \mathcal{T}^d(S \cup T)$ different from S and covering p.

For every $T \in \mathcal{T}^d(W)$ we fix one such T' (choosing arbitrarily if there are several possibilities) and we put $U(T) = T' \setminus S$.

Let us consider the set system $\mathcal{U}_0 = \{U(T) : T \in \mathcal{T}^d(W)\}$. For $U \in \mathcal{U}_0$, let $\overline{U}^S$ be the (unique) transversal T' with $U = T' \setminus S$. The following two properties of $\mathcal{U}_0$ are clear from the construction.

(U1) Every $U \in \mathcal{U}_0$ is a nonempty partial transversal of W such that $\overline{U}^S$ covers p.

(U2) Every transversal $T \in \mathcal{T}^d(W)$ contains some $U \in \mathcal{U}_0$.

Now we will delete some sets from $\mathcal{U}_0$ so that we obtain a system $\mathcal{U}$ still satisfying (U1) and (U2) but *minimal* with respect to (U2); that is,

(U3) for every $U \in \mathcal{U}$ there exists $T \in \mathcal{T}^d(W)$ (a "reason of existence" of U) that contains U but no other set of $\mathcal{U}$.

The deletion procedure works as follows. We begin with $\mathcal{U}_0$ as the current system. If U is a set in the current system such that every $T \in \mathcal{T}^d(W)$ containing it also contains some other set of the current system, we delete U, and we repeat this step as long as we can. The resulting system $\mathcal{U}$ satisfies all of (U1)–(U3).

Each $U \in \mathcal{U}$ corresponds to the transversal $\overline{U}^S$ covering p, so by the assumption of the lemma we have $|\mathcal{U}| < \frac{1}{5}d(d+1)$.

In order to prove the lemma, it suffices to show that there is an i such that at least $\frac{1}{3}(d+1) - 1$ points in $W_i = X_i \cap W$ form singleton sets in $\mathcal{U}$. Indeed, then the points of W_i covered by singletons in $\mathcal{U}$ plus the point s_i form the desired A_S.

First we observe that for every i, we have either $W_i \subseteq \bigcup\mathcal{U}$ or $W_i \cap \bigcup\mathcal{U} = \emptyset$. Indeed, let $U \in \mathcal{U}$ contain a point $w \in W_i$, and let $T \supseteq U$ be a "reason of existence" of U as in (U3) above. Then $R = T \setminus \{w\}$ contains no set of $\mathcal{U}$, and hence every $T' = R \cup \{w'\} \in \mathcal{T}^d(W)$, where $w' \in W_i$, has to contain some $U' \in \mathcal{U}$ with $w' \in U'$.

Let $I = \{i \in \{1, 2, \ldots, d\} : W_i \subseteq \bigcup\mathcal{U}\}$ be the colors covered by $\mathcal{U}$. Let V_i be the part of W_i not covered by singleton sets of $\mathcal{U}$, and let $n_i = |V_i|$. It suffices to show that $n_i \le \frac{7}{15}(d+1)$ for some i, since then at least $|W_i| - |V_i| \ge \frac{4}{5}(d+1) - 1 - \frac{7}{15}(d+1) > \frac{1}{3}(d+1) - 1$ elements of W_i are covered by singletons as needed. So we assume $n_i > \frac{7}{15}(d+1)$ for all $i \in I$. (We note that this implies $|I| \ge 2$, since for $I = \{i\}$ all of W_i is covered by singletons.)

There are $M = \prod_{i \in I} n_i$ transversals of $V = \bigcup_{i \in I} V_i$ (here the transversals have $|I|$ points and they cover only the colors in I). Any $U \in \mathcal{U}$ contained in V has at least two elements (since all singletons have been removed), and hence the number of transversals of V containing it is

$$\frac{M}{\prod_{i:\, U \cap V_i \neq \emptyset} n_i} < \frac{M}{(\frac{7}{15}(d+1))^2}.$$

Since every transversal of V contains some $U \in \mathcal{U}$, we get $|\mathcal{U}| \ge (\frac{7}{15}(d+1))^2 \ge \frac{1}{5}d(d+1)$, contradicting the assumption $|\mathcal{U}| < \frac{1}{5}d(d+1)$. This finishes the proof of Lemma 3.1. □

Now we are ready to finish the proof of Theorem 1.2. For every point $z \in Z$, Corollary 2.2 guarantees the existence of a transversal $S = S(z) \in \mathcal{T}^d(X)$ that contains z and covers some $p = p(z) \in P$. For each such $S(z)$, we apply Lemma 3.1 (of course, we may assume that no $p \in P$ is covered by more than $\frac{1}{5}d(d+1)$ transversals, since otherwise we are done). This yields the system of at least $\frac{1}{3}(d+1)$ transversals $S(z)^a$, $a \in A_{S(z)}$, that all cover p and differ from $S(z)$ in at most one point. Let

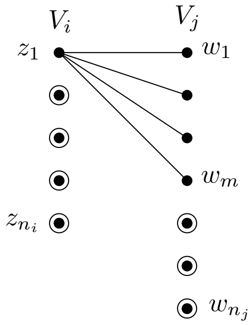


FIG. 2. *The set system $\mathcal{U}(V_i, V_j, m)$.*

us denote this system by $\mathcal{A}(S(z))$ and call it the *system of associated transversals* of $S(z)$.

Let us put $\mathcal{S} = \{S(z) : z \in Z\}$, and let $(S_1, S_2, \dots, S_t)$ be an enumeration of all sets in $\mathcal{S}$ in some arbitrary order (each set occurs only once in the sequence, although the same set may be obtained for many different z).

We observe that if $|S_i \triangle S_j| > 2$ (with $\triangle$ denoting the symmetric difference), then $\mathcal{A}(S_i)$ and $\mathcal{A}(S_j)$ have no transversal in common. Indeed, if both $T \in \mathcal{A}(S_i)$ and $T \in \mathcal{A}(S_j)$, then $|T \triangle S_i| \le 1$ and $|T \triangle S_j| \le 1$, and hence $|S_i \triangle S_j| \le 2$. Moreover, since all S_i have the same size, $|S_i \setminus S_j| \ge 2$ implies $|S_i \triangle S_j| > 2$.

Let us call an index $i \in \{1, 2, \dots, t\}$ a *jump* if $|S_i \setminus S_j| \ge 2$ for every $j < i$, and a *nonjump* otherwise.

If i is a nonjump, then S_i adds at most one point not covered by the union $\bigcup_{j<i} S_j$. For a jump, S_i may add up to d points. If J denotes the number of jumps and N the number of nonjumps, we have $dJ + N \ge |Z| \ge \frac{4}{5}d(d+1)$ (since the S_i cover Z). Now if $N \ge \frac{1}{5}d(d+1)$, we are done since $t \ge N$ and each S_i is a transversal covering some point of P. Otherwise, we have $J \ge \frac{3}{5}(d+1)$. By the above observation, the systems $\mathcal{A}(S_i)$ for all jumps i are disjoint and each contains at least $\frac{1}{3}(d+1)$ transversals, so altogether we have at least $\frac{3}{5}(d+1) \cdot \frac{1}{3}(d+1) > \frac{1}{5}d(d+1)$ transversals. Theorem 1.2 is proved. □

4. Proof of Theorem 1.3.

We use the same notation as before. We begin with a simple lemma about a set system. We let $V_1, \dots, V_d$ be disjoint finite sets, we set $n_i = |V_i|$, and we assume $1 \le n_1 \le \cdots \le n_d$. As before, $\mathcal{T}^d(V)$ denotes the set of all transversals S of $V = V_1 \cup \cdots \cup V_d$; that is, $S \subset V$ with $|S \cap V_i| = 1$ for all i. Finally, let $\mathcal{U}$ be a system of partial transversals of V satisfying conditions (U2)–(U3) as in the proof of Lemma 3.1. (Condition (U1) is not relevant here, since there is no p involved; it would only say that $\mathcal{U}$ is a system of partial transversals, which it is by definition.)

One example of such a $\mathcal{U}$ consists of all the singletons of some V_i. We denote this system by $\mathcal{U}(V_i)$. Another example is the following (Figure 2): Writing $V_i = \{z_1, \dots, z_{n_i}\}$ and $V_j = \{w_1, \dots, w_{n_j}\}$, $i \ne j$, and choosing an integer $m \in \{1, 2, \dots, n_j\}$, we set

$$\mathcal{U}(V_i, V_j, m) = \{\{z_2\}, \dots, \{z_{n_i}\}, \{z_1, w_1\}, \dots, \{z_1, w_m\}, \{w_{m+1}\}, \dots, \{w_{n_j}\}\}.$$

We note that $|\mathcal{U}(V_i, V_j, m)| = n_i + n_j - 1$.

LEMMA 4.1. *Under the above conditions $|\mathcal{U}| \ge n_1$, with equality if and only if $\mathcal{U} = \mathcal{U}(V_i)$ for some i with $n_i = n_1$. Moreover, if $\mathcal{U}$ contains no $\mathcal{U}(V_i)$, then $|\mathcal{U}| \ge$*

n_1+n_2-1 *with equality if and only if* $\mathcal{U}=\mathcal{U}(V_i,V_j,m)$ *for some* i,j *with* $\{n_i,n_j\}=\{n_1,n_2\}$ *and some* m *(with a suitable numbering of the points of* V_i *and* V_j*).*

Proof. The first statement follows easily from the fact that $\mathcal{T}^d(V)$ contains n_1 disjoint transversals.

For the second statement we delete all singletons $\{v\}$ from $\mathcal{U}$, and with every deleted $\{v\}$ we also delete v from the ground set V. The remaining system $\mathcal{U}^*$ satisfies properties (U2) and (U3) on the remaining ground set $V_1^*,\dots,V_d^*$, $|V_i^*|=n_i^*$. No V_i^* is empty and the total number of transversals in $\mathcal{T}^d(V^*)$ is $M=\prod n_k^*$. We also note that each $U\in\mathcal{U}^*$ has at least two elements.

We fix $U\in\mathcal{U}^*$ with $U=\{z_1,w_1,\dots\}$, where $z_1\in V_i^*$ and $w_1\in V_j^*$. Such a U is contained in at most

$$\frac{M}{\prod_{k\colon U\cap V_k^*\neq\emptyset} n_k^*}\le\frac{M}{n_i^*n_j^*}$$

transversals. It follows that $|\mathcal{U}^*|\ge\min n_i^*n_j^*$, where the minimum is taken over all pairs i,j, $i\neq j$. We observe that $n_i^*n_j^*\ge n_i^*+n_j^*-1$, with equality if and only if $n_i^*=1$ or $n_j^*=1$. Adding back the deleted singletons, we get $|\mathcal{U}|\ge\min_{i\neq j}n_i+n_j-1$, and if equality holds, then either $n_i^*=1$ or $n_j^*=1$. It is not hard to check the precise conditions for equality. We omit the details. □

Now we can start the proof of Theorem 1.3. If $k_{\min}\ge 3$, then we even have $3(d+1)$ colorful covering simplices. It follows from Theorem 2.1 that $k_{\min}>0$. So we have $k_{\min}=1$ or $k_{\min}=2$, and we consider these two cases separately.

Case 1. $k_{\min}=1$. Let $p_0\in P$ be a point covered by a single transversal $S\in\mathcal{T}^d(Y)$, and let $p\in S^{d-1}$ be a point not covered by S. We may assume that $S=\{e_1,\dots,e_d\}$, with $e_1,\dots,e_d$ the standard basis of $\mathbf{R}^d$, because the problem is invariant under nondegenerate linear transformations. So a coordinate system is introduced. For a vector $x\in\mathbf{R}^d$ we write $x[j]$ for its jth coordinate.

The octahedron lemma shows that, for every $T\in\mathcal{T}^d(Y)$ disjoint from S, the set $\mathcal{T}^d(S\cup T)$ contains a transversal, to be denoted by T', covering p. We write $U(T)=T'\setminus S$ and we set $\mathcal{U}_0=\{U(T):T\in\mathcal{T}^d(Z)\}$, where $Z=Y\setminus S$. Next we take, in the same way as in the proof of Lemma 3.1, a minimal subsystem $\mathcal{U}\subset\mathcal{U}_0$. The new system $\mathcal{U}$ satisfies conditions (U1)–(U3). Lemma 4.1 implies that $|\mathcal{U}|\ge d$, and so we have $k(p)\ge d$ whenever p is outside $\sigma(S)$. Therefore, if $k(p)=1$, then $p\in\sigma(S)$, or, in other words, if $k(p)=1$, then $p[i]>0$ for each i.

Note that the systems $\mathcal{U}_0$ and $\mathcal{U}$ depend on p and S, and so in case of need we will write $\mathcal{U}=\mathcal{U}(p,S)$.

Claim 4.2. *If* $|\mathcal{U}|=d$*, then* p *has one negative coordinate and* $d-1$ *positive coordinates.*

Proof. Lemma 4.1 shows in this case that $\mathcal{U}=\mathcal{U}(Z_i)$ for some i. For simpler notation we assume $\mathcal{U}=\mathcal{U}(Z_1)$, and $X_i=\{e_1,z_1,\dots,z_d\}$.

We recall that $\sigma(T)$ denotes $S^{d-1}\cap\operatorname{pos}(T)$. For $T=\{x_1,\dots,x_d\}$ we will also use $\sigma(x_1,\dots,x_d)$ to denote $\sigma(T)$. Since $\mathcal{U}=\mathcal{U}(Z_1)$, we have $p\in\sigma(z_i,e_2,\dots,e_d)$ for every $i=1,2,\dots,d$.

Let us suppose that $p[1]>0$. Then, noticing that $\mathcal{U}=\mathcal{U}(Z_1)$ means $p\in\sigma(z_i,e_2,\dots,e_d)$ for every $i=1,2,\dots,d$, we get $z_i[1]>0$ for all i. Consequently, $X_1=\{e_1,z_1,\dots,z_d\}$ lies in the halfspace $\{x\in\mathbf{R}^d:x[1]>0\}$, contradicting the assumption $0\in\operatorname{conv}(X_1)$. Since $p[1]=0$ is impossible by the general position hypothesis, we have $p[1]<0$.

A similar argument shows that $p[j] > 0$ for all $j > 1$. Indeed, if $p[2] < 0$ (say), then $p \in \sigma(z_i, e_2, \dots, e_d)$ implies $z_i[2] < 0$ for all i, and then X_1 would lie in the halfspace $\{x \in \mathbf{R}^d : x[2] < 0\}$, which is again impossible. □

We recall that $k(p)$ denotes the number of transversals covering p. We want to show that $K = \sum_{p \in P} k(p) \geq 3d$.

Subcase 1a. $k(p) > 1$ for at least two $p \in P$. Then

$$K \geq 2d + (d + 1 - 2) = 3d - 1.$$

So $K \geq 3d$ unless equality holds here. If equality holds, then there are exactly two points in P with $k(p) > 1$, let us call them p_{d-1} and p_d, and we have

$$|\mathcal{U}(p_d, S)| = |\mathcal{U}(p_{d-1}, S)| = d.$$

By Claim 4.2 both p_d and p_{d-1} have one negative coordinate and $d - 1$ positive coordinates. Since $d \geq 3$, there is a coordinate j with both $p_d[j] > 0$ and $p_{d-1}[j] > 0$. For all $p \in P \setminus \{p_{d-1}, p_d\}$ we have $k(p) = 1$, which implies $p \in \sigma(S)$ and thus all coordinates of p are positive. Hence P lies completely in the halfspace $x[j] > 0$, and this contradicts the assumption $0 \in \operatorname{conv}(P)$.

Subcase 1b. $k(p) > 1$ for exactly one $p \in P$, say, for $p_d \in P$. Then all other $p \in P$ lie in $\sigma(S)$, and $0 \in \operatorname{conv}(P)$ implies $p_d[j] < 0$ for all j. Claim 4.2 shows that $|\mathcal{U}(p_d, S)| = d$ is impossible, and Lemma 4.1 yields that $|\mathcal{U}(p_d, S)| \geq 2d - 1$. Thus $k(p_d) \geq 2d - 1$ and

$$K \geq (2d - 1) + d = 3d - 1.$$

So $K \geq 3d$ unless equality holds throughout: $|\mathcal{U}(p_d, S)| = 2d - 1$ and $\mathcal{U}(p_d, S)$ is of the type $\mathcal{U}(V_i, V_j, m)$. For simpler notation we assume it is equal to $\mathcal{U}(V_1, V_2, m)$ with $X_1 = \{e_1, z_1, \dots, z_d\}$ and $X_2 = \{e_2, w_1, \dots, w_d\}$, and

$$\mathcal{U}(p_d, S) = \Big\{\{z_2\}, \dots, \{z_{n_i}\}, \{z_1, w_1\}, \dots, \{z_1, w_m\}, \{w_{m+1}\}, \dots, \{w_{n_j}\}\Big\}.$$

Now $p_d \in \sigma(z_i, e_2, \dots, e_d)$ implies $z_i[j] < 0$ for all i, j. Next, $\sigma(w_i, e_2, \dots, e_d)$ contains p_d when $i > m$, showing that all coordinates of w_i are negative. Further, $z_1[j] > 0$ for all $j > 1$ since $z_1[j] < 0$ for some $j > 1$ would imply that X_1 lies in the halfspace $x[j] \leq 0$, and this would contradict $0 \in \operatorname{conv}(X_1)$, by the general position hypothesis. Now $p_d \in \sigma(z_1, w_i, e_3, \dots, e_d)$ holds for $i \leq m$, which yields $w_i[3] < 0$ for all $i \leq m$. But then X_2 lies in the halfspace $x[3] \leq 0$, which is impossible.

So we have $K \geq 3d$ in Case 1.

Case 2. $k_{\min} = 2$. Let $p_0 \in P$ be a point with $k(p_0) = 2$. Thus p_0 is covered by exactly two transversals $S_1, S_2 \in \mathcal{T}^d(Y)$. We set $Z = Y \setminus (S_1 \cup S_2)$. To fix notation we suppose $p_1, p_2, \dots, p_\ell \in \sigma(S_1) \cap \sigma(S_2)$ and $p_{\ell+1}, \dots, p_d \notin \sigma(S_1) \cap \sigma(S_2)$. We observe that $\ell < d$, since otherwise $P \subset \sigma(S_1) \cap \sigma(S_2)$, which would contradict the assumption $0 \in \operatorname{conv}(P)$. For each $p_r \in P$ with $r > \ell$ we construct the set systems $\mathcal{U}_0(p_r, S_1)$ and $\mathcal{U}_0(p_r, S_2)$ and the minimal subsystems $\mathcal{U}(p_r, S_1)$ and $\mathcal{U}(p_r, S_2)$ (where we work with $Z = Y \setminus (S_1 \cup S_2)$ in the construction, with $|Z_i| \geq d - 1$ for all i). Lemma 4.1 shows that

$$k(p_r) \geq |\mathcal{U}(p_r, S_1) \cup \mathcal{U}(p_r, S_2)| \geq |\mathcal{U}(p_r, S_1)| \geq d - 1.$$

Thus

$$K \geq 2\ell + (d - 1)(d - \ell) = d^2 - (d + 1)\ell + 3\ell + 1.$$

In the range $\ell \in \{0, 1, \dots, d-1\}$, the last expression is minimized for $\ell = d-1$, which gives $K \geq d^2 - (d+1)(d-1) + 3(d-1) + 1 = 3d - 1$.

So $K \geq 3d$ unless equality holds here, in which case $\ell = d-1$, and $|\mathcal{U}(p_d, S_1)| = d-1$ and $\mathcal{U}(p_d, S_1) = \mathcal{U}(p_d, S_2)$. The last conditions imply that $|S_1 \cap S_2| = d-1$ and $\mathcal{U}(p_d, S_1)$ is the special system consisting of singletons from Lemma 4.1. As in Case 1, we fix the coordinate system so that $S_1 = \{e_1, e_2, \dots, e_d\}$, and let $S_2 = \{w, e_2, \dots, e_d\}$ and $\mathcal{U}(p_d, S_1) = \{\{z_2\} \dots, \{z_d\}\}$, where $X_1 = \{e_1, w, z_2, \dots, z_d\}$. In this case, of course, $p_r \in \sigma(S_1) \cap \sigma(S_2)$ for all $r < d$.

If $w[1] < 0$, then all of $\sigma(S_2)$ would lie in the halfspace $x[1] \leq 0$, while $\sigma(S_1)$ lies in the halfspace $x[1] > 0$, and this contradicts $p_0 \in \sigma(S_1) \cap \sigma(S_2)$. Hence $w[1] > 0$.

On the other hand, if all coordinates of w are positive, we have $\sigma(S_2) \subset \sigma(S_1)$. So by possibly interchanging the roles of S_1 and S_2, we can make sure that at least one coordinate of w is negative. After renaming the coordinates suitably, we may assume that $w[2] < 0$.

Now it is easy to show that $K = 3d-1$ is impossible. For each $i \geq 2$, $z_i[2] < 0$ must hold since every coordinate of p_d is negative and $p_d \in \sigma(z_i, e_2, \dots, e_d)$ for each $i \geq 2$. But then $X_1 = \{e_1, w, z_2, \dots, z_d\}$ lies in the halfspace $x[2] \leq 0$, which contradicts the assumption $0 \in \operatorname{conv}(X_1)$. □

Remark. It is perhaps interesting to note that $K \geq 3d-1$ is much easier to prove than $K \geq 3d$. In fact, $K \geq 3d$ does not hold when $d = 2$ and we had to use $d > 2$ during the proof.

Acknowledgment. We would like to thank the referees for thoughtful comments.

REFERENCES

[1] I. Bárány, *A generalization of Carathéodory's theorem*, Discrete Math., 40 (1982), pp. 141–152.

[2] I. Bárány and S. Onn, *Carathéodory's theorem, colourful and applicable*, in Intuitive Geometry, Bolyai Soc. Math. Stud. 6, János Bolyai Math. Soc., Budapest, 1997, pp. 11–21.

[3] I. Bárány and S. Onn, *Colourful linear programming and its relatives*, Math. Oper. Res., 22 (1997), pp. 550–567.

[4] A. Deza, S. Huang, T. Stephen, and T. Terlaky, *Colourful simplicial depth*, Discrete Comput. Geom., 35 (2006), pp. 597–615.

[5] K. S. Sarkaria, *Tverberg's theorem via number fields*, Israel J. Math., 79 (1992), pp. 317–320.

[6] T. Stephen and H. Thomas, *A Quadratic Lower Bound for Colourful Simplicial Depth*, preprint, http://www.arxiv.org/abs/math.CO/0512400 (2005).

SIAM J. Discrete Math.
Vol. 21, No. 1, pp. 199–219

ON NEARLY ORTHOGONAL LATTICE BASES AND RANDOM LATTICES[*]

RAMESH NEELAMANI[†], SANJEEB DASH[‡], AND RICHARD G. BARANIUK[§]

Abstract. We study lattice bases where the angle between any basis vector and the linear subspace spanned by the other basis vectors is at least $\frac{\pi}{3}$ radians; we denote such bases as "nearly orthogonal." We show that a nearly orthogonal lattice basis always contains a shortest lattice vector. Moreover, we prove that if the basis vector lengths are "nearly equal," then the basis is the unique nearly orthogonal lattice basis up to multiplication of basis vectors by ±1. We also study random lattices generated by the columns of random matrices with n rows and $m \leq n$ columns. We show that if $m \leq cn$, with $c \approx 0.071$, then the random matrix forms a nearly orthogonal basis for the random lattice with high probability for large n and almost surely as n tends to infinity. Consequently, the columns of such a random matrix contain the shortest vector in the random lattice. Finally, we discuss an interesting JPEG image compression application where nearly orthogonal lattice bases play an important role.

Key words. lattices, shortest lattice vector, random lattice, JPEG, compression

AMS subject classifications. 03G10, 15A52, 94A08

DOI. 10.1137/050635985

1. Introduction. Lattices are regular arrangements of points in space that are studied in numerous fields, including coding theory, number theory, and cryptography [1, 15, 17, 21, 25]. Formally, a lattice $\mathcal{L}$ in $\mathbb{R}^n$ is the set of all linear integer combinations of a finite set of vectors; that is, $\mathcal{L} = \{u_1 b_1 + u_2 b_2 + \cdots + u_m b_m \,|\, u_i \in Z\}$ for some $b_1, b_2, \ldots, b_m$ in $\mathbb{R}^n$. The set of vectors $\mathcal{B} = \{b_1, b_2, \ldots, b_m\}$ is said to *span* the lattice $\mathcal{L}$. An independent set of vectors that spans $\mathcal{L}$ is a *basis* of $\mathcal{L}$. A lattice is said to be m-dimensional (m-D) if a basis contains m vectors.

In this paper we study the properties of lattice bases whose vectors are "nearly orthogonal" to one another. We define a basis to be θ-orthogonal if the angle between any basis vector and the linear subspace spanned by the remaining basis vectors is at least θ. A θ-orthogonal basis is deemed to be *nearly orthogonal* if θ is at least $\frac{\pi}{3}$ radians.

We derive two simple but appealing properties of nearly orthogonal lattice bases.

1. A $\frac{\pi}{3}$-orthogonal basis always contains a shortest nonzero lattice vector.
2. If all vectors of a θ-orthogonal ($\theta > \frac{\pi}{3}$) basis have lengths less than $\frac{\sqrt{3}}{\sin\theta + \sqrt{3}\cos\theta}$ times the length of the shortest basis vector, then the basis is the unique $\frac{\pi}{3}$-orthogonal basis for the lattice (up to multiplication of basis vectors by ±1).

Gauss [13] proved the first property for two-dimensional (2-D) lattices. We prove (generalizations of) the above properties for m-D lattices for arbitrary m.

We also study lattices generated by a set of random vectors; we focus on vectors comprising Gaussian or Bernoulli $(\pm\frac{1}{\sqrt{n}})$ entries. The set of vectors and the generated

[*] Received by the editors July 14, 2005; accepted for publication (in revised form) October 9, 2006; published electronically March 15, 2007. A part of this work was supported by grants from NSF, AFOSR, ONR, DARPA, and the Texas Instruments Leadership Universities program.
http://www.siam.org/journals/sidma/21-1/63598.html

[†] ExxonMobil Upstream Research Company, Houston, TX 77027 (ramesh.neelamani@exxonmobil.com).

[‡] IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 (sanjeebd@us.ibm.com).

[§] Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 (richb@rice.edu).

lattice are henceforth referred to as a *random basis* and a *random lattice*, respectively. Random bases and lattices find applications in coding [7] and cryptography [28]. We prove an appealing property of random lattices.

> *If a random lattice $\mathcal{L}$ in $\mathbb{R}^n$ is generated by $m \leq cn$ ($c \approx 0.071$) random vectors, then the random vectors form a $\frac{\pi}{3}$-orthogonal basis of $\mathcal{L}$ with high probability at finite n and* almost surely *as $n \to \infty$.*

Consequently, the shortest vector in $\mathcal{L}$ is contained by the random basis with high probability.

We also exploit properties of nearly orthogonal bases to solve an interesting digital image processing problem. Digital color images are routinely subjected to compression schemes such as the JPEG standard [26]. The various settings used during JPEG compression of an image—termed as the image's JPEG compression history—are often discarded after decompression. For recompression of images which were earlier in JPEG-compressed form, it is useful to estimate the discarded compression history from their current representation. We call this problem JPEG compression history estimation (CHEst). The JPEG compression step maps a color image into a set of points contained in a collection of related lattices [23]. We show that the JPEG CHEst problem can be solved by estimating the nearly orthogonal bases spanning these lattices. Then, we invoke the derived properties of nearly orthogonal bases in a heuristic to solve the JPEG CHEst problem [23].

Lattices that contain nearly orthogonal bases are somewhat special[1] because there exist lattices without any $\frac{\pi}{3}$-orthogonal basis (see (4) for an example). Consequently, the new properties of nearly orthogonal lattice bases in this paper cannot be exploited in all lattice problems.

This paper is organized as follows. Section 2 provides some basic definitions and well-known results about lattices. Section 3 formally states our results on nearly orthogonal lattice bases, and section 4 furnishes the proofs for the results in section 3. Section 5 identifies new properties of random lattices. Section 6 describes the role of nearly orthogonal bases in solving the JPEG CHEst problem. Section 7 discusses some limitations of our results and future research directions.

2. Lattices.

Consider an m-D lattice $\mathcal{L}$ in $\mathbb{R}^n$, $m \leq n$. By an *ordered basis* for $\mathcal{L}$, we mean a basis with a certain ordering of the basis vectors. We represent an ordered basis by an ordered set and also by a matrix whose columns define the basis vectors and their ordering. We use the braces $(.,.)$ for ordered sets (for example, $(b_1, b_2, \ldots, b_m)$) and $\{.,.\}$ otherwise (for example, $\{b_1, b_2, \ldots, b_m\}$). For vectors $u, v \in \mathbb{R}^n$, we use both $u^T v$ (with T denoting matrix or vector transpose) and $\langle u, v \rangle$ to denote the inner product of u and v. We denote the Euclidean norm of a vector v in $\mathbb{R}^n$ by $\|v\|$.

Any two bases $\mathcal{B}_1$ and $\mathcal{B}_2$ of $\mathcal{L}$ are related (when treated as $n \times m$ matrices) as $\mathcal{B}_1 = \mathcal{B}_2 \mathcal{U}$, where $\mathcal{U}$ is a $m \times m$ *unimodular matrix*; that is, $\mathcal{U}$ is an integer matrix with determinant equal to ± 1.

The *closest vector problem* (CVP) and the *shortest vector problem* (SVP) are two closely related fundamental lattice problems [1, 2, 10, 15]. Given a lattice $\mathcal{L}$ and an input vector (not necessarily in $\mathcal{L}$), CVP aims to find a vector in $\mathcal{L}$ that is closest (in the Euclidean sense) to the input vector. Even finding approximate CVP solutions is known to be NP-hard [10]. The SVP seeks a vector in $\mathcal{L}$ with the shortest (in the Euclidean sense) nonzero length $\lambda(\mathcal{L})$. The decision version of SVP is not known

[1] However, our random basis results suggest nearly orthogonal bases occur frequently in low-dimensional lattices.

to be NP-complete in the traditional sense, but SVP is NP-hard under randomized reductions [2]. In fact, even finding approximately shortest vectors (to within any constant factor) is NP-hard under randomized reductions [16, 20].

A shortest lattice vector is always contained by orthogonal bases. Hence, one approach to finding short vectors in lattices is to obtain a basis that is close (in some sense) to orthogonal and use the shortest vector in such a basis as an approximate solution to the SVP. A commonly used measure to quantify the "orthogonality" of a lattice basis $\{b_1, b_2, \ldots, b_m\}$ is its *orthogonality defect* [17]:

$$\frac{\prod_{i=1}^{m} \|b_i\|}{|\det([b_1, b_2, \ldots, b_m])|},$$

with det denoting the determinant. For rational lattices (lattices comprising rational vectors), the Lovász basis reduction algorithm [17], often called the LLL algorithm, obtains an *LLL-reduced* lattice basis in polynomial time. Such a basis has a small orthogonality defect. There exist other notions of reduced bases due to Minkowski and to Korkine and Zolotarev (KZ) [15]. Both Minkowski-reduced and KZ-reduced bases contain the shortest lattice vector, but it is NP-hard to obtain such bases.

We choose to quantify a basis's closeness to orthogonality in terms of the following new measures.

- *Weak θ-orthogonality:* An *ordered* set of vectors $(b_1, b_2, \ldots, b_m)$ is weakly θ-orthogonal if for $i = 2, 3, \ldots, m$, the angle between b_i and the subspace spanned by $\{b_1, b_2, \ldots, b_{i-1}\}$ lies in the range $\left[\theta, \frac{\pi}{2}\right]$. That is,

$$(1) \qquad \cos^{-1}\left(\frac{|\langle b_i, \sum_{j=1}^{i-1} \alpha_i\, b_i\rangle|}{\|b_i\| \left\|\sum_{j=1}^{i-1} \alpha_i\, b_i\rangle\right\|}\right) \geq \theta \text{ for all } \alpha_j \in \mathbb{R} \text{ with } \sum_j |\alpha_j| > 0.$$

- *θ-orthogonality:* A set of vectors $\{b_1, b_2, \ldots, b_m\}$ is θ-orthogonal if every ordering of the vectors yields a weakly θ-orthogonal set.

A (weakly) θ-orthogonal basis is one whose vectors are (weakly) θ-orthogonal. Babai [4] proved that an n-D LLL-reduced basis is θ-orthogonal where $\sin\theta = (\sqrt{2}/3)^n$; for large n, this value of θ is very small. Thus the notion of an LLL-reduced basis is quite different from that of a weakly $\frac{\pi}{3}$-orthogonal basis.

We will encounter θ-orthogonal bases in random lattices in section 5 and weakly θ-orthogonal bases (with $\theta \geq \frac{\pi}{3}$) in the JPEG CHEst application in section 6.

3. Nearly orthogonal bases: Results. This section formally states the two properties of nearly orthogonal lattice bases that were identified in the introduction. We also identify an additional property characterizing unimodular matrices that relate two nearly orthogonal bases; this property is particularly useful for the JPEG CHEst application.

Obviously, in an orthogonal lattice basis, the shortest basis vector is a shortest lattice vector. More generally, given a lattice basis $\{b_1, b_2, \ldots, b_m\}$, let θ_i be the angle between b_i and the subspace spanned by the other basis vectors. Then

$$(2) \qquad \lambda(\mathcal{L}) \geq \min_{i \in \{1,2,\ldots,m\}} \|b_i\| \sin\theta_i.$$

Therefore, a θ-orthogonal basis has a basis vector whose length is no more than $\lambda(\mathcal{L}) / \sin\theta$; if $\theta = \frac{\pi}{3}$, this bound becomes $\frac{2\,\lambda(\mathcal{L})}{\sqrt{3}}$. This shows that nearly orthogonal lattice bases contain short vectors.

Gauss proved that in $\mathbb{R}^2$ every $\frac{\pi}{3}$-orthogonal lattice basis indeed contains a shortest lattice vector and provided a polynomial time algorithm to determine such a basis in a rational lattice; see [32] for a nice description. We first show that Gauss's shortest lattice vector result can be extended to higher-dimensional lattices.

Theorem 1. *Let $\mathcal{B} = (b_1, b_2, \ldots, b_m)$ be an ordered basis of a lattice $\mathcal{L}$. If $\mathcal{B}$ is weakly $\left(\frac{\pi}{3} + \epsilon\right)$-orthogonal for $0 \le \epsilon \le \frac{\pi}{6}$, then a shortest vector in $\mathcal{B}$ is a shortest nonzero vector in $\mathcal{L}$. More generally,*

$$\min_{j \in \{1,2,\ldots,m\}} \|b_j\| \le \left\| \sum_{i=1}^{m} u_i b_i \right\| \qquad \textit{for all } u_i \in \mathbb{Z} \textit{ with } \sum_{i=1}^{m} |u_i| \ge 1, \tag{3}$$

with equality possible only if $\epsilon = 0$ or $\sum_{i=1}^{m} |u_i| = 1$.

Corollary 1. *If $0 < \epsilon \le \frac{\pi}{6}$, then a weakly $\left(\frac{\pi}{3} + \epsilon\right)$-orthogonal basis contains every shortest nonzero lattice vector (up to multiplication by ± 1).*

Theorem 1 asserts that a θ-orthogonal lattice basis is guaranteed to contain a shortest lattice vector if $\theta \ge \frac{\pi}{3}$. In fact, the bound $\frac{\pi}{3}$ is tight because, for any $\epsilon > 0$, there exist lattices where some θ-orthogonal basis, with $\theta = \frac{\pi}{3} - \epsilon$, does not contain the shortest lattice vector. For example, consider a lattice in $\mathbb{R}^2$ defined by the basis $\{b_1, b_2\}$, with $\|b_1\| = \|b_2\| = 1$, and the angle between them equal to $\frac{\pi}{3} - \epsilon$. Obviously $b_2 - b_1$ has length less than 1.

For a rational lattice defined by some basis $\mathcal{B}_1$, a weakly $\frac{\pi}{3}$-orthogonal basis $\mathcal{B}_2 = \mathcal{B}_1 \mathcal{U}$, with $\mathcal{U}$ polynomially bounded in size, provides a polynomial-size certificate for $\lambda(\mathcal{L})$. However, we do not expect all rational lattices to have such bases because this would imply that NP=co-NP, assuming SVP is NP-complete. For example, the lattice $\mathcal{L}$ spanned by the basis

$$\mathcal{B} = \begin{bmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \tag{4}$$

does not have any weakly $\frac{\pi}{3}$-orthogonal basis. It is not difficult to verify that $[1\ 0\ 0]^T$ is a shortest lattice vector. Thus, $\lambda(\mathcal{L}) = 1$. Now, assume that $\mathcal{L}$ possesses a weakly $\frac{\pi}{3}$-orthogonal basis $\widetilde{\mathcal{B}} = (b_1, b_2, b_3)$. Let θ_1 be the angle between b_2 and b_1, and let θ_2 be the angle between b_3 and the subspace spanned by b_1 and b_2. Since b_1, b_2, and b_3 have length at least 1,

$$\det(\widetilde{\mathcal{B}}) = \|b_1\| \, \|b_2\| \, \|b_3\| \; |\sin\theta_1| \, |\sin\theta_2| \ge \sin^2\frac{\pi}{3} = \frac{3}{4}. \tag{5}$$

But $\det(\mathcal{B}) = \frac{1}{\sqrt{2}} < \det(\widetilde{\mathcal{B}})$, which shows that the lattice $\mathcal{L}$ with basis $\mathcal{B}$ in (4) has no weakly $\frac{\pi}{3}$-orthogonal basis.

Our second observation describes the conditions under which a lattice contains the unique (modulo permutations and sign changes) set of nearly orthogonal lattice basis vectors.

Theorem 2. *Let $\mathcal{B} = (b_1, b_2, \ldots, b_m)$ be a weakly θ-orthogonal basis for a lattice $\mathcal{L}$ with $\theta > \frac{\pi}{3}$. For all $i \in \{1, 2, \ldots, m\}$, if*

$$\|b_i\| < \eta(\theta) \min_{j \in \{1,2,\ldots,m\}} \|b_j\|, \tag{6}$$

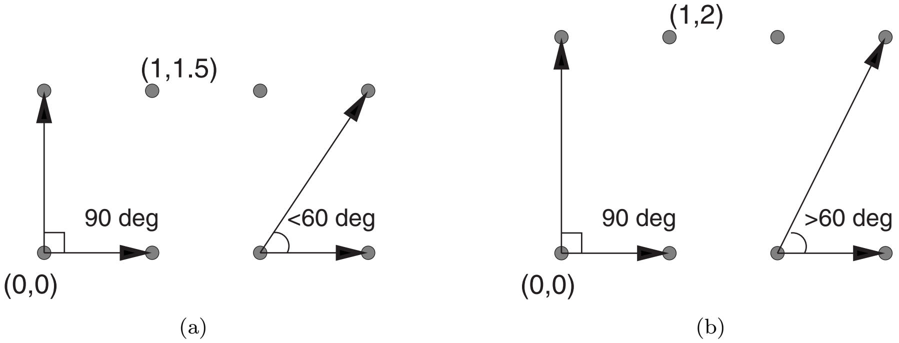


FIG. 1. (a) *The vectors comprising the lattice are denoted by circles. One of the lattice bases comprises two orthogonal vectors of lengths* 1 *and* 1.5*. Since* $1.5 < \eta(\frac{\pi}{2}) = \sqrt{3}$*, the lattice possesses no other basis such that the angle between its vectors is at least* $\frac{\pi}{3}$ *radians.* (b) *This lattice contains at least two* $\frac{\pi}{3}$*-orthogonal bases. One of the lattice bases comprises two orthogonal vectors of lengths* 1 *and* 2*. Here* $2 > \eta(\frac{\pi}{2})$*, and this basis is not the only* $\frac{\pi}{3}$*-orthogonal basis.*

$$\text{with } \eta(\theta) = \frac{\sqrt{3}}{\sin\theta + \sqrt{3}\cos\theta}, \tag{7}$$

then any $\frac{\pi}{3}$*-orthogonal basis comprises the vectors in* $\mathcal{B}$ *multiplied by* ± 1.

In other words, a nearly orthogonal basis is essentially unique when the lengths of its basis vectors are nearly equal. For example, both Figures 1(a) and 1(b) illustrate 2-D lattices that can be spanned by orthogonal basis vectors. For the lattice in Figure 1(a), the ratio of the lengths of the basis vectors is less than $\eta\left(\frac{\pi}{2}\right) = \sqrt{3}$. Hence, there exists only one (modulo sign changes) basis such that the angle between the vectors is greater than $\frac{\pi}{3}$. In contrast, the lattice in Figure 1(b) contains many distinct $\frac{\pi}{3}$-orthogonal bases.

In the JPEG CHEst application [23], the target three-dimensional (3-D) lattice bases in $\mathbb{R}^3$ are known to be weakly $\left(\frac{\pi}{3}+\epsilon\right)$-orthogonal but not $\left(\frac{\pi}{3}+\epsilon\right)$-orthogonal. Theorem 2 addresses the uniqueness of $\frac{\pi}{3}$-orthogonal bases but not weakly $\frac{\pi}{3}$-orthogonal bases. To estimate the target lattice basis, we need to understand how different weakly orthogonal bases are related. The following theorem guarantees that for 3-D lattices a weakly $\left(\frac{\pi}{3}+\epsilon\right)$-orthogonal basis with nearly equal-length basis vectors is related to every weakly orthogonal basis by a unimodular matrix with small entries.

THEOREM 3. *Let* $\mathcal{B} = (b_1, b_2, \ldots, b_m)$ *and* $\widetilde{\mathcal{B}}$ *be two weakly* θ*-orthogonal bases for a lattice* $\mathcal{L}$*, where* $\theta > \frac{\pi}{3}$*. Let* $\mathcal{U} = (u_{ij})$ *be a unimodular matrix such that* $\mathcal{B} = \widetilde{\mathcal{B}}\mathcal{U}$*. Define*

$$\kappa\left(\mathcal{B}\right) = \left(\frac{2}{\sqrt{3}}\right)^{m-1} \times \frac{\max_{i\in\{1,2,\ldots,m\}} \|b_i\|}{\min_{i\in\{1,2,\ldots,m\}} \|b_i\|}. \tag{8}$$

Then, $|u_{ij}| \leq \kappa\left(\mathcal{B}\right)$ *for all* i *and* j.

For example, if $\mathcal{B}$ is a weakly θ-orthogonal basis of a 3-D lattice with $\frac{\max_{i\in\{1,2,3\}}\|b_i\|}{\min_{i\in\{1,2,3\}}\|b_i\|} < 1.5$, then the entries of the unimodular matrix relating another weakly θ-orthogonal basis $\widetilde{\mathcal{B}}$ to $\mathcal{B}$ are either 0 or ± 1.

4. Nearly orthogonal bases: Proofs.

4.1. Proof of Theorem 1. We first prove Theorem 1 for 2-D lattices (Gauss's result) and then tackle the proof for higher-dimensional lattices via induction.

4.1.1. Proof for 2-D lattices. Consider a 2-D lattice with a basis $\mathcal{B} = \{b_1, b_2\}$ satisfying the conditions of Theorem 1. Let θ' denote the angle between b_1 and b_2. Since $\frac{\pi}{3} \leq \theta' \leq \frac{\pi}{2}$ by assumption,

$$|\langle b_1, b_2\rangle| = \|b_1\|\,\|b_2\| \cos\theta' \leq \frac{\|b_1\|\,\|b_2\|}{2}. \tag{9}$$

The squared-length of any nonzero lattice vector $v = u_1 b_1 + u_2 b_2$, with $u_1, u_2 \in \mathbb{Z}$ and $|u_1| + |u_2| > 0$, equals

$$\begin{aligned}
\|v\|^2 &= |u_1|^2\|b_1\|^2 + |u_2|^2\|b_2\|^2 + 2u_1u_2\langle b_1, b_2\rangle \\
&\geq |u_1|^2\|b_1\|^2 + |u_2|^2\|b_2\|^2 - 2|u_1||u_2||\langle b_1, b_2\rangle| \\
&\geq |u_1|^2\|b_1\|^2 + |u_2|^2\|b_2\|^2 - |u_1||u_2|\|b_1\|\,\|b_2\| \qquad \text{(using (9))} \\
&= \left(|u_1|\|b_1\| - |u_2|\|b_2\|\right)^2 + |u_1||u_2|\|b_1\|\,\|b_2\| \qquad (10) \\
&\geq \min\left(\|b_1\|^2,\, \|b_2\|^2\right),
\end{aligned}$$

with equality possible only if either $|u_1| + |u_2| = 1$ or $\theta' = \frac{\pi}{3}$. This proves Theorem 1 for 2-D lattices.

4.1.2. Proof for higher-dimensional lattices. Let $k > 2$ be an integer, and assume that Theorem 1 is true for every $(k-1)$-D lattice. Consider a k-D lattice $\mathcal{L}$ spanned by a weakly $\left(\frac{\pi}{3} + \epsilon\right)$-orthogonal basis $(b_1, b_2, \ldots, b_k)$, with $\epsilon \geq 0$. Any nonzero vector in $\mathcal{L}$ can be written as $\sum_{i=1}^{k} u_i\, b_i$ for integers u_i, where $u_i \neq 0$ for some $i \in \{1, 2, \ldots, k\}$. If $u_k = 0$, then $\sum_{i=1}^{k} u_i\, b_i$ is contained in the $(k-1)$-D lattice spanned by the weakly $\left(\frac{\pi}{3} + \epsilon\right)$-orthogonal basis $(b_1, b_2, \ldots, b_{k-1})$. For $u_k = 0$, by the induction hypothesis, we have

$$\left\|\sum_{i=1}^{k} u_i\, b_i\right\| = \left\|\sum_{i=1}^{k-1} u_i\, b_i\right\| \geq \min_{j \in \{1,2,\ldots,k-1\}} \|b_j\| \geq \min_{j \in \{1,2,\ldots,k\}} \|b_j\|.$$

If $\epsilon > 0$, then the first inequality in the above expression can hold as an equality only if $\sum_{i=1}^{k-1} |u_i| = 1$. If $u_k \neq 0$ and $u_i = 0$ for $i = 1, 2, \ldots, k-1$, then again

$$\left\|\sum_{i=1}^{k} u_i\, b_i\right\| \geq \|b_k\| \geq \min_{j \in \{1,2,\ldots,k\}} \|b_j\|.$$

Again, it is necessary that $|u_k| = 1$ for the equality to hold above.

Assume that $u_k \neq 0$ and $u_i \neq 0$ for some $i \in \{1, 2, \ldots, k-1\}$. Now $\sum_{i=1}^{k} u_i\, b_i$ is contained in the 2-D lattice spanned by the vectors $\sum_{i=1}^{k-1} u_i\, b_i$ and $u_k b_k$. Since the ordered set $(b_1, b_2, \ldots, b_k)$ is weakly $\left(\frac{\pi}{3} + \epsilon\right)$-orthogonal, the angle between the nonzero vectors $\sum_{i=1}^{k-1} u_i\, b_i$ and $u_k b_k$ lies in the interval $\left[\frac{\pi}{3} + \epsilon, \frac{\pi}{2}\right]$. Invoking Theorem 1 for 2-D lattices, we have

$$\begin{aligned}
\left\|\sum_{i=1}^{k} u_i\, b_i\right\| &\geq \min\left(\left\|\sum_{i=1}^{k-1} u_i\, b_i\right\|,\ \|u_k b_k\|\right) \\
&\geq \min\left(\min_{j \in \{1,2,\ldots,k-1\}} \|b_j\|,\ \|u_k b_k\|\right) \\
&\geq \min_{j \in \{1,2,\ldots,k\}} \|b_j\|. \qquad (11)
\end{aligned}$$

Thus, the set of basis vectors $\{b_1, b_2, \ldots, b_k\}$ contains a shortest nonzero vector in the k-D lattice. Also, if $\epsilon > 0$, then equality is not possible in (11), and the second part of the theorem follows. $\square$

4.2. Proof of Theorem 2. Similar to the proof of Theorem 1, we first prove Theorem 2 for 2-D lattices and then prove the general case by induction.

4.2.1. Proof for 2-D lattices. Consider a 2-D lattice in $\mathbb{R}^n$ with basis vectors b_1 and b_2 such that the basis $\{b_1, b_2\}$ is weakly θ-orthogonal with $\theta > \frac{\pi}{3}$. Note that for 2-D lattices, weak θ-orthogonality is the same as θ-orthogonality. Without loss of generality (w.l.o.g.), we can assume that $1 = \|b_1\| \leq \|b_2\|$. Further, by rotating the 2-D lattice, the basis vectors can be expressed as the columns of the $n \times 2$ matrix

$$\begin{bmatrix} 1 & b_{21} \\ 0 & b_{22} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}.$$

Let $\theta' \in \left[\theta, \frac{\pi}{2}\right]$ denote the angle between b_1 and b_2. Clearly,

$$\cos\theta' = \frac{|b_{21}|}{\|b_2\|} \text{ and } \sin\theta' = \frac{|b_{22}|}{\|b_2\|}.$$

Since (6) holds by assumption,

$$\|b_2\| < \frac{\sqrt{3}\|b_1\|}{\sin\theta + \sqrt{3}\cos\theta} \leq \frac{\sqrt{3}\|b_1\|}{\sin\theta' + \sqrt{3}\cos\theta'} = \frac{\sqrt{3}}{\frac{|b_{22}|}{\|b_2\|} + \sqrt{3}\frac{|b_{21}|}{\|b_2\|}},$$

where we have used the fact that $\eta(\theta)$ is a nondecreasing function of θ for $\theta \in \left[\frac{\pi}{3}, \frac{\pi}{2}\right]$. Therefore,

$$|b_{22}| < \sqrt{3}(1 - |b_{21}|). \tag{12}$$

Let $\{\widetilde{b}_1, \widetilde{b}_2\}$ denote another $\frac{\pi}{3}$-orthogonal basis for the same 2-D lattice. Using Theorem 1 and Corollary 1, we infer that $\{b_1, b_2\}$ contains every shortest lattice vector (multiplied by ± 1) and $\{b_1, b_2\}$ and $\left\{\widetilde{b}_1, \widetilde{b}_2\right\}$ contain a common shortest lattice vector. Assume w.l.o.g. that $\widetilde{b}_1 = \pm b_1$ is a shortest lattice vector. Then, we can write

$$\begin{bmatrix} \widetilde{b}_1 & \widetilde{b}_2 \end{bmatrix} = \begin{bmatrix} b_1 & b_2 \end{bmatrix} \begin{bmatrix} \pm 1 & u \\ 0 & \pm 1 \end{bmatrix} \quad \text{with } u \in \mathbb{Z}.$$

To prove Theorem 2, we need to show that $u = 0$.

Let $\widetilde{\theta}$ denote the angle between $\widetilde{b}_1$ and $\pm\widetilde{b}_2$. Then,

$$\begin{aligned} \cos^2\widetilde{\theta} &= \frac{\left|\langle\widetilde{b}_1,\widetilde{b}_2\rangle\right|^2}{\left\|\widetilde{b}_1\right\|^2\left\|\widetilde{b}_2\right\|^2} \\ &= \frac{(u\pm b_{21})^2}{(u\pm b_{21})^2+b_{22}^2} \\ &> \frac{(u\pm b_{21})^2}{(u\pm b_{21})^2+3(1-|b_{21}|)^2} \qquad \text{(using (12))} \\ &= \frac{1}{1+\frac{3(1-|b_{21}|)^2}{(u\pm b_{21})^2}}. \end{aligned} \tag{13}$$

If $u \neq 0$, then

$$|u\pm b_{21}| \geq |u| - |b_{21}| \geq 1-|b_{21}| \geq 0 \qquad \text{(from (12))}.$$

Hence,

$$|u\pm b_{21}|^2 \geq (1-|b_{21}|)^2.$$

Therefore, from (13) we have

$$\cos^2\widetilde{\theta} > \frac{1}{4}, \tag{14}$$

which holds if and only if $\widetilde{\theta} < \frac{\pi}{3}$. Thus, $\{\widetilde{b}_1,\widetilde{b}_2\}$ can be $\frac{\pi}{3}$-orthogonal only if $u=0$. This proves Theorem 2 for 2-D lattices.

4.2.2. Proof for higher-dimensional lattices. Let $\mathcal{B}$ and $\widetilde{\mathcal{B}}$ be two $n\times k$ matrices defining bases of the same k-D lattice in $\mathbb{R}^n$. We can write $\mathcal{B}=\widetilde{\mathcal{B}}\mathcal{U}$ for some integer unimodular matrix $\mathcal{U}=(u_{ij})$. Using induction on k, we will show that if $\mathcal{B}$ is weakly θ-orthogonal with $\frac{\pi}{3}<\theta\leq\frac{\pi}{2}$, if the columns of $\mathcal{B}$ satisfy (6), and if $\widetilde{\mathcal{B}}$ is $\frac{\pi}{3}$-orthogonal, then $\widetilde{\mathcal{B}}$ can be obtained by permuting the columns of $\mathcal{B}$ and multiplying them by ±1. Equivalently, we will show every column of $\mathcal{U}$ has exactly one component equal to ±1 and all others equal to 0 (we call such a matrix a *signed permutation matrix*).

Assume that Theorem 2 holds for all $(k-1)$-D lattices with $k>2$. Let $b_1,b_2,\dots,b_k$ denote the columns of $\mathcal{B}$, and let $\widetilde{b}_1,\widetilde{b}_2,\dots,\widetilde{b}_k$ denote the columns of $\widetilde{\mathcal{B}}$. Since permuting the columns of $\widetilde{\mathcal{B}}$ does not destroy $\frac{\pi}{3}$-orthogonality, we can assume w.l.o.g. that $\widetilde{b}_1$ is $\widetilde{\mathcal{B}}$'s shortest vector. From Theorem 1, $\widetilde{b}_1$ is also a shortest lattice vector. Further, using Corollary 1, $\pm\widetilde{b}_1$ is contained in $\mathcal{B}$. Assume that $b_\ell=\pm\widetilde{b}_1$ for some $\ell\in\{1,2,\dots,k\}$. Then

$$\mathcal{B}=\widetilde{\mathcal{B}}\begin{bmatrix} u_{11} & \dots & u_{1\ell-1} & \pm1 & u_{1\ell+1} & \dots & u_{1k} \\ & & & \vdots & & & \\ & \mathcal{U}_1' & & 0 & & \mathcal{U}_2' & \\ & & & \vdots & & & \end{bmatrix}. \tag{15}$$

Above, $\mathcal{U}_1'$ is a $(k-1)\times(\ell-1)$ submatrix, where as $\mathcal{U}_2'$ is a $(k-1)\times(k-\ell)$ submatrix.

We will show that $u_{1j} = 0$ for all $j \in \{1, 2, \ldots, k\}$ with $j \neq \ell$. Define

$$\mathcal{B}_r = \begin{bmatrix} b_\ell & b_j \end{bmatrix}, \qquad \widetilde{\mathcal{B}}_r = \begin{bmatrix} \widetilde{b}_1 & \sum_{i=2}^{k} u_{ij}\widetilde{b}_i \end{bmatrix}. \tag{16}$$

Then, from (15) and (16),

$$\mathcal{B}_r = \widetilde{\mathcal{B}}_r \begin{bmatrix} \pm 1 & u_{1j} \\ 0 & 1 \end{bmatrix}.$$

Since $\mathcal{B}_r$ and $\widetilde{\mathcal{B}}_r$ are related by a unimodular matrix, they both define bases of the same 2-D lattice. Further, $\mathcal{B}_r$ is weakly θ-orthogonal with $||b_j|| < \eta(\theta)||b_\ell||$, and $\widetilde{\mathcal{B}}_r$ is $\frac{\pi}{3}$-orthogonal. Invoking Theorem 2 for 2-D lattices, we can infer that $u_{1j} = 0$. It remains to be shown that $\mathcal{U}' = [\mathcal{U}_1' \; \mathcal{U}_2']$ is also a signed permutation matrix, where

$$\mathcal{B}' = \widetilde{\mathcal{B}}' \mathcal{U}',$$

with $\mathcal{B}' = [b_1, b_2, \ldots, b_{\ell-1}, \; b_{\ell+1}, \ldots, b_k]$ and $\widetilde{\mathcal{B}}' = \begin{bmatrix} \widetilde{b}_2, \widetilde{b}_3, \ldots, \widetilde{b}_k \end{bmatrix}$. Observe that $\det(\mathcal{U}') = \det(\mathcal{U}) = \pm 1$. Both $\mathcal{B}'$ and $\widetilde{\mathcal{B}}'$ are bases of the same $(k-1)$-D lattice as $\mathcal{U}'$ is unimodular. $\widetilde{\mathcal{B}}'$ is $\frac{\pi}{3}$-orthogonal, whereas $\mathcal{B}'$ is weakly θ-orthogonal, and its columns satisfy (6). By the induction hypothesis, $\mathcal{U}'$ is a signed permutation matrix. Therefore, $\mathcal{U}$ is also a signed permutation matrix. □

4.3. Proof of Theorem 3. Theorem 3 is a direct consequence of the following lemma.

LEMMA 1. *Let $\mathcal{B} = (b_1, b_2, \ldots, b_m)$ be a weakly θ-orthogonal basis of a lattice, where $\theta > \frac{\pi}{3}$. Then, for any integers $u_1, u_2, \ldots, u_m$,*

$$\left\| \sum_{i=1}^{m} u_i b_i \right\| \geq \left(\frac{\sqrt{3}}{2} \right)^{m-1} \times \max_{i \in \{1,2,\ldots,m\}} \| u_i b_i \|. \tag{17}$$

Lemma 1 can be proved as follows. Consider the vectors b_1 and b_2; the angle θ between them lies in the interval $\left(\frac{\pi}{3}, \frac{\pi}{2}\right)$. Recall from (10) that

$$\|u_1 b_1 + u_2 b_2\|^2 \geq (|u_1| \, \|b_1\| - |u_2| \, \|b_2\|)^2 + |u_1||u_2|\|b_1\|\|b_2\|.$$

Consider the expression $(y - x)^2 + yx$ with $0 \leq x \leq y$. For fixed y this expression attains its minimum value of $\left(\frac{3}{4}\right) y^2$ when $x = \frac{y}{2}$. By setting $y = |u_1| \, \|b_1\|$ and $x = |u_2| \, \|b_2\|$ w.l.o.g, we can infer that

$$\|u_1 b_1 + u_2 b_2\| \geq \frac{\sqrt{3}}{2} \max_{i \in \{1,2\}} \|u_i b_i\|.$$

Since $\mathcal{B}$ is weakly θ-orthogonal, the angle between $u_k b_k$ and $\sum_{i=1}^{k-1} u_i b_i$ lies in the interval $\left(\frac{\pi}{3}, \frac{\pi}{2}\right)$ for $k = 2, 3, \ldots, m$. Hence (17) follows by induction. □

We now proceed to prove Theorem 3 by invoking Lemma 1. First, we define $\Delta = (\sqrt{3}/2)^{m-1}$. For any $j \in \{1, 2, \ldots, m\}$, we have

$$\|b_j\| = \left\| \sum_{i=1}^{m} u_{ij} \widetilde{b}_i \right\| \geq \Delta \max_{i \in \{1,2,\ldots,m\}} \left\| u_{ij} \widetilde{b}_i \right\| \geq \Delta \min_{i \in \{1,2,\ldots,m\}} \|\widetilde{b}_i\| \max_{i \in \{1,2,\ldots,m\}} |u_{ij}|.$$

Since $\mathcal{B}$ and $\widetilde{\mathcal{B}}$ are both weakly θ-orthogonal with $\theta > \frac{\pi}{3}$, $\min_{i\in\{1,2,\ldots,m\}} \|\widetilde{b}_i\| = \min_{i\in\{1,2,\ldots,m\}} \|b_i\|$. Therefore,

$$\Delta \max_{i\in\{1,2,\ldots,m\}} |u_{ij}| \leq \frac{\|b_j\|}{\min_{i\in\{1,2,\ldots,m\}} \|\widetilde{b}_i\|} \leq \frac{\max_{i\in\{1,2,\ldots,m\}} \|b_i\|}{\min_{i\in\{1,2,\ldots,m\}} \|b_i\|} = \Delta\kappa\left(\mathcal{B}\right).$$

Thus, $|u_{ij}| \leq \kappa\left(\mathcal{B}\right)$ for all i and j. □

5. Random lattices and SVP. In several applications, the orthogonality of random lattice bases and the length of the shortest vector $\lambda(\mathcal{L})$ in a random lattice $\mathcal{L}$ play an important role. For example, in certain wireless communications applications involving multiple transmitters and receivers, the received message ideally lies on a lattice spanned by a random basis [7]. The random basis models the fluctuations in the communication channel between each transmitter-receiver pair. Due to the presence of noise, the ideal received message is retrieved by solving a CVP. The complexity of this problem is controlled by the orthogonality of the random basis [1]. Random bases are also employed to perform error correction coding [28] and in cryptography [28]. The level of achievable error correction is controlled by the shortest vector in the lattice.

In this section, we determine the θ-orthogonality of random bases. This result immediately lets us identify conditions under which a random basis contains (with high probability) the shortest lattice vector.

Before describing our results on random lattices and bases, we first review some known properties of random lattices and then list some powerful results from random matrix theory.

5.1. Known properties of random lattices. Consider an m-D lattice generated by a random basis with each of the m basis vectors chosen independently and uniformly from the unit ball in $\mathbb{R}^n$ $(n \geq m)$.[2] With m fixed and with $n \to \infty$, the probability that the random basis is Minkowski-reduced tends to 1 [11]. Thus, as $n \to \infty$, the random basis contains a shortest vector in the lattice almost surely. Recently, [3] proved that, as $n - m \to \infty$, the probability that a random basis is LLL-reduced $\to 1$. Further, [3] also showed that a random basis is LLL-reduced with nonzero probability when $n - m$ is fixed with $n \to \infty$.

5.2. Known properties of random matrices. Random matrix theory, a rich field with many applications [6, 12], has witnessed several significant developments over the past few decades [12, 18, 19, 30]. We will invoke some of these results to derive some new properties of random bases and lattices; the paper [6] provides an excellent summary of the results we mention below.

Consider an $n \times m$ matrix $\mathcal{B}$ with each element of $\mathcal{B}$ an independent identically distributed random variable. If the variables are zero-mean Gaussian distributed with variance $\frac{1}{n}$, then we refer to such a $\mathcal{B}$ as a *Gaussian* random basis. If the variables take on values in $\{-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\}$ with equal probability, then we term $\mathcal{B}$ to be a *Bernoulli* random basis. We say that $\mathcal{B}$ is a *scaled* Gaussian (Bernoulli) basis if it is obtained by scaling the columns of a Gaussian (Bernoulli) basis arbitrarily.

Gaussian and Bernoulli random bases enjoy the following properties. Below, ψ_i^2, $i = 1, 2, \ldots, m$, denote the eigenvalues of $\mathcal{B}^T\mathcal{B}$.

[2] The m vectors form a basis because they are linearly independent almost surely.

1. For both Gaussian and Bernoulli $\mathcal{B}$, $\mathcal{B}^T\mathcal{B}$'s smallest and largest eigenvalues, $\psi^2_{\min}$ and $\psi^2_{\max}$, converge almost surely to $(1-\sqrt{c})^2$ and $(1+\sqrt{c})^2$, respectively, as $n, m \to \infty$ and $\frac{m}{n} \to c < 1$ [6,12,30].
2. Let $\epsilon > 0$ be given. Then, there exists an N_ϵ such that, for every $n > N_\epsilon$ and $r > 0$,

$$P\left(|\psi_{\min}| \leq \left(1 - \sqrt{\frac{m}{n}}\right) - (r + \epsilon)\right) \leq e^{-\frac{nr^2}{\rho}}, \tag{18}$$

$$P\left(|\psi_{\max}| \geq \left(1 + \sqrt{\frac{m}{n}}\right) + (r + \epsilon)\right) \leq e^{-\frac{nr^2}{\rho}}, \tag{19}$$

   with $\rho = 2$ for Gaussian $\mathcal{B}$ and $\rho = 16$ for Bernoulli $\mathcal{B}$ [6,18].

In essence, a random matrix's largest and smallest singular values converge, respectively, to $1 \pm \sqrt{\frac{m}{n}}$ almost surely as $n, m \to \infty$ and lie close to $1 \pm \sqrt{\frac{m}{n}}$ with very high probability at finite (but sufficiently large) n.

5.3. New results on random lattices. We now formally state the new properties of random lattices mentioned in the introduction plus several additional corollaries. Our proofs assume that the lattices are generated by Gaussian or Bernoulli random bases (whose column vectors are essentially unit-length). However, our results easily extend to lattices generated by Gaussian or Bernoulli random bases because the θ-orthogonality of a basis does not change upon scaling the basis vectors.

The key step in proving our results is to relate the condition number of a random basis to its θ-orthogonality (see Lemma 2). A matrix's condition number is defined as the ratio of the largest to the smallest singular value. Then we invoke the results in section 5.2 to quantify the θ-orthogonality of random bases. Finally we invoke previously deduced properties of nearly orthogonal lattice bases.

We wish to emphasize that we prove our statements only for lattices which are not full-dimensional. Our computational results suggest these statements are not true for full-dimensional lattices. Further, Sorkin [31] proves that, with high probability, Gaussian random matrices are not nearly orthogonal when $m > n/4$. See the paragraph after Corollary 3 for more details.

LEMMA 2. *Consider an arbitrary $n \times m$ real-valued matrix $\mathcal{B}$, with $m \leq n$, whose largest and smallest singular values are denoted by $\psi_{\max}$ and $\psi_{\min}$, respectively. Then the columns of $\mathcal{B}$ are θ-orthogonal with*

$$\theta = \sin^{-1}\left(\frac{2\,\psi_{\max}\,\psi_{\min}}{\psi^2_{\min} + \psi^2_{\max}}\right). \tag{20}$$

The proof is given in section 5.4. The value of θ in (20) is the best possible in the sense that there is a 2×2 matrix $\mathcal{B}$ with singular values $\psi_{\min}$ and $\psi_{\max}$ such that the angle between the two columns of $\mathcal{B}$ is given by (20). Note that for large $\frac{\psi_{\min}}{\psi_{\max}}$ (that is, for a small condition number), the θ in (20) is close to $\frac{\pi}{2}$. Thus, Lemma 2 quantifies our intuition that a matrix with a small condition number should be nearly orthogonal.

By combining Lemma 2 with the properties of random matrices listed in section 5.2, we can immediately deduce the θ-orthogonality of an $n \times m$ random basis. See section 5.4.2 for the proof.

THEOREM 4. *Let $\mathcal{B}$ denote an $n \times m$ Gaussian or Bernoulli random basis. If $m \leq cn$, $0 \leq c < 1$, then as $n \to \infty$, $\mathcal{B}$ is θ-orthogonal almost surely with*

$$\theta = \sin^{-1}\left(\frac{1-c}{1+c}\right). \tag{21}$$

Further, given an $\epsilon > 0$, there exists an N_ϵ such that, for every $n > N_\epsilon$ and $r > 0$, $\mathcal{B}$ is θ-orthogonal,

$$\theta = \sin^{-1}\left(\frac{1-c}{1+c} - \frac{3\sqrt{3}}{4}(r+\epsilon)\right), \tag{22}$$

with probability greater than $1 - 2e^{-\frac{nr^2}{\rho}}$, where $\rho = 2$ for Gaussian $\mathcal{B}$ and $\rho = 16$ for Bernoulli $\mathcal{B}$.

The value of θ in (21) is not the best possible in the sense that, for a given value of c, a random $n \times m$ Gaussian matrix with $m \leq cn$ would be θ'-orthogonal (with high probability) for some $\theta' > \theta$ (see Figure 2). The reason is that the θ predicted by Lemma 2 is satisfied by *all* matrices. However, Theorem 4 is restricted to random matrices.

Theorem 4 allows us to bound the length of the shortest nonzero vector in a random lattice.

COROLLARY 2. *Let the $n \times m$ matrix $\mathcal{B} = (b_1, b_2, \ldots, b_m)$, with $m \leq cn$ and $0 \leq c < 1$, denote a Gaussian or Bernoulli random basis for a lattice $\mathcal{L}$. Then the shortest vector's length $\lambda(\mathcal{L})$ satisfies*

$$\lambda(\mathcal{L}) \geq \frac{1-c}{1+c}$$

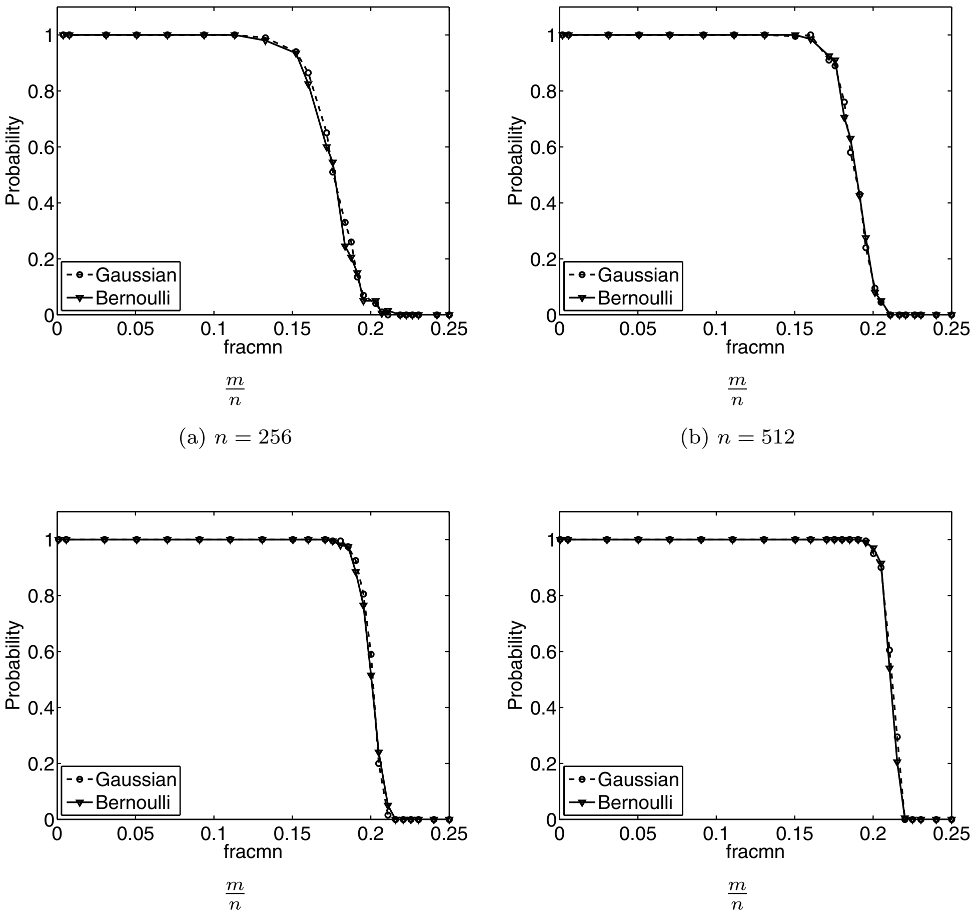
almost surely as $n \to \infty$.

Each column of a Bernoulli $\mathcal{B}$ is unit-length by construction. For Gaussian $\mathcal{B}$, it is not difficult to show that all columns have length 1 almost surely as $n \to \infty$. Hence Corollary 2 is an immediate consequence of Theorem 4 and (2). Corollary 2 implies that, in random lattices that are not full-dimensional, it is easy to obtain approximate solutions to the SVP (within a constant factor). This is because for random lattices in $\mathbb{R}^n$ with dimension $n(1-\epsilon)$, $\lambda(\mathcal{L})$ is greater than ϵ times the length of the shortest basis vector (approximately). Compare this with Daudé's and Vallée's [9] result that in random full-dimensional lattices in $\mathbb{R}^n$, $\lambda(\mathcal{L})$ is at least $O(1/\sqrt{n})$ times the length of the shortest basis vector with high probability.

By substituting $\theta = \frac{\pi}{3}$ into Theorem 4 and then invoking Corollary 1, we can deduce sufficient conditions for a random basis to be $\frac{\pi}{3}$-orthogonal.

COROLLARY 3. *Let the $n \times m$ matrix $\mathcal{B}$ denote a Gaussian or Bernoulli random basis for lattice $\mathcal{L}$. If $\frac{m}{n} \leq c < \left(7 - \sqrt{48}\right)$ (≈ 0.071), then $\mathcal{B}$ is $\frac{\pi}{3}$-orthogonal almost surely as $n \to \infty$. Further, given an $\epsilon > 0$, there exists an N_ϵ such that, for every $n > N_\epsilon$ and $\frac{4(1-c)}{3\sqrt{3}(1+c)} - \epsilon - \frac{2}{3} > r > 0$, $\mathcal{B}$ is $\frac{\pi}{3}$-orthogonal with probability greater than $1 - 2e^{-nr^2/\rho}$, where $\rho = 2$ for Gaussian $\mathcal{B}$ and $\rho = 16$ for Bernoulli $\mathcal{B}$.*

Figure 2 illustrates that, in practice, an $n \times m$ Gaussian and Bernoulli random matrix is nearly orthogonal for much larger values of $\frac{m}{n}$ than our results claim. Our plots suggest that the probability for a random basis to be nearly orthogonal sharply transitions from 1 to 0 for $\frac{m}{n}$ values in the interval $[0.2, 0.25]$. Sorkin [31] has shown us that if the columns of $\mathcal{B}$ represent points chosen uniformly from the unit sphere in



(a) $n = 256$

(b) $n = 512$

(c) $n = 1024$

(d) $n = 2048$

FIG. 2. *Empirical probability that a $n \times m$ Gaussian or Bernoulli random matrix is $\frac{\pi}{3}$-orthogonal. At $n = 256$, 512, 1024, and 2048 and at m indicated by circles (for Gaussian) and triangles (for Bernoulli), we tested* 200 *randomly generated matrices. The empirical probability is the fraction of random matrices that were $\frac{\pi}{3}$-orthogonal.*

$\mathbb{R}^n$ (one can obtain such points by dividing the columns of a Gaussian matrix by their norms), then the best possible $\frac{m}{n}$ value for random $n \times m$ matrices to be $\frac{\pi}{3}$-orthogonal is $\frac{m}{n} = 0.25$. Further, if $m/n > 0.25$, $\mathcal{B}$ is almost surely not $\frac{\pi}{3}$-orthogonal as $n \to \infty$. For large n, the columns of a Gaussian matrix almost surely have length 1 and thus behave like points chosen uniformly from the unit sphere in $\mathbb{R}^n$. Therefore, as $n \to \infty$, random $n \times n/4$ Gaussian matrices are almost surely $\frac{\pi}{3}$-orthogonal.

5.4. Proof of results on random lattices. This section provides the proofs for Lemma 2 and Theorem 4.

5.4.1. Proof of Lemma 2. Our goal is to construct a lower-bound for the angle between any column of $\mathcal{B}$ and the subspace spanned by all the other columns in terms of the singular values of $\mathcal{B}$. Clearly, if $\psi_{\min} = 0$, then the columns of $\mathcal{B}$ are linearly dependent. Hence, (20) holds as $\mathcal{B}$'s columns are θ-orthogonal with $\theta = 0$. For the rest of the proof, we will assume that $\psi_{\min} \neq 0$.

Consider the SVD of $\mathcal{B}$:

$$\mathcal{B} = \mathcal{X}\Psi\mathcal{Y}, \tag{23}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are $n\times m$ and $m\times m$ real-valued matrices, respectively, with orthonormal columns and Ψ is a $m \times m$ real-valued diagonal matrix. Let b_i and x_i denote the ith column of $\mathcal{B}$ and $\mathcal{X}$, respectively, let y_{ij} denote the element from the ith row and jth column of $\mathcal{Y}$, and let ψ_i denote the ith diagonal element of Ψ. Then, (23) can be rewritten as

$$\begin{bmatrix} b_1 & b_2 & \dots & b_m \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \dots & x_m \end{bmatrix} \begin{bmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & \psi_m \end{bmatrix} \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \vdots & & \ddots & \vdots \\ y_{m1} & \cdots & \cdots & y_{mm} \end{bmatrix}.$$

We now analyze the angle between b_1 (w.l.o.g) and the subspace spanned by $\{b_2, b_3, \dots, b_m\}$. Note that

$$b_1 = \sum_{i=1}^{m} \psi_i\, y_{i1}\, x_i.$$

Let $\widetilde{b}_1$ denote an arbitrary nonzero vector in the subspace spanned by $\{b_2, b_3, \dots, b_m\}$. Then,

$$\widetilde{b}_1 = \sum_{k=2}^{m} \alpha_k\, b_k = \sum_{k=2}^{m} \alpha_k \sum_{i=1}^{m} \psi_i\, y_{ik}\, x_i = \sum_{i=1}^{m} x_i\, \psi_i \sum_{k=2}^{m} \alpha_k\, y_{ik}$$

for some $\alpha_k \in \mathbb{R}$ with $\sum_k |\alpha_k| > 0$. Let $\widetilde{y}_{i1} = \sum_{k=2}^{m} \alpha_k\, y_{ik}$. Then,

$$\widetilde{b}_1 = \sum_{i=1}^{m} \psi_i \widetilde{y}_{i1} x_i.$$

Let $\widetilde{\theta} \geq \theta$ denote the angle between b_1 and $\widetilde{b}_1$. Then,

$$\cos\widetilde{\theta} = \frac{\left|\left\langle b_1, \widetilde{b}_1 \right\rangle\right|}{\|b_1\| \left\|\widetilde{b}_1\right\|} = \frac{|\langle \sum_{i=1}^{m} \psi_i\, y_{i1}\, x_i, \sum_{i=1}^{m} \psi_i\, \widetilde{y}_{i1}\, x_i \rangle|}{\|\sum_{i=1}^{m} \psi_i\, y_{i1}\, x_i\| \; \|\sum_{i=1}^{m} \psi_i\, \widetilde{y}_{i1}\, x_i\|} \tag{24}$$

$$= \frac{\left|\sum_{i=1}^{m} \psi_i^2\, y_{i1}\, \widetilde{y}_{i1}\right|}{\sqrt{\sum_{i=1}^{m} \psi_i^2\, y_{i1}^2}\; \sqrt{\sum_{i=1}^{m} \psi_i^2\, \widetilde{y}_{i1}^2}}, \tag{25}$$

where the orthonormality of the $\mathcal{X}$ columns is used to obtain (25) from (24). Let y_i, $i = 1, 2, \dots, m$, and $\widetilde{y}_1$ denote column vectors

$$y_i := \begin{bmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{mi} \end{bmatrix} \text{ and } \widetilde{y}_1 := \begin{bmatrix} \widetilde{y}_{11} \\ \widetilde{y}_{21} \\ \vdots \\ \widetilde{y}_{m1} \end{bmatrix}.$$

Since $\widetilde{y}_1 = \sum_{k=2}^{m} \alpha_k\, y_k$,

$$\widetilde{y}_1^T y_1 = 0.$$

Then (25) can be rewritten using matrix notation as

$$\cos\widetilde{\theta} = \frac{\left|y_1^T\,\Psi^2\,\widetilde{y}_1\right|}{\sqrt{y_1^T\Psi^2\,y_1}\,\sqrt{\widetilde{y}_1^T\Psi^2\,\widetilde{y}_1}}, \tag{26}$$

with $\Psi^2 := \Psi^T\Psi$. The angle $\widetilde{\theta}$ is minimized when the right-hand side of (26) is maximized.

For arbitrary $\mathcal{B}$ with only the singular values known (that is, Ψ is known), the θ-orthogonality of $\mathcal{B}$ is given by solving the following constrained optimization problem:

$$\cos\theta = \max_{y_1,\widetilde{y}_1} \quad \frac{\left|y_1^T\Psi^2\widetilde{y}_1\right|}{\sqrt{y_1^T\Psi^2\,y_1}\,\sqrt{\widetilde{y}_1^T\Psi^2\widetilde{y}_1}} \quad \text{such that} \quad \widetilde{y}_1^T y_1 = 0. \tag{27}$$

Wielandt's inequality [14, Thm. 7.4.34] states that if A is a positive definite matrix, with $\gamma_{\min}$ and $\gamma_{\max}$ denoting its minimum and maximum eigenvalues (both are positive), then

$$|x^TAy|^2 \le \left(\frac{\gamma_{\max}-\gamma_{\min}}{\gamma_{\max}+\gamma_{\min}}\right)^2 (x^TAx)(y^TAy)$$

for every pair of orthogonal vectors x and y (equality holds for some pair of orthogonal vectors). In our problem, $A=\Psi^2$, $x=\widetilde{y}_1$, $y=y_1$, $\gamma_{\max}=\psi^2_{\max}$, and $\gamma_{\min}=\psi^2_{\min}$. Therefore, using Wielandt's inequality and (27), we have

$$\cos\theta = \frac{\psi^2_{\max}-\psi^2_{\min}}{\psi^2_{\max}+\psi^2_{\min}}.$$

Hence

$$\sin\theta = \frac{2\psi_{\max}\psi_{\min}}{\psi^2_{\max}+\psi^2_{\min}}, \tag{28}$$

which proves (20). □

5.4.2. Proof of Theorem 4. The first part of Theorem 4 follows easily. From section 5.2, we can infer that with $m \le cn$, $0 \le c < 1$, both $\psi_{\min} \ge 1-\sqrt{c}$ and $\psi_{\max} \le 1+\sqrt{c}$ almost surely as $n\to\infty$. Invoking Lemma 2 and substituting $\psi_{\min} = 1-\sqrt{c}$ and $\psi_{\max} = 1+\sqrt{c}$ into (20), it follows that, as $n\to\infty$, $\mathcal{B}$ is θ-orthogonal almost surely with θ given by (21).

We now focus on proving the second part of Theorem 4. Let $d=\sqrt{c}$, and define

$$G(d) := \frac{1-d^2}{1+d^2}.$$

We first show that, for $\delta \ge 0$,

$$G(d+\delta) \ge G(d) - \frac{3\sqrt{3}}{4}\delta. \tag{29}$$

Using the mean value theorem,

$$G(d+\delta) = G(d) + G'\left(d+\widetilde{\delta}\right)\delta \quad \text{for some } \widetilde{\delta}\in(0,\delta), \tag{30}$$

with G' denoting the derivative of G with respect to d. Further,

$$G'(d) = \frac{-4d}{\left(1+d^2\right)^2} \geq -\frac{3\sqrt{3}}{4} \qquad \text{for } d > 0. \tag{31}$$

One can verify the inequality above by differentiating $G'(d)$ and observing that $G'(d)$ is minimized when $3d^4 + 2d^2 - 1 = 0$. The only positive root of this quadratic equation is $d^2 = 1/3$ or $d = 1/\sqrt{3}$. Combining (30) and (31), we obtain (29).

From the results in section 5.2, it follows that the probability that both minimum and maximum singular values of $\mathcal{B}$ satisfy

$$|\psi_{\min}| \geq 1 - \left(\sqrt{c} + r + \epsilon\right) \quad \text{and} \quad |\psi_{\max}| \leq 1 + \left(\sqrt{c} + r + \epsilon\right) \tag{32}$$

is greater than $1 - 2e^{-\frac{nr^2}{\rho}}$. When (32) holds, $\mathcal{B}$ is at least $\sin^{-1}\left(G\left(\sqrt{c} + r + \epsilon\right)\right)$-orthogonal. This follows from (20). Invoking (29), we can infer that $\mathcal{B}$ is θ-orthogonal with θ as in (22). □

6. JPEG CHEst.

In this section, we review the JPEG CHEst problem that motivates our study of nearly orthogonal lattices and describe how we use this paper's results to solve this problem. We first touch on the topic of digital color image representation and briefly describe the essential components of JPEG image compression.

6.1. Digital color image representation.

Traditionally, digital color images are represented by specifying the color of each pixel, the smallest unit of image representation. According to the trichromatic theory [29], three parameters are sufficient to specify any color perceived by humans.[3] For example, a pixel's color can be conveyed by a vector $w_{\scriptscriptstyle RGB} = (w_{\scriptscriptstyle R}, w_{\scriptscriptstyle G}, w_{\scriptscriptstyle B}) \in \mathbb{R}^3$, where $w_{\scriptscriptstyle R}$, $w_{\scriptscriptstyle G}$, and $w_{\scriptscriptstyle B}$ specify the intensity of the color's red (R), green (G), and blue (B) components, respectively. Call $w_{\scriptscriptstyle RGB}$ the RGB encoding of a color. RGB encodings are vectors in the vector space where the R, G, and B colors form the standard unit basis vectors; this coordinate system is called the RGB *color space*. A color image with M pixels can be specified using RGB encodings by a matrix $P \in \mathbb{R}^{3\times M}$.

6.2. JPEG compression and decompression.

To achieve color image compression, schemes such as JPEG first transform the image to a color encoding other than the RGB encoding and then perform *quantization*. Such color encodings can be related to the RGB encoding by a *color-transform* matrix $C \in \mathbb{R}^{3\times 3}$. The columns of C form a different basis for the color space spanned by the R, G, and B vectors. Hence an RGB encoding $w_{\scriptscriptstyle RGB}$ can be transformed to the C encoding vector as $C^{-1}w_{\scriptscriptstyle RGB}$; the image P is mapped to $C^{-1}P$. For example, the matrix relating the RGB color space to the ITU.BT-601 $YCbCr$ color space is given by [27]

$$\begin{bmatrix} w_{\scriptscriptstyle Y} \\ w_{\scriptscriptstyle Cb} \\ w_{\scriptscriptstyle Cr} \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.5 \\ 0.5 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} w_{\scriptscriptstyle R} \\ w_{\scriptscriptstyle G} \\ w_{\scriptscriptstyle B} \end{bmatrix}. \tag{33}$$

The quantization step is performed by first choosing a diagonal positive (nonzero entries are positive) integer *quantization* matrix Q and then computing the quantized (compressed) image from $C^{-1}P$ as $P_c = \left\lceil Q^{-1}C^{-1}P \right\rfloor$, where $\lceil . \rfloor$ stands for

[3] The underlying reason is that the human retina has only three types of receptors that influence color perception.

the operation of rounding to the nearest integer. JPEG decompression constructs $P_d = CQP_c = CQ\left\lceil Q^{-1}C^{-1}P\right\rfloor$. Larger Q's achieve more compression but at the cost of greater distortion between the decompressed image P_d and the original image P.

In practice, the image matrix P is first decomposed into different frequency components $P = \{P_1, P_2, \dots, P_k\}$ for some $k > 1$ (usually $k = 64$) during compression. Then, a common color transform C is applied to all the submatrices $P_1, P_2, \dots, P_k$, but each submatrix P_i is quantized with a different quantization matrix Q_i. The compressed image is $P_c = \{P_{c,1}, P_{c,2}, \dots, P_{c,k}\} = \left\{\left\lceil Q_1^{-1}C^{-1}P_1\right\rfloor, \left\lceil Q_2^{-1}C^{-1}P_2\right\rfloor, \dots, \left\lceil Q_k^{-1}C^{-1}P_k\right\rfloor\right\}$, and the decompressed image is $P_d = \{CQ_1P_{c,1}, CQ_2P_{c,2}, \dots, CQ_kP_{c,k}\}$.

During compression, the JPEG-compressed file format stores the matrix C and the matrices Q_i along with P_c. These stored matrices are utilized to decompress the JPEG image but are discarded afterward. Hence we refer to the set $\{C, Q_1, Q_2, \dots, Q_k\}$ as the *compression history* of the image.

6.3. JPEG CHEst problem statement. This paper's contributions are motivated by the following question: *Given a decompressed image $P_d = \left\{CQ_1P_{c,1}, CQ_2P_{c,2}, \dots, CQ_kP_{c,k}\right\}$ and some information about the structure of C and the Q_i's, can we estimate the color transform C and the quantization matrices Q_i?* As $\{C, Q_1, Q_2, \dots, Q_k\}$ comprises the compression history of the image, we refer to this problem as JPEG CHEst. An image's compression history is useful for applications such as JPEG recompression [5, 22, 23].

6.4. Near-orthogonality and JPEG CHEst. The columns of $CQ_iP_{c,i}$ lie on a 3-D lattice with basis CQ_i because $P_{c,i}$ is an integer matrix. The estimation of CQ_i's comprises the main step in JPEG CHEst. Since a lattice can have multiple bases, we must exploit some additional information about practical color transforms to correctly deduce the CQ_i's from the $CQ_iP_{c,i}$'s. Most practical color transforms aim to represent a color using an approximately rotated reference coordinate system. Consequently, most practical color transform matrices C (and, thus, CQ_i) can be expected to be almost orthogonal. We have verified that all C's used in practice are weakly $\left(\frac{\pi}{3} + \epsilon\right)$-orthogonal, with $0 < \epsilon \le \frac{\pi}{6}$.[4] Thus, nearly orthogonal lattice bases are central to JPEG CHEst.

6.5. Our approach. Our approach is to first estimate the products CQ_i by exploiting the near-orthogonality of C and to then decompose CQ_i into C and Q_i. We will assume that C is weakly $\left(\frac{\pi}{3} + \epsilon\right)$-orthogonal, $0 < \epsilon \le \frac{\pi}{6}$.

6.5.1. Estimating the CQ_i's. Let $\mathcal{B}_i$ be a basis of the lattice $\mathcal{L}_i$ spanned by CQ_i. Then, for some unimodular matrix $\mathcal{U}_i$, we have

$$\mathcal{B}_i = CQ_i\mathcal{U}_i. \tag{34}$$

If $\mathcal{B}_i$ is given, then estimating CQ_i is equivalent to estimating the respective $\mathcal{U}_i$.

Thanks to our problem structure, the correct $\mathcal{U}_i$'s satisfy the following constraints. Note that these constraints become increasingly restrictive as the number of frequency components k increases.

1. The $\mathcal{U}_i$'s are such that $\mathcal{B}_i\mathcal{U}_i^{-1}$ is weakly $\left(\frac{\pi}{3} + \epsilon\right)$-orthogonal.
2. The product $\mathcal{U}_i\mathcal{B}_i^{-1}\mathcal{B}_j\mathcal{U}_j^{-1}$ is diagonal with positive entries for any $i, j \in \{1, 2, \dots, k\}$. This is an immediate consequence of (34).

[4] In general, the stronger assumption of $\frac{\pi}{3}$-orthogonality does not hold for some practical color transform matrices.

Table 6.1
Number of unimodular matrices satisfying constraints 3 *and* 4 *for small* κ.

κ	Constraint 4	Constraints 3 and 4
1	6960	5232
2	135408	43248
3	1281648	197616
4	5194416	513264
5	20852976	1324272

If, in addition, $\mathcal{B}_i$ is weakly $\left(\frac{\pi}{3}+\epsilon\right)$-orthogonal, then the following hold.

3. The columns of $\mathcal{U}_i$ corresponding to the shortest columns of $\mathcal{B}_i$ are the standard unit vectors times ±1. This follows from Corollary 1 because the columns of both $\mathcal{B}_i$ and CQ_i indeed contain all shortest vectors in $\mathcal{L}_i$ up to multiplication by ±1.
4. All entries of $\mathcal{U}_i$ are $\leq \kappa(\mathcal{B}_i)$ in magnitude. This follows from Theorem 3.

We now outline our heuristic.

(i) Obtain bases $\mathcal{B}_i$ for the lattices $\mathcal{L}_i$, $i = 1, 2, \ldots, k$. Construct a weakly $\left(\frac{\pi}{3}+\epsilon\right)$-orthogonal basis $\mathcal{B}_\ell$ for at least one lattice $\mathcal{L}_\ell$, $\ell \in \{1, 2, \ldots, k\}$.

(ii) Compute $\kappa(\mathcal{B}_\ell)$.

(iii) For every unimodular matrix $\mathcal{U}_\ell$ satisfying constraints 1, 3, and 4, go to step (iv).

(iv) For $\mathcal{U}_\ell$ chosen in step (iii), test if there exist unimodular matrices $\mathcal{U}_j$ for each $j = 1, 2, \ldots, k, j \neq \ell$, that satisfies constraint 2. If such a collection of matrices exists, then return this collection; otherwise, go to step (iii).

For step (i), we simply use the LLL algorithm to compute LLL-reduced bases for each $\mathcal{L}_i$. Such bases are not guaranteed to be weakly $\left(\frac{\pi}{3}+\epsilon\right)$-orthogonal, but in practice, this is usually the case for a number of the $\mathcal{L}_i$'s. Instead of LLL, the method proposed in [24] could be also employed (as suggested by the referees). In contrast to the LLL, [24] always finds a basis that contains the shortest lattice vector in low-dimensional lattices (up to four dimensions) such as the $\mathcal{L}_i$'s in our problem. In step (iv), for each frequency component $j \neq \ell$, we compute the diagonal matrix D_j with smallest positive entries such that $\widetilde{\mathcal{U}}_j = \mathcal{B}_j^{-1}\mathcal{B}_\ell\mathcal{U}_\ell^{-1}D_j$ is integral, and then we test whether $\widetilde{\mathcal{U}}_j$ is unimodular. If not, then for the given $\mathcal{U}_\ell$ no appropriate unimodular matrix $\mathcal{U}_j$ exists.

The overall complexity of the heuristic is determined mainly by the number of times we repeat step (iv), which equals the number of distinct choices for $\mathcal{U}_\ell$ in step (iii). This number is typically not very large because in step (i) we are usually able to find some weakly $\left(\frac{\pi}{3}+\epsilon\right)$-orthogonal basis $\mathcal{B}_l$ with $\kappa < 2$. In fact, we enumerate all unimodular matrices satisfying constraints 3 and 4 and then test constraint 1. (In practice, one can avoid enumerating the various column permutations of a unimodular matrix). Table 6.1 provides the number of unimodular matrices satisfying constraint 4 alone and also constraints 3 and 4. Clearly, constraints 3 and 4 help us to significantly limit the number of unimodular matrices we need to test, thereby speeding up our search.

Our heuristic returns a collection of unimodular matrices $\{\mathcal{U}_i\}$ that satisfy constraints 1 and 2; of course, they also satisfy constraints 3 and 4 if the corresponding $\mathcal{B}_i$'s are weakly $\left(\frac{\pi}{3}+\epsilon\right)$-orthogonal. From the U_i's, we compute $CQ_i = \mathcal{B}_i\mathcal{U}^{-1}$. If constraints 1 and 2 can be satisfied by another solution $\{\mathcal{U}_i'\}$, then it is easy to see that $\mathcal{U}_i' \neq \mathcal{U}_i$ for *every* $i = 1, 2, \ldots, k$. In section 6.5.3, we will argue (without proof) that constraints 1 and 2 are likely to have a unique solution in most practical cases.

6.5.2. Splitting CQ_i into C and Q_i. Decomposing the CQ_i's into C and Q_i's is equivalent to determining the norm of each column of C because the Q_i's are diagonal matrices. Since the Q_i's are integer matrices, the norm of each column of CQ_i is an integer multiple of the corresponding column norm of C. In other words, the norms of the jth column ($j \in \{1, 2, 3\}$) of different CQ_i's form a sublattice of the one-dimensional lattice spanned by the jth column norm of C. As long as the greatest common divisor of the jth diagonal values of the matrices Q_i is 1, we can uniquely determine the jth column of C; the values of Q_i follow trivially.

6.5.3. Uniqueness. Does JPEG CHEst have a unique solution? In other words, is there a collection of matrices

$$(C', Q'_1, Q'_2, \ldots, Q'_k) \neq (C, Q_1, Q_2, \ldots, Q_k)$$

such that $C'Q'_i$ is a weakly $\left(\frac{\pi}{3} + \epsilon\right)$-orthogonal basis of $\mathcal{L}_i$ for all $i \in \{1, 2, \ldots, k\}$? We believe that the solution can be nonunique only if the Q_i's are chosen carefully. For example, let Q be a diagonal matrix with positive diagonal coefficients. Assume that, for $i = 1, 2, \ldots, k$, $Q_i = \alpha_i Q$, with $\alpha_i \in \mathbb{R}$ and $\alpha_i > 0$. Further, assume that there exists a unimodular matrix $\mathcal{U}$ not equal to the identity matrix I such that $C' = CQ\mathcal{U}$ is weakly $\left(\frac{\pi}{3} + \epsilon\right)$-orthogonal. Define $Q'_i = \alpha_i I$ for $i = 1, 2, \ldots, k$. Then $C'Q'_i$ is also a weakly $\left(\frac{\pi}{3} + \epsilon\right)$-orthogonal basis for $\mathcal{L}_i$. Typically, JPEG employs Q_i's that are not related in any special way. Therefore, we believe that for most practical cases JPEG CHEst has a unique solution.

6.5.4. Experimental results. We tested the proposed approach using a wide variety of test cases. In reality, the decompressed image P_d is always corrupted with some additive noise. Consequently, to estimate the desired compression history, the approach described above was combined with some additional noise mitigation steps. Our algorithm provided accurate estimates of the image's JPEG compression history for all the test cases. We refer the reader to [22, 23] for details on the experimental setup and results.

7. Discussion and conclusions. In this paper, we derived some interesting properties of nearly orthogonal lattice bases and random bases. We chose to directly quantify the orthogonality of a basis in terms of the minimum angle θ between a basis vector and the linear subspace spanned by the remaining basis vectors. When $\theta \geq \frac{\pi}{3}$ radians, we say that the basis is nearly orthogonal. A key contribution of this paper is to show that a nearly orthogonal lattice basis always contains a shortest lattice vector. We also investigated the uniqueness of nearly orthogonal lattice bases. We proved that if the basis vectors of a nearly orthogonal basis are nearly equal in length, then the lattice essentially contains only one nearly orthogonal basis. These results enable us to solve a fascinating digital color imaging problem called JPEG CHEst.

The applicability of our results on nearly orthogonal bases is limited by the fact that every lattice does not necessarily admit a nearly orthogonal basis. In this sense, lattices that contain a nearly orthogonal basis are somewhat special.

However, in random lattices, nearly orthogonal bases occur frequently when the lattice is sufficiently low-dimensional. Our second main result is that an m-D Gaussian or Bernoulli random basis that spans a lattice in $\mathbb{R}^n$, with $m < 0.071\,n$, is nearly orthogonal almost surely as $n \to \infty$ and with high probability at finite but large n. Consequently, a random $n \times 0.071\,n$ lattice basis contains the shortest lattice vector with high probability. In fact, based on [31], the bound 0.071 can be relaxed to 0.25, at least in the Gaussian case.

We believe that analyzing random lattices using some of the techniques developed in this paper is a fruitful area for future research. For example, we have recently realized (using Corollary 3) that a random $n \times 0.071\,n$ lattice basis is Minkowski-reduced with high probability [8].

Acknowledgments. We thank Gabor Pataki for useful comments and for the reference to Gauss's work in Vazirani's book. We also thank the editor Alexander Vardy and the anonymous reviewers for their thorough and thought-provoking reviews; our work on random lattices was motivated by their comments. Finally, we thank Gregory Sorkin who gave us numerous insights into the properties of random matrices.

REFERENCES

[1] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, *Closest point search in lattices*, IEEE Trans. Inform. Theory, 48 (2002), pp. 2201–2214.

[2] M. Ajtai, *The shortest vector problem in L_2 is NP-hard for randomized reductions*, in Proceedings of the 30th Annual ACM Symposium on Theory of Computing, 1998, pp. 10–19.

[3] A. Akhavi, J.-F. Marckert, and A. Rouault, *On the Reduction of a Random Basis*, e-print math 060433, http://arxiv.org/abs/math/06043331 (2006). (2006).

[4] L. Babai, *On lovaász' lattice reduction and the nearest lattice point problem*, Combinatorica, 6 (1986), pp. 1–14.

[5] H. H. Bauschke, C. H. Hamilton, M. S. Macklem, J. S. McMichael, and N. R. Swart, *Recompression of JPEG images by requantization*, IEEE Trans. Image Process., 12 (2003), pp. 843–849.

[6] E. Candès and T. Tao, *Near optimal signal recovery from random projections: Universal encoding strategies?*, IEEE Trans. Inform. Theory, 25 (2006), pp. 5402–5425.

[7] O. Damen, A. Chkeif, and J. Belfiore, *Lattice code decoder for space-time codes*, IEEE Commun. Lett., 4 (2000), pp. 161–163.

[8] S. Dash and R. Neelamani, *Some Properties of SVP in Random Lattices*, manuscript in preparation, 2006.

[9] H. Daudé and B. Vallée, *An upper bound on the average number of iterations of the LLL algorithm*, Theoret. Comput. Sci., 123 (1994), pp. 95–115.

[10] I. Dinur, G. Kindler, R. Raz, and S. Safra, *Approximating CVP to within almost-polynomial factors is NP-hard*, Combinatorica, 23 (2003), pp. 205–243.

[11] J. L. Donaldson, *Minkowski reduction of integral matrices*, Math. Comput., 33 (1979), pp. 201–216.

[12] N. El-Karoui, *Recent results about the largest eigenvalue of random covariance matrices and statistical applications*, Acta Phys. Polon. B, 36 (2005), pp. 2681–2697.

[13] C. F. Gauss, *Disquisitiones Arithmeticae*, A. A. Clark, ed., Springer-Verlag, New York, 1986 (in English).

[14] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[15] R. Kannan, *Algorithmic geometry of numbers*, Ann. Rev. Comput. Sci., 2 (1987), pp. 231–267.

[16] S. Khot, *Hardness of approximating the shortest vector problem in lattices*, J. ACM, 52 (2005), pp. 789–808.

[17] A. K. Lenstra, H. W. Lenstra, Jr., and L. Lovász, *Factoring polynomials with rational coefficients*, Math. Ann., 261 (1982), pp. 515–534.

[18] A. E. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann, *Smallest singular value of random matrices and geometry of random polytopes*, Adv. Math., 195 (2005), pp. 491–523.

[19] V. A. Marchenko and L. A. Pastur, *Distribution of eigenvalues in certain sets of random matrices*, Mat. Sb., 72 (1967), pp. 407–535 (in Russian).

[20] D. Micciancio, *The shortest vector problem is NP-hard to approximate to within some constant*, SIAM J. Comput., 30 (2001), pp. 2008–2035.

[21] D. Micciancio and S. Goldwasser, *Complexity of Lattice Problems: A Cryptographic Perspective*, Kluwer Academic Publishers, Boston, 2002.

[22] R. Neelamani, *Inverse Problems in Image Processing*, Ph.D. dissertation, Rice University, Houston, TX, 2003; also available online from www.dsp.rice.edu/∼neelsh/publications.

[23] R. Neelamani, R. de Queiroz, Z. Fan, S. Dash, and R. G. Baraniuk, *JPEG compression history estimation for color images*, IEEE Trans. Image Process., 15 (2006), pp. 1365–1378.

[24] P. Nguyen and D. Stehlé, *Low-dimensional lattice basis reduction revisited*, in Proceedings of the 6th International Symposium on Algorithmic Number Theory (ANTS VI), Lecture Notes in Comput. Sci. 3076, Springer-Verlag, Berlin, 2004, pp. 338–357.

[25] P. Q. Nguyen and J. Stern, *Lattice reduction in cryptology: An update*, in Proceedings of the 4th International Symposium on Algorithmic Number Theory (ANTS IV), Lecture Notes in Comput. Sci. 1838, Springer-Verlag, Berlin, 2000, pp. 85–112.

[26] W. Pennebaker and J. Mitchell, *JPEG, Still Image Data Compression Standard*, Van Nostrand Reinhold, New York, 1993.

[27] C. Poynton, *A Technical Introduction to Digital Video*, Wiley, New York, 1996.

[28] O. Regev, *On lattices, learning with errors, random linear codes, and cryptography*, in Proceedings of the 37th annual ACM Symposium on Theory of Computing, New York, 2005, pp. 84–93.

[29] G. Sharma and H. Trussell, *Digital color imaging*, IEEE Trans. Image Process., 6 (1997), pp. 901–932.

[30] J. W. Silverstein, *The smallest eigenvalue of a large dimensional Wishart matrix*, Ann. Probab., 13 (1985), pp. 1364–1368.

[31] G. Sorkin, *private communication*, 2006.

[32] V. V. Vazirani, *Approximation Algorithms*, Springer-Verlag, Berlin, 2001.

SIAM J. DISCRETE MATH.
Vol. 21, No. 1, pp. 220–236

A PRIMAL BARVINOK ALGORITHM BASED ON IRRATIONAL DECOMPOSITIONS*

MATTHIAS KÖPPE[†]

Abstract. We introduce variants of Barvinok's algorithm for counting lattice points in polyhedra. The new algorithms are based on irrational signed decomposition in the primal space and the construction of rational generating functions for cones with low index. We give computational results that show that the new algorithms are faster than the existing algorithms by a large factor.

Key words. rational generating functions, irrational decompositions

AMS subject classifications. 05A15, 52C07, 68W30

DOI. 10.1137/060664768

1. Introduction.

Thirteen years have passed since Alexander Barvinok's amazing algorithm for counting lattice points in polyhedra was published [2]. In the meantime, efficient implementations [14, 24] have been designed, which helped to make Barvinok's algorithm a practical tool in many applications in discrete mathematics. The implications of Barvinok's technique, of course, reach far beyond the domain of combinatorial counting problems: For example, De Loera et al. [11] pointed out applications in integer linear programming, and De Loera et al. [12, 13] obtained a fully polynomial-time approximation scheme (FPTAS) for optimizing arbitrary polynomial functions over the mixed-integer points in polytopes of fixed dimension.

Barvinok's algorithm first triangulates the supporting cones of all vertices of a polytope, to obtain simplicial cones. Then, the simplicial cones are recursively decomposed into unimodular cones. It is essential that one uses *signed decompositions* here; triangulating these cones is not good enough to give a polynomiality result. The rational generating functions of the resulting unimodular cones can then be written down easily. Adding and subtracting them according to the inclusion-exclusion principle and Brion's theorem [7] gives the rational generating function of the polytope. The number of lattice points in the polytope can finally be obtained by applying residue techniques on the rational generating function.

The algorithm in the original paper [2] worked explicitly with all the lower-dimensional cones that arise from the intersecting faces of the subcones in an inclusion-exclusion formula. Later it was pointed out that it is possible to simplify the algorithm by computing with full-dimensional cones only, by making use of Brion's "polarization trick" (see [3, Remark 4.3]): The computations with rational generating functions are invariant with respect to the contribution of nonpointed cones (cones containing a nontrivial linear subspace). By operating in the dual space, i.e., by computing with the polars of all cones, lower-dimensional cones can be safely discarded, because

*Received by the editors July 11, 2006; accepted for publication (in revised form) October 11, 2006; published electronically March 22, 2007. This work was supported by a Feodor Lynen Research Fellowship from the Alexander von Humboldt Foundation.

http://www.siam.org/journals/sidma/21-1/66476.html

[†]Department of Mathematics, Institute for Mathematical Optimization (IMO), Otto-von-Guericke-Universität Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany (mkoeppe@mail.math.uni-magdeburg.de).

this is equivalent to discarding nonpointed cones in the primal space. The practical implementations also rely heavily on this polarization trick.

In practical implementations of Barvinok's algorithm, one observes that in the hierarchy of cone decompositions, the index of the decomposed cones quickly descends from large numbers to fairly low numbers. The "last mile," i.e., decomposing many cones with fairly low index, creates a huge number of unimodular cones and thus is the bottleneck of the whole computation in many instances.

The idea of this paper is to stop the decomposition when the index of a cone is small enough, and to compute with generating functions for the integer points in cones of small index rather than unimodular cones. When we try to implement this simple idea in Barvinok's algorithm, as outlined in section 3, we face a major difficulty, however: Polarizing back a cone of small index can create a cone of very large index, because determinants of $d \times d$ matrices are homogeneous of order d.

To address this difficulty, we avoid polarization altogether and perform the signed decomposition in the primal space instead. To avoid having to deal with all the lower-dimensional subcones, we use the concept of *irrational decompositions* of rational polyhedra. Beck and Sottile [6] introduced this notion to give astonishingly simple proofs for three theorems of Stanley on generating functions for the integer points in rational polyhedral cones. Using the same technique, Beck, Haase, and Sottile [4] gave simplified proofs of theorems of Brion and Lawrence–Varchenko. An irrational decomposition of a polyhedron is a decomposition into polyhedra whose proper faces do not contain any lattice points. Counting formulas for lattice points based on irrational decompositions therefore do not need to take any inclusion-exclusion principle into account.

We give an explicit construction of a *uniform irrational shifting vector* $\mathbf{s}$ for a cone $\mathbf{v} + K$ with apex $\mathbf{v}$ such that the shifted cone $(\mathbf{v} + \mathbf{s}) + K$ has the same lattice points and contains no lattice points on its proper faces (section 4). More strongly, we prove that *all cones* appearing in the signed decompositions of $(\mathbf{v} + \mathbf{s}) + K$ in Barvinok's algorithm contain no lattice points on their proper faces. Therefore, discarding lower-dimensional cones is safe. Despite its name, the vector $\mathbf{s}$ only has *rational coordinates*, so after shifting the cone by $\mathbf{s}$, large parts of existing implementations of Barvinok's algorithm can be reused to compute the irrational primal decompositions.

In section 5, we show the precise algorithm. We also show that the same technique can be applied to the "homogenized version" of Barvinok's algorithm that was proposed in [9].

In section 6, we extend the irrationalization technique to nonsimplicial cones. This gives rise to an "all-primal" Barvinok algorithm, where triangulation of nonsimplicial cones is also performed in the primal space. This allows us to handle problems where the triangulation of the dual cones is hard, e.g., in the case of cross polytopes.

Finally, in section 7, we report on computational results. Results on benchmark problems show that the new algorithms are faster than the existing algorithms by orders of magnitude. We also include results for problems that could not previously be solved with Barvinok techniques.

2. Barvinok's algorithm.

Let $P \subseteq \mathbf{R}^d$ be a rational polyhedron. The *generating function* of $P \cap \mathbf{Z}^d$ is defined as the formal Laurent series

$$\tilde{g}_P(\mathbf{z}) = \sum_{\boldsymbol{\alpha} \in P \cap \mathbf{Z}^d} \mathbf{z}^{\boldsymbol{\alpha}} \in \mathbf{Z}[[z_1, \dots, z_d, z_1^{-1}, \dots, z_d^{-1}]],$$

using the multiexponent notation $\mathbf{z}^{\boldsymbol{\alpha}} = \prod_{i=1}^{d} z_i^{\alpha_i}$. If P is bounded, $\tilde{g}_P$ is a Laurent polynomial, which we consider as a rational function g_P. If P is not bounded but is pointed (i.e., P does not contain a straight line), there is a nonempty open subset $U \subseteq \mathbf{C}^d$ such that the series converges absolutely and uniformly on every compact subset of U to a rational function g_P. If P contains a straight line, we set $g_P \equiv 0$. The rational function $g_P \in \mathbf{Q}(z_1, \dots, z_d)$ defined in this way is called the *rational generating function* of $P \cap \mathbf{Z}^d$.

Barvinok's algorithm computes the rational generating function of a polyhedron P. It proceeds as follows. By Brion's theorem [7], the rational generating function of a polyhedron can be expressed as the sum of the rational generating functions of the supporting cones of its vertices. Let $\mathbf{v}_i \in \mathbf{Q}^d$ be a vertex of the polyhedron P. Then the *supporting cone* $\mathbf{v}_i + C_i$ of $\mathbf{v}_i$ is the (shifted) polyhedral cone defined by $\mathbf{v}_i + \operatorname{cone}(P - \mathbf{v}_i)$. Every supporting cone $\mathbf{v}_i + C_i$ can be triangulated to obtain simplicial cones $\mathbf{v}_i + C_{ij}$. Let $K = \mathbf{v} + B\mathbf{R}_+^d$ be a simplicial full-dimensional cone, whose *basis vectors* $\mathbf{b}_1, \dots, \mathbf{b}_d$ (i.e., representatives of its extreme rays) are given by the columns of some matrix $B \in \mathbf{Z}^{d\times d}$. We assume that the basis vectors are primitive vectors of the standard lattice $\mathbf{Z}^d$. Then the *index* of K is defined to be $\operatorname{ind} K = |\det B|$; it can also be interpreted as the cardinality of $\Pi \cap \mathbf{Z}^d$, where Π is the *fundamental parallelepiped* of K, i.e., the half-open parallelepiped

$$\Pi = \mathbf{v} + \left\{ \sum_{i=1}^{d} \lambda_i \mathbf{b}_i : 0 \le \lambda_i < 1 \right\}.$$

We remark that the set $\Pi \cap \mathbf{Z}^d$ can also be seen as a set of representatives of the cosets of the lattice $B\mathbf{Z}^d$ in the standard lattice $\mathbf{Z}^d$; we shall make use of this interpretation in section 3. Barvinok's algorithm now computes a *signed decomposition* of the simplicial cone K to produce other simplicial cones with smaller index. To this end, the algorithm constructs a vector $\mathbf{w} \in \mathbf{Z}^d$ such that

$$\mathbf{w} = \alpha_1 \mathbf{b}_1 + \dots + \alpha_d \mathbf{b}_d \quad \text{with} \quad |\alpha_i| \le |\det B|^{-1/d} \le 1. \tag{1}$$

This can be accomplished using integer programming or lattice basis reduction. The cone is then decomposed into cones spanned by d vectors from the set $\{\mathbf{b}_1, \dots, \mathbf{b}_d, \mathbf{w}\}$; each of the resulting cones then has an index bounded above by $(\operatorname{ind} K)^{(d-1)/d}$. In general, these cones form a signed decomposition of K (see Figure 2); if $\mathbf{w}$ lies inside K, they form a triangulation of K (see Figure 1). The resulting cones and their intersecting proper faces (arising in an inclusion-exclusion formula) are recursively processed until *unimodular* cones, i.e., cones of index 1, are obtained. Finally, for a unimodular cone $\mathbf{v} + B\mathbf{R}_+^d$, the rational generating function can be easily written as

$$\frac{\mathbf{z}^{\mathbf{a}}}{\prod_{j=1}^{d}(1 - \mathbf{z}^{\mathbf{b}_j})}, \tag{2}$$

where $\mathbf{a}$ is the unique integer point in the fundamental parallelepiped of the cone. We summarize Barvinok's algorithm below.

Algorithm 1 (Barvinok's original (primal) algorithm).

Input: A polyhedron $P \subset \mathbf{R}^d$ given by rational inequalities.

Output: The rational generating function for $P \cap \mathbf{Z}^d$ in the form

$$g_P(\mathbf{z}) = \sum_{i \in I} \epsilon_i \frac{\mathbf{z}^{\mathbf{a}_i}}{\prod_{j=1}^{d}(1 - \mathbf{z}^{\mathbf{b}_{ij}})}, \tag{3}$$

where $\epsilon_i \in \{\pm 1\}$, $\mathbf{a}_i \in \mathbf{Z}^d$, and $\mathbf{b}_{ij} \in \mathbf{Z}^d$.

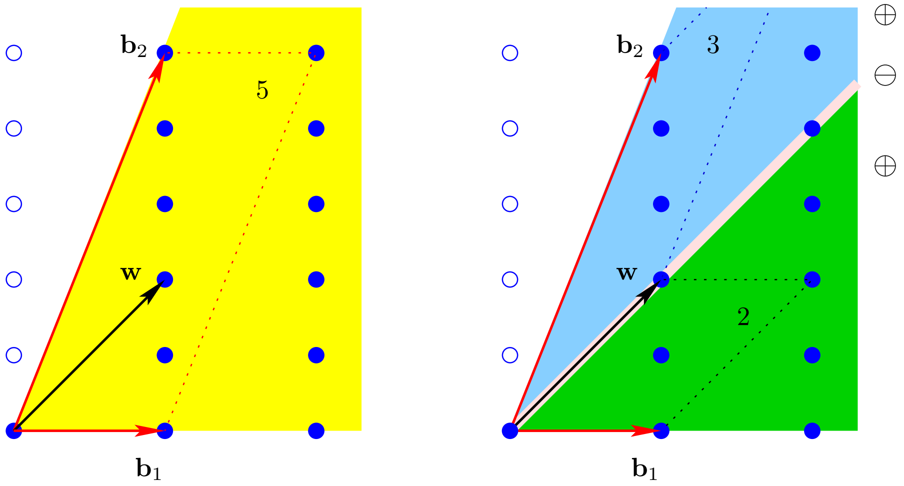


Fig. 1. *A triangulation of the cone of index* 5 *generated by* $\mathbf{b}^1$ *and* $\mathbf{b}^2$ *into the two cones spanned by* $\{\mathbf{b}^1, \mathbf{w}\}$ *and* $\{\mathbf{b}^2, \mathbf{w}\}$*, having an index of* 2 *and* 3*, respectively. We have the inclusion-exclusion formula* $g_{\text{cone}\{\mathbf{b}_1,\mathbf{b}_2\}}(\mathbf{z}) = g_{\text{cone}\{\mathbf{b}_1,\mathbf{w}\}}(\mathbf{z}) + g_{\text{cone}\{\mathbf{b}_2,\mathbf{w}\}}(\mathbf{z}) - g_{\text{cone}\{\mathbf{w}\}}(\mathbf{z})$*; here the one-dimensional cone spanned by* $\mathbf{w}$ *needed to be subtracted.*

1. Compute all vertices $\mathbf{v}_i$ and corresponding supporting cones C_i of P.
2. Triangulate C_i into simplicial cones C_{ij}, keeping track of all the intersecting proper faces.
3. Apply signed decomposition to the cones $\mathbf{v}_i + C_{ij}$ to obtain unimodular cones $\mathbf{v}_i + C_{ijl}$, keeping track of all the intersecting proper faces.
4. Compute the unique integer point $\mathbf{a}_i$ in the fundamental parallelepiped of every resulting cone $\mathbf{v}_i + C_{ijl}$.
5. Write down the formula (3).

The recursive decomposition of cones defines a *decomposition tree*. Due to the descent of the indices in the signed decomposition procedure, the following estimate holds for its depth.

Lemma 2 (see Barvinok [2]). *Let* $B\mathbf{R}^d_+$ *be a simplicial full-dimensional cone, whose basis is given by the columns of the matrix* $B \in \mathbf{Z}^{d\times d}$*. Let* $D = |\det B|$*. Then the depth of the decomposition tree is at most*

$$k(D) = \left\lfloor 1 + \frac{\log_2 \log_2 D}{\log_2 \frac{d}{d-1}} \right\rfloor. \tag{4}$$

Because at each decomposition step at most $\mathrm{O}(2^d)$ cones are created and the depth of the tree is doubly logarithmic in the index of the input cone, Barvinok could obtain a polynomiality result *in fixed dimension*.

Theorem 3 (see Barvinok [2]). *Let* d *be fixed. There exists a polynomial-time algorithm for computing the rational generating function of a polyhedron* $P \subseteq \mathbf{R}^d$ *given by rational inequalities.*

Later the algorithm was improved by making use of Brion's "polarization trick" (see [3, Remark 4.3]): The computations with rational generating functions are invariant with respect to the contribution of nonpointed cones (cones containing a nontrivial linear subspace). The reason is that the rational generating function of every nonpointed cone is zero. By operating in the dual space, i.e., by computing

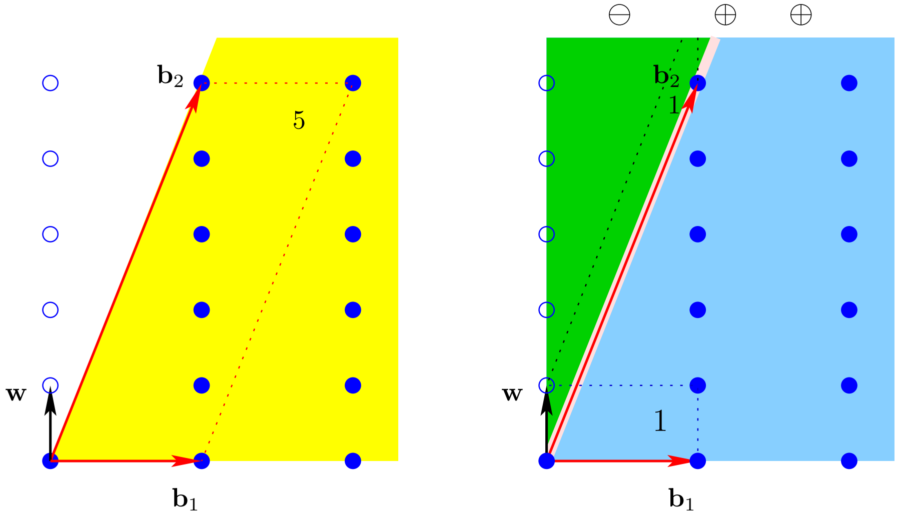


FIG. 2. *A signed decomposition of the cone of index* 5 *generated by* $\mathbf{b}^1$ *and* $\mathbf{b}^2$ *into the two unimodular cones spanned by* $\{\mathbf{b}^1, \mathbf{w}\}$ *and* $\{\mathbf{b}^2, \mathbf{w}\}$*. We have the inclusion-exclusion formula* $g_{\operatorname{cone}\{\mathbf{b}_1,\mathbf{b}_2\}}(\mathbf{z}) = g_{\operatorname{cone}\{\mathbf{b}_1,\mathbf{w}\}}(\mathbf{z}) - g_{\operatorname{cone}\{\mathbf{b}_2,\mathbf{w}\}}(\mathbf{z}) + g_{\operatorname{cone}\{\mathbf{w}\}}(\mathbf{z})$.

with the polars of all cones, lower-dimensional cones can be safely discarded, because this is equivalent to discarding nonpointed cones in the primal space.

ALGORITHM 4 (dual Barvinok algorithm).

Input: A polyhedron $P \subset \mathbf{R}^d$ given by rational inequalities.

Output: The rational generating function for $P \cap \mathbf{Z}^d$ in the form

$$g_P(\mathbf{z}) = \sum_{i\in I} \epsilon_i \frac{\mathbf{z}^{\mathbf{a}_i}}{\prod_{j=1}^{d}(1-\mathbf{z}^{\mathbf{b}_{ij}})}, \tag{5}$$

where $\epsilon_i \in \{\pm 1\}$, $\mathbf{a}_i \in \mathbf{Z}^d$, and $\mathbf{b}_{ij} \in \mathbf{Z}^d$.

1. Compute all vertices $\mathbf{v}_i$ and corresponding supporting cones C_i of P.
2. Polarize the supporting cones C_i to obtain C_i°.
3. Triangulate C_i° into simplicial cones C_{ij}°, discarding lower-dimensional cones.
4. Apply Barvinok's signed decomposition to the cones $\mathbf{v}_i + C_{ij}^\circ$ to obtain cones $\mathbf{v}_i + C_{ijl}^\circ$, stopping decomposition when a unimodular cone is obtained. Discard all lower-dimensional cones.
5. Polarize back C_{ijl}° to obtain cones C_{ijl}.
6. Compute the unique integer point $\mathbf{a}_i$ in the fundamental parallelepiped of every resulting cone $\mathbf{v}_i + C_{ijl}$.
7. Write down the formula (5).

This variant of the algorithm is much faster than the original algorithm because in each step of the signed decomposition at most d, rather than $\mathrm{O}(2^d)$, cones are created. The practical implementations LattE [14], by J. A. De Loera et al., and `barvinok` [24], by S. Verdoolaege, also rely heavily on this polarization trick.

3. The Barvinok algorithm with stopped decomposition. We start out by introducing a first variant of Barvinok's algorithm that stops decomposing cones before unimodular cones are reached. As we will see in the computational results in section 7, the simple modification that we propose can already give a significant improvement of the running time for some problems, at least in low dimension.

ALGORITHM 5 (dual Barvinok algorithm with stopped decomposition).

Input: A polyhedron $P \subset \mathbf{R}^d$ given by rational inequalities; the maximum enumerated cone index ℓ.

Output: The rational generating function for $P \cap \mathbf{Z}^d$ in the form

$$g_P(\mathbf{z}) = \sum_{i \in I} \epsilon_i \frac{\sum_{\mathbf{a} \in A_i} \mathbf{z}^{\mathbf{a}}}{\prod_{j=1}^{d} (1 - \mathbf{z}^{\mathbf{b}_{ij}})}, \tag{6}$$

where $\epsilon_i \in \{\pm 1\}$, $A_i \subseteq \mathbf{Z}^d$ with $|A_i| \leq \ell$, and $\mathbf{b}_{ij} \in \mathbf{Z}^d$.

1. Compute all vertices $\mathbf{v}_i$ and corresponding supporting cones C_i of P.
2. Polarize the supporting cones C_i to obtain C_i°.
3. Triangulate C_i° into simplicial cones C_{ij}°, discarding lower-dimensional cones.
4. Apply Barvinok's signed decomposition to the cones $\mathbf{v}_i + C_{ij}^\circ$ to obtain cones $\mathbf{v}_i + C_{ijl}^\circ$, stopping decomposition when a polarized-back cone $C_{ijl} = (C_{ijl}^\circ)^\circ$ has index at most ℓ. Discard all lower-dimensional cones.
5. Polarize back C_{ijl}° to obtain cones C_{ijl}.
6. Enumerate the integer points in the fundamental parallelepipeds of all resulting cones $\mathbf{v}_i + C_{ijl}$ to obtain the sets A_i.
7. Write down the formula (6).

As mentioned above, the integer points in the fundamental parallelepiped of a cone $\mathbf{v}_i + B_{ijl}\mathbf{R}_+^d$ can be interpreted as representatives of the cosets of the lattice $B_{ijl}\mathbf{Z}^d$ in the standard lattice $\mathbf{Z}^d$. Hence they can be easily enumerated in step 6 by computing the Smith normal form of the generator matrix B_{ijl}; see Lemma 5.2 of [1]. The Smith normal form can be computed in polynomial time, even if the dimension is not fixed [19].

We remark that both triangulation and signed decomposition are done in the dual space, but the stopping criterion is the index of the polarized-back cones (in the primal space). The reason for this stopping criterion is that we wish to control the maximum number of points in the fundamental parallelepipeds that need to be enumerated. Indeed, when the maximum ℓ is chosen as a constant or polynomially in the input size, then Algorithm 5 clearly runs in polynomial time (in fixed dimension).

Each step of Barvinok's signed decomposition reduces the index of the decomposed cones. When the index of a cone C_{ijl}° is Δ, in the worst case the polarized-back cone C_{ijl} has index Δ^{d-1}, where d is the dimension. If the dimension is too large, the algorithm often needs to decompose cones down to a very low index or even index 1, so the speed-up of the algorithm will be very limited. This can be seen from the computational results in section 7.

4. Construction of a uniform irrational shifting vector.

In this section, we will give an explicit construction of an *irrational shifting vector* $\mathbf{s}$ for a simplicial cone $\mathbf{v} + K$ with apex $\mathbf{v}$ such that the shifted cone $(\mathbf{v} + \mathbf{s}) + K$ has the same lattice points and contains no lattice points on its proper faces. The "irrationalization" (or perturbation) will be *uniform* in the sense that also every cone arising during the Barvinok decomposition does not contain any lattice points on its proper faces. This will enable us to perform the Barvinok decomposition in the primal space, discarding all lower-dimensional cones.

To accomplish this goal, we shall first describe a subset of the *stability region* of a cone $\mathbf{v} + K$ with apex at $\mathbf{v}$, i.e., the set of apex points $\tilde{\mathbf{v}}$ such that $\tilde{\mathbf{v}} + K$ contains the same lattice points as $\mathbf{v} + K$; see Figure 3.

LEMMA 6 (stability cube). *Let $\mathbf{v} + B\mathbf{R}_+^d$ be a simplicial full-dimensional cone with apex at $\mathbf{v} \in \mathbf{Q}^d$, whose basis is given by the columns of the matrix $B \in \mathbf{Z}^{d \times d}$.*

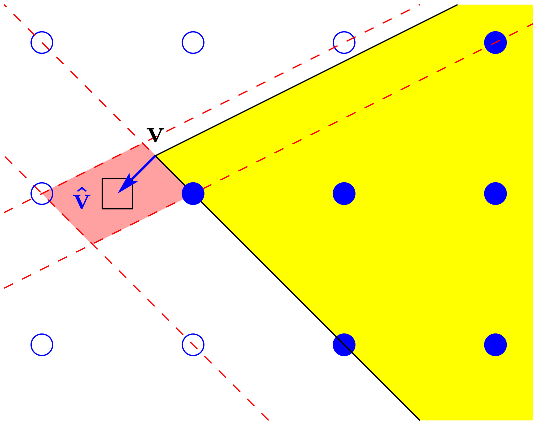


FIG. 3. *The stability region of a cone.*

Let $\mathbf{b}_1^*, \dots, \mathbf{b}_d^*$ *be a basis of the dual cone, given by the columns of the matrix* $B^* = -(B^{-1})^\top$.

Let $D = |\det B|$. *Let* $\boldsymbol{\lambda} \in \mathbf{Q}^d$ *and* $\hat{\boldsymbol{\lambda}} \in \mathbf{Q}^d$ *be defined by*

$$\lambda_i = \langle \mathbf{b}_i^*, \mathbf{v} \rangle \quad \text{and} \quad \hat{\lambda}_i = \frac{1}{D}\left(\lfloor D\lambda_i \rfloor + \frac{1}{2} \right) \quad \text{for } i = 1, \dots, d.$$

Let

$$\hat{\mathbf{v}} = -B\hat{\boldsymbol{\lambda}} \quad \text{and} \quad \rho = \frac{1}{2D \cdot \max_{i=1}^d \|\mathbf{b}_i^*\|_1}.$$

Then, for every $\tilde{\mathbf{v}}$ *with* $\|\tilde{\mathbf{v}} - \hat{\mathbf{v}}\|_\infty < \rho$, *the cone* $\tilde{\mathbf{v}} + B\mathbf{R}_+^d$ *contains the same integer points as the cone* $\mathbf{v} + B\mathbf{R}_+^d$ *and does not have integer points on its proper faces.*

In the proof of the lemma, we will use the H-representation (inequality description) of the simplicial cone $\mathbf{v} + B\mathbf{R}_+^d$. It is given by the basis vectors of the dual cone

$$\mathbf{v} + B\mathbf{R}_+^d = \left\{ \mathbf{x} \in \mathbf{R}^d : \langle \mathbf{b}_i^*, \mathbf{x} \rangle \le \langle \mathbf{b}_i^*, \mathbf{v} \rangle \text{ for } i = 1, \dots, d \right\}. \tag{7}$$

Proof of Lemma 6. Let $\tilde{\boldsymbol{\lambda}}$ be defined by $\tilde{\lambda}_i = \langle \mathbf{b}_i^*, \tilde{\mathbf{v}} \rangle$. Then we have

$$\left| \tilde{\lambda}_i - \hat{\lambda}_i \right| \le \|\mathbf{b}_i^*\|_1 \cdot \|\tilde{\mathbf{v}} - \hat{\mathbf{v}}\|_\infty < \|\mathbf{b}_i^*\|_1 \cdot \rho \le \frac{1}{2D}. \tag{8}$$

By (7), a point $\mathbf{x} \in \mathbf{Z}^d$ lies in the cone $\mathbf{v} + B\mathbf{R}_+^d$ if and only if

$$\langle \mathbf{b}_i^*, \mathbf{x} \rangle \le \langle \mathbf{b}_i^*, \mathbf{v} \rangle = \lambda_i \quad \text{for } i = 1, \dots, d.$$

Likewise, $\mathbf{x} \in \tilde{\mathbf{v}} + B\mathbf{R}_+^d$ if and only if

$$\langle \mathbf{b}_i^*, \mathbf{x} \rangle \le \langle \mathbf{b}_i^*, \tilde{\mathbf{v}} \rangle = \tilde{\lambda}_i \quad \text{for } i = 1, \dots, d.$$

Note that for $\mathbf{x} \in \mathbf{Z}^d$, the left-hand sides of both inequalities are an integer multiple of $\frac{1}{D}$. Therefore, we obtain equivalent statements by rounding down the right-hand sides to integer multiples of $\frac{1}{D}$. For the right-hand side of (4) we have by (8),

$$\tilde{\lambda}_i = \hat{\lambda}_i + (\tilde{\lambda}_i - \hat{\lambda}_i) < \frac{1}{D}\left(\lfloor D\lambda_i \rfloor + \frac{1}{2}\right) + \frac{1}{2D} = \frac{1}{D}\lfloor D\lambda_i \rfloor + 1, \tag{9a}$$

$$\tilde{\lambda}_i = \hat{\lambda}_i + (\tilde{\lambda}_i - \hat{\lambda}_i) > \frac{1}{D}\left(\lfloor D\lambda_i \rfloor + \frac{1}{2}\right) - \frac{1}{2D} = \frac{1}{D}\lfloor D\lambda_i \rfloor, \tag{9b}$$

so λ_i and $\tilde{\lambda}_i$ are rounded down to the same value $\frac{1}{D}\lfloor D\lambda_i \rfloor$. Thus, the cone $\tilde{\mathbf{v}} + B\mathbf{R}^d_+$ contains the same integer points as the cone $\mathbf{v} + B\mathbf{R}^d_+$. Moreover, since the inequalities (9) are strict, the cone $\tilde{\mathbf{v}} + B\mathbf{R}^d_+$ does not have integer points on its proper faces. □

For nonsimplicial cones, we will give an algorithmic construction for a stability cube in section 6.

Next we make use of the estimate for the depth of the decomposition tree in Barvinok's algorithm given in Lemma 2. On each level of the decomposition, the entries in the basis matrices can grow, but not by much. We then obtain the following lemma.

LEMMA 7. *Let $B\mathbf{R}^d_+$ be a simplicial full-dimensional cone, whose basis is given by the columns of the matrix $B \in \mathbf{Z}^{d\times d}$. Let $D = |\det B|$. Let $C \in \mathbf{Z}_+$ be a number such that $|B_{i,j}| \le C$.*

Then all the basis matrices $\bar{B}$ of the cones that appear in the recursive signed decomposition procedure of Barvinok's algorithm applied to $B\mathbf{R}^d_+$ have entries bounded above by $d^{k(D)}C$, where $k(D)$ is defined by (4).

Proof. Given a cone spanned by the columns $\mathbf{b}_1, \dots, \mathbf{b}_d$ of the matrix B, Barvinok's algorithm constructs a vector $\mathbf{w} \in \mathbf{Z}^d$ such that

$$\mathbf{w} = \alpha_1\mathbf{b}_1 + \dots + \alpha_d\mathbf{b}_d \quad \text{with} \quad |\alpha_i| \le |\det B|^{-1/d} \le 1. \tag{10}$$

Thus $\|\mathbf{w}\|_\infty \le dC$. The cone is then decomposed into cones spanned by d vectors from the set $\{\mathbf{b}_1, \dots, \mathbf{b}_d, \mathbf{w}\}$. Thus the entries in the corresponding basis matrices are bounded by dC. The result follows then by Lemma 2. □

If we can bound the entries of an integer matrix with nonzero determinant, we can also bound the entries of its inverse.

LEMMA 8. *Let $B \in \mathbf{Z}^{d\times d}$ be a matrix with $|B_{i,j}| \le C$. Let $D = |\det B|$. Then the absolute values of the entries of B^{-1} are bounded above by*

$$\frac{1}{D}(d-1)!C^{d-1}.$$

Proof. We have $\left|(B^{-1})_{k,l}\right| = \frac{1}{D}\left|\det B_{(k,l)}\right|$, where $B_{(k,l)}$ is the matrix obtained from deleting the kth row and lth column from B. Now the desired estimate follows from a formula for $\det B_{(k,l)}$ and from $|B_{i,j}| \le C$. □

Thus, we obtain a bound on the norm of the basis vectors of the polars of all cones occurring in the signed decomposition procedure of Barvinok's algorithm.

COROLLARY 9 (a bound on the dual basis vectors). *Let $B\mathbf{R}^d_+$ be a simplicial full-dimensional cone, whose basis is given by the columns of the matrix $B \in \mathbf{Z}^{d\times d}$. Let $D = |\det B|$. Let C be a number such that $|B_{i,j}| \le C$.*

Let $\bar{B}^* = -(\bar{B}^{-1})^\top$ be the basis matrix of the polar of an arbitrary cone $\bar{B}\mathbf{R}^d_+$ that appears in the recursive signed decomposition procedure applied to $B\mathbf{R}^d_+$. Then, for every column vector $\bar{\mathbf{b}}^*_i$ of $\bar{B}^*$ we have the estimate

$$\left\|\det \bar{B} \cdot \mathbf{b}^*_i\right\|_\infty \le (d-1)!\left(d^{k(D)}C\right)^{d-1} =: L, \tag{11}$$

where $k(D)$ is defined by (4).

Proof. By Lemma 7, the entries of $\bar{B}$ are bounded above by $d^{k(D)}C$. Then the result follows from Lemma 8. $\square$

The construction of the "irrational" shifting vector is based on the following lemma.

LEMMA 10 (the irrational lemma). *Let $M \in \mathbf{Z}_+$ be an integer. Let*

$$\mathbf{q} = \left(\frac{1}{2M}, \frac{1}{(2M)^2}, \dots, \frac{1}{(2M)^d}\right). \tag{12}$$

Then $\langle \mathbf{c}, \mathbf{q}\rangle \notin \mathbf{Z}$ for every $\mathbf{c} \in \mathbf{Z}^d \setminus \{\mathbf{0}\}$ with $\|\mathbf{c}\|_\infty < M$.

Proof. The proof follows from the principle of representations of rational numbers in a positional system of base $2M$. $\square$

THEOREM 11. *Let $\mathbf{v} + B\mathbf{R}^d_+$ be a simplicial full-dimensional cone with apex at $\mathbf{v} \in \mathbf{Q}^d$, whose basis is given by the columns of the matrix $B \in \mathbf{Z}^{d\times d}$. Let $D = |\det B|$, C be a number such that $|B_{i,j}| \le C$, and let $\hat{\mathbf{v}} \in \mathbf{Q}^d$ and $\rho \in \mathbf{Q}_+$ be the data from* Lemma 6 *describing the stability cube of $\mathbf{v} + B\mathbf{R}^d_+$. Let $0 < r \in \mathbf{Z}$ such that $r^{-1} < \frac{1}{2}\rho$. We define a vector $\mathbf{w}$ by rounding each coordinate of $\hat{\mathbf{v}}$ to the nearest integer multiple of r^{-1}. Using*

$$k = \left\lfloor 1 + \frac{\log_2 \log_2 D}{\log_2 \frac{d}{d-1}} \right\rfloor, \tag{13}$$

$L = (d-1)!(d^k C)^{d-1}$, and $M = 2L$, define

$$\mathbf{s} = \frac{1}{r} \cdot \left(\frac{1}{(2M)^1}, \frac{1}{(2M)^2}, \dots, \frac{1}{(2M)^d}\right).$$

Finally, let $\tilde{\mathbf{v}} = \mathbf{w} + \mathbf{s}$.

(i) *We have $(\tilde{\mathbf{v}} + B\mathbf{R}^d_+) \cap \mathbf{Z}^d = (\mathbf{v} + B\mathbf{R}^d_+) \cap \mathbf{Z}^d$; i.e., the shifted cone has the same set of integer points as the original cone.*

(ii) *The shifted cone $\tilde{\mathbf{v}} + B\mathbf{R}^d_+$ contains no lattice points on its proper faces.*

(iii) *More strongly,* all cones *appearing in the signed decompositions of the shifted cone $\tilde{\mathbf{v}} + B\mathbf{R}^d_+$ in Barvinok's algorithm contain no lattice points on their proper faces.*

Proof. Part (i). This follows from Lemma 6 because $\tilde{\mathbf{v}}$ clearly lies in the open stability cube.

Parts (ii) *and* (iii). Every cone appearing in the course of Barvinok's signed decomposition algorithm has the same apex $\tilde{\mathbf{v}}$ as the input cone and a basis $\bar{B} \in \mathbf{Z}^{d\times d}$ with $\left|\det \bar{B}\right| \le D$. Let such a $\bar{B}$ be fixed and denote by $\bar{\mathbf{b}}^*_i$ the columns of the dual basis matrix $\bar{B}^* = -(\bar{B}^{-1})^\top$. Let $\mathbf{z} \in \mathbf{Z}^d$ be an arbitrary integer point. We shall show that $\mathbf{z}$ is not on any of the facets of the cone, i.e.,

$$\langle \bar{\mathbf{b}}^*_i, \mathbf{z} - \tilde{\mathbf{v}}\rangle \ne 0 \qquad \text{for } i = 1, \dots, d. \tag{14}$$

Let $i \in \{1, \dots, d\}$ arbitrary. We will show (14) by proving that

$$\langle \det \bar{B} \cdot \bar{\mathbf{b}}^*_i, \tilde{\mathbf{v}}\rangle \notin \mathbf{Z}. \tag{15}$$

Clearly, if (15) holds, we have $\langle \bar{\mathbf{b}}^*_i, \tilde{\mathbf{v}}\rangle \notin (\det \bar{B})^{-1}\mathbf{Z}$. But since $\langle \bar{\mathbf{b}}^*_i, \mathbf{z}\rangle \in (\det \bar{B})^{-1}\mathbf{Z}$, we have $\langle \bar{\mathbf{b}}^*_i, \mathbf{z} - \tilde{\mathbf{v}}\rangle \notin \mathbf{Z}$; in particular it is nonzero, which proves (14).

To prove (15), let $\mathbf{c} = \det \bar{B} \cdot \bar{\mathbf{b}}_i^*$. By Corollary 9, we have $\|\mathbf{c}\|_\infty \leq L < M$. Now Lemma 10 gives $\langle \mathbf{c}, \mathbf{s} \rangle \notin \frac{1}{r}\mathbf{Z}$. By the construction of $\mathbf{w}$, we have

$$\langle \mathbf{c}, \mathbf{w} \rangle \in \tfrac{1}{r}\mathbf{Z}.$$

Therefore, we have $\langle \mathbf{c}, \tilde{\mathbf{v}} \rangle = \langle \mathbf{c}, \mathbf{w} + \mathbf{s} \rangle \notin \frac{1}{r}\mathbf{Z}$. This proves (15), and thus completes the proof. □

5. The irrational algorithms. The following is our variant of the Barvinok algorithm.

Algorithm 12 (primal irrational Barvinok algorithm).

Input: A polyhedron $P \subset \mathbf{R}^d$ given by rational inequalities; the maximum enumerated cone index ℓ.

Output: The rational generating function for $P \cap \mathbf{Z}^d$ in the form

$$g_P(\mathbf{z}) = \sum_{i \in I} \epsilon_i \frac{\sum_{\mathbf{a} \in A_i} \mathbf{z}^{\mathbf{a}}}{\prod_{j=1}^{d} (1 - \mathbf{z}^{\mathbf{b}_{ij}})}, \tag{16}$$

where $\epsilon_i \in \{\pm 1\}$, $A_i \subseteq \mathbf{Z}^d$ with $|A_i| \leq \ell$, and $\mathbf{b}_{ij} \in \mathbf{Z}^d$.

1. Compute all vertices $\mathbf{v}_i$ and corresponding supporting cones C_i of P.
2. Polarize the supporting cones C_i to obtain C_i°.
3. Triangulate C_i° into simplicial cones C_{ij}°, discarding lower-dimensional cones.
4. Polarize back C_{ij}° to obtain simplicial cones C_{ij}.
5. Irrationalize all cones by computing new apex vectors $\tilde{\mathbf{v}}_{ij} \in \mathbf{Q}^d$ from $\mathbf{v}_i$ and C_{ij} as in Theorem 11.
6. Apply Barvinok's signed decomposition to the cones $\tilde{\mathbf{v}}_{ij} + C_{ij}$, discarding lower-dimensional cones, until all cones have index at most ℓ.
7. Enumerate the integer points in the fundamental parallelepipeds of all resulting cones to obtain the sets A_i.
8. Write down the formula (16).

Theorem 13. Algorithm 12 *is correct and runs in polynomial time when the dimension d is fixed and the maximum index ℓ is bounded by a polynomial in the input size.*

Proof. This is an immediate consequence of the analysis of Barvinok's algorithm. The irrationalization (step 5 of the algorithm) increases the encoding length of the apex vector only by a polynomial amount, because the dimension d is fixed and the depth k only depends doubly logarithmic on the initial index of the cone. □

The same technique can also be applied to the "homogenized version" of Barvinok's algorithm that was proposed in [9]; see also [14, Algorithm 11].

Algorithm 14 (irrational homogenized Barvinok algorithm).

Input: A polyhedron $P \subset \mathbf{R}^d$ given by rational inequalities in the form $A\mathbf{x} \leq \mathbf{b}$; the maximum index ℓ.

Output: A rational generating function in the form (16) for the integer points in the homogenization of P, i.e., the cone

$$C = \{\, (\xi\mathbf{x}, \xi) : \mathbf{x} \in P,\ \xi \in \mathbf{R}_+ \,\}. \tag{17}$$

1. Consider the inequality description for C; it is given by $A\mathbf{x} - \mathbf{b}\xi \leq 0$. The polar C° then has the rays $(A_{i,\cdot}, -b_i)$, $i = 1, \dots, m$.
2. Triangulate C° into simplicial cones C_j°, discarding lower-dimensional cones.
3. Polarize back the cones C_j° to obtain simplicial cones C_j.

4. Irrationalize the cones C_j to obtain shifted cones $\tilde{\mathbf{v}}_j + C_j$.
5. Apply Barvinok's signed decomposition to the cones $\tilde{\mathbf{v}}_j + C_j$, discarding lower-dimensional cones, until all cones have index at most ℓ.
6. Write down the generating function.

6. Extension to the nonsimplicial case. For polyhedral cones with few rays and many facets, it is usually much faster to perform triangulation in the primal space than in the dual space; cf. [8]. In this section, we show how to perform both Barvinok decomposition and triangulation in the primal space.

The key idea is to use linear programming to compute a subset of the stability region of the nonsimplicial cones.

LEMMA 15. *There is a polynomial-time algorithm that, given the vertex $\mathbf{v} \in \mathbf{Q}^d$ and the facet vectors $\mathbf{b}_i^* \in \mathbf{Z}^d$, $i = 1, \dots, m$, of a full-dimensional polyhedral cone $C = \mathbf{v} + B\mathbf{R}_+^n$, where $n \geq d$, computes a point $\hat{\mathbf{v}} \in \mathbf{Q}^d$ and a positive scalar $\rho \in \mathbf{Q}$ such that for every $\tilde{\mathbf{v}}$ in the open cube with $\|\tilde{\mathbf{v}} - \hat{\mathbf{v}}\|_\infty < \rho$, the cone $\tilde{\mathbf{v}} + B\mathbf{R}_+^n$ has no integer points on its proper faces and contains the same integer points as $\mathbf{v} + B\mathbf{R}_+^n$.*

Proof. We maximize ρ subject to the linear inequalities

$$\langle \mathbf{b}_i^*, \hat{\mathbf{v}} \rangle + \|\mathbf{b}_i^*\|_1 \, \rho \leq \lfloor \langle \mathbf{b}_i^*, \mathbf{v} \rangle \rfloor + 1, \tag{18a}$$

$$-\langle \mathbf{b}_i^*, \hat{\mathbf{v}} \rangle + \|\mathbf{b}_i^*\|_1 \, \rho \leq - \lfloor \langle \mathbf{b}_i^*, \mathbf{v} \rangle \rfloor , \tag{18b}$$

where $\hat{\mathbf{v}} \in \mathbf{R}^d$ and $\rho \in \mathbf{R}_+$. We can solve this linear optimization problem in polynomial time. Let $(\hat{\mathbf{v}}, \rho)$ be an optimal solution. Let $\tilde{\mathbf{v}} \in \mathbf{R}^d$ with $\|\tilde{\mathbf{v}} - \hat{\mathbf{v}}\|_\infty < \rho$. Let $\mathbf{x} \in (\tilde{\mathbf{v}} + B\mathbf{R}_+^d) \cap \mathbf{Z}^d$. Then we have for every $i \in \{1, \dots, m\}$

$$\begin{aligned} \langle \mathbf{b}_i^*, \mathbf{x} \rangle &\leq \langle \mathbf{b}_i^*, \tilde{\mathbf{v}} \rangle \\ &= \langle \mathbf{b}_i^*, \hat{\mathbf{v}} \rangle + \langle \mathbf{b}_i^*, \tilde{\mathbf{v}} - \hat{\mathbf{v}} \rangle \\ &\leq \langle \mathbf{b}_i^*, \hat{\mathbf{v}} \rangle + \|\mathbf{b}_i^*\|_1 \, \|\tilde{\mathbf{v}} - \hat{\mathbf{v}}\|_\infty \\ &< \langle \mathbf{b}_i^*, \hat{\mathbf{v}} \rangle + \|\mathbf{b}_i^*\|_1 \, \rho \\ &\leq \lfloor \langle \mathbf{b}_i^*, \mathbf{v} \rangle \rfloor + 1 \qquad \text{by (18a).} \end{aligned}$$

Because $\langle \mathbf{b}_i^*, \mathbf{x} \rangle$ is an integer, we actually have $\langle \mathbf{b}_i^*, \mathbf{x} \rangle \leq \lfloor \langle \mathbf{b}_i^*, \mathbf{v} \rangle \rfloor$. Thus, $\mathbf{x}$ lies in the cone $\mathbf{v} + B\mathbf{R}_+^d$. Conversely, let $\mathbf{x} \in (\mathbf{v} + B\mathbf{R}_+^d) \cap \mathbf{Z}^d$. Then, for every $i \in \{1, \dots, m\}$, we have $\langle \mathbf{b}_i^*, \mathbf{x} \rangle \leq \langle \mathbf{b}_i^*, \mathbf{v} \rangle$. Since $\mathbf{x} \in \mathbf{Z}^d$, we can round down the right-hand side and obtain

$$\begin{aligned} \langle \mathbf{b}_i^*, \mathbf{x} \rangle &\leq \lfloor \langle \mathbf{b}_i^*, \mathbf{v} \rangle \rfloor \\ &\leq \langle \mathbf{b}_i^*, \hat{\mathbf{v}} \rangle - \|\mathbf{b}_i^*\|_1 \, \rho \qquad \text{by (18b)} \\ &< \langle \mathbf{b}_i^*, \hat{\mathbf{v}} \rangle - \|\mathbf{b}_i^*\|_1 \, \|\tilde{\mathbf{v}} - \hat{\mathbf{v}}\|_\infty \\ &\leq \langle \mathbf{b}_i^*, \hat{\mathbf{v}} \rangle + \langle \mathbf{b}_i^*, \tilde{\mathbf{v}} - \hat{\mathbf{v}} \rangle \\ &= \langle \mathbf{b}_i^*, \tilde{\mathbf{v}} \rangle. \end{aligned}$$

Thus, $\mathbf{x} \in \tilde{\mathbf{v}} + B\mathbf{R}_+^d$. Moreover, since the inequality is strict, $\mathbf{x}$ does not lie on the face $\langle \mathbf{b}_i^*, \mathbf{x} \rangle = \langle \mathbf{b}_i^*, \tilde{\mathbf{v}} \rangle$ of the cone $\tilde{\mathbf{v}} + B\mathbf{R}_+^d$. □

LEMMA 16 (bound on the index of all subcones). *Let $\mathbf{b}_i \in \mathbf{Z}^d$, $i = 1, \dots, n$, be the generators of a full-dimensional polyhedral cone $K \subseteq \mathbf{R}^d$. Then the cones of any triangulation of K have an index bounded by*

$$D = \left(\max\nolimits_{i=1}^n \|\mathbf{b}_i\|^2 \right)^{n/2} . \tag{19}$$

Proof. Let $B \in \mathbf{Z}^{d\times d}$ be the generator matrix of a full-dimensional cone of a triangulation of K; then the columns B form a subset $\{\mathbf{b}_{i_1}, \dots, \mathbf{b}_{i_d}\} \subseteq \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$. Therefore

$$|\det B| \leq \prod_{k=1}^{d} \|\mathbf{b}_{i_k}\| \leq \left(\max_{i=1}^{n} \|\mathbf{b}_i\|^2\right)^{n/2},$$

giving the desired bound. □

With these preparations, the following corollary is immediate.

COROLLARY 17. *Let $\mathbf{v} + B\mathbf{R}^n_+$ be a full-dimensional polyhedral cone with apex at $\mathbf{v} \in \mathbf{Q}^d$, whose basis is given by the columns of the matrix $B \in \mathbf{Z}^{d\times d}$. Let D be defined by* (19). *Let $\hat{\mathbf{v}} \in \mathbf{Q}^d$ and $\rho \in \mathbf{Q}_+$ be the data from Lemma* 15 *describing the stability cube of $\mathbf{v} + B\mathbf{R}^n_+$. Using these data, construct $\tilde{\mathbf{v}}$ as in Theorem* 11. *Then the assertions of Theorem* 11 *hold.*

ALGORITHM 18 (all-primal irrational Barvinok algorithm).

Input: A polyhedron $P \subset \mathbf{R}^d$ given by rational inequalities; the maximum enumerated cone index ℓ.

Output: The rational generating function for $P \cap \mathbf{Z}^d$ in the form (16).

1. Compute all vertices $\mathbf{v}_i$ and corresponding supporting cones C_i of P.
2. Irrationalize all cones by computing new apex vectors $\tilde{\mathbf{v}}_i \in \mathbf{Q}^d$ from $\mathbf{v}_i$ by Corollary 17.
3. Triangulate $\tilde{\mathbf{v}}_i + C_i$ into simplicial cones $\tilde{\mathbf{v}}_i + C_{ij}$, discarding lower-dimensional cones.
4. Apply Barvinok's signed decomposition to the cones $\tilde{\mathbf{v}}_i + C_{ij}$, until all cones have index at most ℓ.
5. Enumerate the integer points in the fundamental parallelepipeds of all resulting cones to obtain the sets A_i.
6. Write down the formula (16).

7. Computational experiments.

Algorithms 12 and 18 have been implemented in a new version of the software package LattE, derived from the official LattE release 1.2 [10]. The new version, called LattE macchiato, is freely available on the Internet [20]. In this section, we discuss some implementation details and show the results of the first computational experiments.

7.1. Two substitution methods.

When the generating function g_P has been computed, the number of lattice points can be obtained by evaluating $g_P(\mathbf{1})$. However, $\mathbf{1}$ is a pole of every summand of the expression

$$g_P(\mathbf{z}) = \sum_{i\in I} \epsilon_i \frac{\sum_{\mathbf{a}\in A_i} \mathbf{z}^{\mathbf{a}}}{\prod_{j=1}^{d}(1 - \mathbf{z}^{\mathbf{b}_{ij}})}.$$

The method implemented in LattE 1.2 [14] is to use the *polynomial substitution*

$$\mathbf{z} = ((1+s)^{\lambda_1}, \dots, (1+s)^{\lambda_d})$$

for a suitable vector $\boldsymbol{\lambda}$. Then the constant coefficient of the Laurent expansion of every summand about $s = 0$ is computed using polynomial division. The sum of all the constant coefficients finally gives the number of lattice points.

Another method from the literature (see, for instance, [3]) is to use the *exponential substitution*

$$\mathbf{z} = (\exp\{\tau\lambda_1\}, \dots, \exp\{\tau\lambda_d\})$$

for a suitable vector $\boldsymbol{\lambda}$. By letting $\tau \to 0$, one then obtains the formula

$$(20) \qquad |P \cap \mathbf{Z}^d| = \sum_{i \in I} \epsilon_i \sum_{k=0}^{d} \frac{\mathrm{td}_{d-k}(\langle \boldsymbol{\lambda}, \mathbf{b}_{i1} \rangle, \dots, \langle \boldsymbol{\lambda}, \mathbf{b}_{id} \rangle)}{k! \cdot \langle \boldsymbol{\lambda}, \mathbf{b}_{i1} \rangle \cdots \langle \boldsymbol{\lambda}, \mathbf{b}_{id} \rangle} \sum_{\mathbf{a} \in A_i} \langle \boldsymbol{\lambda}, \mathbf{a} \rangle^k,$$

where td_{d-k} is the so-called Todd polynomial. In LattE macchiato, the exponential substitution method has been implemented in addition to the existing polynomial substitution; see [15] for implementation details.

7.2. Implementation details. We enumerate the lattice points in the fundamental parallelepiped by computing the Smith normal form of the generator matrix B; see Lemma 5.2 of [1].[1] For computing Smith normal forms, we use the library LiDIA, version 2.2.0 [21]. For solving the linear program in Lemma 15, we use the implementation of the revised dual simplex method in exact rational arithmetic in `cddlib`, version 0.94a [16]. All other computations are done using the libraries NTL, version 5.4 [22], and GMP, version 4.1.4 [17] for providing exact integer and rational arithmetic.

7.3. Evaluation of variants of the algorithms. We compare the variants of the algorithms using test instances that can also be solved without the proposed irrationalization techniques. We consider the test instances `hickerson-12`, `hickerson-13`, and `hickerson-14`, related to the manuscript [18]. They describe simplices in $\mathbf{R}^6$ and $\mathbf{R}^7$ that contain 38, 14, and 32 integer points, respectively. The examples are good test cases for our algorithms because the vertices and cones are trivially computed, and all computation time is spent in the Barvinok decomposition. We show the results in Tables 1, 2, and 3. The tables show results for the following methods.

1. Methods without irrationalization, using polarization to avoid computing with lower-dimensional cones:
   (a) LattE 1.2 [10], decomposing down to unimodular cones in the dual space (Algorithm 5 with $\ell = 1$).
   (b) Likewise, but using the implementation in the library `barvinok` 0.21 [23].
   (c) LattE macchiato, decomposing cones in the dual space, until all cones in the primal space have at most index ℓ (Algorithm 5), then using polynomial substitution. We show the results for different values of ℓ.
   (d) Likewise, but using exponential substitution.
2. Methods with irrationalization, performing triangulation in the dual space and Barvinok decomposition in the primal space (Algorithm 12):
   (a) LattE macchiato with polynomial substitution.
   (b) LattE macchiato with exponential substitution.

The tables show computation times in CPU seconds on a PC with a Pentium M processor with 1.4 GHz. They also show the total number of simplicial cones created in the decomposition, using the different variants of LattE; note that we did not measure the number of simplicial cones that the library `barvinok` produced.

We can make the following observations.

1. By stopping Barvinok decomposition before the cones are unimodular, it is possible to significantly reduce the number of simplicial cones. This effect is much stronger with irrational decomposition in the primal space than with decomposition in the dual space.

[1] The author wishes to thank Susan Margulies for prototyping the enumeration code.

TABLE 1
Results for hickerson-12.

	Without irrationalization					With irrationalization		
		Time (s)					Time (s)	
Max.		LattE	`barv.`	LattE macchiato			LattE macchiato	
index	Cones	v 1.2	v 0.21	Poly	Exp	Cones	Poly	Exp
1	11625	17.9	11.9	10.0	16.7	7929	7.8	12.7
10	4251			6.9	7.0	803	1.9	1.6
100	980			6.9	2.1	84	**1.3**	0.3
200	550			9.1	1.5	76	1.3	0.3
300	474			9.9	1.4	58	1.4	0.3
500	410			11.7	1.3	42	1.6	0.3
1000	130			7.2	0.7	22	1.7	**0.2**
2000	7			**2.2**	0.2	22	1.8	0.2
5000	7			2.8	**0.2**	7	2.8	0.2

TABLE 2
Results for hickerson-13.

	Without irrationalization					With irrationalization		
		Time (s)					Time (s)	
Max.		LattE	`barv.`	LattE macchiato			LattE macchiato	
index	Cones	v 1.2	v 0.21	Poly	Exp	Cones	Poly	Exp
1	466 540	793	589	421	707	483 507	479	770
10	272 922			**345**	428	55 643	117	109
100	142 905			489	249	9 158	**83**	22
200	122 647			625	222	6 150	93	17
300	98 654			903	199	4 674	105	14
500	90 888			1056	193	3 381	137	13
1000	73 970			1648	**190**	2 490	174	**13**
2000	66 954			2166	201	1 857	237	14
5000	49 168			5040	286	1 488	354	18
10000	43 511			7278	370	1 011	772	34

2. The newly implemented exponential substitution has a computational overhead compared to the polynomial substitution that was implemented in LattE 1.2.
3. However, when we compute with simplicial, nonunimodular cones, the exponential substitution becomes much more efficient than the polynomial substitution. Hence the break-even point between enumeration and decomposition is reached at a larger cone index. The reason is that the inner loops are shorter for the exponential substitution; essentially, only a sum of powers of scalar products needs to be evaluated in the formula (20). This can be done very efficiently.
4. The best results are obtained with the irrational primal decomposition down to an index of about 500 to 1000 and exponential substitution.

7.4. Results for challenge problems. In Table 4 we show the results for some larger test cases related to [18]. We compare LattE 1.2 with our implementation of irrational primal decomposition (Algorithm 12) with maximum index 500. The computation times are given in CPU seconds. The computations with LattE 1.2 were done on a PC Pentium M, 1.4 GHz; the computations with LattE macchiato were

Table 3
Results for hickerson-14.

	Without irrationalization					With irrationalization		
		Time (s)					Time (s)	
				LattE macchiato			LattE macchiato	
Max. index	Cones	LattE v 1.2	`barv.` v 0.21	Poly	Exp	Cones	Poly	Exp
1	1 682 743	4 017	15 284	2 053	3 466	552 065	792	1 244
10	1 027 619			**1736**	2 177	49 632	168	143
100	455 474			2 294	1 089	8 470	**128**	29
200	406 491			2 791	990	5 554	157	22
300	328 340			4 131	875	4 332	187	19
500	303 566			4 911	842	3 464	235	**18**
1000	236 626			8 229	**807**	2 384	337	18
2000	195 368			12 122	817	1 792	481	21
5000	157 496			22 972	1 034	1 276	723	27
10000	128 372			31 585	1 270	956	1 095	38

Table 4
Results for larger Hickerson problems.

			LattE v 1.2		LattE macchiato	
n	d	Lattice points	Cones	Time	Cones	Time
15	7	20	293 000	10 min 55 s	2 000	22 s
16	8	54	3 922 000	3 h 35 min	19 000	3 min 56 s
17	8	18			2 655 000	7 h 59 min
18	9	44	61 500 000	77 h 00 min	200 000	49 min 12 s
20	10	74			2 742 000	13 h 05 min

done on a slightly slower machine, a Sun Fire V890 with UltraSPARC-IV processors, 1.2 GHz.

Both the traditional Barvinok algorithm (Algorithm 5 with $\ell = 1$) and the homogenized variant of Barvinok's algorithm [9] do not work well for cross polytopes. The reason is that triangulation is done in the dual space, so hypercubes need to be triangulated. We show the performance of the traditional Barvinok algorithm in Table 5. We also show computational results for the all-primal irrational algorithm (Algorithm 18 with $\ell = 500$), using exponential substitution. The computation times are given in CPU seconds on a Sun Fire V440 with UltraSPARC-IIIi processors, 1.6 GHz.

A challenge problem related to the paper [5], case $m = 42$, could be solved using the all-primal irrational decomposition algorithm (Algorithm 18) with exponential substitution. The method decomposed the polyhedron to a total of 1.1 million simplicial cones of index at most 500. The computation took 66,000 CPU seconds on a Sun Fire V440 with UltraSPARC-IIIi processors, 1.6 GHz. The problem could not be solved previously because the traditional algorithms first tried to triangulate the polar cones, which does not finish within 17 days of computation.

8. Conclusions and future work. The above computational results with our preliminary implementation have shown that the proposed irrationalization techniques can speed up the Barvinok algorithm by large factors.

A further speed-up can be expected from a refined implementation. For example, the choice of the irrational shifting vector is based on worst-case estimates. It may be

Table 5
Results for cross polytopes.

	Without irrationalization		All-primal irrational	
d	Cones	Time (s)	Cones	Time (s)
4	384	1.1		0.9
5	3 840	6.5		1.4
6	46 264	91.7		2.7
7	653 824	1688.7		5.5
8			1 000	12.3
9			2 000	29.6
10			5 000	74.8
11			11 000	189.1
12			24 000	483.0
13			53 000	1 231.2
14			114 000	3 145.6
15			245 000	8 180.9

worthwhile to implement a randomized choice of the shifting vector (within the stability cube), using shorter rational numbers than those constructed in the paper. The randomized choice, of course, would not give the same guarantees as our deterministic construction. However, it is easy and efficient to check, during the decomposition, if the generated cones are all irrational; when they are not, one could choose a new random shifting vector (or resort to the one constructed in this paper) and restart the computation.

Acknowledgments. The author is grateful to David Einstein for pointing out a mistake in an earlier version of this paper. The author also acknowledges the hospitality of Jesús De Loera and the Department of Mathematics of the University of California, Davis.

REFERENCES

[1] A. I. Barvinok, *Computing the volume, counting integral points, and exponential sums*, Discrete Comput. Geom., 10 (1993), pp. 123–141.

[2] A. I. Barvinok, *Polynomial time algorithm for counting integral points in polyhedra when the dimension is fixed*, Math. Oper. Res., 19 (1994), pp. 769–779.

[3] A. I. Barvinok and J. E. Pommersheim, *An algorithmic theory of lattice points in polyhedra*, in New Perspectives in Algebraic Combinatorics, Math. Sci. Res. Inst. Publ. 38, Cambridge University Press, Cambridge, UK, 1999, pp. 91–147.

[4] M. Beck, C. Haase, and F. Sottile, *Theorems of Brion, Lawrence, and Varchenko on Rational Generating Functions for Cones*, http://arxiv.org/abs/math.CO/0506466 (2006).

[5] M. Beck and S. Hoşten, *Cyclotomic polytopes and growth series of cyclotomic lattices*, Math. Res. Lett., 13 (2006), pp. 607–622.

[6] M. Beck and F. Sottile, *Irrational Proofs for Three Theorems of Stanley*, http://arxiv.org/abs/math.CO/0501359 (2005).

[7] M. Brion, *Points entiers dans les polyédres convexes*, Ann. Sci. École Norm. Sup., 21 (1988), pp. 653–663.

[8] B. Büeler, A. Enge, and K. Fukuda, *Exact volume computation for polytopes: A practical study.*, in Polytopes—Combinatorics and Computation, G. Kalai and G. M. Ziegler, eds., DMV Sem. 29, Birkhäuser, Basel, 2000, pp. 131–154.

[9] J. A. De Loera, D. Haws, R. Hemmecke, P. Huggins, B. Sturmfels, and R. Yoshida, *Short rational functions for toric algebra and applications*, J. Symbolic Comput., 38 (2004), pp. 959–973.

[10] J. A. De Loera, D. Haws, R. Hemmecke, P. Huggins, J. Tauzer, and R. Yoshida, *LattE, Version* 1.2, http://www.math.ucdavis.edu/~latte/ (2005).

[11] J. A. De Loera, D. Haws, R. Hemmecke, P. Huggins, and R. Yoshida, *A computational*

study of integer programming algorithms based on Barvinok's rational functions, Discrete Optim., 2 (2005), pp. 135–144.

[12] J. A. De Loera, R. Hemmecke, M. Köppe, and R. Weismantel, *FPTAS for mixed-integer polynomial optimization with a fixed number of variables*, in Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (Miami, FL), SIAM, Philadelphia, 2006, pp. 743–748.

[13] J. A. De Loera, R. Hemmecke, M. Köppe, and R. Weismantel, *Integer polynomial optimization in fixed dimension*, Math. Oper. Res., 31 (2006), pp. 147–153.

[14] J. A. De Loera, R. Hemmecke, J. Tauzer, and R. Yoshida, *Effective lattice point counting in rational convex polytopes*, J. Symbolic Comput., 38 (2004), pp. 1273–1302.

[15] J. A. De Loera and M. Köppe, *New LattE Flavors*, manuscript, 2006.

[16] K. Fukuda, *cddlib, Version* 0.94a, http://www.cs.mcgill.ca/˜fukuda/soft/cdd_home/cdd.html (2005).

[17] *GMP, Version* 4.1.4*, the GNU Multiple Precision Arithmetic Library*, http://www.swox.com/gmp/ (2004).

[18] D. Hickerson, *Relatives of the Triple and Quintuple Product Identities*, manuscript, 1991.

[19] R. Kannan and A. Bachem, *Polynomial algorithms for computing the Smith and Hermite normal forms of an integer matrix*, SIAM J. Comput., 8 (1979), pp. 499–507.

[20] M. Köppe, *LattE macchiato, Version* 1.2-mk-0.6, http://www.math.uni-magdeburg.de/˜mkoeppe/latte/ (2006).

[21] *LiDIA, Version* 2.2.0*, a C++ Library for Computational Number Theory*, http://www.informatik.tu-darmstadt.de/TI/LiDIA/ (2006).

[22] V. Shoup, *NTL, a Library for Doing Number Theory*, http://www.shoup.net/ntl/ (2005).

[23] S. Verdoolaege, barvinok, *Version* 0.21, http://freshmeat.net/projects/barvinok/ (2006).

[24] S. Verdoolaege, R. Seghir, K. Beyls, V. Loechner, and M. Bruynooghe, *Counting integer points in parametric polytopes using Barvinok's rational functions*, Algorithmica, to appear.

SIAM J. Discrete Math.
Vol. 21, No. 1, pp. 237–250

ADJACENT VERTEX DISTINGUISHING EDGE-COLORINGS[*]

P. N. BALISTER[†], E. GYŐRI[†], J. LEHEL[†], AND R. H. SCHELP[†]

Abstract. An adjacent vertex distinguishing edge-coloring of a simple graph G is a proper edge-coloring of G such that no pair of adjacent vertices meets the same set of colors. The minimum number of colors $\chi'_a(G)$ required to give G an adjacent vertex distinguishing coloring is studied for graphs with no isolated edge. We prove $\chi'_a(G) \le 5$ for such graphs with maximum degree $\Delta(G) = 3$ and prove $\chi'_a(G) \le \Delta(G) + 2$ for bipartite graphs. These bounds are tight. For k-chromatic graphs G without isolated edges we prove a weaker result of the form $\chi'_a(G) = \Delta(G) + O(\log k)$.

Key words. proper edge-colorings, chromatic number, bipartite graphs

AMS subject classification. 05C15

DOI. 10.1137/S0895480102414107

1. Introduction.

Let G be a simple graph. We say a proper edge-coloring of G is *adjacent vertex distinguishing*, or an *avd*-coloring, if for any pair of adjacent vertices x and y, the set of colors incident to x is not equal to the set of colors incident to y. It is clear that an avd-coloring exists provided G contains no isolated edge. A *k-avd-coloring* is an avd-coloring using at most k colors. Let $\chi'_a(G)$ be the minimum number of colors in an avd-coloring of G. In [7] the following conjecture was made.

Conjecture 1. *If G is a simple connected graph on at least 3 vertices and $G \neq C_5$ (a 5-cycle), then $\Delta(G) \le \chi'_a(G) \le \Delta(G) + 2$.*

Since $\chi'_a(G)$ is at least as large as the edge-chromatic number of G it is clear that $\chi'_a(G) \ge \Delta(G)$, where $\Delta(G)$ is the maximum degree of any vertex in G. There are many examples of graphs for which $\chi'_a(G) > \Delta(G)+1$. For example, consider a graph of the form $G = K_{n,n} - H$, where H is a 2-factor of the complete bipartite graph $K_{n,n}$ containing no C_4. Assume we have an avd-coloring of G using $\Delta(G) + 1$ colors. Then each vertex is not incident to precisely one color, and assigning this missing color to each vertex gives a proper vertex-coloring of G with $\Delta(G)+1$ colors. Since G is bipartite with equal class sizes, the set of edges of a given color must miss the same number of vertices in each class. Hence each color occurs the same number of times on the vertices of each class. Since $\Delta(G) + 1 = n - 1$ there is a color that occurs at least twice in each class, but the vertices with this color do not form an independent set in G. Hence $\chi'_a(G) > \Delta(G) + 1$.

More generally, if G is regular, then both $\chi'_a(G)$ and the total chromatic number $\chi_T(G)$ are at least $\Delta + 1$, and the above argument shows that $\chi'_a(G) = \Delta + 1$ if and only if $\chi_T(G) = \Delta+1$. Hence any regular graph with $\chi_T(G) > \Delta+1$ gives an example of a graph with $\chi'_a(G) > \Delta + 1$.

We shall prove the following upper bounds for $\chi'_a(G)$.

[*] Received by the editors September 5, 2002; accepted for publication (in revised form) August 25, 2005; published electronically April 6, 2007.

http://www.siam.org/journals/sidma/21-1/41410.html

[†] Department of Mathematical Sciences, University of Memphis, Memphis, TN 38152-3240 (pbalistr@memphis.edu, gyori@renyi.hu, jlehel@memphis.edu, rschelp@memphis.edu). The work of the first author was partially supported by NSF grant EIA-0130352. The second author is on leave from the Rényi Institute of Mathematics, Hungarian Academy of Sciences and was partially supported by OTKA grant T34702. The third author is on leave from the Computer and Automation Research Institute of the Hungarian Academy of Sciences.

Theorem 1.1. *If G is a graph with no isolated edges and $\Delta(G) = 3$, then $\chi'_a(G) \leq 5$.*

Theorem 1.2. *If G is a bipartite graph with no isolated edges, then $\chi'_a(G) \leq \Delta(G) + 2$.*

Theorem 1.3. *If G is a k-chromatic graph with no isolated edges, then $\chi'_a(G) \leq \Delta(G) + O(\log k)$.*

In particular, Conjecture 1 holds for all bipartite graphs and all graphs with $\Delta(G) \leq 3$. Note that even for bipartite graphs, Conjecture 1 is best possible, as the example above shows. Theorem 1.3 is not best possible; indeed, Hatami [5] has recently shown using probabilistic methods that $\chi'_a(G) \leq \Delta(G) + 300$ for sufficiently large $\Delta(G)$, which is stronger than Theorem 1.3 for graphs with an extremely high chromatic number. Theorem 1.1 will be proved in section 2, Theorem 1.2 will be proved in section 3, and Theorem 1.3 will be proved in section 4.

Adjacent vertex distinguishing colorings are related to vertex distinguishing colorings in which *every* pair of vertices sees distinct color sets. This concept has been studied in many papers; see, for example, [1, 2, 3, 4, 5, 6].

2. Graphs with $\boldsymbol{\Delta(G) = 3}$. We start with the special case of regular graphs having a Hamiltonian cycle. Our coloring scheme is based on the idea of using the four elements of the Klein group $\mathbb{Z}_2 \times \mathbb{Z}_2$ to color the Hamiltonian cycle, defining the colors used algebraically, and a new fifth color for the chords forming a 1-factor. Local adjustments will be made to complete the colorings.

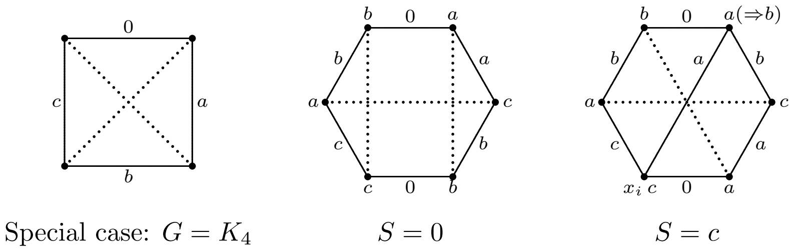
Lemma 2.1. *If G is a 3-regular Hamiltonian graph, then G has a 5-avd-coloring.*

Proof. Let the five colors be the elements $\{0, a, b, c\}$ of the Klein group $K = \mathbb{Z}_2 \times \mathbb{Z}_2$ together with the extra color 5. We have a commutative and associative addition defined on K such that $x + x = 0$ for all x and $a + b = c$. Let $C = x_1 \dots x_n$ be a Hamiltonian cycle of G and let I be the remaining 1-factor of G. We may assume $G \neq K_4$ (see Figure 1 for a 5-avd-coloring of K_4), so by Brooks' theorem, G has a vertex 3-coloring $f\colon V(G) \to \{a, b, c\}$. We may also assume that each of the three colors occurs at least once on G; otherwise a single vertex can be recolored to introduce the missing color. Let $S = \sum_{i=1}^{n} f(x_i) \in K$.

If $S = 0$, then label $x_n x_1$ with 0 and inductively label $x_i x_{i+1}$ for $i = 1, \dots, n-1$ so that $f(x_i)$ is the sum (in the group K) of the colors on $x_{i-1}x_i$ and $x_i x_{i+1}$. Equivalently, the color on $x_i x_{i+1}$ is the sum of the color on $x_{i-1}x_i$ and $f(x_i)$. Then $f(x_n)$ is the sum of the colors on $x_n x_1$ and $x_{n-1}x_n$. Color the 1-factor I with color 5. Each vertex x sees color 5 and two colors from K summing to $f(x)$. Since $f(x) \neq 0$ these two colors from K are distinct, and since $f(x) \neq f(y)$ for any two adjacent vertices x and y, the color sets at x and y must be distinct. Thus the coloring is a 5-avd-coloring of G.

Now suppose $S \neq 0$. Without loss of generality we may assume $S = c$. Pick any vertex x_i with $f(x_i) = c$. Let $x_i x_j \in I$. Then $f(x_j)$ is either a or b. Recolor x_j with b or a, respectively. Now $S = 0$ and we can recolor the edges of the Hamiltonian cycle as above (see Figure 1). Coloring I with 5 gives a proper edge-coloring that distinguishes adjacent vertices, except possibly at x_j. Since $f(x_i) \neq f(x_j)$ the pair of colors from K meeting x_i cannot be disjoint from the pair meeting x_j. Hence there must be some color of K missing from the edges incident to x_i or x_j. Recoloring the edge $x_i x_j$ with this missing color gives a 5-avd-coloring of G. The vertices x_i and x_j are distinguished from each other since $f(x_i) \neq f(x_j)$ and are distinguished from all other vertices since all other vertices meet color 5. ◻

We shall now assume that G is 3-regular with a 1-factor, but is not necessarily



Special case: $G = K_4$ $S = 0$ $S = c$

FIG. 1. *Colorings in Lemma* 2.1. *Dotted edges are colored* 5.

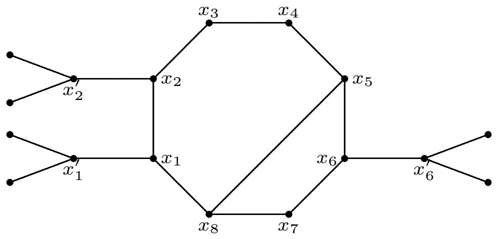


FIG. 2. *Graph H with $V_Y = \{x_1, x_2, x_6\}$, $V_C = \{x_5, x_8\}$, $V_S = \{x_3, x_4, x_7\}$.*

Hamiltonian. Since G has a 1-factor, G can be written as a union of this 1-factor and a collection of cycles. We shall show that under certain conditions we can extend a partial coloring to each cycle in turn.

We shall first find suitable colorings of graphs H of the following form. Let H be a cycle $C = x_1 \dots x_n$ with some extra 3-stars and chords added. To be precise, partition $V(C)$ as $V_Y \cup V_C \cup V_S$. For $x_i \in V_Y$, H will contain an edge $x_i x_i'$, $x_i' \notin V(C)$, where x_i' is joined to two degree 1 vertices. For $x_i \in V_C$, H will contain a chord $x_i x_j$, where $x_j \in V_C$. For $x_i \in V_S$, $d_H(x_i) = 2$ (see Figure 2).

We shall color such graphs so that adjacent degree 3 vertices are distinguished. We shall specify the colors incident to the x_i' for all $x_i \in V_Y$, and try to extend the coloring to the rest of H.

LEMMA 2.2. *Let H be a graph as above with $|V_S| \geq 2$. Suppose the edges incident to each x_i' with $x_i \in V_Y$ are properly colored with colors from $K \cup \{5\}$ and $x_i x_i'$ is colored* 5. *Then we can properly color the remaining edges of H with colors from $K \cup \{5\}$ so that adjacent degree* 3 *vertices are distinguished. Moreover, if $x_i \in V_S$, then we can ensure that either x_i meets color* 5 *or both neighbors of x_i meet color* 5.

Proof. We partition V_S into two sets V_I and V_M as follows. If $x, y \in V_S$ are adjacent on C, color the edge xy with color 5 and place x and y in the set V_M. Repeat with other adjacent pairs of V_S (that have not been used already) until $V_I = V_S \setminus V_M$ is an independent set. We shall now 3-color the degree 3 vertices of H with $\{a, b, c\}$. For $x_i \in V_Y$, set the color of x_i' to be the sum of the two colors of K incident to x_i'. Extend this vertex-coloring to a proper vertex-coloring of $V(H) \setminus V_S$ using a greedy algorithm—proceed around C, starting at any vertex immediately after a vertex of V_S, coloring each vertex of $V_Y \cup V_C$ in turn with any color from $\{a, b, c\}$ that ensures that the coloring is still proper.

If $V_M = \emptyset$, then $|V_I| \geq 2$. By coloring vertices of V_I (not necessarily properly) with colors from $\{a, b, c\}$, we can ensure that the sum of the vertex colors on C is $0 \in K$. If $V_M \neq \emptyset$, color V_I arbitrarily with $\{a, b, c\}$. Color the uncolored edges around C as in Lemma 2.1. At each vertex we add the vertex color in the Klein group to get the color of the next edge. The edge after any pair of vertices from V_M can be colored arbitrarily with any color from K. Color each chord of C with color 5. The resulting coloring satisfies the conditions of the lemma. □

Note that if we add an edge $x_i x_i'$ to H for some $x_i \in V_S$, then we can color $x_i x_i'$ with some color from K so that the new coloring is still proper and distinguishes degree 3 vertices. Indeed, if x_i meets color 5 in a coloring given by Lemma 2.2, then there are three colors which make the coloring proper and at most two of these will fail to distinguish x_i from x_{i+1} or x_{i-1}. If x_i does not meet color 5, then $x_i x_i'$ may be colored with either of the remaining colors of K since both x_{i+1} and x_{i-1} meet color 5.

Lemma 2.3. *Let H be a graph as above with $V_S = \emptyset$ and $x_1 \in V_Y$. Suppose the edges incident to each x_i' with $x_i \in V_Y \setminus \{x_1\}$ are properly colored with colors from $K \cup \{5\}$, $x_i x_i'$ is colored 5, and either of the following two conditions holds:*

(a) *All the edges incident to x_1' are colored, and one of the two edges that are incident to x_1' but not x_1 is colored 5.*

(b) *The edges incident to x_1' are colored, except for $x_1 x_1'$ which remains uncolored.*

Then the coloring can be completed to form a 5-avd-coloring of H. Moreover, in this coloring, $x_1 x_1'$ is not colored 5, but either x_1 meets color 5, or both x_2 and x_n meet color 5.

Proof. We shall provisionally color all chords $x_i x_j$ of C with color 5. As in the proof of Lemma 2.1 we shall 3-color the vertices of H with $\{a, b, c\}$. Each x_i' for $x_i \in V_Y$ is assigned the sum of the colors of K meeting it in H. We 3-color the vertices $x_2, \ldots, x_n$ in turn so that the coloring is proper using a greedy algorithm. The vertex x_1 will remain uncolored. Let this coloring be denoted by f and write $S = \sum_{i=2}^{n} f(x_i)$. If $S \neq 0$, then assign $x_1 x_2$ any color of K, and color the edges around the cycle as in the proof of Lemma 2.1. This gives four possible avd-colorings of $H - x_1'$, depending on the choice of color for $x_1 x_2$, and yields either $\{0, S\}$ or $K \setminus \{0, S\}$ as the pair of colors on $x_n x_1$ and $x_1 x_2$.

Assume that there is a chord $x_i x_j$ of C which does not meet either x_2 or x_n. Suppose without loss of generality that $f(x_i) = a$ and $f(x_j) = b$. Recolor either x_i or x_j with c and change the color of $x_i x_j$ to some color of K as in the proof of Lemma 2.1 so as to keep the coloring proper. Note that the coloring distinguishes x_i and x_j from all their neighbors, since their neighbors all meet color 5. In this way we can construct colorings with three distinct values of S (the original coloring, the coloring changing $f(x_i)$, and the coloring changing $f(x_j)$). At least two of these will have $S \neq 0$, and by varying the choice of color on $x_1 x_2$ as above, we obtain colorings with four possible values for the pair of colors on $x_n x_1$ and $x_1 x_2$. These four pairs form the edges of a C_4 inside K_K—the complete graph on the color set K. Moreover, both x_2 and x_n meet color 5, so are distinguished from x_1 regardless of the color (in K) of $x_1 x_1'$. In case (a) we are done since we can choose a coloring for which the pair of colors on $x_n x_1$ and $x_1 x_2$ avoids the color of $x_1 x_1'$. In case (b) we are done since we can choose a coloring for which the pair of colors on $x_n x_1$ and $x_1 x_2$ is neither equal nor disjoint from the pair that meet x_1'. Then there is at least one remaining color of K with which to color $x_1 x_1'$.

Assume now that $x_2 x_j$ is a chord of C. In case (b) color $x_1 x_1'$ arbitrarily with

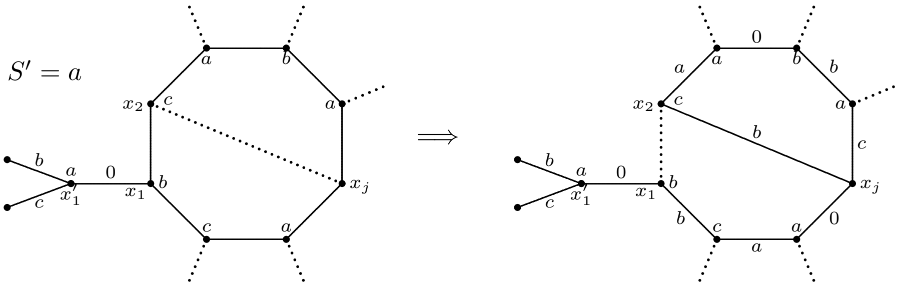


FIG. 3. *The case when x_2x_j is a chord of H.*

TABLE 1
Values of S.

$f(x_2)$	$f(x_3)$	S (j even)				S (j odd)			
a	c	a	0	c	b	b	c^*	0	a
a	b	a	0	c	b	c	b	a	0
b	c	b	c	0	a	a	0	c	b
c	b	c	b	a	0	a	0	c	b

some color of K so that the coloring is proper at x_1'. Restart from scratch and 3-color the vertices of H with $\{a, b, c\}$ as follows. As before, x_i' gets the sum of the colors of K meeting it when $x_i \in V_Y$. For the cycle C, we start the coloring at x_{j-1}, working backwards greedily until we reach x_2. The vertex x_2 can be colored in two ways. We pick one that ensures that $S' = \sum_{i=2}^{j-1} f(x_i) \neq 0$. Now continue coloring greedily with x_1, and then $x_n, \ldots, x_{j+1}$. The vertex x_j will remain uncolored.

Starting at x_1 and working backwards around C, color the edges so that $f(x_i)$ is the sum of the colors of K meeting x_i. For the edge x_jx_{j-1} pick either color of K that is not the same as the color of $x_{j+1}x_j$, or the sum of this color and S'. We continue coloring the edges of C as in Lemma 2.1 until we get to x_2. Color x_1x_2 with color 5 and color x_2x_j with the sum of $f(x_2)$ and the color on x_2x_3 (see Figure 3). This color will be the sum of S' and the color on x_jx_{j-1}, so it is distinct from the colors on $x_{j+1}x_j$ and x_jx_{j-1}. The resulting coloring satisfies the conditions of the lemma.

A similar argument deals with the case when there is a chord of the form x_nx_j, so we may now assume there are no chords, $V_C = \emptyset$. Restart by coloring the vertices of $C - x_1$ with $\{a, b, c\}$ as follows. Assume $f(x_3') = \cdots = f(x_j') = a \neq f(x_{j+1}')$ (or $j = n$). Color x_{j+1} with a and greedily color x_i for $i > j + 1$. The vertices $x_3, \ldots, x_j$ can be colored alternately by b and c, starting with either b or c. The vertex x_2 will be colored a, b, or c (possibly equal to the color of x_2', but not equal to the color of x_3). Let $S = \sum_{i=2}^{n} f(x_i)$.

We now list the possible colorings. For each choice of colorings of x_2 and x_3, there are four possible values of S depending on the value of $S' = \sum_{i=j+1}^{n} f(x_i)$ and j. Table 1 lists the various possibilities. Each value of S' and j has a column in the table for S. Since $f(x_2)$ and $f(x_3)$ can be changed independently of S' we have several choices for the vertex-coloring for each S' and j. We describe several cases in which we can find a suitable corresponding edge-coloring.

Case A. $f(x_2) \neq f(x_2')$, $S \neq 0$.

As in Lemma 2.1, we can edge color C starting at x_1x_2. Since $S \neq 0$, the colors on x_1x_n and x_1x_2 will be distinct. Depending on the choice of x_1x_2, the pair of colors meeting x_1 can be chosen to be either $\{0, S\}$ or $K \setminus \{0, S\}$. (We assume x_1x_1' is uncolored for now.)

Case B. $f(x_2) = f(x_2') \neq S$, $S \neq 0$.

As before we color the edges of C. However, this time only two choices for x_1x_2 are allowed since we must ensure that x_2 is distinguished from x_2' (either color not meeting x_2' will do for x_1x_2). These choices differ by the addition of $f(x_2)$ to every edge of C, and since $f(x_2) \notin \{0, S\}$, this swaps the pairs of colors $\{0, S\}$ and $K \setminus \{0, S\}$ on x_nx_1 and x_1x_2.

Case C. $f(x_2) = f(x_2') = S$, $S \neq 0$.

Unfortunately, both choices above of the color for x_1x_2 give the same pair of colors on x_nx_1 and x_1x_2. Hence we can only guarantee that colorings exist making x_1 meet *one* of the pairs $\{0, S\}$ or $K \setminus \{0, S\}$.

For each S' and j (corresponding to a column in Table 1) there are always at least two possible nonzero values for S. Moreover, for any choice of $f(x_2')$, we can find two choices of $f(x_2)$ and $f(x_3)$ with distinct values of $S \neq 0$, at least one of which has either $f(x_2) \neq f(x_2')$ or $f(x_2) = f(x_2') \neq S$ (the second case occurs only in the column indicated by *). Hence the set of pairs of colors meeting x_1 can be chosen as any edge of a path of edge length 3 in K_K (one value of S gives a matching in K_K, the other value gives at least one more edge in K_K).

In case (a) we are now done, since there is always a choice of the pair of colors that does not include the color on x_1x_1'. Also, x_1 and x_1' are distinguished since only x_1' meets color 5. In case (b) there is some choice for this pair of colors that is not equal or disjoint from the pair of colors meeting x_1'. Hence there is a choice of color in K for x_1x_1' which makes the coloring proper and distinguishes x_1 and x_1'. □

THEOREM 2.4. *If G is a 3-regular graph containing a 1-factor, then there exists a 5-avd-coloring of G.*

Proof. Without loss of generality we may assume G is connected. Decompose G as a 1-factor I and a union of cycles C_i. If there is only one cycle, then G is Hamiltonian and we are done by Lemma 2.1. Otherwise construct a new graph M with vertex set $V(M)$ equal to the set of cycles C_i and edges joining C_i and C_j when there is an edge of I joining some vertex of C_i to some vertex of C_j. Since G is connected, M is also connected. Pick a spanning tree T of M. Decompose T as a vertex disjoint union of stars S_j, $|V(S_j)| \geq 2$. For each S_j let G_j be the subgraph of G with an edge set made up of the edges of the cycles C_i of S_j, together with their chords in G and one edge of I joining C_i and $C_{i'}$ for each edge $C_iC_{i'}$ of S_j. Color G in the following way. Each edge (of I) that does not lie in any G_j will be colored 5. Now color each G_j in turn. If the star S_j has at least 3 vertices in M, use Lemma 2.2 to color the central cycle C_{i_0} of S_j. The graph H of Lemma 2.2 consists of C_{i_0}, its chords in G, and some 3-stars. The vertices of C_{i_0} incident to an edge joining C_{i_0} to another cycle in G_j will be placed in V_S, and we attach a 3-star to each remaining vertex of C_{i_0} that does not meet a chord of C_{i_0}. The edges of this 3-star correspond in an obvious way to some of the edges of G (although the degree 1 vertices and the edges incident with degree 1 vertices of H may not necessarily be distinct in G). We color the edges of the 3-stars with the corresponding colors already assigned in G, or arbitrarily (but properly) if no color has been assigned yet. Note that the edge of a 3-star incident to C_{i_0} will be colored 5. Lemma 2.2 now extends the coloring to the edges and chords of C_{i_0}. Now color the edges x_iy_i joining C_{i_0} to the other cycles C_i

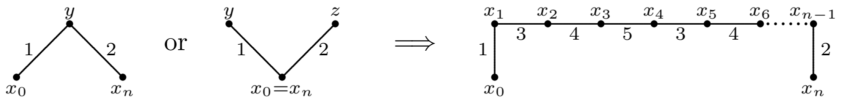


FIG. 4. *The case when G contains adjacent degree 2 vertices.*

of G_j with some color of K if x_i meets color 5 on C in such a way that the coloring is avd on C (see note after Lemma 2.2). Otherwise leave x_iy_i uncolored. We now color the other cycles C_i of G_j using Lemma 2.3 in a similar manner using the edge x_iy_i as the edge x_1x_1' of Lemma 2.3. The conditions of Lemma 2.3 ensure that we can find a coloring that is a 5-avd-coloring regardless of the choices of colors on the edges already colored. If the star S_j consists of just two vertices, use Lemma 2.3 on both constituent cycles. For the first cycle we use case (b) of Lemma 2.3. This will result in the edge x_1x_1' between the cycles being colored. If x_1 does not meet color 5, then uncolor x_1x_1'. Now color the other cycle using case (a) or (b) of Lemma 2.3. If x_1x_1' is recolored with a new color, then x_1 does not meet color 5, but both its neighbors on the first cycle do. Hence the coloring is still avd on the first cycle. □

Proof of Theorem 1.1. We shall prove Theorem 1.1 by induction on $|E(G)|$. Paths and cycles on at least 3 vertices have 5-avd-colorings [7], so we may assume that G is connected with maximum degree 3.

Assume x is a vertex of degree 1 in G. Let y be the neighbor of x. Then y is of degree 2 or 3. Since $G \neq P_3$ we can find a 5-avd-coloring of $G' = G - x$ by induction. In G', y has degree at most 2, so there are at least three colors not incident to y. At most two of these colors cannot be used to color xy, as they may result in y meeting the same set of colors as some neighbor in G'. However, there is still at least one color that can be given to xy so that the coloring is avd. Hence we may assume G contains no degree 1 vertex.

Assume two vertices of degree 2 are adjacent in G. Let $x_0x_1x_2 \dots x_n$, $n > 2$, be a *suspended trail* in G, i.e., a trail with $d_G(x_0) = d_G(x_n) = 3$ and $d_G(x_i) = 2$ for $0 < i < n$. If $x_0 \neq x_n$, let G' be the graph obtained by contracting this path to x_0yx_n. If $x_0 = x_n$, let G' be the graph obtained by deleting the vertices $x_1, \dots, x_{n-1}$ and adding two degree 1 vertices y, z to $x_0 = x_n$ (see Figure 4). By induction G' has a 5-avd-coloring. We may assume without loss of generality that the edge x_0y has color 1 and x_ny (or x_nz) has color 2. The edges x_ix_{i+1} of G can be colored with 1 for $i = 0$, 2 for $i = n - 1$, and cyclically with the colors $\{3, 4, 5\}$ for other values of i.

Hence we can assume that any vertex of degree 2 is adjacent only to vertices of degree 3. If G contains a bridge xy, let G_1 and G_2 be components of $G - xy$ with $x \in V(G_1)$ and $y \in V(G_2)$. Give $G_1 \cup xy$ and $G_2 \cup xy$ 5-avd-colorings by induction. (These graphs have smaller edge counts than G since G has no degree 1 vertices.) By permuting the colors on $G_2 \cup xy$, we can assume the edge xy receives the same color in each coloring and the color set incident to x in $G_1 \cup xy$ is not the same as the color set incident to y in $G_2 \cup xy$. This now gives a 5-avd-coloring of G (see Figure 5).

Hence we can assume that G is a graph with maximum degree 3, no vertices of degree 1, no pair of adjacent degree 2 vertices, and is bridgeless. By Tutte's 1-factor theorem, any cubic graph without a 1-factor must contain at least three bridges, so if G contains no degree 2 vertices, we are done by Theorem 2.4. If G contains degree 2 vertices, then let G' be the graph obtained by taking two copies of G and joining their corresponding degree 2 vertices by an edge. Then G' is 3-regular and contains at most

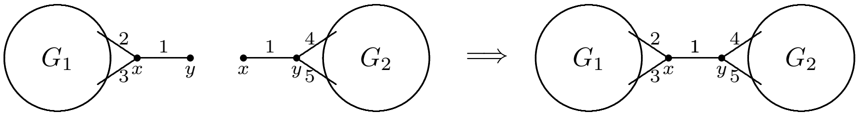


FIG. 5. *The case when G contains a bridge.*

one bridge. Hence G' has a 1-factor and so by Theorem 2.4 G' has a 5-avd-coloring. This coloring of G' induces a 5-avd-coloring of G since no two vertices of degree 2 are adjacent in G. □

3. Bipartite graphs. If G has an edge-coloring with colors $c_1, \dots, c_k$, write $G\{c_1, \dots, c_r\}$ for the subgraph of G consisting of all the vertices of G together with the edges of G that are colored with a color in $\{c_1, \dots, c_r\}$. Write $S(v)$ for the set of colors incident to v and $\chi' = \chi'(G)$ for the edge-chromatic number of G.

The bound $\chi'_a(G) \leq \Delta(G) + 3$ for regular bipartite graphs comes rather easily using the 1-factorization of regular bipartite graphs. To see this, observe that a 2-regular bipartite graph H with bipartition $V(H) = A \cup B$ has a straightforward 5-avd-coloring along each cycle such that $S(a) \in \{\{1,2\},\{3,4\},\{3,5\}\}$ and $S(b) \in \{\{1,4\},\{2,3\},\{4,5\},\{1,3\}\}$ for every $a \in A$ and $b \in B$. For $\Delta(G) > 2$ use this coloring for a 2-factor $H \subseteq G$ and give $G \setminus H$ any proper coloring with the remaining $\Delta(G) - 2$ colors. To obtain the bound $\chi'_a(G) \leq \Delta(G) + 2$ for any bipartite graph, however, much more effort will be required.

LEMMA 3.1. *If G is a bipartite graph with no isolated edges, then there exists a proper edge-coloring with colors $\{1, \dots, \chi'(G)\}$ such that*

A. *if $uv \in E(G) \setminus E(G\{1,2\})$, then either $\{1,2\} \subseteq S(u)$ or $\{1,2\} \subseteq S(v)$;*
B. *if C is a cycle in $G\{1,2\}$ which does not meet color 3 in G, then $\{1,2,3\} \subseteq S(v)$ for every neighbor v in $G \setminus C$ of any vertex of C;*
C. *if C is a cycle in $G\{1,2\}$ which does meet color 3 in G, then there exists a $u \in V(C)$ and $uv \in E(G\{3\})$ with $\{1,2\} \subseteq S(v)$ (we allow $v \in V(C)$); and*
D. *if uv is an isolated edge in $G\{1,2,3\}$, then $S(u) \neq S(v)$.*

Proof. Consider the set of edge-colorings of G with χ' colors. For all such colorings pick one such that

(1) $G\{1,2\}$ has maximal edge count;
(2) subject to (1), $G\{1,2\}$ has the minimum number of components (counting isolated vertices as components);
(3) subject to (1)–(2), $G\{3\}$ has maximal edge count; and
(4) subject to (1)–(3), the number of edges uv in G failing condition D is minimal.

We shall show that such a coloring satisfies conditions A–D.

Condition A. Let $uv \in E(G) \setminus E(G\{1,2\})$ be an edge with $\{1,2\} \not\subseteq S(u), S(v)$. Then u and v are either isolated vertices or the end-vertices of paths in $G\{1,2\}$. By recoloring uv with either color 1 or 2 (and possibly interchanging colors 1 and 2 in the component of v in $G\{1,2\}$) we obtain a proper edge-coloring with more edges colored $\{1,2\}$, contradicting (1). Note that if u and v are end-vertices of the same path in $G\{1,2\}$, then since G is bipartite, the edge uv can be recolored without changing any colors on this path.

Condition B. Assume $uv \in E(G)$ with $u \in V(C)$, $v \notin V(C)$, and $\{1,2,3\} \not\subseteq S(v)$. Note that uv is not colored with 1, 2, or 3. If $3 \notin S(v)$, then we can recolor uv with 3, contradicting (3). Hence without loss of generality $1 \notin S(v)$. Recolor uv with 1 and

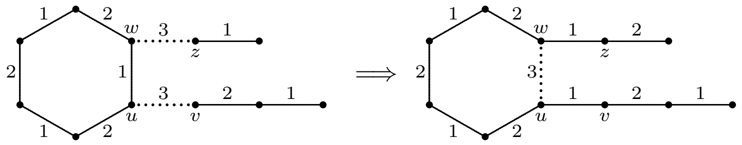


Fig. 6. *Proof of Condition C.*

recolor the color 1 edge on C meeting u with color 3. This contradicts condition (2).

Condition C. Suppose $u \in V(C)$ meets color 3 on an edge uv with $1 \notin S(v)$. Clearly $v \notin V(C)$. Let w be the neighbor of u on C with uw colored 1. If $3 \notin S(w)$, then recolor uw with 3 and uv with 1. This gives a coloring contradicting (2). If $3 \in S(w)$, let zw be the edge incident to w colored 3. If $\{1, 2\} \subseteq S(z)$, then we are done; otherwise we can assume z is either an isolated vertex or the end of a path in $G\{1, 2\}$. Recolor uv and wz with 1, uw with 3, and, if necessary, swap colors 1 and 2 on the path from z in $G\{1, 2\}$ so as to make the coloring proper (see Figure 6). If the paths in $G\{1, 2\}$ meeting v and z are the same, then recoloring this path will be unnecessary since G is bipartite. We now have a new coloring with more edges in $G\{1, 2\}$, contradicting (1).

Condition D. Let u_1v_1 be an isolated edge of $G\{1, 2, 3\}$. By Condition A, u_1v_1 is colored with either 1 or 2. Since G contains no isolated edge, we can assume that $d_G(u_1) \geq 2$ and that u_1 meets another color $k > 3$ on some edge of G. Swap colors 3 and k along a Kempe chain (component path of $G\{3, k\}$) starting at u_1 in G. By condition (3) we may assume that the last edge of this chain is recolored k. This reduces the number of edges failing condition D unless after the recoloring the other end-vertex v_2 of this chain lies in some isolated edge u_2v_2 of $G\{1, 2, 3\}$ and $S(u_2) = S(v_2)$. In this case u_2 also meets color k, so we can form a new Kempe chain starting at u_2 using colors 3 and k, disjoint from the u_1-v_2 chain. Repeating this process we get a sequence of Kempe chains on colors 3 and k from u_i to v_{i+1}. Note that properties (1)–(3) still hold after these recolorings. Eventually this process must terminate with a coloring reducing the number of edges failing condition D. Note that all the Kempe chains are vertex disjoint, and none end at v_1 since otherwise some recoloring would increase the number of edges colored 3, contradicting (3). □

Proof of Theorem 1.2. Color G as in Lemma 3.1. We shall recolor the edges of $G\{1, 2, 3\}$ with the five colors from $K \cup \{3\}$, where $K = \{0, a, b, c\}$ is the Klein group. This will give an avd-coloring with $\chi'(G) + 2 = \Delta(G) + 2$ colors. In addition, a vertex v will meet color 3 in the new coloring only if it met 3 in the original coloring, so $|S(v) \cap K|$ will be at least as large as the original degree of v in $G\{1, 2\}$.

The edges of $G\{1, 2\}$ form a set of vertex disjoint paths and even cycles. Construct a new graph M with a vertex set equal to the nonsingleton components C_i of $G\{1, 2\}$ and edges joining C_i and C_j when either

1. there is an edge of $G\{3\}$ joining a vertex of degree 2 in C_i to a vertex of degree 2 in C_j; or
2. either C_i or C_j is a single edge and there is an edge of $G\{3\}$ joining any vertex of C_i to any vertex of C_j.

As in the proof of Theorem 2.4, we take a star decomposition $\{S_j\}$ of a spanning forest of M and consider a corresponding subgraph G' of $G\{1, 2, 3\}$ in G consisting of the induced subgraphs in $G\{1, 2, 3\}$ of each cycle C_i and a choice of edges from $G\{3\}$

as above, joining C_i and $C_{i'}$ when $C_iC_{i'}$ is an edge of one of the stars in the star decomposition. Note that the graph M may contain isolated vertices, so some of the stars may be isolated vertices as well. We shall color every edge that does not lie in G' with its original color in G. The colors 1 and 2 do not appear on these edges. The subgraph G' will be colored with colors from $K \cup \{3\}$ so as to obtain an avd-coloring of G using at most two more colors.

We say a component C_i of $G\{1,2\}$ is *bad* if it is either a single edge where the end-vertices are not distinguished in the coloring of G, or a cycle of length congruent to 2 mod 4 that meets color 3, but has no color 3 chord. All other C_i's will be called *good*.

If C_i is a bad cycle, then by condition C, C_i is adjacent by an edge of $G\{3\}$ to a vertex of degree 2 in $G\{1,2\}$. In particular, C_i is not isolated in M. If C_i is a bad edge, then by condition D it meets an edge of $G\{3\}$, and so once again C_i is not isolated. Thus all isolated components are good.

Now we consider the stars S_j. Suppose we have a star with central component C_0 and end-components $C_1, \ldots, C_r$. If $r \geq 2$, delete the edge of G' from C_0 to good components C_i in S_j, $i > 0$, until either $r = 1$ or all C_i, $i > 0$, are bad. If $r = 1$ and C_0 and C_1 are good, delete the edge joining them in S_j. If C_0 is bad and C_1 is good, we consider C_1 to be the center of the star. Furthermore, if C_0 is an edge, then C_1 is not an edge (otherwise we would have two adjacent vertices of degree 1 in $G\{1,2\}$, contradicting condition A). In this case also we swap C_0 and C_1, so we can assume without loss of generality that C_0 is not a single edge when $r = 1$ (or $r > 2$). Any edge deleted from G' will remain colored 3 in our final coloring.

Hence we may assume each star S_j is either an isolated good C_i or a star with all end-components either bad or single edges. Also, the color 3 edges in G' joining C_0 to the end-components are incident to degree 2 vertices of C_0 except in the case when $r = 2$ and C_0 is a single edge.

We now recolor G' with colors from $K \cup \{3\}$. Let G have bipartition $V(G) = A \cup B$. We shall provisionally color the vertices of A with $a \in K$ and the vertices of B with $b \in K$. We shall color the edges of G in such a way that (with a few exceptions) each $v \in A$ with $d_{G'}(v) \geq 2$ will be colored so that $S(v) \cap K \in S_A$, where

$$S_A = \{\ \{0,a\},\ \{a,b\},\ \{b,c\},\ \{0,a,c\},\ \{0,b,c\}\ \},$$

while for $v \in B$, $S(v) \cap K \in S_B$, where

$$S_B = \{\ \{0,b\},\ \{0,c\},\ \{a,c\},\ \{0,a,b\},\ \{a,b,c\}\ \}.$$

This is sufficient, since if $uv \in E(G)$, $u \in A$, $v \in B$, and $S(u) = S(v)$, then $S(u) \cap K = S(v) \cap K$. But $S_A \cap S_B = \emptyset$, so $d_{G'}(u) < 2$, say. But then $|S(v) \cap K| = |S(u) \cap K| < 2$, so $S(v) \cap K \notin S_A, S_B$ and $d_{G'}(v) < 2$. However, $E(G') \supseteq E(G\{1,2\})$, so if $uv \notin E(G\{1,2\})$, then by condition A we can't have $d_{G'}(u), d_{G'}(v) < 2$. Finally, if $uv \in E(G\{1,2\})$, then $S(u) \cap \{4, \ldots, \chi'\} \neq S(v) \cap \{4, \ldots, \chi'\}$ by condition D and the fact that we do not recolor any edges of $G\{4, \ldots, \chi'\}$.

We shall now color each component of G' independently.

Case 1. *Good isolated paths.* Using the elements of K, color the edges of a good path arbitrarily so that the sum of the two colors meeting a degree 2 vertex of the path is equal to the color of this vertex. Any degree 2 vertex v will have $S(v) \cap K \in \{\{0,a\},\{b,c\}\} \subseteq S_A$ if $v \in A$ and $S(v) \cap K \in \{\{0,b\},\{a,c\}\} \subseteq S_B$ if $v \in B$.

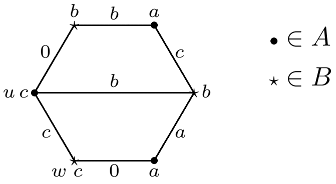


FIG. 7. *Cycle of length* 2 *mod* 4 *with chord.*

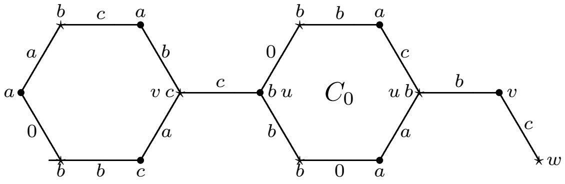


FIG. 8. *Stars of components.*

Case 2. *Good isolated cycles.* If the cycle length is divisible by 4, then we can color the edges from K so that the sum of the two colors meeting a vertex v is equal to the vertex color in K. Any color 3 chord will remain colored 3. If the cycle length is not divisible by 4 and there are no color 3 chords, then none of the vertices meets color 3 in G. However, by condition B of Lemma 3.1 all the neighbors of vertices of the cycle meet all three colors $\{1, 2, 3\}$ in G. If we give the cycle any avd-coloring using colors from K, we are done since every vertex on the cycle will meet only two colors from $K \cup \{3\}$, whereas their neighbors off the cycle will meet three such colors. (This is one case where we do not insist that $S(v) \cap K$ lies in S_A or S_B.) Finally, if the cycle has a color 3 chord uv, recolor u and a neighbor w of u on C with color c. Now color the edges around the cycle so that u meets $\{0, c\}$ if $u \in A$ or $\{a, b\}$ if $u \in B$. Then v is still labeled with a or b so the chord uv can be recolored by some color of K, making the coloring on C proper (see Figure 7). It is easily checked that $S(v) \cap K$, $S(u) \cap K$, and $S(w) \cap K$ lie in the correct set S_A or S_B as required.

Case 3. *Stars of components.* Remove any edges from G' that are chords of some component cycle C_i of the star. These edges will remain colored 3. If the central component C_0 is a cycle of length 2 mod 4, relabel one (and only one) vertex $u \in C_0$ that is adjacent to an end-component with b if $u \in A$ and a if $u \in B$. Assume now that C_0 is not a single edge. Color the central component so that the sum of two colors meeting a degree 2 vertex of C_0 is the vertex color of this vertex. Now recolor the color 3 edges uv from C_0 to C_i ($u \in C_0$, $v \in C_i$) with either 0 or c if $u \in A$, or a or b if $u \in B$. Now for each degree 2 or 3 vertex u of C_0, $S(u) \cap K \in S_A$ if $u \in A$ and $S(u) \cap K \in S_B$ if $u \in B$ (see Figure 8).

Each end-component is either a bad cycle of length 2 mod 4, or a single edge. For each end-component that is a cycle C, let v be the vertex of C joined to the central component C_0. Recolor v and a neighbor of v on C with color $c \in K$. Now color the edges of C so that the colors of K meeting v on C are $\{0, c\}$ if $v \in A$ and $\{a, b\}$ if $v \in B$. Now for each degree 2 or 3 vertex w of C, $S(w) \cap K \in S_A$ if $w \in A$ and $S(w) \cap K \in S_B$ if $w \in S_B$.

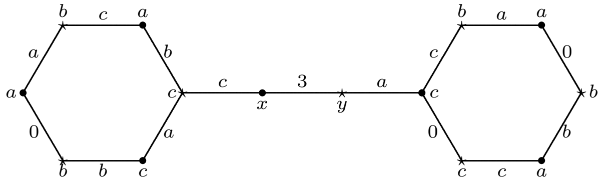


FIG. 9. *Central component is a single edge.*

For each end-component vw that is a single edge, let uv be the edge joining vw to C_0 (see Figure 8). If $v \in A$, then uv has been colored a or b. For either choice there is a choice of 0 or c on edge vw for which $S(v) \cap K \in S_A$. Similarly, if $v \in B$, then uv has been colored 0 or c. For either choice there is a choice of a or b on edge vw for which $S(v) \cap K \in S_B$.

Finally, assume C_0 is a single edge xy. Then C_0 is joined to two components, which by condition A must be cycles (see Figure 9). Recolor the edge xy with color 3. (Note that both x and y meet color 3 in the original coloring.) Now if we color the edges to the end-components and the edges of the end-components as before, we obtain a coloring with x distinguished from y. In this case $S(x) \cap K$ and $S(y) \cap K$ are not in S_A or S_B, so we need to check that x and y are distinguished from all neighbors in G. Clearly x and y are distinguished from their neighbors in G'. If, say, $zx \in E(G)$, then by condition A, $\{1, 2\} \subseteq S(z)$ in the original coloring. Hence in the final coloring $|S(z) \cap K| \geq 2 > |S(x) \cap K|$, so $S(z) \neq S(x)$. $\square$

Note that in the proof of Lemma 3.1 we only recolored the edges colored 1, 2, or 3, and for each edge uv, either the vertices u and v are distinguished by the colors in $K \cup \{3\}$, or uv is one of the isolated edges of $G\{1, 2, 3\}$ in condition D of Lemma 3.1.

4. General graphs.

The bound in Theorem 1.3 will be obtained by decomposing a general graph into bipartite graphs (Lemma 4.1), and by using an extended version of Lemma 3.1 that makes it possible to color these bipartite graphs "simultaneously" (Lemma 4.2).

LEMMA 4.1. *If G is a k-chromatic graph with no isolated edge or isolated K_3, then G can be written as the edge disjoint union of $\lceil \log_2 k \rceil$ bipartite graphs, each of which has no isolated edge.*

Proof. Let $r = \lceil \log_2 k \rceil$. Then $k \leq 2^r$. We first show that G is the union of r bipartite graphs without the restriction on isolated edges. For $r = 1$ this is clear. For $r > 1$ write $V(G)$ as the union of k independent color classes $V_1, \ldots, V_k$. Partition the classes into two groups $V_1, \ldots, V_{\lceil k/2 \rceil}$ and $V_{\lceil k/2 \rceil + 1}, \ldots, V_k$. Let G_1 be the bipartite graph formed by taking all edges from the first set of color classes to the second. Then $G \setminus E(G_1)$ has chromatic number at most $\lceil k/2 \rceil \leq 2^{r-1}$. Hence, by induction, $G \setminus E(G_1)$ can be written as the edge disjoint union of $r-1$ bipartite graphs $G_2, \ldots, G_r$. Thus G is the union of r bipartite graphs as required.

Write G as a union of r bipartite graphs in such a way that the total number of isolated edges in the subgraphs G_i is minimized. Suppose there is an isolated edge xy in G_1, say. Since there are no isolated edges in G, there must be some other bipartite graph G_2, with some edge incident to x, say. If we can add xy to G_2 without creating an odd cycle, then by removing xy from G_1 and adding it to G_2 we reduce the number of isolated edges. Hence we may assume there is an even length path from x to y in G_2.

If there are edges xz of G_2 with $d_{G_2}(z) = 1$, then remove one such edge from G_2 and add it to G_1. Since there is an even length path from x to y, no isolated edges are formed in G_2, but there are fewer isolated edges now in G_1. Similarly we are done if there are edges yz of G_2 with $d_{G_2}(z) = 1$. If no such edge xz or yz exists, remove an edge of an even length path from x to y in G_2 and add it to G_1. Use the edge of this path incident to y if $d_{G_2}(x) > 1$; otherwise use the edge incident to x. This will reduce the total number of isolated edges, except in the case when G_2 contains a component consisting of a path xzy of length 2 from x to y.

Since G does not contain an isolated K_3 there must be some other edge meeting $\{x, y, z\}$ in G. Suppose such an edge is incident to either x or y. Then this edge must lie in some other bipartite subgraph, say G_3. Considering G_3 in place of G_2 we may assume G_3 has a component xwy which is a path of length 2 from x to y. In this case put edge wx in G_1 and wy in G_2. Both G_1 and G_2 remain bipartite and G_3 loses a component. The number of isolated edges in G_1 decreases, contradicting our choice of decomposition into bipartite graphs.

Hence we may assume G has some other edge meeting z, but $d_G(x) = d_G(y) = 2$. The edge meeting z lies in G_i where $i = 1$ or $i > 2$. In this case we can move zx to G_i and xy to G_2. Both G_i and G_2 remain bipartite and G_1 loses the isolated edge xy. This reduces the number of isolated edges and contradicts the assumption that there is an isolated edge in some G_j. Hence there is a decomposition into r bipartite graphs, each of which has no isolated edge. □

LEMMA 4.2. *Assume G is a graph which is the edge disjoint union of bipartite graphs $G_1, \dots, G_r$, each of which has no isolated edge. Then there exists a proper edge-coloring with colors $\{1_1, \dots, 1_r, 2_1, \dots, 2_r, 3_1, \dots, 3_r, 4, \dots, \chi'\}$ such that colors 1_i, 2_i, and 3_i occur only on the edges of G_i and*

A. *if $uv \in E(G_i) \setminus E(G_i\{1_i, 2_i\})$, then either $\{1_i, 2_i\} \subseteq S(u)$ or $\{1_i, 2_i\} \subseteq S(v)$;*

B. *if C is a cycle in $G\{1_i, 2_i\}$ which does not meet color 3_i in G, then $\{1_i, 2_i, 3_i\} \subseteq S(v)$ for every neighbor v in $G_i \setminus C$ of any vertex of C;*

C. *if C is a cycle in $G\{1_i, 2_i\}$ which does meet color 3_i in G, then there exists a $u \in V(C)$ and $uv \in E(G\{3_i\})$ with $\{1_i, 2_i\} \subseteq S(v)$; and*

D. *if uv is an isolated edge in $G\{1_i, 2_i, 3_i\}$, then either $S(u) \cap \{4, \dots, \chi'\} \neq S(v) \cap \{4, \dots, \chi'\}$ or there is an edge in G incident to u colored with color 4.*

Proof. By coloring G with $\{1, \dots, \chi'\}$ and splitting colors 1, 2, and 3 into 1_i, 2_i, and 3_i according to which G_i the edge belongs to, we can find a coloring with the given set of colors so that edges colored 1_i, 2_i, or 3_i occur only in G_i. For all such colorings pick one such that

(1) $G\{1_1, \dots, 1_r, 2_1, \dots, 2_r\}$ has maximal edge count;

(2) subject to (1), the sum over i of the number of components of $G\{1_i, 2_i\}$ is minimal;

(3) subject to (1)–(2), $G\{3_1, \dots, 3_r\}$ has maximal edge count; and

(4) subject to (1)–(3), the number of edges uv failing condition D (for any i) is minimal.

As in the proof of Lemma 3.1, we see that conditions A–C hold for each i. It remains to prove condition D. Let u_1v_1 be an isolated edge of $G\{1_i, 2_i, 3_i\}$. Since G_i contains no isolated edge, we can assume that $d_{G_i}(u_1) \geq 2$ and that u_1 meets another color $k > 4$ on some edge of G_i. Swap colors 4 and k along a Kempe chain (in G) starting at u_1. This will reduce the number of edges failing condition D unless the other end-vertex v_2 of this chain lies in some isolated edge u_2v_2 of $G\{1_j, 2_j, 3_j\}$ and after the recoloring u_2v_2 fails condition D. In this case u_2 also meets color k, so we can form a new Kempe chain starting at u_2 using colors 4 and k. Repeating this process we get a sequence of Kempe chains on colors 4 and k from u_i to v_{i+1}. Eventually this process must terminate with a coloring reducing the number of edges failing condition D, or with some $v_r = v_1$. However, in this last case recoloring all these Kempe chains makes both v_1 and u_1 meet color 4. $\square$

Proof of Theorem 1.3. Since K_3 has a 3-avd-coloring, we can assume G contains no K_3 component. Decompose G using Lemma 4.1 and color G as in Lemma 4.2. Now recolor each bipartite subgraph G_i, replacing 1_i, 2_i, 3_i with a set of five colors $K_i = \{0_i, a_i, b_i, c_i, 3_i\}$, disjoint for each i, as in the proof of Theorem 1.2. Some edges uv of G_i may be isolated in $G_i\{1_i, 2_i, 3_i\}$, so u and v will not necessarily be distinguished in G_i; however, for all other edges $uv \in E(G_i)$, $S(u) \cap K_i \neq S(v) \cap K_i$ by the comment at the end of section 3. Also, this recoloring does not change any of the colors $4, \ldots, \chi'$ on G. Hence by condition D, if $uv \in E(G_i)$ and $S(u) \cap K_i = S(v) \cap K_i$, then either $S(u) \cap \{4, \ldots, \chi'\} \neq S(v) \cap \{4, \ldots, \chi'\}$ or $4 \in S(u)$. Let H be the subgraph of edges $uv \in E(G)$ such that u and v are not distinguished by the colors in $K_i \cup \{4, \ldots, \chi'\}$, where $uv \in E(G_i)$. Let H_I be the subgraph of H consisting of all the isolated edges of H. Each nonisolated vertex in H meets color 4, so $G\{4\} \cup H_I$ forms a collection of paths and cycles with all edges of H_I on the interior of any path or cycle. Split color 4 into three colors 4_A, 4_B, and 4_C. By alternately changing 4 into 4_A or 4_B along the paths and cycles of $G\{4\} \cup H_I$ we can distinguish the end-vertices of each edge of H_I. If a cycle of length 2 mod 4 occurs, we shall also need to color some of the color 4 edges of this cycle with 4_C. All other color 4 edges in G may become 4_C without loss of generality. This increases the number of colors used by 2 and distinguishes u and v for all $uv \in E(H_I)$. The graph $H_C = H \setminus H_I$ has no isolated edge, and $\Delta(H_C) \leq r \leq \lceil \log_2 k \rceil$. Pick $\chi'_a(H_C)$ new colors and recolor H_C so that it has an avd-coloring using these colors. The resulting coloring is avd. To see this, pick any edge uv of G. If $uv \in E(G_i)$ and $uv \notin E(H)$, then $S(u) \cap K_i \neq S(v) \cap K_i$ or $S(u) \cap \{4_A, 4_B, 4_C, 5, \ldots, \chi'\} \neq S(v) \cap \{4_A, 4_B, 4_C, 5, \ldots, \chi'\}$ since the recoloring of H_C removes elements from $S(u) \cap K_i$ only when u is in an isolated edge of $G_i\{1_i, 2_i, 3_i\}$. But in this case $|S(v) \cap K_i| \geq 2$ (by condition A) and $|S(u) \cap K_i| = 0$. If $uv \in E(H_I)$, then $S(u) \cap \{4_A, 4_B, 4_C\} \neq S(v) \cap \{4_A, 4_B, 4_C\}$, and if $uv \in E(H_C)$, then u and v are distinguished by the $\chi'_a(H_C)$ new colors.

Thus $\chi'_a(G) \leq \chi'(G) - 3 + 5r + 2 + \chi'_a(H_C)$. Finally, $\Delta(H_C) \leq \chi'(H_C) \leq r < \Delta(G)$. So by induction on $\Delta(G)$ we may assume $\chi'_a(H_C) = r + O(\log r)$, and $\chi'_a(G) = \Delta(G) + O(r) = \Delta(G) + O(\log k)$. $\square$

REFERENCES

[1] M. Aigner, E. Triesch, and Z. Tuza, *Irregular assignments and vertex-distinguishing edge-colorings of graphs*, in Proceedings of Combinatorics '90, A. Barlotti et al., eds., North–Holland, Amsterdam, 1992, pp. 1–9.

[2] P. N. Balister, O. M. Riordan, and R. H. Schelp, *Vertex-distinguishing edge colorings of graphs*, J. Graph Theory, 42 (2003), pp. 95–109.

[3] A. C. Burris and R. H. Schelp, *Vertex-distinguishing proper edge-colorings*, J. Graph Theory, 26 (1997), pp. 70–82.

[4] J. Černý, M. Hořňák, and R. Soták, *Observability of a graph*, Math. Slovaca, 46 (1996), pp. 21–31.

[5] H. Hatami, $\Delta + 300$ *is a bound on the adjacent vertex distinguishing edge chromatic number*, J. Combin. Theory Ser. B, 95 (2005), pp. 246–256.

[6] O. Favaron, H. Li, and R. H. Schelp, *Strong edge colorings of graphs*, Discrete Math., 159 (1996), pp. 103–109.

[7] Z. Zhongfu, L. Liu, and J. Wang, *Adjacent strong edge coloring of graphs*, Appl. Math Lett., 15 (2002), pp. 623–626.

SIAM J. DISCRETE MATH.
Vol. 21, No. 1, pp. 251–257

ON THE MINIMUM ORDER OF EXTREMAL GRAPHS TO HAVE A PRESCRIBED GIRTH[*]

C. BALBUENA[†] AND P. GARCÍA–VÁZQUEZ[‡]

Abstract. We show that any n-vertex extremal graph G without cycles of length at most k has girth exactly $k+1$ if $k \geq 6$ and $n > (2(k-2)^{k-2}+k-5)/(k-3)$. This result provides an improvement of the asymptotical known result by Lazebnik and Wang [*J. Graph Theory*, 26 (1997), pp. 147–153] who proved thatthe girth is exactly $k+1$ if $k \geq 12$ and $n \geq 2^{a^2+a+1}k^a$, where $a = k-3-\lfloor (k-2)/4 \rfloor$. Moreover, we prove that the girth of G is at most $k+2$ if $n > (2(t-2)^{k-2}+t-5)/(t-3)$, where $t = \lceil (k+1)/2 \rceil \geq 4$. In general, for $k \geq 5$ we show that the girth of G is at most $2k-4$ if $n \geq 2k-2$.

Key words. extremal graphs, girth, forbidden cycles, cages

AMS subject classification. 05C35

DOI. 10.1137/060656747

1. Introduction.

Throughout this paper, only undirected simple graphs without loops or multiple edges are considered. Unless otherwise stated, we follow [2] for terminology and definitions.

Let $V(G)$ and $E(G)$ denote the set of vertices and the set of edges of a graph G, respectively. The order of G is denoted by $|V(G)| = n$ and the size by $|E(G)| = e(G)$. The minimum length of a cycle contained in G is the *girth* $g(G)$ of G. A cycle of minimum length is said to be *a girdle* and if G does not contain a cycle, we set $g(G) = \infty$. By C_r we will denote a cycle of length r, $r \geq 3$.

Let $\mathcal{F}$ be a family of graphs. The extremal number $ex(n, \mathcal{F})$ is the maximum number of edges in a graph of order n that does not contain any graph of $\mathcal{F}$ as a subgraph. The graphs of order n and size $ex(n, \mathcal{F})$ not containing any $F \in \mathcal{F}$ as a subgraph are the extremal graphs and are denoted by $EX(n, \mathcal{F})$. We refer to graphs from $EX(n, \mathcal{F})$ as *extremal* $\mathcal{F}$-free graphs of order n, or just *extremal*.

By $ex(n; \{C_3, C_4, \ldots, C_k\})$ we denote the maximum number of edges in a graph of order n and girth at least $k+1$, and by $EX(n; \{C_3, C_4, \ldots, C_k\})$ we denote the set of all graphs of order n, girth at least $k+1$, and with $ex(n; \{C_3, C_4, \ldots, C_k\})$ edges. Erdös and Sachs [3] showed that an r-regular graph of girth at least $k+1$ with the least possible number of vertices has girth equal to $k+1$. (A proof of this result can be found in Lovász [7, pp. 66, 384, 385, and the references therein].) These graphs are called $(r; k+1)$-*cages*.

In this paper we consider a similar question asked by Garnick and Nieuwejaar in [5] on extremal graphs with a relatively large girth. Is there a constant c such that for all $k \geq 5$ and all $n \geq ck$, the girth of any extremal graph with girth $\geq k+1$ is $k+1$? They give an affirmative answer for $k = 4$. Lazebnik and Wang [6] showed that the

[*] Received by the editors April 7, 2006; accepted for publication (in revised form) October 11, 2006; published electronically April 6, 2007. This research was supported by the Ministry of Education and Science, Spain, and the European Regional Development Fund (ERDF) under project MTM2005-08990-C02-02.

http://www.siam.org/journals/sidma/21-1/65674.html

[†] Departament de Matemàtica Aplicada III, Universitat Politècnica de Catalunya, Barcelona, Spain (m.camino.balbuena@upc.edu).

[‡] Departamento de Matemática Aplicada I, Universidad de Sevilla, Sevilla, Spain (pgvazquez@us.es).

answer is negative for $c = 2$ and affirmative if $k = 5$ or if n is large in comparison with k. More precisely they proved the following result.

Theorem A. *Let* $k \geq 12$, $a = k - 3 - \lfloor (k-2)/4 \rfloor$, $n \geq 2^{a^2+a+1} k^a$, *and* $G \in EX(n; \{C_3, C_4, \ldots, C_k\})$. *Then the girth* $g(G) = k+1$.

In order to prove Theorem A, Lazebnik and Wang used the following result, which they also stated in [6].

Theorem B. *Let* $k \geq 3$, $G \in EX(n; \{C_3, C_4, \ldots, C_k\})$, *and the maximum degree be* $\Delta(G) \geq k$. *Then* $g(G) = k+1$.

Our main contribution to this problem is to provide an improvement of Theorem A. More precisely we prove that the girth of $G \in EX(n; \{C_3, C_4, \ldots, C_k\})$ is $k+1$ if either $k = 3$ and $n \geq 5$; or $k = 4$ and $n \geq 9$; or $k = 5$ and $n \geq 8$; or $k = 6$ and $n \geq 171$; or $k \geq 7$ and

$$n \geq \frac{2(k-2)^{k-2} + k - 5}{k-3} + 1.$$

This contribution contains the known results for $k = 3, 4, 5$; see [4, 5, 6]. Furthermore, it gives an answer to the problem for $k = 6$ posed by Lazebnik and Wang [6], who asked to prove the girth of an extremal $\{C_3, C_4, C_5, C_6\}$-free graph is 7.

Moreover, we show that the girth of $G \in EX(n; \{C_3, C_4, \ldots, C_k\})$ is at most $2k - 4$ provided that $k \geq 5$ and $n \geq 2k - 2$. This clearly implies that for $k = 6$ the girth of an extremal graph is at most 8 for $10 \leq n \leq 170$.

Let $t = \lceil (k+1)/2 \rceil$. We also prove that the girth of $G \in EX(n; \{C_3, C_4, \ldots, C_k\})$ is at most $k+2$ if $k \geq 7$ and

$$n \geq \frac{2(t-2)^{k-2} + t - 5}{t-3} + 1.$$

From this result it follows for $k = 7$ that if $n \geq 64$, then $g(G) \leq 9$.

2. Main results. The set of neighbors of $u \in V(G)$ is denoted by $N_G(u)$. The number of neighbors of u is the degree $d_G(u)$ of u in G, or briefly $d(u)$ when it is clear which graph is meant. The distance $d_G(x, y)$ in G of two vertices x, y is the length of a shortest $x - y$ path in G. The greatest distance between any two vertices in G is the diameter $D(G)$ of G. Diameter and girth are related by $g(G) \leq 2D(G) + 1$. Let $e = xy$ be an edge of G. As usual we will denote by $G/\{e\} = G/e$ the graph obtained from G by contracting the edge e into a new vertex v_e, which becomes adjacent to all the former neighbors of x and y. Taking into account that we dealt with simple graphs of girth at least 4 the resultant graph by any edge contraction remains simple.

Throughout the paper $k \geq 3$ is an integer. We begin by proving a technical and useful lemma.

Lemma 2.1. *Let* $G \in EX(n; \{C_3, \ldots, C_k\})$ *have two distinct edges* e_1 *and* e_2 *such that every cycle of* G *containing both of them has a length of at least* $k+3$. *Then the girth is* $g(G) = k+1$ *if the diameter is* $D(G/\{e_1, e_2\}) \geq k - 2$.

Proof. Let $G \in EX(n; \{C_3, \ldots, C_k\})$ satisfy the hypothesis of the lemma and suppose that the girth is $g(G) \geq k+2$. The graph $G' = G/\{e_1, e_2\}$ has $g(G') \geq k+1$ because by hypothesis any cycle passing through both edges e_1 and e_2 has a length of at least $k+3$. Let u', v' be two vertices of G' such that $d_{G'}(u', v') = D(G')$; then by hypothesis $d_{G'}(u', v') = D(G') \geq k - 2$. Let us consider the graph G^* obtained from G' by adding two new vertices x_1, x_2 and the three edges $u'x_1, x_1x_2$, and x_2v'. We have $g(G^*) = \min\{g(G'), D(G') + 3\} \geq k + 1$, $|V(G^*)| = |V(G')|$

$+ 2 = n$, and $e(G^*) = e(G) + 1$, which contradict the maximality of G. Therefore $g(G) = k + 1$. □

As a first consequence of the above lemma, we obtain in the next theorem an upper bound for the girth of any extremal graph which contains the known result $g = k + 1$ for $k = 5$; see [6].

THEOREM 2.2. *Let $G \in EX\,(n; \{C_3, \ldots, C_k\})$ be for $k \geq 5$ and $n \geq 2k-2$. Then G has a girth of $g(G) \leq 2k - 4$.*

Proof. Let $G \in EX\,(n; \{C_3, \ldots, C_k\})$ satisfy the hypothesis of the theorem, and assume the girth of G is $g \geq 2k - 2$. Let $C : u_0 u_1 \cdots u_{g-1} u_0$ be a girdle in G, and notice that $g \geq k+3$ because $k \geq 5$. The graph $G' = G/\{u_0u_1, u_1u_2\}$ clearly has girth $g(G') \geq 2k-4$; hence the diameter is $D(G') \geq \lfloor g(G')/2 \rfloor \geq \lfloor (2k-4)/2 \rfloor = k-2$. By Lemma 2.1 we have $g = g(G) = k+1$, yielding $2k-2 \leq k+1$, which is a contradiction because $k \geq 5$. Therefore the girth of G is $g \leq 2k-3$. Assume the girth of G is exactly $g = 2k-3$. As $n \geq 2k-2$ the graph G must contain a vertex y not belonging to C. Without loss of generality, suppose that u_0y is an edge of G. Notice that u_{k-2} and u_{k-1}, both belonging to C, satisfy that $d_C(u_0, u_{k-2}) = d_C(u_0, u_{k-1}) = k-2$. Then both $u_0 - u_{k-2}$ and $u_0 - u_{k-1}$ paths contained in C must be the unique shortest $u_0 - u_{k-2}$ and $u_0 - u_{k-1}$ paths in G, because $k-2 = (g-1)/2$. This implies that $d_G(y, u_{k-2}) \geq k-2$ and $d_G(y, u_{k-1}) \geq k-2$ so that every cycle, if any, containing both edges u_0y and $u_{k-2}u_{k-1}$ must have a length of at least $g+1 = 2k-2$, which is at least $k+3$ because $k \geq 5$. Now let $G'' = G/\{u_0y, u_{k-2}u_{k-1}\}$. Clearly, $D(G'') \geq d_{G''}(u_1, u_k) = d_G(u_1, u_k) = k-2$. By Lemma 2.1 we obtain $g(G) = g = k+1$, i.e., $2k-3 \leq k+1$, which is impossible for $k \geq 5$. Hence the girth of G is at most $2k-4$ and the theorem is valid. □

Next, we obtain the following result which is an improvement of Theorem A and also contains the known results for $k = 3, 4, 5$; see [4, 5, 6].

THEOREM 2.3. *Let $G \in EX\,(n; \{C_3, \ldots, C_k\})$. Then $g(G) = k+1$ if either $k = 3$ and $n \geq 5$; or $k = 4$ and $n \geq 9$; or $k = 5$ and $n \geq 8$; or $k = 6$ and $n \geq 171$; or $k \geq 7$ and*

$$n \geq \frac{2(k-2)^{k-2} + k - 5}{k-3} + 1.$$

Proof. From Theorem 2.2 it follows that any graph $G \in EX\,(n; \{C_3, C_4, C_5\})$ for $n \geq 8$ has girth of 6. Therefore we can assume $k = 3, 4$ or $k \geq 6$. Let $G \in EX\,(n; \{C_3, \ldots, C_k\})$ and suppose that its girth is $g(G) \geq k+2$. Then, by Theorem B we have $\Delta \leq k-1$, where Δ denotes the maximum degree of G. Let D be the diameter of G and let us take two vertices x, y at distance $d_G(x,y) = D$. Then $D \leq k-1$ because otherwise by adding the edge xy to G we would obtain a graph G' of order n having girth $g(G') \geq k+1$ and more edges than G, which contradicts the maximality of G. Let us consider the two cases $D = k-1$ and $D \leq k-2$ separately.

Case 1. $D = k-1$. Define the set $N_G^{k-1}(x) = \{y \in V(G) : d_G(x,y) = k-1\}$. Clearly, $|N_G^{k-1}(x)| \geq 1$, because $y \in N_G^{k-1}(x)$. Let us see that $|N_G^{k-1}(x)| = 1$.

Let $W = \{w \in V(G) : d_G(x,w) + d_G(w,y) = k-1\}$ and suppose that there exists a vertex $u \in V(G) \setminus W$. Then $d_G(x,u) + d_G(u,y) \geq k$ or, in other words, all the possible paths passing through u that connect x with y have a length of at least k. Take any vertex $v \in N_G(u)$ and consider the graph G' resulting by contracting the edge uv in G. The girth of this new graph is $g(G') \geq k+1$ and the diameter $D(G') = D = k-1$. So let $x', y' \in V(G')$ be such that $d_{G'}(x', y') = k-1$, and denote by G^* the graph obtained from G' by adding a new vertex x^* and the edges

$x'x^*$ and x^*y'. Clearly, $|V(G^*)| = |V(G')| + 1 = n$, and girth $g(G^*) = k + 1$, but $e(G^*) = e(G') + 2 = e(G) + 1$, which contradicts the maximality of G. Hence, $V(G) = W$, which readily implies that y is the only vertex at distance $D = k-1$ from x and the number of vertices at distance $D - 1 = k - 2$ from x is at most Δ, since these vertices must be neighbors of y.

Therefore, if $k = 3$, then $n \le 1+\Delta+1 \le 1+k = 4$, contradicting the hypothesis for this case. If $k = 4$, then $n \le 1 + \Delta + \Delta + 1 \le 2k = 8$, contradicting again the hypothesis for this case. So assume that $k \ge 6$. As for $1 \le i \le D - 2 = k - 3$, the maximum number of vertices at distance i from x is $\Delta(\Delta - 1)^{i-1}$, we obtain

$$\begin{aligned} n &\le 1 + \Delta \sum_{i=0}^{k-4} (\Delta - 1)^i + \Delta + 1 \le 1 + (k - 1) \sum_{i=0}^{k-4} (k - 2)^i + k \\ &= \frac{(k - 1)(k - 2)^{k-3} - 2}{k - 3} + k \\ &< \frac{(k - 1)(k - 2)^{k-3} - 2}{k - 3} + (k - 2)^{k-3} = \frac{2(k - 2)^{k-2} - 2}{k - 3}. \end{aligned}$$

This contradicts the hypothesis of the theorem, so $g(G) = k+1$ in the case $D = k-1$.

Case 2. $D \le k - 2$. Notice that $k = 3, 4$ are impossible for this case because $D \ge \lfloor g/2\rfloor \ge \lfloor (k + 2)/2\rfloor$. So we have $k \ge 6$.

Let x^* be a vertex of G with degree $d_G(x^*) = \delta$, where δ is the minimum degree of G, and let us denote by $\epsilon(x^*) = \max\{d_G(x^*, y) : y \in V(G)\}$ the eccentricity of x^*. As the diameter is the maximum of the eccentricities we have $\epsilon(x^*) \le D \le k - 2$. Suppose first that $\epsilon(x^*) \le k - 3$. As for $1 \le i \le k - 3$, the maximum number of vertices at distance i from x^* is $\delta(\Delta - 1)^{i-1}$, it is immediate that

$$n \le 1 + \delta \sum_{i=0}^{k-4} (\Delta - 1)^i \le 1 + (k - 1) \sum_{i=0}^{k-4} (k - 2)^i \le \frac{(k - 1)(k - 2)^{k-3} - 2}{k - 3},$$

which is a contradiction. Therefore $\epsilon(x^*) = k - 2$, which means $D = k - 2$. Let us consider the set $N_G^{k-2}(x^*) = \{y \in V(G) : d_G(x^*, y) = k - 2\}$. Let us prove the following claim.

Claim. Given any vertex $y \in N_G^{k-2}(x^*)$, every neighbor of vertex y is at a distance of $k - 3$ from x^*.

Otherwise suppose that there exists a vertex $y_1 \in N_G^{k-2}(x^*) \cap N_G(y)$. Let us denote by $x^* = x_0x_1x_2 \cdots x_{k-2} = y$ any shortest $x^* - y$ path. Clearly, every cycle containing both edges x^*x_1 and yy_1, if any, has a length of at least $k + 3$ because $k \ge 6$. Then we consider the new graph G' obtained from G by contracting the edges x^*x_1 and yy_1. If the diameter of G' is $D(G') = k - 2$, then by Lemma 2.1 we would have $g(G) = k + 1$, which is a contradiction with our assumption $g(G) \ge k + 2$. Therefore $D(G') = k - 3$, which implies that for all $z \in N(x^*)$, $d_G(z, y') = k - 3$ for all $y' \in N_G^{k-2}(x^*)$. Consequently, the edge yy_1 and any vertex $z \in N_G(x^*)$ lies on a cycle in G of length at most $2k - 5$, which is impossible for $k = 6$ because $g \ge k + 2$. Hence every neighbor of vertex y is at a distance of $k - 3$ from x^* when $k = 6$ and the claim is true for this case.

Furthermore, for $k \ge 7$ we have $d_{G'}(v_{x^*x_1}, v_{yy_1}) = k - 3$, where $v_{x^*x_1}$ and v_{yy_1} denote the newly arising vertices by the contraction of the edges x^*x_1 and yy_1. Besides, $d_{G'}(v_{x^*x_1}) = d_G(x^*) + d_G(x_1) - 2 \le \delta + \Delta - 2 \le 2(\Delta - 1)$ and $d_{G'}(v_{yy_1}) =$

$d_G(y) + d_G(y_1) \leq 2(\Delta - 1)$. Therefore,

$$V(G') = \{v_{x^*x_1}\} \cup \bigcup_{i=1}^{k-3} N^i_{G'}(v_{x^*x_1}),$$

where $N^i_{G'}(v_{x^*x_1})$ denotes the set of vertices of G' at a distance of i from vertex $v_{x^*x_1}$. Thus $\left|N^i_{G'}(v_{x^*x_1})\right| \leq 2(\Delta-1)(\Delta-1)^{i-1} = 2(\Delta-1)^i$, for $i = 1, \ldots, k-3$, and we get

$$\begin{aligned} n = 2 + |V(G')| &\leq 3 + 2\sum_{i=1}^{k-3}(\Delta-1)^i \\ &\leq 3 + 2\sum_{i=1}^{k-3}(k-2)^i \\ &= 3 + \frac{2(k-2)^{k-2} - 2(k-2)}{k-3} = \frac{2(k-2)^{k-2} + k - 5}{k-3}, \end{aligned}$$

contradicting the hypothesis of the theorem. Thus, every vertex $y \in N_G^{k-2}(x^*)$ has all its neighbors at distance $k-3$ from x^* and the claim holds.

Hence, $|N^i_G(x^*)| \leq \delta(\Delta-1)^{i-1}$, for $i = 1, \ldots, k-3$, and $|N_G^{k-2}(x^*)| \leq (\Delta-1)^{k-3}$. Then, for $k \geq 6$ we have

$$\begin{aligned} n &\leq 1 + \delta\sum_{i=0}^{k-4}(\Delta-1)^i + (\Delta-1)^{k-3} \\ &\leq 1 + \delta\sum_{i=0}^{k-4}(k-2)^i + (k-2)^{k-3} \\ &\leq \frac{(k-1)(k-2)^{k-3} - 2}{k-3} + (k-2)^{k-3} = \frac{2(k-2)^{k-2} - 2}{k-3}. \end{aligned}$$

This contradicts the hypothesis of the theorem, so we conclude that $g(G) = k+1$. □

Next, the goal is to provide a lower bound on n in order to guarantee that the girth is at most $k+2$ for $k \geq 7$. To do that first we state that an extremal $\{C_3, \ldots, C_k\}$-free graph with maximum degree $\Delta \geq \lceil (k+1)/2 \rceil$ has necessarily a girth of at most $k+2$.

THEOREM 2.4. *Let $k \geq 7$ be an integer. Let G be a graph belonging to the family $EX\,(n; \{C_3, \ldots, C_k\})$ with a minimum degree of at least 2 and maximum degree Δ. Then $g(G) \leq k+2$ if $\Delta \geq \lceil (k+1)/2 \rceil$.*

Proof. Let $G \in EX\,(n; \{C_3, \ldots, C_k\})$ satisfy the hypothesis of the theorem, and assume $g(G) \geq k+3$. Let x be a vertex of maximum degree Δ and let $y_1, y_2, \ldots, y_\Delta$ be all the neighbors of x. Since $d_G(y_i) \geq 2$, for each $i = 1, \ldots, \Delta$, there exists $x_i \in V(G) - x$ adjacent to y_i. Notice also that $x_i \neq x_j$ for all $i \neq j$, since $g(G) > 4$. Taking into account that $g(G) \geq k+3$, we deduce that $d_{G-x}(x_i, x_j) \geq g(G) - 4 \geq k-1$, $d_{G-x}(y_i, y_j) \geq g(G) - 2 \geq k+1$, and $d_{G-x}(x_i, y_j) \geq g(G) - 3 \geq k$ for all $i, j = 1, \ldots, \Delta$ with $i \neq j$. Let G^* be the graph obtained from G by first deleting the $\Delta-1$ edges $xy_2, \ldots, xy_\Delta$ and second adding the new Δ edges $y_1x_2, \ldots, y_{\Delta-1}x_\Delta, y_\Delta x_1$. Then G^* has order n and size $e(G^*) = e(G) + 1$. Since G is extremal, G^* must contain a cycle of length at most k. Let us denote by C^* a shortest cycle in G^* (notice that

$x \notin V(C^*)$, since x has degree 1 in G^*). We denote by C the cycle $x_1y_1x_2y_2\cdots x_\Delta y_\Delta x_1$ which has length $2\Delta \geq k+1$. Observe that C is an induced cycle of G^*, since x_i is nonadjacent to y_j in G, for any $i \neq j$ and the only newly introduced edges are y_ix_{i+1} for $i = 1, \ldots, \Delta - 1$ and $y_\Delta x_1$. Moreover, $C^* \neq C$, since $g(C) \geq k+1$ and $g(C^*) \leq k$. So, we may express $C^* = P_1 \cup P_2$, where P_1 is the longest path whose edges belong to the set $E(C^*)\backslash E(C) \subseteq E(G-x)$, and P_2 is the rest of C^*. Notice that the endvertices of P_1 must belong to $\{x_1, \ldots, x_\Delta\} \cup \{y_1, \ldots, y_\Delta\}$ by the construction of P_1. Observe also that P_2 contains at least one edge of $E(C)$, because otherwise the cycle C^* would be contained in G against the assumption $g(G) \geq k+3$. If the endvertices of P_1 are x_i and x_j for certain $i, j \in \{1, 2 \ldots, \Delta\}$, then the edge $y_{i-1}x_i$ or x_iy_i and the edge $y_{j-1}x_j$ or x_jy_j must be contained in P_2 and then $e(P_2) \geq 2$. This implies that $|V(C^*)| = e(C^*) = e(P_1) + e(P_2) \geq d_{G-x}(x_i, x_j) + 2 \geq k - 1 + 2 = k + 1$; a contradiction. If the endvertices of P_1 are x_i and y_i, for some $i \in \{1, \ldots, \Delta\}$, then $e(P_1) \geq d_{G-x-\{x_iy_i\}}(x_i, y_i) \geq g(G) - 1 \geq k+2$ and hence $|V(C^*)| = e(C^*) = e(P_1) + e(P_2) \geq k+3$, again a contradiction. Otherwise,

$$e(P_1) \geq \min\{d_{G-x}(y_i, y_j), d_{G-x}(x_i, y_j) : i, j = 1, \ldots, \Delta \text{ and } i \neq j\} \geq k,$$

which implies $|V(C^*)| = e(C^*) = e(P_1) + e(P_2) \geq k+1 > k$, arriving at a contradiction. Hence, $g(G) \leq k+2$. □

From Theorem 2.4 we derive the following sufficient condition in terms of the order for an extremal $\{C_3, \ldots, C_k\}$-free graph to have girth at most $k+2$.

Theorem 2.5. *Let $G \in EX\,(n; \{C_3, \ldots, C_k\})$ be of a minimum degree of at least 2. Then the girth is $g(G) \leq k+2$ if $k \geq 7$ and*

$$n \geq \frac{2(t-2)^{k-2} + t - 5}{t-3} + 1,$$

where $t = \lceil (k+1)/2 \rceil$.

Proof. If $\Delta \geq \lceil (k+1)/2 \rceil$, then $g(G) \leq k+2$ for $k \geq 7$ because of Theorem 2.4 and the theorem holds. Hence assume $\Delta \leq \lceil (k+1)/2 \rceil - 1$ and $g(G) \geq k+3$. Let $t = \lceil (k+1)/2 \rceil$. As in the proof of Theorem 2.3 we consider two cases $D = k-1$ and $D \leq k-2$ separately and repeat this proof but taking into account that now $\Delta \leq t-1$ instead of $\Delta \leq k-1$. In this way we arrive at a contradiction, which implies $g(G) \leq k+2$, and the theorem holds. □

As an immediate consequence of Theorems 2.3 and 2.5, the following information about the girth of any extremal $\{C_3, \ldots, C_7\}$-free graph is provided.

Corollary 2.6. *Let G be a graph belonging to the family $EX\,(n; \{C_3, \ldots, C_7\})$. Then the girth $g(G) = 8$ if $n \geq 783$, and the girth is $g(G) \leq 9$ if $n \geq 64$.*

3. Conclusions. Theorem 2.3 can be compared with Theorem A. Both results give a sufficient condition on the order of an extremal graph to contain a cycle of minimum length $k+1$. Recall that $a = k - 3 - \lfloor (k-2)/4 \rfloor$; then for $k \geq 12$ we have $2^a > (k-2)^2$. Hence $2^{a^2+a+1} > 2(k-2)^{2a+2} \geq 2(k-2)^{(3k-6)/2}$, and thus $n \geq 2^{a^2+a+1}k^a > 2(k-2)^{(3k-6)/2}k^a$ (which is much larger than the requirement obtained in Theorem 2.3), $n > (2(k-2)^{k-2} + k - 5)/(k-3)$.

Moreover, Theorems 2.2 and 2.3 provide information on the girth of any extremal $\{C_3, C_4, C_5, C_6\}$-free graph G. The girth is $g(G) = 7$ if $n \geq 171$, and the girth is $g(G) \leq 8$ if $n \geq 10$. It is known for $r = 3, 4, 5$ that each $(r; 8)$-cage is the incidence graph of a projective geometry called *generalized quadrangle*; see the survey by Wong [8]. The order of each of these graphs is $30, 80, 170$, respectively. As a referee suggests, it appears that a result of Alon, Hoory, and Linial [1] can be used to show these cages do belong to $EX(n;\{C_3, C_4, C_5, C_6, C_7\})$. The question is if these cages are also $\{C_3, C_4, C_5, C_6\}$-free extremal. We would like to suggest the following open problems.

PROBLEM 1. *Prove or disprove that each* $(r; 8)$*-cage for* $r = 3, 4, 5$ *is a graph belonging to* $EX\,(n;\{C_3, C_4, C_5, C_6\})$*, for* $n = 30, 80, 170$.

PROBLEM 2. *Is it possible to improve the lower bound on* n *in Theorem* 2.3 *for* $k \geq 7$*?*

Acknowledgments. We would like to express our thanks to the referees for their helpful comments and suggestions.

REFERENCES

[1] N. ALON, S. HOORY, AND N. LINIAL, *The Moore bound for irregular graphs*, Graphs Combin., 18 (2002), pp. 53–57.

[2] R. DIESTEL, *Graph Theory*, 3rd ed., Springer-Verlag, Berlin, 2005.

[3] P. ERDÖS AND H. SACHS, *Reguläre Graphen gegebener Taillenweite mit minimaler Knotenzahl*, Wiss. Z. Martin-Luther-Univ. Halle-Wittenberg Math.-Natur. Reihe, 12 (1963), pp. 251–257.

[4] D. K. GARNICK, Y. H. H. KWONG, AND F. LAZEBNIK, *Extremal graphs without three-cycles or four-cycles*, J. Graph Theory, 17 (1993), pp. 633–645.

[5] D. K. GARNICK AND N. A. NIEUWEJAAR, *Nonisomorphic extremal graphs without three-cycles or four-cycles*, J. Combin. Math. Combin. Comput., 12 (1993), pp. 33–56.

[6] F. LAZEBNIK AND P. WANG, *On the structure of extremal graphs of high girth*, J. Graph Theory, 26 (1997), pp. 147–153.

[7] L. LOVÁSZ, *Combinatorial Problems and Exercises*, North–Holland, Amsterdam, 1979.

[8] P. K. WONG, *Cages—A survey*, J. Graph Theory, 6 (1982), pp. 1–22.

SIAM J. DISCRETE MATH.
Vol. 21, No. 1, pp. 258–263

PRECOLORING EXTENSION FOR 2-CONNECTED GRAPHS[*]

MARGIT VOIGT[†]

Abstract. Let $G = G(V, E)$ be a simple graph, $\mathcal{L}$ a list assignment with $|L(v)| = \Delta(G)$ for all $v \in V$, and $W \subseteq V$ an independent subset of the vertex set. Define $d(W) := \min\{d(v, w) \mid v, w \in W\}$ to be the minimum distance between two vertices of W. In this paper it is shown that if G is 2-connected with $\Delta(G) \geq 4$ and G is not the complete graph $K_{\Delta(G)+1}$, then every precoloring of W is extendable to a proper list coloring of G provided that $d(W) \geq 4$. An example shows that the bound is sharp. This extends a result of Axenovich [*Electron. J. Combin.*, 10 (2003), note 1] and Albertson, Kostochka, and West [*SIAM J. Discrete Math.*, 18 (2004), pp. 542–553], who proved that $d(W) \geq 8$ guarantees such an extension for all G with $\Delta(G) \geq 3$ not containing $K_{\Delta(G)+1}$.

Key words. list coloring, precoloring extension, distance constraints

AMS subject classification. 05C15

DOI. 10.1137/05064312X

1. Introduction.

Let us consider simple graphs $G = (V, E)$ with maximum degree $k = \Delta(G) \geq 3$. The well-known theorem of Brooks [10] states that such a graph is k-colorable if it does not contain K_{k+1} as a component. The aim of this paper is the generalization of this theorem in two directions.

First, we consider the list version of this problem. That means every vertex has a set $L(v)$ of available colors. The set $L(v)$ is also called a *list* of v and the collection of all lists is called a list assignment $\mathcal{L}$ of G. The graph G is *$\mathcal{L}$-list colorable* if a proper coloring of the vertices exists where every vertex gets a color from its list in $\mathcal{L}$. This concept was introduced by Vizing [14] and independently by Erdős, Rubin, and Taylor [11] at the end of the seventies. A *k-assignment* is a list assignment $\mathcal{L}$ where $|L(v)| = k$ for all $v \in V$. Among others, in [11, 13] a Brooks-type theorem is proved which says that a graph G with maximum degree $k = \Delta(G) \geq 3$ is $\mathcal{L}$-list colorable for every k-assignment $\mathcal{L}$ if G does not contain K_{k+1}. With regard to this theorem it is natural to ask what happens if $\mathcal{L}$ is a list assignment with $|L(v)| = d_G(v)$ for all $v \in V(G)$, where $d_G(v)$ is the degree of v in G. Let us define a *supervalent list assignment* $\mathcal{L}$ being a list assignment with $|L(v)| \geq d_G(v)$ for all $v \in V(G)$. Investigating the mentioned question we need a special class of graphs, the Gallai trees. A *Gallai tree* is a connected graph in which every block is a complete graph or an odd cycle. Now let H be a connected graph. A *leaf block* of H is a block of H containing at most one cut vertex and the *block-cutpoint graph* T of H has a vertex for each block in H and a vertex for each cut vertex of H, and a cut vertex v is adjacent to a block B if $v \in V(B)$. Note that the block-cutpoint graph of a connected graph H is a tree and every leaf of T corresponds to a leaf block of H.

An important tool for the proof of the main result of this paper is the following theorem (see [9] and [11]). A short proof is given in [13].

THEOREM 1 (see [9, 11, 13]). *If $\mathcal{L}$ is a supervalent list assignment for a connected graph G and there is no $\mathcal{L}$-coloring of G, then*

(a) *$|L(v)| = d(v)$ for every vertex $v \in V(G)$.*

[*] Received by the editors October 19, 2005; accepted for publication (in revised form) August 23, 2006; published electronically April 6, 2007.
http://www.siam.org/journals/sidma/21-1/64312.html

[†] University of Applied Sciences, Dresden 01008, Germany (m.voigt@htw-dresden.de).

(b) *G is a Gallai tree.*

Now we assume additionally that there is a subset $W \subseteq V$ of the vertex set which is already precolored. Denote by $d(W)$ the minimum distance between two components of W in G. We would like to extend the precoloring of W to a proper coloring of the whole vertex set. Clearly the existence of such an extension depends on $d(W)$ and either the number of available colors or the length of the lists of the list assignment, respectively. First, results in this direction were given by Albertson in 1998 [1] answering a question of Thomassen from 1997. He stated that if G is k-colorable and W is independent with $d(W) \geq 4$, then every (k+1)-coloring of W can be extended to a proper $(k+1)$-coloring of V. There were several papers in the past few years dealing with this topic from different points of view; see, for example, [2, 3, 4, 5, 6, 7, 8] and [12].

In this paper we ask for the extension of a precoloring of W to a proper list coloring if every vertex has a list of $k = \Delta(G) \geq 4$ colors. Axenovich [8] and Albertson, Kostochka, and West [5] proved that for independent W, $k = \Delta(G) \geq 3$, and $d(W) \geq 8$, such an extension is always possible if G does not contain a K_{k+1} as subgraph. Furthermore, they give an example showing that the bound 8 is sharp. Remarkably, the mentioned example is a 1-connected graph. So let us consider in this paper graphs which are 2-connected, which means that the removal of at most one vertex does not disconnect the graph. For these graphs the bound for $d(W)$ will be reduced in this paper to $d(W) \geq 4$ provided that $k = \Delta(G) \geq 4$ and G is not a K_{k+1}. Furthermore, an example shows that this bound is sharp.

For 2-connected graphs with $\Delta(G) = 3$ there is a recent analogous result saying that $d(W) \geq 6$ ensures an extension of a precoloring (see [15]). An example at the end of this paper shows that this bound is sharp.

If G is a 2-connected graph with $\Delta(G) = 2$, then G is a cycle. If we forbid odd cycles as in the original version of Brooks' theorem, then G is an even cycle C_{2k}. In this case a bad precoloring of two vertices of distance k does not allow an extension to a coloring with two colors. Thus there does not exist a similar result for $\Delta(G) = 2$.

Additionally, let us mention that if we allow one color more in the lists, which means $|L(v)| = \Delta(G)+1$ for every $v \in V(G)$ and the precolored set W is independent, then $d(W) \geq 3$ already guarantees the extension to a proper list coloring. This fact follows from a result of Albertson [1] and the mentioned Brooks-type theorem for list colorings.

2. Extension of Brooks' theorem.

The main result of this paper is given in the following theorem.

THEOREM 2. *Let $G = (V, E)$ be a 2-connected graph with $k = \Delta(G) \geq 4$ which is not K_{k+1}, $W \subseteq V$ an independent subset of the vertex set, $d(W) \geq 4$, and $\mathcal{L}$ a list assignment with $|L(v)| = k$ for all $v \in V$. Then every precoloring of W extends to a proper $\mathcal{L}$-list coloring of V.*

Proof. The first part of the proof is similar to the proof of [5].

Assume that the statement of Theorem 2 is not true and G is a counterexample with the smallest number of vertices.

Delete the colors of the precoloring of W from the lists of the corresponding neighbors. Denote the new list assignment by $\mathcal{L}'$ and the graph induced by $V(G) \setminus W$ by H. Because of $d(W) \geq 4$ we know that

$$|L'(v)| \geq |L(v)| - 1. \tag{1}$$

Note, furthermore, that $\mathcal{L}'$ is a supervalent list assignment for H since $|L(v)| = \Delta(G) \geq d_G(v)$ and therefore $L'(v) \geq d_H(v)$ for all $v \in V(H)$. By the minimality of G we may assume that H is connected.

Since G is a counterexample to the statement of Theorem 2 it follows that H is not list colorable from the lists of $\mathcal{L}'$. Thus H fulfills the assumptions of Theorem 1 and H and $\mathcal{L}'$ have the properties stated there. Hence we have the following claim.

Claim 1. *For H and $\mathcal{L}'$ the following holds:*

(a) *H is a Gallai tree and $|L(v)| = d_H(v)$ for all $v \in V(H)$.*
(b) *$d_G(v) = k$ for all $v \in V(H)$.*
(c) *Each vertex of H has at most one neighbor in W.*
(d) *For all $v \in V(H)$ we have $k-1 \leq d_H(v) \leq k$ and $d_H(v) = k-1$ if and only if v has a neighbor in W.*

These properties can be derived easily from the assumptions and Theorem 1 and are proved in [5] too.

In the following claim, denote the set of the noncut vertices of a block B of H by B'.

Claim 2. *Let B be a leaf block of H. Then the following holds:*

(a) *$B = K_k$.*
(b) *H has more than one block.*
(c) *There exists a unique vertex $w_B \in W$ which is adjacent to all noncut vertices of B. Thus w_B has exactly $k-1$ neighbors in B'.*
(d) *w_B has exactly one neighbor, y_B, in $V(H) \setminus \bigcup_{B \in \mathcal{B}_\ell} V(B)$, where $\mathcal{B}_\ell$ is the set of all leaf blocks of H. For two different leaf blocks, B_1 and B_2, we have $y_{B_1} \neq y_{B_2}$.*

Proof.

(a) Since H is a Gallai tree, B has to be a complete graph or an odd cycle. Assume B is an odd cycle or a complete graph K_s with $s \leq k-1$. Then for $v \in B'$ we have $d_H(v) < k-1$. Using Claim 1 and inequality (1), it follows for $v \in B'$: $k = |L(v)| \leq |L'(v)| + 1 = d_H(v) + 1 < k$; a contradiction. Thus $B = K_k$ since G cannot contain K_{k+1}.
(b) Assume H has only one block B. Since $B = K_k$ and $\Delta(G) = k$, there exists a vertex in W which is adjacent to all vertices of B because the lists $L'(v)$ for all $v \in V(B) = H$ have the same cardinality $k-1$ (see Claim 1(a)). Thus $G = K_{k+1}$; a contradiction.
(c) Let v_B be the cut vertex of H from B which exists because of (b). Using Claim 1 we obtain $d_H(v_B) = |L(v_B)| = k$ and $d_H(v) = |L(v)| = k-1$ for all $v \in B'$. Consequently, there has to be a vertex $w_B \in W$ adjacent to all $k-1$ vertices of B'. Note that a second neighbor of vertices of B' in W would violate $d(W) \geq 4$.
(d) Because of (c) w_B can have at most one neighbor outside B' since $d_G(w_B) \leq k$. If w_B has no neighbor outside B', the cut vertex v_B of H is a cut vertex of G too, contradicting the 2-connectedness of G. So w_B has exactly one neighbor, say y_B, outside B'. Note that $y_B \notin W$ since otherwise W is not independent. Assume y_B belongs to a second leaf block $B_1 \neq B$. If $w_B = w_{B_1}$, then $d_G(w_B) \geq 2(k-1) = 2k-2 > k$; a contradiction. Otherwise we have $w_B \neq w_{B_1}$ and both are adjacent to some vertex in B_1' contradicting Claim 2(c).

Denote the set of all blocks of H which are not leaf blocks by $\mathcal{B}^*$. This set cannot be empty because of Claim 2(d).

Claim 3. *Let B be a leaf block of H and y_B be the neighbor of the corresponding*

w_B outside B' (see Claim 2). Furthermore, let B^* be an element of $\mathcal{B}^*$ containing y_B. If y_B belongs to more than one block in $\mathcal{B}^*$, then let B^* be such a block with largest order.

(a) *If $B^* = K_s$ with $s \geq 3$, then B^* contains at least three cut vertices of H.*

(b) *If $B^* = C_{2s+1}$ $(s \geq 2)$, then all vertices of B^* are cut vertices of H. At most s of the vertices of C_{2s+1} have neighbors in W.*

(c) *If $B^* = K_2$, then y_b is a cut vertex in H.*

Proof.

(a) First, note that B^* cannot contain another cut vertex beside y_B which is adjacent to a vertex of W since $d(W) \geq 4$. Thus for all $v \neq y_B \in V(B^*)$ we have $k = |L(v)| = d_H(v)$. Since $B^* \neq K_{k+1}$ all of these vertices have neighbors outside B^* in H, which means they are cut vertices of H. If $s \geq 4$ we are done. If $s = 3$ and y_B is not a cut vertex, then $d_G(y_B) = 2+1 = 3 < 4 \leq k$; a contradiction. Thus for $s = 3$ the vertex y_B is also a cut vertex.

(b) The degree of the vertices in the cycle is 2, so every vertex v must have neighbors outside B^* in H because of $d_H(v) \geq d_G(v) - 1 \geq 3$. Thus all vertices of the cycle are cut vertices in H.

Clearly, at most s of the vertices of the cycle may have neighbors in W since otherwise $d(W) \leq 3$.

(c) If the biggest block containing y_B is K_2, then y_B is incident to $k-1$ blocks K_2 of H because of $k - 1 = |L'(y_B)| = d_H(y_B)$ by Claim 1. Thus y_B is a cut vertex in H.

We would like to show now that if T is the block-cutpoint graph of H and T has ℓ leaves, then because of the above claims there is a certain set of inner vertices of the tree with "high" degree. Let $L(T)$ denote the set of leaves of T. Then we estimate the number of leaves using the well-known equality

$$|L(T)| = 2 + \sum_{v \in V(T) \setminus L(T)} (d_T(v) - 2) \tag{2}$$

for trees with $|V(T)| \geq 2$. Finally we show that T contains more than ℓ leaves, giving the contradiction we are looking for.

We know (see Claim 2) that every leaf block $B \in \mathcal{B}_\ell$ corresponds with a vertex y_B not belonging to a leaf block. Let B^* be defined as in Claim 3. Denote

- $Y_1 := \{y_B \mid B \in \mathcal{B}_\ell \text{ and } B^* = K_s, s \geq 3\}$,
  $V_1 := \{v \in V(T) \mid v \text{ represents } B^* \text{ for a } y_B \in Y_1\}$;
- $Y_2 := \{y_B \mid B \in \mathcal{B}_\ell \text{ and } B^* = C_{2s+1}, s \geq 2\}$,
  $V_2 := \{v \in V(T) \mid v \text{ represents } B^* \text{ for at least one } y_B \in Y_2\}$;
- $Y_3 := \{y_B \mid B \in \mathcal{B}_\ell \text{ and } B^* = K_2\}$,
  $V_3 := \{v \in V(T) \mid v \text{ represents a cut vertex } y_B \text{ with } y_B \in Y_3\}$.

From Claim 3 we know

- $d_T(v) \geq 3$ for $v \in V_1$,
- $d_T(v) = 2s + 1$ for $v \in V_2$, and
- $d_T(v) = k - 1 \geq 3$ for $v \in V_3$.

Note that $|Y_1| = |V_1|$ and $|Y_3| = |V_3|$ because of $d(W) \geq 4$. Because of equality (2) we are looking for an estimation of $d_T(v) - 2$ for $v \in V_2$. We obtain $d_T(v) - 2 = 2s - 1 > s$, where s is the maximum number of vertices of Y_2 belonging to $B^* = C_{2s+1}$ because of $d(W) \geq 4$. Thus it follows that $\sum_{v \in V_2}(d_T(v) - 2) > |Y_2|$.

Furthermore, because of the definition we have $L(T) \cap (V_1 \cup V_2 \cup V_3) = \emptyset$. Thus $\ell := |L(T)| = |\mathcal{B}_\ell|$ and because of Claim 2 we have $|Y_1 \cup Y_2 \cup Y_3| = |Y_1| + |Y_2| + |Y_3| = \ell$.

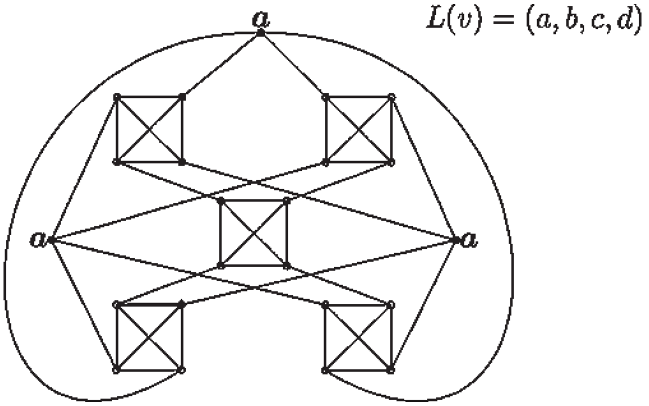


FIG. 1. *Nonextendable precoloring for $k = \Delta(G) = 4$ and $d(W) = 3$.*

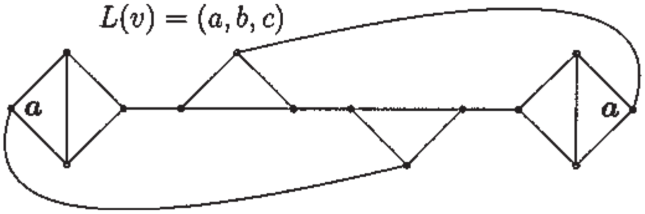


FIG. 2. *Nonextendable precoloring for $\Delta(G) = 3$ and $d(W) = 5$.*

With equality (2) we obtain

$$\ell \geq 2 + \sum_{v \in V_1 \cup V_2 \cup V_3} (d_T(v) - 2) \geq 2 + |V_1| + |V_3| + \sum_{v \in V_2} (d_T(v) - 2) > |Y_1| + |Y_2| + |Y_3| = \ell.$$

Obviously the last line gives a contradiction, the counterexample G does not exist, and the statement of Theorem 2 is proved. □

To see that the distance constraint of the theorem is sharp look for the following example.

The 4-connected example in Figure 1 shows that for $d(W) = 3$ the given precoloring is not extendable. It is obvious that Claim 2(c) is not true in this case. Note that there are analogous examples for arbitrary $k = \Delta(G) \geq 5$. Construct a Gallai tree of maximum degree k consisting of $k+1$ copies of K_k and k copies of K_2, where k of the copies of K_k are leaf blocks (see Figure 1). Then add $k - 1$ vertices $\{w_1, \ldots, w_{k-1}\}$ belonging to W, where every w_i is adjacent to a private neighbor in each leaf block of the Gallai tree such that the constructed graph is k-regular (see Figure 1). If the vertices of W are precolored by the same color a and all other vertices have identical lists containing a, then the precoloring is not extendable.

In fact there are smaller 2-connected examples where Claim 2(b) does not hold: let $k = 4$, H be the K_4 with $V(H) = \{v_1, v_2, v_3, v_4\}$, and $W = \{w_1, w_2\}$, where w_1 is adjacent to v_1 and v_2 and w_2 is adjacent to v_3 and v_4. Furthermore, assume $L(v_i) = (a, b, c, d)$ for $i = 1, \ldots, 4$ and w_1 and w_2 are precolored by d. Thus $d(W) = 3$ and the precoloring is not extendable. Note that these examples are examples for ordinary colorings. Thus the bound $d(W) \geq 4$ given in Theorem 2 is sharp even in this case.

Moreover, for $k = \Delta(G) = 3$ there are examples that $d(W) = 5$ does not guarantee such an extension (see Figure 2).

However, an extension of every precoloring is possible if G is a 2-connected graph, $\Delta(G) = 3$, G is not K_4, and $d(W) \geq 6$ (see [15]). Thus the above example shows that the bound is sharp and the problem is completely solved.

REFERENCES

[1] M. O. Albertson, *You can't paint yourself into a corner*, J. Combin. Theory Ser. B, 73 (1998), pp. 189–194.

[2] M. O. Albertson and J. P. Hutchinson, *Extending colorings of locally planar graphs*, J. Graph Theory, 36 (2001), pp. 105–116.

[3] M. O. Albertson and J. P. Hutchinson, *Graph color extensions: When Hadwiger's conjecture and embeddings help*, Electron. J. Combin., 9 (2002), research paper R37.

[4] M. O. Albertson and J. P. Hutchinson, *Extending precolorings of subgraphs of locally planar graphs*, European J. Combin., 25 (2004), pp. 863–871.

[5] M. O. Albertson, A. V. Kostochka, and D. B. West, *Precoloring extension of Brooks' theorem*, SIAM J. Discrete Math., 18 (2004), pp. 542–553.

[6] M. O. Albertson and E. H. Moore, *Extending graph colorings*, J. Combin. Theory Ser. B, 77 (1999), pp. 83–95.

[7] M. O. Albertson and E. H. Moore, *Extending graph colorings using no extra colors*, Discrete Math., 234 (2001), pp. 125–132.

[8] M. Axenovich, *A note on graph coloring extensions and list-colorings*, Electron. J. Combin., 10 (2003), note 1.

[9] O. V. Borodin, *Criterion of chromaticity of a degree description*, in Abstracts of IV All-Union Conf. on Theoretical Cybernetics (Novosibirsk), 1977, pp. 127–128 (in Russian).

[10] R. L. Brooks, *On colouring the nodes of a network*, Proc. Cambridge Philos. Soc., 37 (1941), pp. 194–197.

[11] P. Erdős, A. L. Rubin, and H. Taylor, *Choosability in graphs*, in Proceedings of the West Coast Conference on Combinatorics, Graph Theory and Computing (Humboldt State Univ., Arcata, Calif. 1979), Congress. Numer. XXVI, Utilitas Math., Winnipeg, Manitoba, 1980, pp. 125–157.

[12] J. P. Hutchinson and E. H. Moore, *Distance Constraints in Graph Color Extensions*, manuscript, 2005.

[13] A. V. Kostochka, M. Stiebitz, and B. Wirth, *The colour theorems of Brooks and Gallai extended*, Discrete Math., 162 (1996), pp. 299–303.

[14] V. G. Vizing, *Coloring the vertices of a graph in prescribed colors*, Diskret. Analiz. No. 29 Metody Diskret. Anal. v Teorii Kodov i Shem, (1976), pp. 3–10, 101 (in Russian).

[15] M. Voigt, *Precoloring Extension for 2-connected Graphs with $\Delta(G) = 3$*, in preparation.

SIAM J. Discrete Math.
Vol. 21, No. 1, pp. 264–272

GRAPHS HAVING SMALL NUMBER OF SIZES ON INDUCED k-SUBGRAPHS*

MARIA AXENOVICH[†] AND JÓZSEF BALOGH[‡]

Abstract. Let ℓ be any positive integer, let n be a sufficiently large number, and let G be a graph on n vertices. Define, for any k, $\nu_k(G) = |\{|E(H)| : H$ is an induced subgraph of G on k vertices$\}|$. We show that if there exists a k, $2\ell \leq k \leq n-2\ell$, such that $\nu_k(G) \leq \ell$, then G has a complete or an empty subgraph on at least $n-\ell+1$ vertices and a homogeneous set of size at least $n-2\ell+2$. These results are sharp.

Key words. induced subgraphs, reconstruction, homogeneous sets

AMS subject classifications. 05C35, 05C60, 05C69

DOI. 10.1137/05064357X

1. Introduction.

As is customary in graph theory, the *order* of a graph is the number of its vertices, and the *size* of a graph is the number of its edges. Following standard notation, K_n denotes the complete graph and E_n the empty (or edgeless) graph of order n. For a k integer, $1 \leq k \leq n$, a *k-subgraph* of G is an induced subgraph of order k. A *trivial* set in G is a subset of vertices of G inducing either an empty or a complete graph. Let $t(G)$ denote the number vertices of a largest trivial set of G. We define a relation $\sim$ on $V(G)$ such that $u \sim v$ iff $N(u) - \{v\} = N(v) - \{u\}$. It is easy to check that $\sim$ is an equivalence relation. The equivalence classes are called the *homogeneous sets*. Let $h(G)$ denote the maximum size of a homogeneous set in G. Note that a homogeneous set is also a trivial set, and therefore $h(G) \leq t(G)$. Recently, the homogeneous classes of certain graphs were shown to be a powerful tool handling questions on hereditary graph properties [3, 4, 5].

Vertex graph reconstruction problems are concerned with the conditions on induced subgraphs necessary to determine the original graph. In particular, the graph reconstruction conjecture, see, for example, [13], states that if one knows all $(n-1)$-subgraphs of a graph G, then a graph G itself can be reconstructed. One of the examples when a graph can be "almost" reconstructed knowing some facts about its k-subgraphs is the following result of Akiyama, Exoo, and Harary [1] and Bosák [8].

Proposition 1. *Let G be a graph on n vertices. If there is a k, $2 \leq k \leq n-2$, such that all k-subgraphs of G have the same size, then G is either the complete graph K_n or the empty graph E_n.*

In this work, we investigate the following question. How much information about the structure of a graph G can we retrieve knowing the sizes of k-subgraphs of G? We know from Proposition 1 that if all k-subgraphs have the same size, then a graph can be determined almost uniquely. What if the number of sizes of the k-subgraph of G is two, or three, or ten? Here we answered this question by showing the existence of a

*Received by the editors October 26, 2005; accepted for publication (in revised form) October 26, 2006; published electronically April 10, 2007.

http://www.siam.org/journals/sidma/21-1/64357.html

[†]Department of Mathematics, Iowa State University, Ames, IA 50011 (axenovic@math.iastate.edu).

[‡]Department of Mathematics, University of Illinois at Urbana-Champaign, Urbana, IL 61801 (jobal@math.uiuc.edu). This author's research was partly supported by NSF grants DMS-0302804 and DMS-0603769.

large homogeneous subset in G, which allows us to "almost" reconstruct the structure of G, with the exception of the subgraph induced by a small number of vertices.

Let $\nu_k(G)$ denote the number of sizes of k-subgraphs of G, i.e.,

$$\nu_k(G) = |\{|E(H)| : H \text{ is a } k\text{-subgraph of } G\}|.$$

Let $i(G)$ be the total number of isomorphism classes on induced subgraphs of G, loosely speaking, the number of induced subgraphs of G.

The parameter $i(G)$ was investigated in multiple papers in attempts to find the maximum of $i(G)$ over all graphs on n vertices; see Korshunov [10, 11]. It has been shown by Alon and Bollobás [2] and Erdős and Hajnal [9] that graphs with "small" $i(G)$ have a large trivial subset of vertices; in particular, if $\varepsilon < 10^{-21}$ and $i(G) \leq \varepsilon n^2$, then G has a trivial subset of vertices of size at least $(1 - 4\varepsilon)n$.

Here we show that graphs G for which $\nu_k(G)$ is "small" exhibit a behavior similar to graphs for which $i(G)$ is "small." In particular, we show that in this case, G must have large trivial and large homogeneous subsets, where "large" is $|V(G)| - c$, for a constant $c = c(\nu_k(G))$. Of course, in order to make such a conclusion, we must require that k is not too small or too large. For example, when $\nu_2(G) = 2$, we cannot draw any conclusions about the structure of G. Our main result is the following.

Theorem 1. *Let $\ell \geq 2$ be any positive integer. Then there is an $n(\ell)$ such that the following holds. If $n = |V(G)| \geq n(\ell)$ and there exists a k, $2\ell \leq k \leq n - 2\ell$, such that $\nu_k(G) \leq \ell$, then G has a trivial vertex set of cardinality at least $n - \ell + 1$ and a homogeneous vertex set of cardinality at least $n - 2\ell + 2$. These results are sharp.*

The graph $M_{n,\ell-1}$ of order n, of size $\ell - 1$, and of maximal degree 1 (or its complement) shows that the bounds on the orders of trivial and homogeneous sets are best possible, as $t(M_{n,\ell-1}) = n - \ell + 1$, $h(M_{n,\ell-1}) = n - 2\ell + 2$, and $\nu_k(G) = \ell$ for $2\ell - 2 \leq k \leq n - 2\ell + 2$.

The graph $M_n = M_{n,\lfloor n/2 \rfloor}$ shows that the condition $2\ell \leq k \leq n - 2\ell$ is necessary, as $t(M_n) = \lceil n/2 \rceil$, $h(M_n) = 2$, and $\nu_{2\ell-1}(M_n) = \ell$.

When $\ell = 2$, Theorem 1 gives a precise structural result for large n. We prove the same result in section 3 for all n.

Theorem 2. *Let $n \geq 8$ be any integer. Let G be a graph of order n such that $\nu_k(G) = 2$ for some k, $4 \leq k \leq n - 4$. Then G is either a star, a disjoint union of an edge and $n - 2$ vertices, or the complement of one of these graphs.*

2. Proofs. For two disjoint sets A, B of vertices of a graph G, we denote by (A, B) the bipartite subgraph of G containing all edges of G with one endpoint in A and another in B. For two disjoint sets A, B, we write $A \sim B$ if (A, B) is a complete bipartite and $A \not\sim B$ if (A, B) is an empty bipartite graph. If either $A \sim B$ or $A \not\sim B$ holds, we say that the pair (A, B) is *trivial.* If X and Y are either trivial sets or trivial pairs of sets of vertices, we say that X and Y are of different types if one of them induces an empty graph (or an empty bipartite subgraph) and the other one induces a complete graph (or a complete bipartite subgraph). Recall that a homogeneous set must span a trivial graph; furthermore, if A, B are two homogeneous sets, then either $A \sim B$ or $A \not\sim B$. A *q-skewchain* is a bipartite graph with parts $A = \{a_1, \dots, a_q\}$ and B such that $N(a_i) \subsetneq N(a_{i+1})$, $i = 1, 2, \dots, q - 1$. Our main tool is the following reformulation of a result of Balogh, Bollobás, and Weinreich [3] (for different variants of this theorem, see [6] or [7]).

Theorem 3. *There is a function $g(t)$ such that for every positive integer t, the following holds. Let G be a bipartite graph with partite sets A, B, $|A| = |B| = n$,*

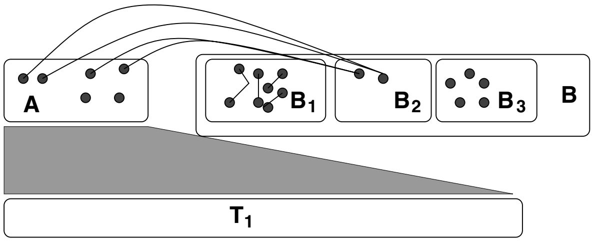


FIG. 1. *Sets T_1, A, and $B = B_1 \cup B_2 \cup B_3$.*

where $n \geq g(t)$. Suppose that the vertices in A all have distinct neighborhoods. Then G has either

(i) *an induced matching of size t or*

(ii) *a bipartite complement of an induced matching of size t or*

(iii) *an induced t-skewchain.*

Let

$$f(\ell) = 2\ell R(g(R(8\ell^3))),$$

where $R(n) = R(n,n)$ is the classical symmetric Ramsey number [14] and $g(t)$ is the function from Theorem 3. Let the parameters n, k, ℓ satisfy the conditions of Theorem 1 with $n \geq 3f(\ell)$, and let G be a graph of order n. The proof of Theorem 1 will be based on three cases—according to whether G has at least one "very large" homogeneous set, G has two "relatively large" homogeneous sets, or G has many "small" distinct homogeneous sets. We shall consider corresponding lemmas, Lemmas 1–3, and complete the proof based on them. Note that we shall need to use Theorem 3 only in the proof of Lemma 3.

LEMMA 1. *If G, of order $n > 3f(\ell)$, has a homogeneous set of order at least $n - f(\ell)$ and $\nu_k(G) \leq \ell$, then G has a homogeneous set of cardinality at least $n - 2\ell + 2$ and a trivial set of cardinality at least $n - \ell + 1$.*

Proof. Let T_1 be the (maximal) homogeneous set of G of order at least $n - f(\ell)$. Without loss of generality, T_1 is an independent set. For some sets A and B, we have that $V - T_1 = A \cup B$, where $T_1 \sim A$ and $T_1 \not\sim B$. Let $B = B_1 \cup B_2 \cup B_3$ such that in $G[B]$, $B_2 \cup B_3$ is the set of isolated vertices, and B_3 is the set of isolated vertices of G which are in B. Note that each vertex of $B_2 \subset B - B_1$ is adjacent only to some but not all vertices of A (otherwise, that vertex would have belonged to T_1), and $|B_1| \neq 1$. Denote r the number of components of the graph $G[B_1]$. Since $G[B_1]$ does not have isolated vertices,

$$r \leq \lfloor |B_1|/2 \rfloor. \tag{1}$$

For an illustration, see Figure 1.

First, assume that $A = \emptyset$. Then $B_2 = B_3 = \emptyset$, and G consists of edges induced by B_1 and T_1, the set of isolated vertices of G. Denote the components of B_1 by $C_1, \ldots, C_r$, and order the vertices of C_i as $v_{i,1}, v_{i,2}, \ldots, v_{i,q_i}$, such that $C_i - \{v_{i,j}, v_{i,j+1}, \ldots, v_{i,q_i}\}$ is connected for every j, $1 \leq j \leq q_i$, where $|C_i| = q_i$. If $k \leq n - f(\ell) \leq n - |B_1|$, then starting with a k-subset of T_1 and changing the vertices one by one by the vertices of B_1 in the order $v_{1,1}, v_{1,2}, \ldots, v_{1,q_1}, v_{2,1}, v_{2,2}, \ldots$ creates $|B_1| - r + 1$ different sizes. If $k > n - f(\ell)$, then start with a k-set containing B_1, and change the vertices of B_1 to the vertices of T_1 in the order $v_{r,q_r}, v_{r,q_r-1}, \ldots, v_{r,2}, v_{r-1,q_{r-1}}, \ldots,$

until it is possible, such that either there is no other vertex from B_1 to delete, or there is no other vertex from T_1 to add. Note that we leave out from the changing $v_{1,1}, v_{2,1}, \ldots, v_{r,1}$. This way $\min\{|B_1| - r + 1,\, n - k + 1\}$ distinct sizes are generated. Using that $|T_1| + |B_1| = n$ and (1), we obtain that $|T_1| \geq n - 2\ell + 2$, and G has a trivial set of order at least $|T_1| + r \geq n - \ell + 1$.

Now we assume that $A \neq \emptyset$. We shall use the following definition. For sets F, C, D of vertices, we say that the sets $F_1, \ldots, F_t$ are obtained from $F = F_0$ by (C, D)-*exchange in t-steps*, for $t \leq s := \min\{|C \cap F|, |D - F|\}$, if F_i is obtained from F_{i-1} by deleting a vertex from $F_{i-1} \cap C$ and adding a vertex from $D - F_{i-1}$. If the number of steps t in an exchange is equal to s, we say that the exchange is *full*.

Case 1. $n - 2f(\ell) \leq k \leq n - 2\ell$.

Let F_0 be a k-set containing all vertices of A, as many vertices from T_1 and as few vertices from $B_2 \cup B_3$ as possible. Since $n - |F_0| \geq 2\ell$, we can create $\min\{|B_3| + |B_2| + 1, 2\ell\}$ k-sets of distinct sizes on corresponding subgraphs by $(T_1, B_3 \cup B_2)$-exchange performed on F_0. Here we used that each vertex of B_2 is adjacent to some, but not all, vertices of A. Therefore, $|B_2| + |B_3| \leq \ell - 1$.

This makes it possible to choose a k-set F_1 such that $A \cup B_1 \subseteq F_1$, $(B_2 \cup B_3) \cap F_1 = \emptyset$. First, perform a full (T_1, B_2)-exchange on F_1. Let the resulting set be F_2. Next, perform an (A, T_1)-exchange on F_2 for $|A| - 1$ steps. Let the last set obtained be F_3. We have that $B_1 \cup B_2 \subseteq F_3$ and $F_3 \cap A = \{a\}$. Next, perform a full (T_1, B_3)-exchange on F_3. Finally, do an $(\{a\}, T_1)$-exchange on F_3, followed by a (B_1, T_1)-exchange producing as many distinct sizes of the resulting subgraphs as possible. As a result of these exchanges, we have obtained k-subgraphs with decreasing sizes. The total number of such distinct sizes is at least $|B_2| + (|A| - 1) + (|B_3| + 1) + (|B_1| - r + 1) \leq \nu_k(G) \leq \ell$. Using (1), we have $|A| + |B_1|/2 + |B_2| + |B_3| + 1 \leq \ell$ and $|A| + |B_1| - r \leq \ell - 1$. Applying the first inequality, we conclude that $|A| + |B_1| + |B_2| + |B_3| \leq 2\ell - 2$, giving that $|T_1| \geq n - 2\ell + 2$, and using the second inequality, we conclude that $t(G) \geq n - |A| - |B_1| + r \geq n - \ell + 1$.

Case 2. $2\ell \leq k < n - 2f(\ell)$.

First, we shall prove that

$$|B_1| + |A| \leq \ell + r - 1. \tag{2}$$

Let $F_0 \subseteq T_1$ be a set of order k. First, construct sets from F_0 by (T_1, B_1)-exchange so that as many distinct sizes as possible are obtained on the corresponding k-subgraphs, and denote the last k-set by F_1. Note that we have at least $(|B_1| - r + 1)$ distinct sizes (including F_0's size), which implies that $|B_1| - r + 1 \leq \ell$ and, using the inequality $2r \leq |B_1|$, that $|B_1| \leq 2\ell - 2 < k$, and hence $B_1 \subseteq F_1$. Observe that by (1), it also yields $r \leq \ell - 1$, which with (2) implies $|B_1| + |A| \leq 2\ell - 2 < k$. Next, construct as many sets as possible from F_1 by (T_1, A)-exchange so that the sizes of the corresponding k-subgraphs are increasing. Note that we can always do this until the number of vertices of a considered set F in A is at most the number of vertices of F in T_1. Thus, this (T_1, A)-exchange creates at least $\min\{|A|, \lfloor (k - |B_1|)/2 \rfloor\}$ distinct sizes on k-subgraphs. Therefore, the total number of distinct sizes created so far is at least $x = |B_1| - r + 1 + \min\{|A|, \lfloor (k - |B_1|)/2 \rfloor\} \geq |B_1| - r + 1 + \min\{|A|, \lfloor (2\ell - |B_1|)/2 \rfloor\}$. It cannot be that $|A| > \lfloor (2\ell - |B_1|)/2 \rfloor$; otherwise, using (1), $x \geq |B_1| - r + 1 + \lfloor (2\ell - |B_1|)/2 \rfloor = \lceil |B_1|/2 \rceil - r + 1 + \ell > \ell$, a contradiction. Thus, $|A| \leq \lfloor (2\ell - |B_1|)/2 \rfloor$ and $x \geq |B_1| - r + 1 + |A|$. Since $x \leq \ell$, (2) follows. Also, we obtain that

$$|B_1|/2 + |A| \leq \ell - 1. \tag{3}$$

Inequality (2) implies that $t(G) \geq n - |A| - |B_1| + r \geq n - \ell + 1$.

Next, we create a sequence of k-subsets with decreasing sizes on the corresponding subgraphs as follows. Since $|B_2| + |B_3| < f(\ell)$, using (2), we can find a k-set, F_0, containing $A \cup B_1$ and being disjoint from $B_2 \cup B_3$. Let $H_0 = F_0 \cap T_1$. Consider (T_1, B_2)-exchange on F_0 in s steps, where

$$s = \min\{|H_0| - 1, |B_2|\}.$$

This gives us $s + 1$ sets $F_0, F_1, \ldots, F_s$ with decreasing sizes on the corresponding k-subgraphs.

Subcase 2a. $s = |T_1 \cap F_0| - 1$.

We have that $F_0, F_1, \ldots, F_s$ induce $s + 1 = |H_0| = k - |A| - |B_1|$ k-subgraphs of distinct sizes. Let $F_{s+1} = (F_s - A) \cup H_1$, where $|H_1| = |A|$ and $H_1 \subseteq T_1 - F_s$. Note that F_{s+1} spans fewer edges than F_s. Create sets $F_{s+2}, F_{s+3}, \ldots$ from F_{s+1} by (B_1, T_1)-exchange such that as many distinct sizes on corresponding subgraphs occur as possible. The number of k-subgraphs with distinct sizes constructed so far is at least $x = (k - |A| - |B_1|) + |B_1| - r = k - |A| - r$. Using (3), (1), and $k \geq 2\ell$, we have $x \geq 2\ell - |A| - r \geq 2\ell - \ell + |B_1| - 2r + 1 > \ell$, a contradiction.

Subcase 2b. $s = |B_2|$.

We have that $F_0, F_1, \ldots, F_s$ induce $|B_2| + 1$ k-subgraphs of distinct sizes. Note that $F_s = A \cup B_1 \cup B_2 \cup H_2$ for some $H_2 \subseteq T_1$. Let us perform a full (T_1, B_3)-exchange on F_s. Let the last resulting set be F_p. Let $F_{p+1} = (F_p - A) \cup H_3$, where $H_3 \subseteq T_1 - F_p$ and $F_{p+1} = k$. Finally, we perform a full (B_1, T_1)-exchange on F_{p+1}. With some work, it can be checked that in each exchange the number of the edges spanned by the obtained k-sets is strictly decreasing. We have two cases to check.

If $|H_2| = |T_1 \cap F_s| = k - |A| - |B_1| - |B_2| \leq |B_3|$, then we have obtained all together at least $|B_2| + 1 + |H_2| + |B_1| - r + 1 = 2 + |B_2| + k - |A| - |B_1| - |B_2| + |B_1| - r = 2 + k - |A| - r > \ell$ distinct sizes (here the last inequality follows from (2), (1), and $k \geq 2\ell$), a contradiction.

If $|F_2| = |T_1 \cap F_s| = k - |A| - |B_1| - |B_2| > |B_3|$, then we have obtained at least $|B_2| + 1 + |B_3| + 1 + |B_1| - r \leq \ell$ distinct sizes on the corresponding subgraphs. Adding (2) gives us that $|A| + 2|B_1| + |B_2| + |B_3| + 2 \leq 2\ell + 2r - 1$. Thus, $|A| + |B_1| + |B_2| + |B_3| \leq 2\ell + 2r - 3 - |B_1| \leq 2\ell - 3$. Thus, $|T_1| \geq n - (|A| + |B_1| + |B_2| + |B_3|) \geq n - 2\ell + 2$. □

Lemma 2. *If G has two distinct maximal homogeneous sets of orders at least 2ℓ each, then $\nu_k(G) > \ell$.*

Proof. Let T_1, T_2 be distinct maximal homogeneous sets, $|T_i| \geq 2\ell$, $i = 1, 2$. Consider sets $A_1 \subseteq T_1$, $A_2 \subseteq T_2$ such that $|A_i| = 2\ell$, $i = 1, 2$. Let $R_i \subseteq A_1$, $S_i \subseteq A_2$, $|R_i| = |S_i| = i$, $i = 0, 1, \ldots, 2\ell$. Let $X \subseteq V(G) - (A_1 \cup A_2)$ such that $|X| = k - 2\ell$. Note that such set X exists since $|V - (A_1 \cup A_2)| = n - 4\ell \geq k - 2\ell$ for $k \leq n - 2\ell$. Let $X_1 \subseteq X$, $X_2 \subseteq X$ such that $X_i \sim A_i$, $(X - X_i) \not\sim A_i$, $i = 1, 2$. Let $|X_1| = r$, $|X_2| = s$. We distinguish three cases according to the types of A_1, A_2, and (A_1, A_2).

Case (i). $G[A_1 \cup A_2]$ is trivial.

Without loss of generality, we may assume that $G[A_1 \cup A_2]$ is empty. Since T_1 and T_2 are distinct (maximal) homogeneous sets, there is a vertex v such that, without loss of generality, $\{v\} \sim A_1$ and $\{v\} \not\sim A_2$ and $v \notin A_1 \cup A_2$. (Note that this is the only point where the maximality of T_1 and T_2 is used.) We may assume that $v \in X$. Let

$$(4) \quad F_i = G[R_i \cup S_{2\ell - i} \cup X], \quad H_i = G[R_{i+1} \cup S_{2\ell - i} \cup X - \{v\}], \quad i = 0, \ldots, 2\ell - 1.$$

Then

$$|E(F_i)| = ir + (2\ell - i)s + |E(G[X])|,$$
$$|E(H_i)| = (i+1)(r-1) + (2\ell - i)s + |E(G[X - \{v\}])|,$$

$i = 0, \ldots, 2\ell - 1$. Simplifying these expressions, we get $|E(F_i)| = i(r-s) + (2\ell s + |E(G[X])|)$ and $|E(H_i)| = i(r-1-s) + (r-1+2\ell s + |E(G[X-\{v\}])|)$. Either $r - s \neq 0$ or $r - s - 1 \neq 0$, and therefore either the sets H_i or the sets F_i for $i = 0, 1, \ldots, 2\ell - 1$ give $2\ell \geq \ell + 1$ distinct sizes, a contradiction.

Case (ii). A_1, A_2 are trivial of different types.

We may assume, without loss of generality, that A_1 induces a complete graph, A_2 induces an empty graph, and (A_1, A_2) is an empty bipartite graph. Define the sets F_i as in (4) for $0 \leq i \leq 2\ell$. Now we have that

$$|E(F_i)| = ir + (2\ell - i)s + |E(G[X])| + \binom{i}{2}, \quad i = 0, \ldots, 2\ell.$$

$|E(F_i)|$ is a quadratic function of i, and thus for $2\ell + 1$ arguments, it takes at least $\ell + 1$ different values, a contradiction.

Case (iii). A_1, A_2 are trivial of the same types, and (A_1, A_2) is trivial of a type different from the type of A_1. Define the sets F_i as in (4) for $0 \leq i \leq 2\ell$. If A_1 and A_2 are empty, then

$$|E(F_i)| = |E(G[X])| + ir + (2\ell - i)s + i(2\ell - i), \quad i = 0, \ldots, 2\ell.$$

If A_1 and A_2 are complete, then

$$|E(F_i)| = |E(G[X])| + \binom{i}{2} + \binom{2\ell - i}{2} + ir + (2\ell - i)s, \quad i = 0, \ldots, 2\ell.$$

Each of these expressions is a quadratic function of i producing at least $\ell + 1$ different values for $i = 0, 1, \ldots, 2\ell$. □

LEMMA 3. *If G has at least $f(\ell)/2\ell$ distinct maximal homogeneous sets, then $\nu_k(G) > \ell$.*

Proof. Let $T_1, T_2, \ldots, T_m$ be the distinct maximal homogeneous sets of G with $m \geq f(\ell)/2\ell = R(g(R(8\ell^2)))$. Let $v_i \in T_i$ for $i = 1, \ldots, m$. Consider a largest trivial subset Q of $\{v_1, \ldots, v_m\}$. Ramsey's theorem guarantees that $|Q| \geq g(R(8\ell^3))$. Note that the vertices of Q are from different homogeneous sets; hence they have different neighborhoods, and we can apply Theorem 3 to the bipartite subgraph G' of G with partite sets $Q, V - Q$ and edges of G with one endpoint in Q and another in $V - Q$. Then Theorem 3 implies that there are subsets $Q' \subseteq Q$ and $P' \subseteq V - Q$, $|Q'| = |P'| = R(8\ell^3)$, such that $Q' \cup P'$ induces in G' either a matching or the bipartite complement of a matching or a q-skewchain, where $q = |Q'|$. By applying Ramsey's theorem to $G[P']$, we can find a trivial subset $P'' \subset P'$ with $|P''| = 8\ell^3$. Let $B = P''$, and let A be the set of vertices of Q', of order $8\ell^3$, corresponding to P''; i.e., A is a set such that (A, B) is either a matching, the bipartite complement of a matching, or a skewchain, respectively.

Let $A = \{u_1, \ldots, u_{8\ell^3}\}$ and $B = \{v_1, \ldots, v_{8\ell^3}\}$. By taking graph complements and relabeling the vertices, we have the following possible structure induced by A and B: A and B are trivial and either

(a) (A, B) is an induced matching $\{u_i\} \sim \{v_i\}$, $i = 1, \ldots, 8\ell^3$, or

(b) (A, B) is an induced skewchain with $\{u_i\} \sim \{v_i, v_{i+1}, \dots, v_{8\ell^3}\}$, $i = 1, \dots, 8\ell^3$.

For any k satisfying $2\ell \leq k \leq 16\ell^3 - 2\ell - 1$, we shall find $\ell + 1$ k-subgraphs of $G[A \cup B]$ with distinct sizes as follows. Let $a = \lfloor k/2 \rfloor$, $b = \lceil k/2 \rceil$. Let $F_0 = \{u_1, \dots, u_a\} \cup \{v_1, \dots, v_b\}$, and let $F_i = F_{i-1} - \{v_i\} \cup \{v_{8\ell^3+1-i}\}$ for $i = 1, \dots, \ell$. It is easy to check that in case (a) the sets $F_0, \dots, F_\ell$ and in case (b) the sets $F_0, \dots, F_{\ell-1}$ with $\{u_2, \dots, u_{a+1}, v_1, \dots, v_b\}$ span distinct sizes on the corresponding k-subgraphs.

From now on, we can assume that $k \geq 16\ell^3 - 2\ell$. Let $X \subseteq V - A - B$ with $|X| = k - 16\ell^3 + 2\ell$. Let a_i be the number of neighbors of u_i in $X \cup B$, and let b_i be the number of neighbors of v_i in $X \cup A$ for $i = 1, \dots, 8\ell^3$.

By the pigeonhole principle, we have one of the cases (i) or (ii) as follows.

(i) $|\{a_1, \dots, a_{8\ell^3}\}| > 2\ell$, or $|\{b_1, \dots, b_{8\ell^3}\}| > 2\ell$.

Without loss of generality, $a_1, \dots, a_{2\ell+1}$ are all distinct integers. Let

$$F_j := X \cup A \cup B - (\{u_1, \dots, u_{2\ell+1}\} - \{u_j\}), \quad j = 1, \dots, 2\ell + 1.$$

The k-graphs induced by F_js have $2\ell + 1$ distinct sizes.

(ii) There is a subset of 2ℓ indices, without loss of generality $\{1, 2, \dots, 2\ell\}$, such that $a_1 = \cdots = a_{2\ell}$ and $b_1 = \cdots = b_{2\ell}$.

Let $M = \{v_1, u_1, v_2, u_2, \dots, v_{2\ell}, u_{2\ell}\}$. Now let

$$F_j := (X \cup A \cup B - M) \cup \{u_1, \dots, u_\ell, v_1, \dots, v_j, v_{\ell+1}, \dots, v_{2\ell-j}\},$$

$j = 1, \dots, \ell - 1$. Let $F_0 = (X \cup A \cup B - M) \cup \{u_1, \dots, u_\ell, v_{\ell+1}, \dots, v_{2\ell}\}$, and $F_\ell = (X \cup A \cup B - M) \cup \{u_1, \dots, u_\ell, v_1, \dots, v_\ell\}$. In case (a), the k-graphs induced by the sets F_j, $j = 0, \dots, \ell$, and in case (b), the sets $F_0, \dots, F_{\ell-1}$ with $\{u_{\ell+1}, \dots, u_{2\ell}, v_1, \dots, v_\ell\}$ have $\ell + 1$ distinct sizes. □

Proof of Theorem 1. Consider a graph G on n vertices with $\nu_k(G) \leq \ell$. Let $T_1, T_2, \dots, T_m$ be the maximal homogeneous sets of G such that $|T_1| \geq |T_2| \geq \cdots \geq |T_m|$.

Case 1. $|T_1| > n - f(\ell)$.

In this case, the conclusions of the theorem follow immediately from Lemma 1.

Case 2. $|T_2| \geq 2\ell + 1$.

In this case, we arrive at a contradiction using Lemma 2 with homogeneous sets T_1 and T_2.

Case 3. $|T_1| \leq n - f(\ell)$ and $|T_2| \leq 2\ell$.

The conditions $|T_2 \cup T_3 \cup \cdots \cup T_m| \geq f(\ell)$ and $|T_i| \leq 2\ell$ for $i = 2, \dots, m$ imply that $m \geq f(\ell)/2\ell$. Therefore, we arrive at a contradiction using Lemma 3. □

3. Appendix—Proof of Theorem 2.

Let G be a graph on n vertices such that each k-subgraph has size i_1 or i_2 for some integers i_1, i_2. We suppose that both values appear; otherwise, we are done by Proposition 1.

Case 1. $i_1 = 0$ or $i_1 = \binom{k}{2}$.

We may assume, by taking the complement of G if necessary, that $i_1 = 0$. We have that some of the k-subgraphs are empty and others have size $i = i_2$. Consider the largest independent set S of order at least k. Let $v \notin S$; then $N(v) \cap S = S$ or $|N(v) \cap S| = 1$; otherwise, it is easy to find two nonempty k-subgraphs with distinct sizes containing v and $k - 1$ vertices from S. We also have that $i \leq k - 1$, and, if $|N(v) \cap S| = 1$ for some v, then $i = 1$. It is obvious that if $i = 1$ and $k \geq 4$, then G must have exactly one edge. Thus, we may assume that for each $v \notin S$, $N(v) \cap S = S$. If there are two vertices $u, u' \notin S$, then consider u, u' and $k - 2$ vertices of S. These k vertices induce a subgraph with at least $2(k-2) > k-1$ edges for $k \geq 4$, a contradiction. Thus, there is exactly one vertex not in S, and G is a star.

Case 2. $i_1, i_2 \notin \{0, \binom{k}{2}\}$.

Let $i_1 < i_2$ and $i_2 - i_1 = \ell$. Since there are only two sizes of the k-subgraphs of G, there are two k-sets A and B, $|A \cap B| = k-1$, inducing subgraphs of sizes i_1 and i_2, respectively. Let $b \in B - A$, $a \in A - B$. Then $i_2 - i_1$ corresponds to the difference between the number of edges sent by b to $A \cap B$ and the number of edges sent by a to $A \cap B$. Since $|A \cap B| = k-1$, this difference is at most $k-1$, and thus $\ell \leq k-1$.

Case 2.1. There are vertices u, v, such that $|(N(u) - N(v)) \cap S| \geq 2$, for $S = V - \{u, v\}$.

Let $Q = Q(u,v) = S - (N(u) \Delta N(v))$. Assume that $|(N(u) - N(v)) \cap S| \geq |(N(v) - N(u)) \cap S|$. Let us find subsets $U', U'' \subseteq (N(u) - N(v)) \cap S$, $U' \subseteq U''$, $V' \subseteq (N(v) - N(u)) \cap S$ such that $|V'| + 1 \leq |U'| = |U''| - 1$ (note that V' might be empty). Consider largest such subsets such that $|V'| + |U'| + 1 \leq k$. Then choose $Q', Q'' \subseteq Q$ such that $|Q'| + |V'| + |U'| + 1 = k$ and $|Q''| + |V'| + |U''| + 1 = k$. Note that $|Q'| = |Q''| + 1$.

Note that these subsets can be chosen if Q' and Q'' can be chosen, which is always possible if $Q \neq \emptyset$. We have that the subgraphs induced by u, V', U', Q' and by v, V', U', Q' differ in size by $t = |U'| - |V'|$, $t > 0$, and the subgraphs induced by u, V', U'', Q'' and by v, V', U'', Q'' differ in size by $t' = |U''| - |V'| > t > 0$. Thus, we have that $i_2 - i_1 = t$ and $i_2 - i_1 = t'$, a contradiction.

If $Q = \emptyset$ and $(N(v) - N(u)) \cap S = \emptyset$, then $\nu_{k-1}(G[S]) = 1$, and thus by Proposition 1, S induces a trivial set. Thus, G is one of the following: (a) a star or its complement; (b) a star and an isolated vertex; (c) a complement of a star and an isolated vertex. Note that (b) and (c) are impossible since, in that case, $\nu_k(G) \geq 3$.

Finally, if $Q = \emptyset$ and $(N(v) - N(u)) \cap S \neq \emptyset$, then consider a set $A \subseteq S$, $|A| = k-1$, containing as many vertices of $N(u)$ as possible. Consider $B = A - \{a\} \cup \{b\}$, where $a \in N(u) \cap S$, $b \in (N(v) - A) \cap S$. Then the k-subgraphs induced by u, A and v, A differ in size by t, and k-subgraphs induced by u, B and v, B differ in size by t', where $t > t'$, a contradiction.

Case 2.2. For any two vertices $u, v \in V(G)$, if $S = V - \{u, v\}$, then $|(N(u) - N(v)) \cap S| \leq 1$.

Then, in particular, it implies that the degrees of any two vertices differ by at most 1. Thus, $V(G) = V_d \cup V_{d+1}$ such that for each $v \in V_d$, $deg(v) = d$ and for each $v \in V_{d+1}$, $deg(v) = d+1$. Note also that

$$\text{(5)} \qquad \text{if} \quad u \in V_d, \quad v \in V_{d+1}, \quad \text{then} \quad N(u) - \{v\} \subseteq N(v) - \{u\}.$$

Therefore, if $A \subseteq V_d$ induces a nontrivial connected graph in $G[V_d]$, then (A, V_{d+1}) forms a complete bipartite subgraph of G. Consider $A, B \subseteq V_d$ inducing two nontrivial components in $G[V_d]$. Let $a \in A$, $b \in B$. Then, since for any $u, v \in V$, $|N(u) - (N(v) \cup \{u, v\})| \leq 1$, we have that $|N(a) \cap V_d| \leq 1$ and $|N(b) \cap V_d| \leq 1$. Therefore, either $G[V_d]$ is connected, or each nontrivial connected component in $G[V_d]$ has maximum degree 1 and thus is an edge. Note that V_d cannot induce both edges and isolated vertices. Indeed, the degrees of vertices incident to edges in V_d are $|V_{d+1}| + 1$ (since for an edge xy in V_d, $(\{x, y\}, V_{d+1})$ induces a complete bipartite graph), and the degrees of vertices isolated in $G[V_d]$ are at most $|V_{d+1}|$, which is impossible since all vertices in V_d have the same degree d.

Subcase a. V_d induces an empty set in G.

Let $v \in V_d$, $u \in N(v)$. We have by (5) that each $w \in V_{d+1}$ is adjacent to u. Thus, $d + 1 = deg(u) \geq |V_{d+1}|$. We also have that $d = deg(v) \leq |V_{d+1}|$. Therefore,

$d = |V_{d+1}|$ or $d = |V_{d+1}| - 1$. In the first case, we have that (V_d, V_{d+1}) forms a complete bipartite subgraph; thus for any $u \in V_d$, $V_{d+1} \subseteq N(u)$, and thus for any $v \in V_{d+1}$, $V_{d+1} \subseteq N(v) \cup \{v\}$ by (5). Therefore, V_{d+1} induces a complete graph. Thus, $|V_d| = 2$ and $G = K_n - \{e\}$ for an edge e. If $|V_{d+1}| = d+1$, then each vertex from V_d is adjacent to all but one vertex in V_{d+1}, and thus V_{d+1} induces a clique. Since the degree of each vertex in $|V_{d+1}|$ is $d+1$, each vertex in V_{d+1} is adjacent to exactly one vertex in V_d. Thus, the number of nonedges between V_d and V_{d+1} is $|V_d|$, and the number of edges between V_d and V_{d+1} is $|V_{d+1}|$. Thus, $|V_d||V_{d+1}| = |V_d| + |V_{d+1}|$, which is possible only when $|V_d| = |V_{d+1}| = 2$, a contradiction to the assumption that $n \geq 8$.

Subcase b. V_d induces a matching.

In this case, we have as before that V_{d+1} induces a complete graph and (V_d, V_{d+1}) forms a complete bipartite subgraph of G. Then $d = |V_{d+1}| + 1$, $d + 1 = n - 1$. Therefore, $|V_{d+1}| = n - 3$, $|V_d| = 3$, a contradiction since then V_d cannot induce a matching.

Subcase c. V_d is connected.

Then (V_d, V_{d+1}) forms a complete bipartite graph, V_{d+1} induces a complete graph, and thus $d + 1 = n - 1$ and $d = n - 2$. Therefore, V_d must induce the complement of a matching, and the complement of G is a disjoint union of isolated edges and vertices. Thus, for $n \geq 8$, either there is exactly one edge in a complement of G, or $\nu_k(G) > 2$. $\square$

Acknowledgment. The authors are thankful for the careful work of the anonymous referees, which helped a lot to improve the presentation of the results.

REFERENCES

[1] J. Akiyama, G. Exoo, and F. Harary, *The graphs with all induced subgraphs isomorphic*, Bull. Malaysian Math. Soc. (2), 2 (1979), pp. 43–44.

[2] N. Alon and B. Bollobás, *Graphs with a small number of distinct induced subgraphs*, Discrete Math., 75 (1989), pp. 23–30.

[3] J. Balogh, B. Bollobás, and D. Weinreich, *A jump to the Bell number for hereditary graph properties*, J. Combin. Theory Ser. B, 95 (2005), pp. 29–48.

[4] J. Balogh, B. Bollobás, and R. Morris, *Hereditary properties of ordered graphs*, in Topics in Discrete Mathematics, Springer, Berlin, 2006, pp. 179–213.

[5] J. Balogh, B. Bollobás, M. Saks, and V. T. Sós, *The Unlabelled Speed of a Hereditary Graph Property*, preprint.

[6] J. Balogh and B. Bollobás, *Unavoidable traces of set systems*, Combinatorica, 25 (2005), pp. 633–643.

[7] J. Balogh, P. Keevash, and B. Sudakov, *Disjoint representability of sets and their complements*, J. Combin. Theory Ser. B, 95 (2005), pp. 12–28.

[8] J. Bosák, *Induced subgraphs*, in Finite and Infinite Sets, Vols. I, II (Eger, 1981), Colloq. Math. Soc. János Bolyai 37, North–Holland, Amsterdam, 1984, pp. 109–118.

[9] P. Erdős and A. Hajnal, *On the number of distinct induced subgraphs of a graph*, Discrete Math., 75 (1989), pp. 145–154.

[10] A. D. Korshunov, *For what k in almost every n-vertex graph do there exist all nonisomorphic k-vertex subgraphs?*, Diskretn. Anal. Issled. Oper. Ser. 1, 8 (2001), pp. 54–67 (in Russian).

[11] A. D. Korshunov, *The number of nonisomorphic subgraphs in an n-vertex graph*, Mat. Zametki, 9 (1971), pp. 263–273 (in Russian).

[12] H. Maehara, *On the number of induced subgraphs of a random graph*, Discrete Math., 64 (1987), pp. 309–312.

[13] C. St. J. A. Nash-Williams, *The reconstruction problem*, in Selected Topics in Graph Theory, Academic Press, London, 1978, pp. 205–236.

[14] F. Ramsey, *On a problem of formal logic*, Proc. London Math. Soc., 30 (1927), pp. 264–286.

[15] S. Shelah, *Erdős and Rényi conjecture*, J. Combin. Theory Ser. A, 82 (1998), pp. 179–185.

SIAM J. Discrete Math.
Vol. 21, No. 2, pp. 273–302

PACKING ODD CIRCUITS*

MICHELE CONFORTI[†] AND BERT GERARDS[‡]

Abstract. We determine the structure of a class of graphs that do not contain the complete graph on five vertices as a "signed minor." The result says that each graph in this class can be decomposed into elementary building blocks in which maximum packings by odd circuits can be found by flow or matching techniques. This allows us to actually find a largest collection of pairwise edge disjoint odd circuits in polynomial time (for general graphs this is NP-hard). Furthermore it provides an algorithm to test membership of our class of graphs.

Key words. odd circuits, packing, excluded minors, decomposition, signed graphs

AMS subject classifications. 05C22, 05C70, 05C75, 05C83, 90C27

DOI. 10.1137/S0895480198345405

1. Introduction.

The *odd circuit packing problem*, finding in a graph a largest collection of pairwise edge disjoint odd circuits, is NP-hard. In this paper we will present a class of graphs in which this problem can be solved in polynomial time. We prove that each graph in this class can be decomposed into planar graphs, graphs with a vertex meeting all odd circuits, and graphs containing at most six vertices. In such building blocks a maximum packing by odd circuits can be found by flow or matching techniques. Given a graph G in our class, our decomposition theorem allows us to combine such packings for the building blocks of G to a maximum packing by odd circuits in G. With some extra work our decomposition theorem gives an algorithm to test membership of our class.

We present everything in terms of signed graphs. The results can be stated and proved in terms of ordinary graphs without any loss of generality, but in those terms the proofs require extra maneuvering that can be avoided when speaking the language of signed graphs. A *signed graph* is a pair (G, Σ) consisting of an undirected graph G and a collection Σ of its edges. A collection F of edges in G is called *odd* in (G, Σ) if $|F \cap \Sigma|$ is odd; otherwise, F is called *even*. In particular, we speak of odd and even edges, paths, and circuits. We call (G, Σ) *Eulerian* if G is Eulerian, so if each vertex has even degree.

Theorem 1. *The odd circuit packing problem is polynomially solvable for Eulerian signed graphs with no* $\widetilde{K}_5$-, $K_{3,3}^{1,1}$-, $K_{3,3}^{1,2}$-, *or* $K_{3,3}^{2}$-*minor.*

We explain the notions used in this result. A *minor* of (G, Σ) is the result of a series of the following three operations: deletion of an edge or an isolated vertex, contraction of an even edge, and resigning. *Resigning* (*on* $U \subseteq V(G)$) means replacing Σ by the symmetric difference $\Sigma \triangle \delta_G(U)$ of Σ with the cut $\delta_G(U) := \{uv \in E(G) | u \in U, v \notin U\}$. Clearly, the collection $\Omega(G, \Sigma)$ of odd circuits in (G, Σ) is invariant under

*Received by the editors October 2, 1998; accepted for publication (in revised form) August 16, 2006; published electronically April 13, 2007. This research was carried out with financial support from Laboratoire LEIBNIZ, Université de Grenoble and the research training networks DONET (ERB TMRX-CT98-0202) and ADONET (MRTN-CT-2003-504438) of the European Union.

http://www.siam.org/journals/sidma/21-2/34540.html

[†] Dipartimento di Matematica Pura ed Applicata, Università di Padova, Via Belzoni 7, 35131 Padova, Italy (conforti@math.unipod.it).

[‡] Centrum voor Wiskunde en Informatica, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands and Faculteit Wiskunde en Informatica, Technische Universiteit Eindhoven, Postbus 513, 5600 MB, Eindhoven, The Netherlands (bert.gerards@cwi.nl).

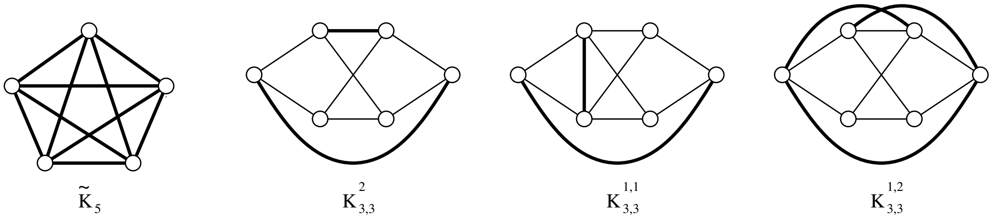


FIG. 1. *Bold edges are odd; thin edges are even.*

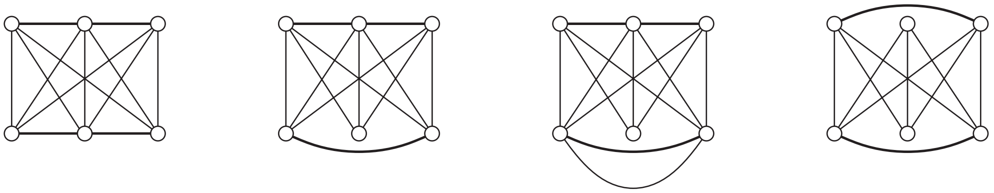


FIG. 2. *Bold edges are odd; thin edges are even.*

resigning. Two signed graphs are *isomorphic* if they are related through resigning and graph-isomorphism. We say that (G, Σ) has a (H, Θ)-*minor* or *contains* (H, Θ) if it has a minor isomorphic to (H, Θ).

The definition of the four signed graphs "excluded" in Theorem 1 can be understood from the following (see Figure 1). If G is a graph, then $\widetilde{G} := (G, E(G))$, so $\widetilde{K}_5$ consists of the complete graph on five vertices with all edges odd. $K_{3,3}^{i} := (K_{3,3}, M)$, where M is a matching of size i. Finally, $K_{3,3}^{1,1}$ and $K_{3,3}^{1,2}$ are the two extensions of $K_{3,3}^{1}$ given in Figure 1.

In addition to Theorem 1 we prove that the signed graph property described there can be recognized in polynomial time.

THEOREM 2. *There exists a polynomial time algorithm that decides whether or not a given signed graph has a $\widetilde{K}_5$-, $K_{3,3}^{1,1}$-, $K_{3,3}^{1,2}$-, or $K_{3,3}^{2}$-minor.*

As we shall see in sections 3 and 5 both Theorems 1 and 2 are a consequence of the following decomposition theorem. It is the main result of this paper.

THEOREM 3. *Let (G, Σ) be a 3-connected signed graph with no improper 3-vertex cutset and no $K_{3,3}^{2}$-minor.*

(i) *If (G, Σ) has no $K_{3,3}^{1}$-minor and no $\widetilde{K}_5$-minor, then $|V(G)| = 5$ or G is planar or (G, Σ) is isomorphic to one of the signed graphs in Figure 2 or (G, Σ) has a blockvertex.*

(ii) *If (G, Σ) has a $K_{3,3}^{1}$ minor, but no $K_{3,3}^{1,1}$ or $K_{3,3}^{1,2}$ minor, then (G, Σ) has a blockvertex.*

Here are the notions used in this result: A *blockvertex* of (G, Σ) is a vertex that is contained in every odd circuit. We call (G, Σ) 3-*connected* if any two vertices in G are connected by two internally vertex disjoint paths; this allows parallel edges. (G, Σ) has an *improper* 3-*vertex cutset* means that it contains signed graphs (G_1, Σ_1) and (G_2, Σ_2) such that $E(G_1)$ and $E(G_2)$ are nonempty and partition $E(G)$, $|V(G_1) \cap V(G_2)| = 3$ and (G_2, Σ_2) has no odd circuits and at least four edges. The proof of (i) is in section 6, and the proof of (ii) is in sections 7–11.

We obtain not only an algorithm for the odd circuit packing problem but also a min-max relation.

THEOREM 4. *Let (G, Σ) be a signed graph with no $\widetilde{K}_5$-, $K^{1,1}_{3,3}$-, $K^{1,2}_{3,3}$-, or $K^2_{3,3}$-minor. If G is Eulerian, then the maximum number of pairwise edge disjoint odd circuits in (G, Σ) is equal to the minimum number of edges needed to cover all odd circuits in (G, Σ).*

This result has been generalized extensively by Geelen and Guenin [2], who proved the min-max relation for all Eulerian signed graphs with no $\widetilde{K}_5$-minor. This was stated as a conjecture in an earlier version of the present article. Geelen and Guenin do not use decompostions, and their methods do not seem to provide a polynomial time algorithm for finding maximum odd circuits packings. However, it does follow from their result and in fact also from the earlier characterization of "weakly bipartite graphs" by Guenin [5] that by linear programming techniques one can find in polynomial time a smallest collection of edges that cover all odd circuits in a signed graph with no $\widetilde{K}_5$-minor. Note that in $\widetilde{K}_5$ itself, which is Eulerian, the min-max relation in Theorem 4 does not hold, so the Geelen–Guenin theorem is in a certain sense as strong as possible.

The min-max relation stated in Theorem 4 may fail to be true if we drop the condition that the graph is Eulerian; $\widetilde{K}_4$ is an example. Actually it follows from a general result of Seymour [10] that the min-max relation does hold for signed graphs with no $\widetilde{K}_4$-minor, even if they are not Eulerian.

Theorem 3 also has consequences for the chromatic number of the graphs involved. In combination with the 4-color theorem it can be used to prove that if $\widetilde{G}$ has none of the forbidden minors of Theorem 1, then G is 4-colorable. (It has been conjectured by one of the authors that G is 4-colorable if $\widetilde{G}$ has no $\widetilde{K}_5$-minor, see Jensen and Toft [8]. Recently Guenin [6] announced a proof of this conjecture.)

Theorem 3 can be regarded as a first step towards a constructive characterization of graphs with no $\widetilde{K}_5$-minor, a small step though; there are quite a few other infinite families of "highly connected" graphs with no $\widetilde{K}_5$-minor known that are not covered by Theorem 3 (see Gerards [4]). The exclusion of $K^2_{3,3}$, $K^{1,1}_{3,3}$, and $K^{1,2}_{3,3}$ is quite restrictive. For each $\Sigma \subseteq E(K_{3,3})$, the signed graph $(K_{3,3}, \Sigma)$ is isomorphic to exactly one of $K^0_{3,3}$, $K^1_{3,3}$, and $K^2_{3,3}$. For instance, $\widetilde{K}_{3,3}$ is isomorphic to $K^0_{3,3}$ and $K^3_{3,3}$ to $K^2_{3,3}$. So up to isomorphism $K^2_{3,3}$ is the only signed $K_{3,3}$ with a $\widetilde{K}_4$ minor. $K^{1,1}_{3,3}$ and $K^{1,2}_{3,3}$ are the smallest 3-connected signed graphs that contain both $K^1_{3,3}$ and $\widetilde{K}_4$ as minors.

2. Odd circuits in signed graphs. We mention some elementary facts on signed graphs that are good to keep in mind while reading this paper. Note that they are all known and not just for odd circuits in graphs but for general binary clutters, which are just collections of odd circuits in signed binary matroids.

A signed graph (G, Σ) is *bipartite* if $\Sigma = \delta_G(U)$ for some $U \subseteq V(G)$. So clearly, (G, Σ) is bipartite if and only if it is isomorphic to $(G, \emptyset)$. Hence, if (G, Σ) is bipartite it has no odd circuits. Actually the converse is also true. To see this, we may assume that G is connected and that we have resigned (G, Σ) such that Σ is as small as possible. That means that Σ does not contain a nonempty cut $\delta_G(U)$ (otherwise resigning on U replaces Σ by $\Sigma \setminus \delta_G(U)$, which then is smaller). Therefore the even edges in (G, Σ) form a connected spanning subgraph of G. Now, if (G, Σ) is nonbipartite there is an odd edge uv in Σ and, as u and v are connected by a path with all edges even, that edge is in an odd circuit. So a signed graph is bipartite if and only if it has no odd circuit.

A subset S of $E(G)$ is a *signature* of (G, Σ) if (G, S) has exactly the same odd circuits as (G, Σ). Clearly, S is a signature if and only if all circuits are even in

$(G, S \triangle \Sigma)$. In other words, the signatures are exactly the sets $\Sigma \triangle \delta_G(U)$ for some $U \subseteq V(G)$. Each signature meets all odd circuits. Conversely, if $F \subseteq E(G)$ meets all odd circuits it contains a signature. Indeed, let H be obtained from G by deleting all edges in F. Then $(H, \Sigma \setminus F)$ has no odd circuits and so is bipartite. Thus there exists a set $U \subseteq V(H) = V(G)$ with $\Sigma \setminus F = \delta_H(U)$. In other words $\Sigma \triangle \delta_G(U) \subseteq F$, so F contains a signature, as claimed. In other words the signatures are exactly the inclusionwise minimal edge sets that meet all odd circuits, and the smallest signatures are exactly the the sets attaining the minimum in Theorem 4.

3. Packing odd circuits—algorithm and min-max relation. We actually consider a "capacitated version" of packing odd circuits, because it is slightly more convenient to work with. If G is a graph and $w \in \mathbb{Z}_+^{E(G)}$, then a *w-packing* is a collection of subsets of $E(G)$, repetition allowed, such that each edge e is in at most $w(e)$ members of the collection. So the maximum size of a w-packing of odd circuits in (G, Σ) is equal to

$$\nu_w(G,\Sigma) := \max\left\{ \sum_{C \in \Omega(G,\Sigma)} \lambda_C \;\middle|\; \lambda \in \mathbb{Z}_+^{\Omega(G,\Sigma)} \text{ and } \sum_{C \in \Omega(G,\Sigma), C \ni e} \lambda_C \leq w(e) \text{ for each } e \in E \right\}.$$

Clearly, $\nu_w(G, \Sigma)$ is bounded from above by

$$\tau_w(G,\Sigma) := \min\{w(S) \mid S \text{ is a signature of } (G,\Sigma)\},$$

where $w(S)$ is short for $\sum_{e \in S} w(e)$.

We call a function $w \in \mathbb{Z}_+^{E(G)}$ *Eulerian* if $w(\delta_G(v))$ is even for each vertex $v \in V(G)$. Theorem 4 is equivalent with the following result:

(1) *If (G, Σ) is a signed graph with no $\widetilde{K}_5$-, $K_{3,3}^{-1,1}$-, $K_{3,3}^{-1,2}$-, or $K_{3,3}^{2}$-minor, then $\nu_w(G,\Sigma) = \tau_w(G,\Sigma)$ for each Eulerian $w \in \mathbb{Z}_+^{E(G)}$.*

Indeed, as the excluded minor condition is invariant under addition of even edges parallel to even edges and of odd edges parallel to odd edges and under deleting edges, (1) follows from Theorem 4, which in turn is the special case of (1) when w is the all-one function.

Now we show that Theorem 3 implies (1) hence also Theorem 4. We first consider the basic building blocks of our decomposition. For these there exist standard constructions, by Barahona and Seymour, to reduce the odd circuit packing problem to flow problems and odd cut packing problems.

(2) *If (G, Σ) has a blockvertex then $\nu_w(G,\Sigma) = \tau_w(G,\Sigma)$ for each $w \in \mathbb{Z}_+^{E(G)}$. Moreover then we can find a maximum w-packing of odd circuits in polynomial time.*

To see this let s be a blockvertex. As the signed graph obtained by deleting s from (G, Σ) is bipartite, we may resign such that $\Sigma \subseteq \delta_G(s)$. Now construct a new graph H by adding a new vertex t and replacing each odd edge us of G with an edge ut in H. Then there is a one-to-one correspondence between odd circuits in (G, Σ) and st-paths in H. Thus (2) follows from network flow theory.

Next we discuss how to deal with the signed graphs in Figure 2 and with signed graphs (K_5, Σ) that are not isomorphic to $\widetilde{K}_5$. In either of these cases (G, Σ) contains a *blocking pair.* This is a pair of vertices such that each odd circuit contains at least one of these two vertices. So we can then apply the following fact:

(3) *If (G, Σ) has a blocking pair, then $\nu_w(G, \Sigma) = \tau_w(G, \Sigma)$ for each Eulerian $w \in \mathbb{Z}_+^{E(G)}$. Moreover then we can find a maximum w-packing of odd circuits in polynomial time.*

To see this we use the same approach, due to Barahona, as in the blockvertex case. Let $\{s_1, s_2\}$ be a blocking pair. By resigning we may assume that each odd edge is incident with at least one of s_1 and s_2. Now construct a new graph H by adding new vertices t_1 and t_2 and by replacing each odd edge us_1 of G with $u \neq s_2$ with an edge ut_1 in H; by replacing each odd edge us_2 of G with $u \neq s_1$ with an edge ut_2 in H; and by replacing an odd edge between s_1 and s_2 (if such edge exists) with an edge $t_1 s_2$ in H. Then there is a one-to-one correspondence between the odd circuits in (G, Σ) and the $s_1 t_1$-paths and $s_2 t_2$-paths in H. Thus we translate the maximum w-packing of odd circuits problem into the integer 2-commodity flow problem. Note that the latter problem does not really change if we would add an edge $s_1 t_1$ with $w(s_1 t_1) = 1$ or an edge $s_2 t_2$ with $w(s_2 t_2) = 1$ or both. Hence we may assume that w is Eulerian on H. Thus (3) follows from the integer 2-commodity flow theorem of Rothschild and Whinston [9].

(4) *If G is planar, then $\nu_w(G, \Sigma) = \tau_w(G, \Sigma)$ for each Eulerian $w \in \mathbb{Z}_+^{E(G)}$. Moreover then we can find a maximum w-packing of odd circuits in polynomial time.*

We use a construction by Seymour [12], and for ease of exposition we restrict ourselves to the case that w is the all-one function, so G is Eulerian. Hence the planar dual G^* of some embedding of G in the plane is bipartite in the ordinary graph sense. Let Σ^* be the edges of G^* corresponding to the edges in Σ. Let T denote the set of vertices of G^* that meet an even number of edges in Σ^*. We call a collection F of odd edges in G^* a T-join if and only if every vertex in T meets an odd number of edges in F and every vertex outside T meets an even number of edges in F. A cut $\delta_{G^*}(U)$ in G^* is a *T-cut* if $|T \cap U|$ is odd. By the relation between circuits in a plane graph and cuts in its plane dual, we see that there is a one-to-one correspondence between T-joins in G^* and signatures in (G, Σ) and between inclusionwise minimal T-cuts in G^* and odd circuits in (G, Σ). Hence the min-max relation in (4) follows from a min-max relation by Seymour [12] that says that in any ordinary (not signed) bipartite graph the minimum size of a T-join is equal to the maximum size of a collection of pairwise disjoint T-cuts. See Barahona [1] for a polynomial algorithm for finding such a maximum collection of disjoint T-cuts; it also allows general Eulerian functions $w \in \mathbb{Z}^{E(G)}$, other than the all-one function. Thus (4) follows.

The following two results, Lemmas 5 and 6, say that all signed graphs that do not satisfy the min-max relation in (1) and are minor-minimal in this respect are 3-connected and have no improper 3-vertex cutsets.

LEMMA 5. *If (G, Σ) does not satisfy $\nu_w(G, \Sigma) = \tau_w(G, \Sigma)$ for each Eulerian $w \in \mathbb{Z}_+^{E(G)}$ and is minor-minimal in this respect, then (G, Σ) is* 3*-connected and has no parallel edges.*

Proof. Let (G, Σ) be a counterexample. We clearly may assume G to be 2-connected, so there exist two vertices u_1 and u_2 in G and two connected graphs G_1

and G_2 with $V(G_1) \cap V(G_2) = \{u_1, u_2\}$ such that $E(G_1)$ and $E(G_2)$ both have at least two elements and partition $E(G)$. For $i = 1, 2$, we define $\Sigma_i := \Sigma \cap E(G_i)$. Let $w \in \mathbb{Z}_+^{E(G)}$ be Eulerian with $\tau_w(G, \Sigma) > \nu_w(G, \Sigma)$.

For each signed graph (H, Θ) containing u_1 and u_2 and for $i = 0, 1$, we define

$$\tau_w(H, \Theta)_i := \min\{w(\Theta \triangle \delta_H(U)) \mid |U \cap \{u_1, u_2\}| = i\}. \tag{5}$$

Then,

$$\tau_w(H, \Theta) = \min\{\tau_w(H, \Theta)_0, \tau_w(H, \Theta)_1\}, \tag{6}$$

and

$$\tau_w(G, \Sigma)_i = \tau_w(G_1, \Sigma_1)_i + \tau_w(G_2, \Sigma_2)_i \quad \text{for } i = 0, 1. \tag{7}$$

Also note that if $U \subseteq V(H)$ with $u_1 \in U$ and $u_2 \notin U$, then

$$\tau_w(H, \Theta)_i = \tau_w(H, \Theta \triangle \delta_H(U))_{1-i} \quad \text{for } i = 0, 1. \tag{8}$$

So by resigning (G, Σ) if necessary we may assume that

$$\tau_w(G_1, \Sigma_1)_1 \geq \tau_w(G_1, \Sigma_1)_0. \tag{9}$$

Let $\omega := \tau_w(G_1, \Sigma_1)_1 - \tau_w(G_1, \Sigma_1)_0$. If $\omega = 0$, let $\widehat{G}_2 := G_2$; if $\omega > 0$, let $\widehat{G}_2$ be obtained from G_2 by adding a new even edge e_2 between u_1 and u_2 with weight $w(e_2) := \omega$.

$$\tau_w(\widehat{G}_2, \Sigma_2) = \tau_w(G, \Sigma) - \tau_w(G_1, \Sigma_1)_0. \tag{10}$$

To see this, note that it follows from (7) that $\tau_w(\widehat{G}_2, \Sigma_2)_0 = \tau_w(G_2, \Sigma_2)_0 = \tau_w(G, \Sigma)_0 - \tau_w(G_1, \Sigma_1)_0$ and $\tau_w(\widehat{G}_2, \Sigma_2)_1 = \tau_w(G_2, \Sigma_2)_1 + \omega = \tau_w(G_2, \Sigma_2)_1 + \tau_w(G_1, \Sigma_1)_1 - \tau_w(G_1, \Sigma_1)_0 = \tau_w(G, \Sigma)_1 - \tau_w(G_1, \Sigma_1)_0$. By (6), this implies (10).

$$(\widehat{G}_2, \Sigma_2) \textit{ is a proper minor of } (G, \Sigma). \tag{11}$$

Suppose this is not true. Then G_1 has no even u_1u_2-path, and $\omega > 0$. We first prove that (G_1, Σ_1) is bipartite. Let C be a circuit in G_1. As G is 2-connected there exist two disjoint paths from $V(C)$ to $\{u_1, u_2\}$. As the union of these paths and C does not contain an even u_1u_2-path, C has to be even. So (G_1, Σ_1) is bipartite indeed. Hence $\Sigma_1 = \delta_{G_1}(U)$ for some $U \subseteq V(G_1)$. We may assume $u_1 \in U$. Then, as there is no even u_1u_2-path, $u_2 \notin U$. Hence as $w(\Sigma_1 \triangle \delta_{G_1}(U)) = w(\emptyset) = 0$, we have that $\tau_w(G_1, \Sigma_1)_1 = 0$. So $\omega = 0$, which is a contradiction. This proves (11).

$$w(\delta_{\widehat{G}_2}(v)) \textit{ is even for each } v \in V(\widehat{G}_2). \tag{12}$$

Indeed, as $w(\delta_G(v))$ is even for each $v \in V(G)$, (12) holds for all $v \notin \{u_1, u_2\}$. So, as there is an even number of vertices v with $w(\delta_{\widehat{G}_2}(v))$ odd, we may restrict ourselves to proving that $w(\delta_{\widehat{G}_2}(u_1))$ is even. Let $U_1 \subseteq V(G_1)$ with $U_1 \cap \{u_1, u_2\} = \{u_1\}$ such that $w(\Sigma_1 \triangle \delta_{G_1}(U_1)) = \tau_w(G_1, \Sigma_1)_1$, and let $U_0 \subseteq V(G_1)$ with $U_0 \cap \{u_1, u_2\} = \emptyset$ such that $w(\Sigma_1 \triangle \delta_{G_1}(U_0)) = \tau_w(G_1, \Sigma_1)_0$. Then we get the following ("$\equiv$" denotes equivalence modulo 2):

$$\begin{aligned}
w(\delta_{\widehat{G}_2}(u_1)) &= w(\delta_{G_2}(u_1)) + w(e_2)\\
&= w(\delta_{G_2}(u_1)) + \tau_w(G_1,\Sigma_1)_1 - \tau_w(G_1,\Sigma_1)_0\\
&= w(\delta_{G_2}(u_1)) + w(\Sigma_1 \triangle \delta_{G_1}(U_1)) - w(\Sigma_1 \triangle \delta_{G_1}(U_0))\\
&\equiv w(\delta_{G_2}(u_1)) + w(\Sigma_1) + w(\delta_{G_1}(U_1)) + w(\Sigma_1) + w(\delta_{G_1}(U_0))\\
&\equiv w(\delta_{G_2}(u_1)) + w(\delta_{G_1}(U_1) \triangle \delta_{G_1}(U_0))\\
&\equiv w(\delta_{G_2}(u_1)) + w(\delta_{G_1}(U_1 \triangle U_0))\\
&= w(\delta_G(U_1 \triangle U_0)) \equiv 0.
\end{aligned}$$

So (12) follows.

By (11) and (12) there exists a w-packing $\mathcal{C}^2 = \{C_1^2, \ldots, C^2_{\tau_w(\widehat{G}_2,\Sigma_2)}\}$ of odd circuits in $(\widehat{G}_2, \Sigma_2)$. For each $e \in E(\widehat{G}_2)$ let $c(e)$ denote the number of members of $\mathcal{C}^2$ that use edge e; abbreviate $\gamma := w(e_2)$. Assume that $C_1^2, \ldots, C_\gamma^2$ are the members of $\mathcal{C}^2$ containing e_2. The function $w - c$ is Eulerian on $\widehat{G}_2$, and as $\mathcal{C}^2$ is a maximum w-packing of odd circuits, the set of edges $e \in E(\widehat{G}_2)$ with $w(e) - c(e) > 0$ contains no odd circuits. Hence, by Euler's theorem on Euler tours and since $(w-c)(e_2) = \omega - \gamma$, there exists a $(w-c)$-packing $\mathcal{D} = \{D_1^2, \ldots, D^2_{\omega-\gamma}\}$ of even circuits in $(\widehat{G}_2, \Sigma_2)$ that all contain e_2.

(13) *We may assume that* $\gamma = 0$ *or* $\omega - \gamma = 0$.

If both are positive, then C_1^2 contains e_2 and D_1^2 exists; by definition D_1^2 also contains e_2. The set $C_1^2 \triangle D_1^2$ contains an odd circuit, C say. As $C_1^2 \triangle D_1^2$ does not contain e_2, neither does C. Replacing in $\mathcal{C}^2$ the odd circuit C_1^2 with C yields a w-packing of the same size as $\mathcal{C}^2$ that has only $c(e_2) - 1$ members using e_2. This proves (13).

If $\omega = 0$, let $\widehat{G}_1 := G_1$. If $\gamma = \omega > 0$, let $\widehat{G}_1$ be obtained from G_1 by adding an odd edge e_1 between u_1 and u_2 with $w(e_1) := \omega$. If $\omega > 0 = \gamma$, let $\widehat{G}_1$ be obtained from G_1 by adding an even edge f_1 between u_1 and u_2 with $w(f_1) := \omega$. If e_1 is included in $\widehat{G}_1$, we define $\widehat{\Sigma}_1 := \Sigma_1 \cup \{e_1\}$; otherwise, $\widehat{\Sigma}_1 := \Sigma_1$.

(14) $(\widehat{G}_1, \widehat{\Sigma}_1)$ *is a proper minor of* (G, Σ).

If e_1 exists in $(\widehat{G}_1, \widehat{\Sigma}_1)$, then $\gamma > 0$, so there exists an odd circuit using e_2 in (G_2, Σ_2), for instance, C_1^2. So in that case there is an odd u_1u_2-path in (G_2, Σ_2). If f_1 exists in $(\widehat{G}_1, \widehat{\Sigma}_1)$, then $\omega - \gamma > 0$, so there exists an even circuit using e_2 in (G_2, Σ_2), for instance, D_1^2. Hence, in that case there is an even u_1u_2-path in (G_2, Σ_2). This proves (14).

(15) $w(\delta_{\widehat{G}_1}(v))$ *is even for each* $v \in V(\widehat{G}_1)$.

This is obvious as the weight of the added edge is $w(e_2)$ and as w is Eulerian on G and on $\widehat{G}_2$.

By (14) and (15) there exists a w-packing $\mathcal{C}^1 = \{C_1^1, \ldots, C^1_{\tau_w(\widehat{G}_1,\widehat{\Sigma}_1)}\}$ of odd circuits in $(\widehat{G}_1, \widehat{\Sigma}_1)$.

(16) $\tau_w(\widehat{G}_1, \widehat{\Sigma}_1) = \tau_w(G_1, \Sigma_1)_0 + \gamma = \tau_w(\widehat{G}_1, \widehat{\Sigma}_1)_0$.

Note that $\gamma = \omega = w(e_1)$ if e_1 exists and $\gamma = 0$ if e_1 does not exist. Hence, $\tau_w(\widehat{G}_1, \widehat{\Sigma}_1)_0 = \tau_w(G_1, \Sigma_1)_0 + \gamma$. Similarly, $\omega - \gamma = \omega = w(f_1)$ if f_1 exists, and

$\omega - \gamma = 0$ if f_1 does not exist. Hence, by the definition of ω we get $\tau_w(\widehat{G}_1, \widehat{\Sigma}_1)_1 = \tau_w(G_1, \Sigma_1)_1 + \omega - \gamma = \tau_w(G_1, \Sigma_1)_0 + 2\omega - \gamma \geq \tau_w(G_1, \Sigma_1)_0 + \gamma$. By (6), this proves (16).

As $\tau_w(\widehat{G}_1, \widehat{\Sigma}_1) = \tau_w(\widehat{G}_1, \widehat{\Sigma}_1)_0$ there exists a minimum weight signature containing e_1 as soon as e_1 exists, that is as soon as $\gamma > 0$. Hence, by "complementary slackness" there are exactly γ odd circuits in $\mathcal{C}^1$ that contain e_1. Assume that $C^1_1, \ldots, C^1_\gamma$ contain e_1 and that $C^1_{\gamma+1}, \ldots, C^1_{\gamma+k}$ are the members of $\mathcal{C}^1$ containing f_1. Note that $k \leq \omega - \gamma$.

Now let $\mathcal{C}$ be the collection of the following odd circuits:

$$\begin{aligned}
&(C^1_i \setminus \{e_1\}) \cup (C^2_i \setminus \{e_2\}) \text{ for } i = 1, \ldots, \gamma,\\
&(C^1_i \setminus \{f_1\}) \cup (D^2_{i-\gamma} \setminus \{e_2\}) \text{ for } i = \gamma+1, \ldots, \gamma+k,\\
&C^1_i \text{ for } i = \gamma+k+1, \ldots, \tau_w(\widehat{G}_1, \widehat{\Sigma}_1),\\
&C^2_i \text{ for } i = \gamma+1, \ldots, \tau_w(\widehat{G}_2, \Sigma_2).
\end{aligned}$$

Clearly, $\mathcal{C}$ is a w-packing in G. Its size is $\tau_w(\widehat{G}_1, \widehat{\Sigma}_1) + \tau_w(\widehat{G}_2, \Sigma_2) - \gamma$. By (10) and (16) this is equal to $\tau_w(G, \Sigma)$. Hence, $\nu_w(G, \Sigma) \geq \tau_w(G, \Sigma)$, contrary to our assumption. This proves the lemma. □

LEMMA 6. *If (G, Σ) does not satisfy $\nu_w(G, \Sigma) = \tau_w(G, \Sigma)$ for each Eulerian $w \in \mathbb{Z}^{E(G)}_+$ and is minor-minimal in this respect, then (G, Σ) has no improper* 3-*vertex cutset.*

Proof. Let (G, Σ) be a counterexample; by Lemma 5 it is 3-connected. Then (G, Σ) contains a signed graph (G_1, Σ_1) and a bipartite signed graph (G_2, Σ_2) such that $E(G_1)$ and $E(G_2)$ partition $E(G)$, $V(G_1) \cap V(G_2) = \{u_1, u_2, u_3\}$, and $|E(G_2)| \geq 4$. By resigning, we may assume that $\Sigma_2 = \emptyset$. Let $w \in \mathbb{Z}^{E(G)}_+$ be Eulerian with $\tau_w(G, \Sigma) > \nu_w(G, \Sigma)$.

For each signed graph (H, Θ) containing $\{u_1, u_2, u_3\}$, we define

$$\tau_w(H, \Theta)_0 := \min\{w(\Theta \triangle \delta_H(U)) \mid U \cap \{u_1, u_2, u_3\} = \emptyset\}, \tag{17}$$

and, for each $i = 1, 2, 3$,

$$\tau_w(H, \Theta)_i := \min\{w(\Theta \triangle \delta_H(U)) \mid U \cap \{u_1, u_2, u_3\} = \{u_i\}\}. \tag{18}$$

Then,

$$\tau_w(H, \Theta) = \min\{\tau_w(H, \Theta)_0, \tau_w(H, \Theta)_1, \tau_w(H, \Theta)_2, \tau_w(H, \Theta)_3\}. \tag{19}$$

Moreover, we define

$$\begin{aligned}
\omega_1 &:= \tfrac{1}{2}[\tau_w(G_2, \emptyset)_2 + \tau_w(G_2, \emptyset)_3 - \tau_w(G_2, \emptyset)_1],\\
\omega_2 &:= \tfrac{1}{2}[\tau_w(G_2, \emptyset)_1 + \tau_w(G_2, \emptyset)_3 - \tau_w(G_2, \emptyset)_2],\\
\omega_3 &:= \tfrac{1}{2}[\tau_w(G_2, \emptyset)_1 + \tau_w(G_2, \emptyset)_2 - \tau_w(G_2, \emptyset)_3].
\end{aligned} \tag{20}$$

Then,

(21) *ω_1, ω_2, and ω_3 are nonnegative.*

To prove that, for ω_1, choose for $i = 2, 3$ a set $U_i \subseteq V(G_2)$ with $U_i \cap \{u_1, u_2, u_3\} = \{u_i\}$ and $w(\delta_{G_2}(U_i)) = \tau_w(G_2, \emptyset)_i$. Then, as $(V(G_2) \setminus (U_2 \cup U_3)) \cap \{u_1, u_2, u_3\} = \{u_1\}$, we get that $\tau_w(G_2, \emptyset)_1 \leq w(\delta_{G_2}(V(G_2) \setminus (U_2 \cup U_3))) = w(\delta_{G_2}(U_2 \cup U_3)) \leq w(\delta_{G_2}(U_2)) + w(\delta_{G_2}(U_3)) = \tau_w(G_2, \emptyset)_2 + \tau_w(G_2, \emptyset)_3$. So indeed, $\omega_1 \geq 0$ and (21) follows.

Moreover,

$$\text{(22)} \qquad \omega_1,\ \omega_2,\ \textit{and}\ \omega_3\ \textit{are integers.}$$

To see that note that the fact that $w(\delta_{G_2}(v))$ is even for each $v \in V(G_2) \setminus \{u_1, u_2, u_3\}$ has the following two consequences: $w(\delta_{G_2}(u_1)) + w(\delta_{G_2}(u_2)) + w(\delta_{G_2}(u_3))$ is even and, for $i = 1, 2, 3$, $w(\delta_{G_2}(U_i)) - w(\delta_{G_2}(u_i))$ is even if $U_i \cap \{u_1, u_2, u_3\} = \{u_i\}$. Hence, by the definition of $\tau_w(G_2, \emptyset)_i$, the number $\tau_w(G_2, \emptyset)_1 + \tau_w(G_2, \emptyset)_2 + \tau_w(G_2, \emptyset)_3$ is even. So (22) follows.

We define both $\widehat{G}_1$ and $\widehat{G}_2$ by adding to G_1 and to G_2 the edges $e_1 := u_2u_3, e_2 := u_1u_3$, and $e_3 := u_1u_2$. Moreover, we define $w(e_i) = \omega_i$ for $i = 1, 2, 3$. Similar calculations as in the proof of Lemma 5 show that

$$\text{(23)} \qquad w(\delta_{\widehat{G}_j}(v))\ \textit{is even for each}\ v \in V(\widehat{G}_j)\ \textit{and}\ j = 1, 2.$$

Next we define $\widehat{\Sigma}_2 := \{e_1, e_2, e_3\}$. Straightforward calculations show that

$$\text{(24)} \quad \tau_w(\widehat{G}_1, \Sigma_1)_i = \tau_w(G, \Sigma)_i\ \textit{and}\ \tau_w(\widehat{G}_2, \widehat{\Sigma}_2)_i = \tau_w(G_2, \Sigma_2)_i + \omega_i = \omega_1 + \omega_2 + \omega_3$$

for each $i = 0, 1, 2, 3$ and thus that

$$\text{(25)} \qquad \tau_w(\widehat{G}_1, \Sigma_1) = \tau_w(G, \Sigma)\ \textit{and}\ \tau_w(\widehat{G}_2, \widehat{\Sigma}_2) = \omega_1 + \omega_2 + \omega_3.$$

From the facts that $|E(G_2)| \geq 4$ and that G is 3-connected, it easily follows that $(\widehat{G}_1, \Sigma_1)$ is a proper minor of (G, Σ). Hence, $\nu_w(\widehat{G}_1, \Sigma_1) = \tau_w(\widehat{G}_1, \Sigma_1)$. So by (25), there exists a w-packing $\mathcal{C}^1$ in $(\widehat{G}_1, \Sigma_1)$ consisting of $\tau_w(G, \Sigma)$ odd circuits.

As $\{u_1, u_2\}$ is a blocking pair of $(\widehat{G}_2, \widehat{\Sigma}_2)$, it follows from (3) and (23) that $\nu_w(\widehat{G}_2, \widehat{\Sigma}_2) = \tau_w(\widehat{G}_2, \widehat{\Sigma}_2)$. Thus by (25) there exists a w-packing $\mathcal{C}^2$ in $(\widehat{G}_2, \widehat{\Sigma}_2)$ consisting of $\omega_1 + \omega_2 + \omega_3$ odd circuits.

As $\{e_1, e_2, e_3\}$ is a minimum weight signature of $(\widehat{G}_2, \widehat{\Sigma}_2)$, there are by complementary slackness for each i exactly ω_i members of $\mathcal{C}^2$ that intersect $\{e_1, e_2, e_3\}$ in exactly e_i. So there exists a w-packing $\mathcal{P}^1 \cup \mathcal{P}^2 \cup \mathcal{P}^3$ in (G_2, Σ_2) such that each $\mathcal{P}^i$ is a collection of ω_i even paths connecting the ends of e_i. Using the paths in $\mathcal{P}^i$ to replace occurrences of e_i in the members of $\mathcal{C}^1$, we can turn $\mathcal{C}^1$ into a w-packing consisting of $\tau_w(G, \Sigma)$ odd circuits in (G, Σ), contradicting our assumption that $\tau_w(G, \Sigma) > \nu_w(G, \Sigma)$. This proves the lemma. □

Proof of Theorem 4 *(from Theorem* 3*).* We prove (1), which implies Theorem 4. From Lemmas 5 and 6 and from (2) and (4), we see that we may assume that $|V(G)| = 5$ or that (G, Σ) is one of the signed graphs in Figure 2. In the latter case (G, Σ) has a blocking pair; thus, (3) applies. So we may assume $|V(G)| = 5$. By Lemma 5 we may assume that G has no parallel edges. This means that G is isomorphic to a subgraph of K_5. As (G, Σ) is not isomorphic to $\widetilde{K}_5$, (G, Σ) has a blocking pair. So again (3) applies. This proves Theorem 4. □

Proof of Theorem 1 *(from Theorem* 3*).* Clearly, if (G, Σ) has a blockvertex or a blocking pair or if G is planar, we can find a maximum w-packing of odd circuits by (2), (3), and (4). So it remains to explain how we can algorithmically deal with 2-separations and improper 3-separations.

First consider an improper 3-separation $(G_1, \Sigma_1), (G_2, \Sigma_2)$ of (G, Σ) as in the proof of Lemma 6. We follow that proof. So we assume that $\Sigma_2 = \emptyset$. Finding $\omega_1, \omega_2, \omega_3$ amounts to calculating $\tau_w(G_2, \Sigma_2)_i$ for $i = 1, 2, 3$, which is just the minimum weight of a cut in G_2 separating u_i from $\{u_1, u_2, u_3\} \setminus \{u_i\}$, so that can be solved by flow techniques. As $\{u_1, u_2\}$ is a blocking pair in $(\widehat{G}_2, \widehat{\Sigma}_2)$ finding a maximum w-packing of odd circuits in $(\widehat{G}_2, \widehat{\Sigma}_2)$ can be done by solving an integer 2-commodity flow problem. As explained in the proof of Lemma 6 the solution of that gives a collection of paths in G_2 that can be used to transform a maximum w-packing of odd circuits in $(\widehat{G}_1, \Sigma_1)$ to a maximum w-packing of odd circuits in (G, Σ). As all this can be done in polynomial time, we have a polynomial time reduction from the odd circuit packing problem in (G, Σ) to the odd circuit packing problem in $(\widehat{G}_1, \Sigma_1)$, which is a proper minor of (G, Σ).

So there exists a polynomial time algorithm for the odd circuit packing problem in 3-connected signed graphs with no $\widetilde{K}_5$-, $K^{1,1}_{3,3}$-, $K^{1,2}_{3,3}$-, or $K^2_{3,3}$-minor. Next we consider the case that the signed graph is not 3-connected. Here there are certain issues involved that need extra care. Consider a 2-separation $(G_1, \Sigma_1), (G_2, \Sigma_2)$ of (G, Σ) as in the proof of Lemma 5. If we can find such separation with (G_1, Σ_1) and (G_2, Σ_2) both bipartite, then u_1 is a blockvertex of (G, Σ), and we can solve the odd circuit packing problem by flow techniques. So we assume that no such 2-separations exist. Therefore as of now we assume that we selected (G_1, Σ_1) and (G_2, Σ_2) such that (G_2, Σ_2) is nonbipartite and under that condition $E(G_1)$ is inclusionwise minimal.

Let (G^1_1, Σ^1_1) be obtained from (G_1, Σ_1) by adding an odd edge e_1 connecting u_1 and u_2, and let (G^0_1, Σ^0_1) be obtained from (G_1, Σ_1) by adding an even edge f_1 connecting u_1 and u_2. Then as (G_2, Σ_2) is nonbipartite both (G^1_1, Σ^1_1) and (G^0_1, Σ^0_1) are proper minors of (G, Σ). Moreover, by minimality of $E(G_1)$ these graphs are 3-connected so we do have a polynomial time algorithm for solving any odd circuit packing problem in (G^1_1, Σ^1_1) or (G^0_1, Σ^0_1). This is important since as we will see we need to solve three such problems in these signed graphs.

For both $i = 0$ and $i = 1$, we can find $\tau_w(G_1, \Sigma_1)_i$ in polynomial time as it amounts to finding a minimum weight signature in (G^i_1, Σ^i_1) where the extra edge between u_1 and u_2 gets a very high weight. Thus we can calculate ω in polynomial time. Now solve the odd circuit packing problem in the signed graph $(\widehat{G}_2, \Sigma_2)$ constructed in the proof of Lemma 5. We do this recursively, so we may use 2-separations again. We also find the collection of even circuits $\mathcal{D}_2$ (which is just a flow problem) and adjust the solution such that γ is either 0 or ω, as in (13). Now we solve the odd circuit packing problem on $(\widehat{G}_1, \widehat{\Sigma}_1)$. Since $\widehat{G}_1$ is 3-connected, we can do this without recursively using 2-separations. Now we combine the optimal packing of odd circuits in $(\widehat{G}_1, \widehat{\Sigma}_1)$ with the optimal packing of odd circuits in $(\widehat{G}_2, \Sigma_2)$ and with the collection $\mathcal{D}^2$ of even circuits to a solution for the odd circuit packing problem in (G, Σ).

This recursive method using 2-separations calls itself only in $(\widehat{G}_2, \Sigma_2)$ and for just a single function w. Hence, it runs in polynomial time. □

4. Subdivisions, homeomorphs, and minors; links and bridges.

If P is a path containing vertices u and v, then P_{uv} denotes the uv-subpath of P.

Subdividing an edge uv of (G, Σ) is replacing it with a uv-path P that is internally vertex disjoint with G and replacing Σ with $(\Sigma \setminus \{uv\}) \cup \Sigma_P$, where Σ_P is any subset of $E(P)$ with the same parity as $\Sigma \cap \{uv\}$. A (G, Σ)-*subdivision* is the result of a series of subdivisions of edges in (G, Σ). If G is just a graph, so with no signing, *subdividing an edge* and G-*subdivision* are defined similarly.

A (G, Σ)-*homeomorph* is a signed graph that is isomorphic to a (G, Σ)-subdivision. Clearly, if a signed graph has a (G, Σ)-homeomorph it has a (G, Σ)-minor. If G has maximum degree 3, the converse is true as well. In particular, for $i = 0, 1, 2$, (G, Σ) has a $K^i_{3,3}$-minor if and only if it has a $K^i_{3,3}$-homeomorph.

Let G be a graph; a *leg* of G is a path such that all of its internal vertices have degree 2 in G and its ends have degree at least 3. Let H be a subgraph of G, and let u and v be two of its vertices. A *uv-link of H*, or just *link of H*, is a uv-path that intersects H exactly in $\{u, v\}$.

If G is a graph and X is a set of vertices, then $G - X$ is the graph obtained from G by deleting the vertices in X and the edges incident to them; if X is a set of edges (or a subgraph with edges), then $G - X$ is obtained by deleting only the edges in X.

A subgraph B of G is called a *bridge* of H if either B consists of a single edge not in $E(H)$ that has both ends in $V(H)$ or B consists of a component of $G - V(H)$ together with the edges from this component to H and their ends in H.

5. Recognizing if a graph has a $\widetilde{K}_5$-, $K_{3,3}^{1,1}$-, $K_{3,3}^{1,2}$-, or $K_{3,3}^{2}$-minor. We describe how to decide in polynomial time if a graph has a $\widetilde{K}_5$-, $K_{3,3}^{1,1}$-, $K_{3,3}^{1,2}$-, or $K_{3,3}^{2}$-minor or not. The algorithm is based on the decomposition in Theorem 3. The idea is standard: we can check in polynomial time if G is planar or if (G, Σ) has a blockvertex or is one of the signed graphs in Figure 2, so we need only recursive procedures for the cases that (G, Σ) is not 3-connected or has improper 3-vertex cutsets. In case (G, Σ) is not 3-connected such a procedure is straightforward, but dealing with decompositions along improper 3-vertex cutsets needs some extra care. So we describe that in detail.

Assume (G, Σ) is 3-connected and contains an improper 3-vertex cutset $\{u_1, u_2, u_3\}$. So, after resigning if necessary, we may assume that G contains graphs G_1 and G_2 with $\Sigma \cap E(G_2) = \emptyset$ such that $E(G_1)$ and $E(G_2)$ partition $E(G)$, $V(G_1) \cap V(G_2) = \{u_1, u_2, u_3\}$, and $|E(G_2)| \geq 4$. Let G^+ be defined by adding to G_1 a new vertex u^+ and three new even edges u^+u_1, u^+u_2, and u^+u_3. Then (G^+, Σ) is a minor of (G, Σ). So if it has a $\widetilde{K}_5$-, $K_{3,3}^{1,1}$-, $K_{3,3}^{1,2}$-, or $K_{3,3}^{2}$-minor, then so does (G, Σ). Also if (G, Σ) has a $\widetilde{K}_5$-, $K_{3,3}^{1,2}$-, or $K_{3,3}^{2}$-minor, (G^+, Σ) will have such a minor. But, as $K_{3,3}^{1,1}$ has improper 3-vertex cutsets, (G, Σ) may have a $K_{3,3}^{1,1}$-minor whereas (G^+, Σ) does not. Fortunately, it can be checked in polynomial time if this happens, as we will explain now. Let G^- be obtained from G_2 by by adding a new vertex u^- and three new edges u^-u_1, u^-u_2, and u^-u_3. The following observation is straightforward.

(26) *(G, Σ) has a $K_{3,3}^{1,1}$-minor if and only if one of the following holds:*

(i) G^- has a $K_{3,3}$-subdivision in which u^- has degree 3 and (G^+, Σ) has a $\widetilde{K}_4$-homeomorph in which u^+ has degree 3 and at least one of u_1, u_2, and u_3 has degree 2.

(ii) G^- has a $K_{3,3}$-subdivision in which u^- has degree 3 and at least one of u_1, u_2, and u_3 has degree 2 and (G^+, Σ) has a $\widetilde{K}_4$-homeomorph in which u^+ has degree 3.

(iii) (G^+, Σ) has a $K_{3,3}^{1,1}$-minor.

So when we encounter an improper 3-separation, we first check if (26i) or (26ii) applies. If so we decide that our signed graph has a $K_{3,3}^{1,1}$-minor. If not we just replace (G, Σ) with (G^+, Σ) and search for the existence of a $\widetilde{K}_5$-, $K_{3,3}^{1,1}$-, $K_{3,3}^{1,2}$-, or $K_{3,3}^{2}$-minor in (G^+, Σ) recursively. To check if (26i) or (26ii) applies we use the following two results:

(27) *If v is a degree 3 vertex in a simple 3-connected graph H,
then v is a degree 3 vertex in some $K_{3,3}$-subdivision in H if and only if
H is nonplanar (Seymour [11]).*

(28) *If v is a degree vertex in a simple 3-connected signed graph (H, Θ), then v is a degree 3 vertex in some $\widetilde{K}_4$-homeomorph in (H, Θ) if and only if (H, Θ) has a $\widetilde{K}_4$-homeomorph.*

We will prove (28) below; (27) is immediate from (11.2) in Seymour [11]. By (27), we can check the condition on G^- in (26i) by checking if G^- is nonplanar. For checking the condition on G^- in (26ii), we construct for each $i = 1, 2, 3$ and each neighbor of $x \neq u^-$ of u_i the graph $G^-_{i,x}$ by deleting from G^- all edges incident with u_i except $u^- u_i$ and $u_i x$. If $G^-_{i,x}$ is nonplanar for some i and some x, the condition on G^- in (26ii) is satisfied; otherwise, it is not.

By (28), we can check the condition on (G^+, Σ) in (26ii) by checking if (G^+, Σ) contains a $\widetilde{K}_4$-homeomorph. This can be done in polynomial time by an algorithm by Gerards, Lovász, Schrijver, Seymour, Shih, and Truemper based on decomposing signed graphs with no $\widetilde{K}_4$-homeomorph (see Gerards [3]; actually the algorithm amounts to applying Truemper's algorithm [13] for recognizing if a binary clutter has a Q_6-minor to the clutter of odd circuits in (G^+, Σ)). Finally to check if (G^+, Σ) satisfies the condition in (26i), we construct for each $i = 1, 2, 3$ and each neighbor of $x \neq u^+$ of u_i the graph $G^+_{i,x}$ by deleting from G^+ all edges incident with u_i except $u^- u_i$ and $u_i x$. If $G^-_{i,x}$ contains a $\widetilde{K}_4$-homeomorph for some i and some x, the condition on G^- in (26ii) is satisfied; otherwise, it is not.

So to see that we can decide in polynomial time if a signed graph has a $\widetilde{K}_5$-, $K^{1,1}_{3,3}$-, $K^{1,2}_{3,3}$-, or $K^{2}_{3,3}$-minor, it remains only to prove (28).

Proof of (28). Suppose it is false; let (H, Θ) be a minimal counterexample.

(29) *Each $\widetilde{K}_4$-homeomorph K satisfies $V(K) \supseteq V(H) \setminus \{u\}$.*

Suppose it is not true; let K be a $\widetilde{K}_4$-homeomorph and x be a vertex not in $V(K) \cup \{u\}$. As H is 3-connected, x has a neighbor y such that $\{x, y\} \not\subseteq \{u, u_1, u_2, u_3\}$. Then $H \backslash xy$ contains K. So if $H \backslash xy$ is a subdivision of a simple 3-connected graph H', it follows, as (H, Θ) is a minimal counterexample, that H' contains a $\widetilde{K}_4$-homeomorph containing u. As H itself does not contain such a homeomorph, this is impossible. So $H \setminus xy$ is not a subdivision of a simple 3-connected graph. Then, as $|V(H)| \geq |V(K) \cup \{x\}| \geq 5$, (11.1) in Seymour [11] says that H/xy is 3-connected. H/xy may have parallel edges though. Let H'' be a subgraph of H/xy consisting of one edge from each parallel class of H/xy. We may choose H'' such that it contains K. Note that u has also degree 3 in H''. Hence, as (H, Θ) is a minimal counterexample, H'' contains a $\widetilde{K}_4$-homeomorph containing u. But then also H contains such a $\widetilde{K}_4$-homeomorph; this contradiction proves (29).

(30) *(H, Θ) contains a $\widetilde{K}_4$-homeomorph $\bar{K}$ with $V(\bar{K}) = V(H)$.*

Indeed, let K be a $\widetilde{K}_4$-homeomorph in H. If $u \notin V(K)$, then, by (29), u has all three neighbors on K. From this it is straightforward to check that the union of K and the three edges incident with u contains a $\widetilde{K}_4$-homeomorph $\bar{K}$ using u. By (29), $V(\bar{K}) = V(H)$. So (30) follows.

Take $\bar{K}$ as in (30). Then as u does not have degree 3 in K, we may assume that uu_1 and uu_2 are edges of the same leg, say, P, of $\bar{K}$. By (28), u_3 lies on $\bar{K}$. If u_3 does not lie on P, then it is straightforward to find in $K \cup \{uu_3\}$ a $\widetilde{K}_4$-homeomorph in which u has degree 3. So u_3 lies on P as well, see Figure 3 (left). As indicated there,

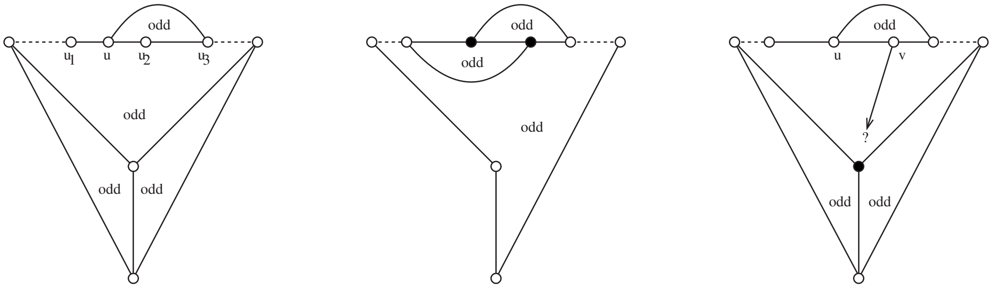


FIG. 3. *The word "odd" indicates that the corresponding face is bounded by an odd circuit. Dashed edges may have length zero.*

the circuit $P_{u_3u} \cup \{uu_3\}$ is odd as otherwise $(\bar{K} - P_{u_3u}) \cup \{uu_3\}$ is a $\widetilde{K}_4$-homeomorph that misses u_2, contradicting (29). As H is 3-connected, $P_{u_1u_3} - u_1 - u_3$ contains a vertex v that is adjacent to a vertex $w \in V(\bar{K}) \setminus V(P_{u_1u_3})$. As u had degree 3 in H, $v \neq u$.

First consider the case that w lies on P. Then the circuit $P_{vw} \cup \{vw\}$ is odd as otherwise $(\bar{K} - P_{vw}) \cup \{vw\}$ is a $\widetilde{K}_4$-homeomorph that misses either u_1 or u_3, contradicting (29). So $\bar{K} \cup \{uu_3, vw\}$ contains a subgraph as indicated in the middle picture in Figure 3, where u is one of the two black vertices. That subgraph is a $\widetilde{K}_4$-homeomorph, and u is a degree 3 vertex of it. This contradicts our assumption that no such homeomorph exists. So we may assume that w is not on P.

Then upto symmetry w lies on a leg of $\bar{K}$ that has the black vertex as an end, as indicated in Figure 3 (right). From the fact that $\bar{K}$ is a $\widetilde{K}_4$-homeomorph, it is again a straightforward case check that $\bar{K} \cup \{uu_3, vw\}$ contains a $\widetilde{K}_4$-homeomorph in which u has degree 3. This concludes the proof of (28). □

6. Nonbipartite subdivisions of $K_{3,3}$: Proof of Theorem 3(i).

We now prove Theorem 3(i). We denote the six degree-3 vertices of a $K_{3,3}$-subdivision K by $r_1^K, r_2^K, r_3^K, r_4^K, r_5^K$, and r_6^K, where the numbering is such that there is a leg between r_i^K and r_j^K if and only if $i = 1, 3, 5$ and $j = 2, 4, 6$. We denote such a leg by P_{ij}^K.

Proof of Theorem 3(i). Suppose the theorem is false. Let (G, Σ) be a minor-minimal counterexample. As G is 3-connected, has no parallel edges, and is not planar and not isomorphic to K_5, it follows from Kuratowski's theorem and a well-known and easy result of Hall [7] that G contains a $K_{3,3}$-subdivision. No $K_{3,3}$-subdivision in G contains odd circuits, as otherwise there would be a $K_{3,3}^1$- or a $K_{3,3}^2$-homeomorph.

Let K be any $K_{3,3}$-subdivision. By resigning, we may assume that all edges in K are even.

(31) *Each odd link of* K *has both ends in* $\{r_1^K, r_3^K, r_5^K\}$ *or both ends in* $\{r_2^K, r_4^K, r_6^K\}$.

Suppose there is a link P contradicting (31). Then $K \cup P$ contains a $K_{3,3}$-subdivision using P as part of one of its legs. As P is odd and all edges in K are even, this is a $K_{3,3}^1$-subdivision; this contradiction proves (31).

(32) *Each odd link of* K *is an edge.*

Suppose this is not true; let P be a link of K contradicting (32). By (31), we may assume that the ends of P are r_1^K and r_3^K. As P is not an edge and G is 3-connected, there exists a link Q of $K \cup P$ with one end in $V(P) \setminus \{r_1^K, r_3^K\}$ and one end, say, r

in $V(K) \setminus \{r_1^K, r_3^K\}$. Clearly, $P \cup Q$ contains an odd link of K with end r. So, by (31), r has to be r_5^K. Now $(K \cup P \cup Q) - P_{21}^K - P_{23}^K - P_{25}^K$ is a $K_{3,3}^1$-homeomorph; this contradiction proves (32).

G has at least seven vertices, as otherwise Theorem 3(i) is easily verified. It is straightforward to derive from that and the fact that G is 3-connected that (G, Σ) has a $K_{3,3}$-subdivision with at least seven vertices. Fix such a $K_{3,3}$-subdivision, and call it K. Let F be the edges of G that form the odd links of K. So each edge in F has both ends in $\{r_1^K, r_3^K, r_5^K\}$ or both ends in $\{r_2^K, r_4^K, r_6^K\}$. For each edge uv of F, there are three internally vertex disjoint uv paths in K. Hence, $G - F$ is 3-connected. Moreover, $G - F$ has no odd circuits because if it had, then by the 3-connectivity of $G - F$ there would exist an odd link of K that is not an edge of F, contradicting (32). So we may resign (G, Σ) such that the edges in F are odd and the edges in $G - F$ are even.

(33) *If $i = 1, 3, 5$ and $j = 2, 4, 6$ and if r_i^K and r_j^K are both ends of some edge in F, then P_{ij}^K consists of a single edge.*

Suppose this is false. Then, as $G - F - r_i^K - r_j^K$ is connected, K has an even link Q with one end in $P_{ij}^K - r_i^K - r_j^K$ and one end not in P_{ij}^K. Then Q is contained in a $K_{3,3}$-subdivision in $K \cup Q$. This $K_{3,3}$-subdivision has an odd link contradicting (32). So (33) follows.

(34) *We may assume that $r_1^K r_3^K$ and $r_2^K r_4^K$ are in F and that $r_1^K r_5^K$, $r_2^K r_6^K$ and $r_4^K r_6^K$ are not in F.*

If no edge in F has its end in $\{r_1^K, r_3^K, r_5^K\}$, then $\{r_2^K, r_4^K, r_6^K\}$ is an improper 3-vertex cutset. Hence, by symmetry, we may assume that $r_1^K r_3^K$ and $r_2^K r_4^K$ are in F. As K has at least seven vertices, it follows from (33) that at least one of $r_1^K, r_2^K, \ldots, r_6^K$ is not an end of an edge in F. So, again by symmetry, we may assume that $r_2^K r_6^K$ and $r_4^K r_6^K$ are not in F. Now if both $r_1^K r_5^K$ and $r_3^K r_5^K$ are in F, then $r_1^K r_3^K, r_1^K r_5^K, r_3^K r_5^K$, $r_2^K r_4^K$, and K contains a $\widetilde{K}_5$-homeomorph. Thus (34) follows.

$$F = \{r_1^K r_3^K, r_2^K r_4^K\}. \tag{35}$$

If not, then by (34), $F = \{r_1^K r_3^K, r_3^K r_5^K, r_2^K r_4^K\}$. Now as F has at least seven vertices it follows from (33) that $P_{61}^K \cup P_{63}^K \cup P_{65}^K$ has at least four edges. Since $\{r_1^K, r_3^K, r_5^K\}$ is not an improper 3-vertex cutset this means that (G, Σ) has the signed graph in Figure 4 as a minor (possibly with r_2^K and r_4^K interchanged). That signed graph has a $K_{3,3}^1$-subdivision, so (35) follows.

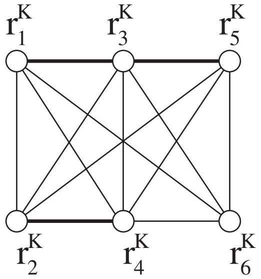


FIG. 4. *Bold edges are odd; thin edges are even.*

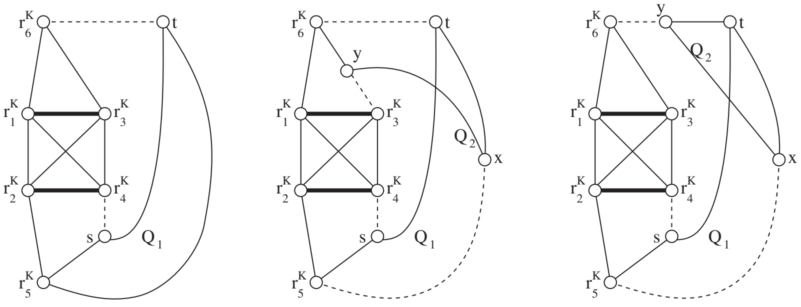


FIG. 5. *Bold edges are odd paths; thin edges are even paths; and dashed edges may have length zero.*

By (33) and (35), each of P_{12}^K, P_{14}^K, P_{32}^K, and P_{34}^K is a single edge. Hence, by symmetry, we may assume that $P_{61}^K \cup P_{63}^K \cup P_{65}^K$ has at least four edges. Since $\{r_1^K, r_3^K, r_5^K\}$ is not an improper 3-vertex cutset, that means that K has an st-link Q_1 with s on $(P_{52}^K \cup P_{54}^K) - r_5^K$ and t on $(P_{61}^K \cup P_{63}^K \cup P_{65}^K) - r_1^K - r_3^K - r_5^K$. Choose K and Q_1 such that t is as close as possible to $P_{61}^K \cup P_{63}^K$ in $P_{61}^K \cup P_{63}^K \cup P_{65}^K$. We may assume that s lies on P_{54}^K.

(36) *t lies on P_{65}^K.*

If not, $K \cup Q_1$ contains a $K_{3,3}$-subdivision that has an odd link contradicting (31).

So we have a situation as depicted in Figure 5 (left). Since $\{r_2^K, r_4^K, t\}$ is not an improper 3-vertex cutset, $K \cup Q_1$ has an xy-link Q_2 with x on $(P_{52}^K \cup P_{54}^K \cup Q_1 \cup (P_{56}^K)_{r_5^K t}) - r_2^K - r_4^K - t$ and y on $(P_{61}^K \cup P_{63}^K \cup (P_{65}^K)_{r_6^K t}) - t$. As K and Q_1 are chosen such that t is as close as possible to $P_{61}^K \cup P_{63}^K$ the end x of Q_2 has to lie on $(P_{56}^K)_{r_5^K t} - t$. If y lies on $P_{63}^K - r_6^K$ (see Figure 5 (middle)) then $(K \cup Q_1 \cup Q_2) - r_2^K r_3^K - r_1^K r_4^K - (P_{63}^K)_{r_6^K y} - (P_{52}^K)_{r_5^K s}$ is a $K_{3,3}^2$-subdivision. Hence, y does not lie on $P_{63}^K - r_6^K$ and, by symmetry, also does not lie on $P_{61}^K - r_6^K$. So y lies on $(P_{65}^K)_{r_6^K t} - t$ (see Figure 5 (right)). Now replacing K with $(K \cup Q_2) - (P_{65}^K)_{xy}$ and Q_1 with $Q_1 \cup (P_{65}^K)_{ty}$ yields a contradiction against the fact that K and Q_1 are chosen such that t is as close as possible to $P_{61}^K \cup P_{63}^K$. This proves Theorem 3(i). □

7. $K_{3,3}^1$-subdivisions and $K_{3,3}^1$-extensions. As of now, if K is a $K_{3,3}^1$-subdivision in (G, Σ), we will assume that the unique odd leg is P_{12}^K. In that case, we can always resign (G, Σ) such that the only odd edge in K is the edge in P_{12}^K with end r_1^K; unless stated otherwise, we will assume that if we call a $K_{3,3}^1$-subdivision K, it has such a canonical signing. Under these assumptions we define $T_1^K := P_{14}^K \cup P_{16}^K$, $T_2^K := P_{23}^K \cup P_{25}^K$, $\textsc{cage}(K) := P_{34}^K \cup P_{36}^K \cup P_{45}^K \cup P_{56}^K$, and $\textsc{core}(K) := V(\textsc{cage}(K)) \setminus \{r_3^K, r_4^K, r_5^K, r_6^K\}$ (see Figure 6).

Clearly, these labelings of vertices and legs of a $K_{3,3}^1$-subdivision and the indicated canonical signing are not unique. For instance if we interchange index 1 with index 2, interchange index pair $\{4, 6\}$ with index pair $\{3, 5\}$, and resign (G, Σ) on the internal vertices of P_{12}^F, we obtain another labeling and canonical signing as indicated above. When we use this symmetry, we refer to it as *left-right symmetry.* Simpler symmetries are 35-*symmetry*, that is interchanging index 3 with index 5, and 46-*symmetry.*

Our strategy in proving Theorem 3(ii) is to start with a $K_{3,3}^1$-subdivision in (G, Σ). Such a $K_{3,3}^1$-subdivision has blockvertices and improper 3-vertex cutsets. So, assuming

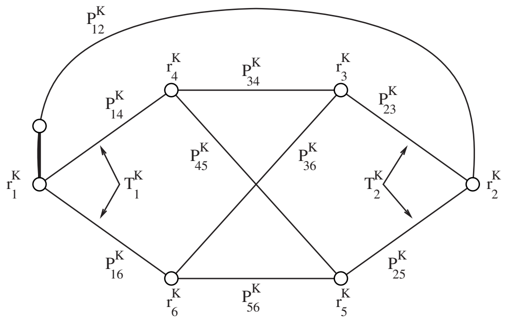


FIG. 6. *A $K^1_{3,3}$-subdivision F.*

(G,Σ) does not have these features, more structure should be available. We try to grasp that structure by studying the links of the $K^1_{3,3}$-subdivision. Ideally such a link, or a combination of a few of them, provides a contradiction by establishing one of the forbidden minors in Theorem 3(ii). There are other links, however, that do not provide any extra structure other then some extra $K^1_{3,3}$-subdivisions, for instance, even links with no end on the unique odd leg of the $K^1_{3,3}$-subdivision. To avoid chasing such useless links, we include many of them in our initial structure; that is, we start with a "$K^1_{3,3}$-extension" rather than with just a $K^1_{3,3}$-subdivision.

Consider a signed graph F consisting of

- six special vertices, $r^F_1, r^F_2, r^F_3, r^F_4, r^F_5$, and r^F_6,
- five internally vertex disjoint paths, $P^F_{12}, P^F_{14}, P^F_{16}, P^F_{23}$, and P^F_{25}, where P^F_{ij} is an $r^F_i r^F_j$-path whose edges are all even, except for the edge of P^F_{12} adjacent to r^F_1 which is odd,
- a 2-connected subgraph $\textsc{cage}(F)$ with even edges only that shares with these paths exactly the vertices r^F_3, r^F_4, r^F_5, and r^F_6.

We define $T^F_1 := P^F_{14} \cup P^F_{16}$, $T^F_2 := P^F_{23} \cup P^F_{25}$, and $\textsc{core}(F) := V(\textsc{cage}(F)) \setminus \{r^F_3, r^F_4, r^F_5, r^F_6\}$.

The set of $K^1_{3,3}$-subdivisions K in F with $P^K_{12} = P^F_{12}$ and $\textsc{cage}(K) \subseteq \textsc{cage}(F)$ is denoted by $\mathcal{K}(F)$. Note that for each $K^1_{3,3}$-subdivision K in $\mathcal{K}(F)$ we can choose the numbering such that: $r^K_1 = r^F_1$, $r^K_2 = r^F_2$, $P^K_{14} \supseteq P^F_{14}$, $P^K_{16} \supseteq P^F_{16}$, $P^K_{23} \supseteq P^F_{23}$, and $P^K_{25} \supseteq P^F_{25}$.

For $u \in V(F)$ we define the following

- If $u \notin \textsc{core}(F)$, then $\mathcal{K}_u(F) := \mathcal{K}(F)$.
- If $u \in \textsc{core}(F)$, then $\mathcal{K}_u(F)$ consists of those $K^1_{3,3}$-subdivisions $K \in \mathcal{K}(F)$ with $u \in \textsc{core}(K)$.

We call F a *$K^1_{3,3}$-extension* if $\mathcal{K}(F) \neq \emptyset$ and for each $u \in \textsc{core}(F)$ there exists a $K^1_{3,3}$-subdivision K in F with $u \in \textsc{core}(K)$ and (after resigning) $P^K_{12} = P^F_{12}$ (see Figure 7).

Note that each $K^1_{3,3}$-subdivision is a $K^1_{3,3}$-extension. A $K^1_{3,3}$-extension F is called *extreme* in (G,Σ) if, even after resigning, there is no $K^1_{3,3}$-extension F' with $P^{F'}_{12} \subsetneq P^F_{12}$ or with $P^{F'}_{12} = P^F_{12}$ and $\textsc{cage}(F') \supsetneq \textsc{cage}(F)$.

8. Links of $K^1_{3,3}$-extensions. As of now we call signed graphs with no $K^2_{3,3}$-, $K^{1,1}_{3,3}$-, or $K^{1,2}_{3,3}$-minor *clean*. In this section we characterize the type of links an extreme $K^1_{3,3}$-extension in a clean signed graph can have (see Figure 8).

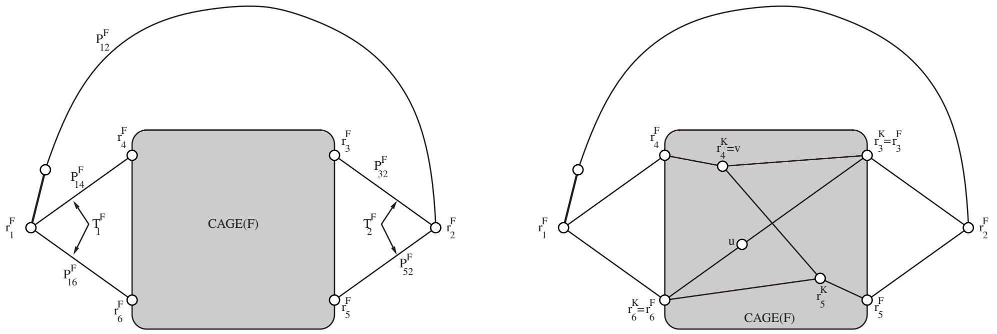


FIG. 7. *Left: a $K^1_{3,3}$-extension F. Right: a $K^1_{3,3}$-extension F with a $K^1_{3,3}$-subdivision K that lies in $\mathcal{K}(F)$ and in $\mathcal{K}_u(F)$ but not in $\mathcal{K}_v(F)$.*

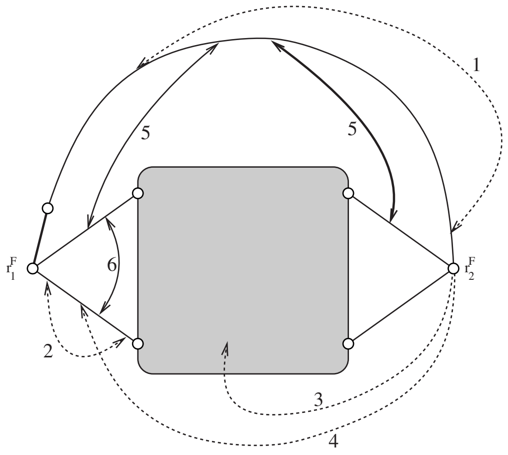


FIG. 8. *A $K^1_{3,3}$-extension F with all possible links (upto symmetry, numbers indicate types, thin lines are even links, bold lines are odd links, and dotted lines have either parity).*

LEMMA 7. *Let F be an extreme $K^1_{3,3}$-extension in a clean signed graph, and let P be a link of F. Then P is exactly one of the following types:*

Type 1. *Both ends of P lie on P^F_{12}.*

Type 2. *Both ends of P lie on P^F_{ij}, where (i,j) is $(1,4),(1,6),(2,3)$, or $(2,5)$.*

Type 3. *P connects r^F_i with a vertex in* CORE(F)*, where $i=1$ or $i=2$.*

Type 4. *P connects r^F_i with a vertex in $T^F_{3-i}-r^F_{3-i}$, where $i=1$ or $i=2$.*

Type 5. *P connects a vertex of $P^F_{12}-r^F_1-r^F_2$ with a vertex on $T^F_i-r^F_i$, where $i=1$ or $i=2$.*

Type 6. *P connects the two components of $T^F_i-r^F_i$, where $i=1$ or $i=2$.*

Moreover, a link P of Type 5 *is even when $i=1$ and odd when $i=2$; all links of Type* 6 *are even.*

We denote the collection of type t links of F by $\mathcal{L}^F_t$. If $t=2,5,6$, $\mathcal{L}^F_{t,i}$ denotes the collection of links in $\mathcal{L}^F_t$ with an end in T^F_i. If $t=1,3,4$, $\mathcal{L}^F_{t,i}$ denotes the collection of links in $\mathcal{L}^F_t$ with r^F_i as an end. So if $t\neq 1$, $\mathcal{L}^F_{t,1}$ and $\mathcal{L}^F_{t,2}$ partition $\mathcal{L}^F_t$. The set of even links in $\mathcal{L}^F_t$ is denoted by $\mathcal{E}^F_t$, and the set of odd links is denoted by $\mathcal{O}^F_t$. Similarly, we define $\mathcal{E}^F_{t,i}$ and $\mathcal{O}^F_{t,i}$.

It is the statement of Lemma 7 that the collection of links of an extreme $K^1_{3,3}$-extension F in a clean signed graph is equal to

$$\mathcal{L}^F_1 \cup \mathcal{L}^F_2 \cup \mathcal{L}^F_3 \cup \mathcal{L}^F_4 \cup \mathcal{E}^F_{5,1} \cup \mathcal{O}^F_{5,2} \cup \mathcal{E}^F_6 .$$

Mind that $\mathcal{E}^F_{5,1}$ corresponds to $\mathcal{O}^F_{5,2}$ under left-right symmetry, and $\mathcal{O}^F_{5,1}$ corresponds to $\mathcal{E}^F_{5,2}$.

Proof of Lemma 7. Suppose the theorem is false; let F and P form a counterexample. Note that as (G,Σ) has no $K^{1,1}_{3,3}$-minor, $\mathcal{O}^F_6 = \emptyset$. So

$$P \notin \mathcal{L}^F_1 \cup \mathcal{L}^F_2 \cup \mathcal{L}^F_3 \cup \mathcal{L}^F_4 \cup \mathcal{E}^F_{5,1} \cup \mathcal{O}^F_{5,2} \cup \mathcal{L}^F_6 . \tag{37}$$

We first prove

$$\textit{P has no end on } P^F_{12}. \tag{38}$$

If not, then as $P \notin \mathcal{L}^F_1 \cup \mathcal{L}^F_2 \cup \mathcal{L}^F_3 \cup \mathcal{L}^F_4$, one end of P, say, u, lies on $P^F_{12} - r^F_1 - r^F_2$ and the other end, say, v, does not lie on P^F_{12}. With u and v in those positions we may assume, by left-right symmetry, that P is even. So as $P \notin \mathcal{E}^F_{5,1}$, v does not lie on T^F_1. Let $K \in \mathcal{K}_v(F)$. Then v is not on T^K_1. By 35-symmetry and 46-symmetry we may assume that v lies on $P^F_{43} \cup P^F_{32} - r^F_4 - r^F_2$. If v lies on P^F_{43}, let $S := P^F_{32}$; if v lies on P^F_{32}, let S be the vr^F_2-subpath of P^F_{32}. Then $K' = (K \cup P) - S$ is a $K^1_{3,3}$-extension with $P^{K'}_{12}$ strictly contained in P^F_{12}; this contradicts that F is extreme. So (38) follows.

$$\textit{Both ends of P lie in the core of F.} \tag{39}$$

Suppose this is not true; then by symmetry we may assume that P has an end u in $P^F_{14} - r^F_1$. Then by (37) and (38) the other end, say, v of P lies on $T^F_2 - r^F_2$ or in the core of F. Let $K \in \mathcal{K}_v(F)$. Then by 35-symmetry, we may assume that v lies on $(P^K_{43} \cup P^K_{63} \cup P^K_{32}) - r^K_2 - r^K_4 - r^K_6$. If v lies on P^K_{43}, let S be the $r^K_4 v$-subpath of P^K_{43}; otherwise, $S := P^K_{43}$. If v lies on P^K_{23}, let R be the $r^K_3 v$-subpath of P^K_{23}; otherwise, $R := \{v\}$. Let Q be the ur^F_4-subpath of P^F_{14}. Then $K' := (K \cup P) - S$ is a $K_{3,3}$-subdivision with odd leg P^F_{12}. Moreover, the leg of K' containing P shares no end with P^F_{12}. Hence, as (G,Σ) has no $K^2_{3,3}$-minor, that leg is even. So K' is a $K^1_{3,3}$-subdivision. The vertices of $(P \cup Q \cup R) - u - v$ lie in $\textsc{core}(K')$. Hence, $F \cup P$ is a $K^1_{3,3}$-extension that has a larger core than F has, a contradiction. So (39) follows.

Let u and v be the two ends of P. Let $K \in \mathcal{K}_u(F)$. As $\textsc{cage}(F) - u$ is connected, it contains a path from v to K. Let P' be the union of this path with P, then P' is a leg of K with one end in $\textsc{core}(K)$ and the other end not in P^K_{12}. Hence, as (G,Σ) has no $K^2_{3,3}$-minor, P' is even. So P' is contained in the cage of a (unique) $K^1_{3,3}$-subdivision in $K \cup P$. Hence, $F \cup P$ is a $K^1_{3,3}$-extension with a larger core than F, a contradiction. □

9. Pairs of links of $K^1_{3,3}$-extensions. We study the occurrence of pairs of links of $K^1_{3,3}$-extensions of different types, but first we give an easy fact.

LEMMA 8. *Let a, b_1, b_2 be vertices in a 3-connected signed graph. Each nonbipartite bridge of a, b_1, b_2 contains an odd ab_1-path disjoint from b_2 or an odd ab_2-path disjoint from b_1.*

Proof. Let C be an odd circuit in the bridge. As the graph is 3-connected, there exist three vertex disjoint paths from C to $\{a, b_1, b_2\}$. So the bridge contains an odd path P with ends in $\{a, b_1, b_2\}$. Assume P is not as claimed. Then it is a b_1b_2-path.

As $\{b_1, b_2\}$ is not a 2-vertex cutset, there exists a path Q from a to P that is disjoint from $\{b_1, b_2\}$. Clearly $P \cup Q$ contains an odd ab_1-path or an odd ab_2-path; it obviously misses one of b_1 and b_2. $\square$

If F is an $K^1_{3,3}$-extension, then $\Lambda^F_i := \mathcal{O}^F_{2,i} \cup \mathcal{O}^F_{3,i} \cup \mathcal{O}^F_{4,i} \cup \mathcal{L}^F_{5,i}$ for $i = 1, 2$.

LEMMA 9. *Let F be an extreme $K^1_{3,3}$-extension in a 3-connected clean signed graph with no blockvertex and no improper 3-vertex cutset. If Λ^F_1 and Λ^F_2 are nonempty, then either $\Lambda^F_1 = \mathcal{O}^F_{2,1} \cup \mathcal{L}^F_{5,1}$ and $\Lambda^F_2 = \mathcal{O}^F_{4,2}$ or $\Lambda^F_1 = \mathcal{O}^F_{4,1}$ and $\Lambda^F_2 = \mathcal{O}^F_{2,2} \cup \mathcal{L}^F_{5,2}$.*

Proof. First we prove some easy facts. In items (40)–(45), K is a $K^1_{3,3}$-subdivision in a clean signed graph.

(40) *If $Q_1 \in \mathcal{O}^K_{2,1}$ and $Q_2 \in \mathcal{O}^K_{2,2}$, then they intersect.*

Indeed, if Q_1 and Q_2 did not intersect, then the unique $K_{3,3}$-subdivision in $K \cup Q_1 \cup Q_2$ that contains both Q_1 and Q_2 would be a $K^2_{3,3}$-subdivision.

(41) *If $Q_1 \in \mathcal{O}^K_{2,1}$ and $Q_2 \in \mathcal{O}^K_{3,2} \cup \mathcal{L}^K_{5,2}$, then they intersect.*

By contracting edges in the cage of F and along P^F_{12}, we can turn K into a $K^1_{3,3}$-subdivision K' so that $Q_2 \in \mathcal{O}^{K'}_{2,2}$. As Q_1 is also in $\mathcal{O}^{K'}_{2,1}$ it follows from (40) that Q_1 and Q_2 intersect after these contractions. As these intersections cannot lie on K', the paths also intersected before the contractions were carried out. So (41) holds indeed.

(42) *If $Q_1 \in \mathcal{O}^K_{4,1}$ and $Q_2 \in \mathcal{O}^K_{4,2}$, then they intersect.*

If not, $K \cup Q_1 \cup Q_2$ contains a $K^{1,2}_{3,3}$-minor.

(43) *If $Q_1 \in \mathcal{O}^K_{3,1}$ and $Q_2 \in \mathcal{O}^K_{4,2}$, then they intersect.*

If not, we can contract edges in the cage of K such that Q_1 and Q_2 stay disjoint and K turns into a $K^1_{3,3}$-subdivision K' with $Q_1 \in \mathcal{O}^{K'}_{4,1}$ and $Q_2 \in \mathcal{O}^{K'}_{4,2}$, contradicting (42).

By a similar contraction argument we derive the following from (41):

(44) *If $Q_1 \in \mathcal{O}^K_{3,1}$ and $Q_2 \in \mathcal{L}^K_{5,2}$, then they intersect.*

Note that (41), (43), and (44) have "left-right symmetrical" versions obtained by swapping the second subscripts 1 and 2. We will not list all such versions but just refer to them by mentioning left-right symmetry.

(45) *If $Q_1 \in \mathcal{O}^K_{3,1}$ and $Q_2 \in \mathcal{O}^K_{3,2}$, then they intersect outside K.*

If Q_1 and Q_2 do not intersect at all, it is possible to contract edges in the cage of K such that K turns into a $K^1_{3,3}$-subdivision K' with $Q_1 \in \mathcal{O}^{K'}_{2,1} \cup \mathcal{O}^{K'}_{4,1}$ and Q_2 still in $\mathcal{O}^{K'}_{3,2}$. If $Q_1 \in \mathcal{O}^{K'}_{2,1}$ this contradicts (41); if $Q_1 \in \mathcal{O}^{K'}_{4,1}$ this contradicts (43), by left-right symmetry. If Q_1 and Q_2 meet only in the cage of K, so at their ends, we can contract edges in $\textsc{cage}(K)$ such that we obtain the signed graph in Figure 9(a) as a minor. As is illustrated in that figure, that signed graph has a $K^{1,1}_{3,3}$-homeomorph, a contradiction. So (45) follows indeed.

Now let F be an extreme $K^1_{3,3}$-extension in a clean signed graph (G, Σ) with no blockvertex and no improper 3-vertex cutset.

(46) *At least one of $\mathcal{O}^F_{2,1} \cup \mathcal{O}^F_{3,1}$ and $\mathcal{O}^F_{2,2} \cup \mathcal{L}^F_{5,2}$ is empty.*

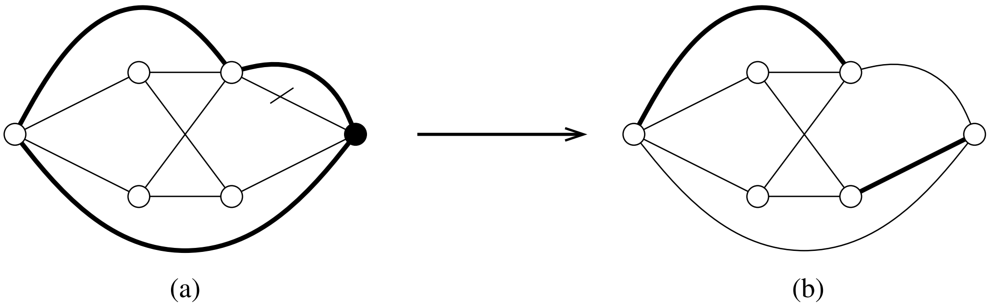


FIG. 9. *Bold edges are odd; thin edges are even. To obtain* (b) *from* (a)*, resign on the black vertex and delete the "crossed" edge.*

Suppose this is false; let $P_1 \in \mathcal{O}^F_{2,1} \cup \mathcal{O}^F_{3,1}$ and $P_2 \in \mathcal{O}^F_{2,2} \cup \mathcal{L}^F_{5,2}$. If $P_1 \in \mathcal{O}^F_{3,1}$, let u be its end in the core of F; otherwise, let u be any vertex of F. Choose $K \in \mathcal{K}_u(F)$. Then $P_1 \in \mathcal{O}^K_{2,1} \cup \mathcal{O}^K_{3,1}$ and $P_2 \in \mathcal{O}^K_{2,2} \cup \mathcal{L}^K_{5,2}$. Hence, it follows from (40), (41), (44), and left-right symmetry that P_1 and P_2 intersect. Clearly this intersection lies outside F. Hence, $P_1 \cup P_2$ contains a link of F that has one end in $(T^F_1 \cup \text{core}(F)) - r^F_1$ and one end in $T^F_2 - r^F_2$. As this contradicts Lemma 7, (46) follows.

(47) *At least one of* $\mathcal{O}^F_{3,1} \cup \mathcal{O}^F_{4,1}$ *and* $\mathcal{O}^F_{4,2}$ *is empty.*

Suppose this is false; let $P_1 \in \mathcal{O}^F_{3,1} \cup \mathcal{O}^F_{4,1}$ and $P_2 \in \mathcal{O}^F_{4,2}$. If $P_1 \in \mathcal{O}^F_{3,1}$, let u be its end in the core of K; otherwise, let u be any vertex of F. Choose $K \in \mathcal{K}_u(F)$. Then $P_1 \in \mathcal{O}^K_{3,1} \cup \mathcal{O}^K_{4,1}$ and $P_2 \in \mathcal{O}^K_{4,2}$. Hence, it follows from (42) and (43) that P_1 and P_2 intersect. Clearly this intersection lies outside F. Hence, $P_1 \cup P_2$ contains a link of F that has one end in $T^F_1 - r^F_1$ and one end in $(T^F_2 \cup \text{core}(F)) - r^F_2$. As this contradicts Lemma 7, (47) follows.

Now assume that the lemma is false and that F is a counterexample. Hence, Λ^F_1 and Λ^F_2 are both nonempty.

(48) $\mathcal{O}^F_2$ *is empty.*

Suppose this is false; assume $\mathcal{O}^F_{2,1} \neq \emptyset$. Then by (46) and left-right symmetry $\Lambda^F_2 = \mathcal{O}^F_{4,2}$. So $\mathcal{O}^F_{4,2} \neq \emptyset$. Hence, (47) implies that $\Lambda^F_1 = \mathcal{O}^F_{2,1} \cup \mathcal{L}^F_{5,1}$. This contradicts that F is a counterexample, so (48) follows.

We consider two cases.

Case 1. $\mathcal{O}^F_3$ is empty.

(49) $\mathcal{L}^F_{5,1}$ *and* $\mathcal{L}^F_{5,2}$ *are not empty.*

If $\mathcal{L}^F_{5,1} = \emptyset$, then, by (48) and as $\mathcal{O}^F_3$ is empty, $\Lambda^F_1 = \mathcal{O}^F_{4,1}$ and $\Lambda^F_2 = \mathcal{O}^F_{4,2} \cup \mathcal{L}^F_{5,2}$. Hence, as F falsifies the lemma, both $\mathcal{O}^F_{4,1}$ and $\mathcal{O}^F_{4,2}$ are nonempty, contradicting (47).

(50) $\mathcal{O}^F_4$ *is empty.*

Suppose this is false; assume $Q \in \mathcal{O}^F_{4,1}$. Let $P_1 \in \mathcal{L}^F_{5,1}$ and $P_2 \in \mathcal{L}^F_{5,2}$. By Lemma 7, Q and P_1 are vertex disjoint and P_1 and P_2 are internally vertex disjoint. Let P'_2 be the link of $F \cup Q$ that is contained in P_2 and has one end on P^K_{12}. Let P''_2 be the link of F in $\mathcal{L}^F_{5,2}$ contained in $P'_2 \cup Q$. By symmetry, we may assume that P_1 has an end on P^F_{14}

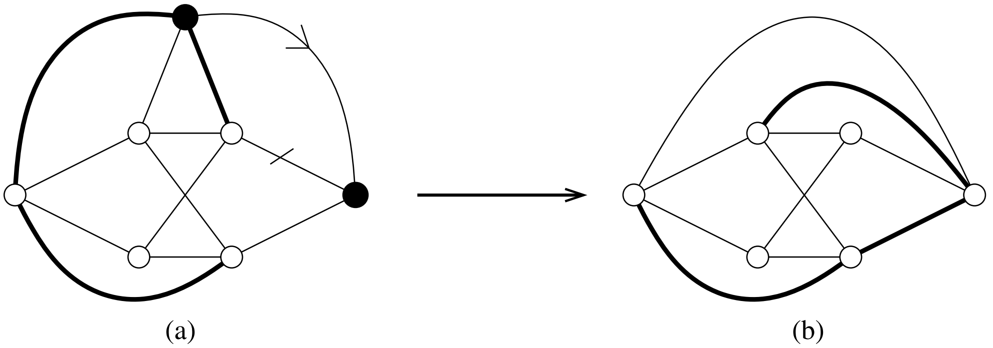


FIG. 10. *Bold edges are odd; thin edges are even. To obtain* (b) *from* (a)*, resign on the black vertices, delete the "crossed" edge, and contract the "directed" edge.*

and that P''_2 has an end on P^F_{23}. Note that, by Lemma 7, $P_1 \in \mathcal{E}^F_{5,1}$ and $P''_2 \in \mathcal{O}^F_{5,2}$. If Q has an end in P^F_{25}, then by construction of P''_2 links Q and P''_2 are disjoint. In that case, $K \cup Q \cup P_1 \cup P''_2$ contains the signed graph in Figure 10(a) as a minor, and as illustrated in Figure 10 that signed graph has a $K^{1,2}_{3,3}$-minor. So Q has an end in P^F_{23}. If Q and P''_2 share edges, resign (if necessary) to make them even, and contract them. Now it it easy to see that $K \cup Q \cup P''_2$ has the signed graph in Figure 9(a) as a minor, hence also a $K^{1,1}_{3,3}$-minor. That contradicts the cleaness of (G, Σ), so (50) follows indeed.

(51) *There exists a vertex $v \in P^F_{12}$ such that each path in $\mathcal{L}^F_5$ has v as one of its ends.*

By (49), it suffices to prove that if $P_1 \in \mathcal{L}^F_{5,1}$ has end p_1 on P^F_{12} and $P_2 \in \mathcal{L}^F_{5,2}$ has end p_2 on P^F_{12}, then $p_1 = p_2$. Suppose this is not the case. Choose $K \in \mathcal{K}(F)$. By Lemma 7, P_1 and P_2 are vertex disjoint. If p_1 lies between r^K_1 and p_2 along P^K_{12}, then the unique $K_{3,3}$-subdivision in $K \cup P_1 \cup P_2$ that contains P^K_{12}, P_1, and P_2 is a $K^2_{3,3}$-subdivision. So p_1 lies between p_2 and r^K_2 along P^K_{12}. Then $K \cup P_1 \cup P_2$ is a subdivision of the signed graph in Figure 11(a). Hence, as illustrated in Figure 11(b), it contains a $K^1_{3,3}$-extension F' with $P^{F'}_{12} = (P^F_{12})_{p_1 p_2}$. That contradicts the extremeness of F, so (51) follows.

As G is 3-connected, $\{r^F_1, r^F_2\}$ is not a 2-vertex cutset of $G - v$. Hence, it follows from (51) that P^F_{12} consists of only two edges: $r^F_1 v$ and $v r^F_2$. Fix $P_1 \in \mathcal{E}^F_{5,1}$ and $P_2 \in \mathcal{O}^F_{5,2}$. Resign on the internal vertices of P_1 and P_2 so that all edges on P_1 and on $P_2 - v$ are even. As (G, Σ) has no blockvertex, $(G, \Sigma) - v$ contains an odd circuit. Hence, as $G - v$ is 2-connected, $(F \cup P_1 \cup P_2) - v$ has an odd link Q contained in $G - v$. By Lemma 7, (48), (50), and (51), and as $\mathcal{O}^F_3$ is empty, Q is disjoint with P_1 and P_2, and $Q \in \mathcal{O}^F_1$. So the ends of Q are r^F_1 and r^F_2. Consider the $K^1_{3,3}$-subdivision $(F - P^F_{12}) \cup Q$; it is extreme in $F \cup P_1 \cup P_2 \cup Q$. The union of P_1 and P_2 is a link of that $K^1_{3,3}$-subdivision that contradicts Lemma 7. So Case 1 cannot apply.

Case 2. $\mathcal{O}^F_3$ is not empty.

If $\mathcal{O}^F_{3,1}$ is not empty, then by (46) and (47), $\Lambda^F_2 = \mathcal{O}^F_{3,2}$, so $\mathcal{O}^F_{3,2}$ is nonempty as well. Hence, by left-right symmetry it follows from $\mathcal{O}^F_3 \neq \emptyset$ that $\mathcal{O}^F_{3,1} = \Lambda^F_1 \neq \emptyset$ and $\mathcal{O}^F_{3,2} = \Lambda^F_2 \neq \emptyset$.

(52) *Each link in $\mathcal{O}^F_{3,1}$ intersects each link in $\mathcal{O}^F_{3,2}$ outside F.*

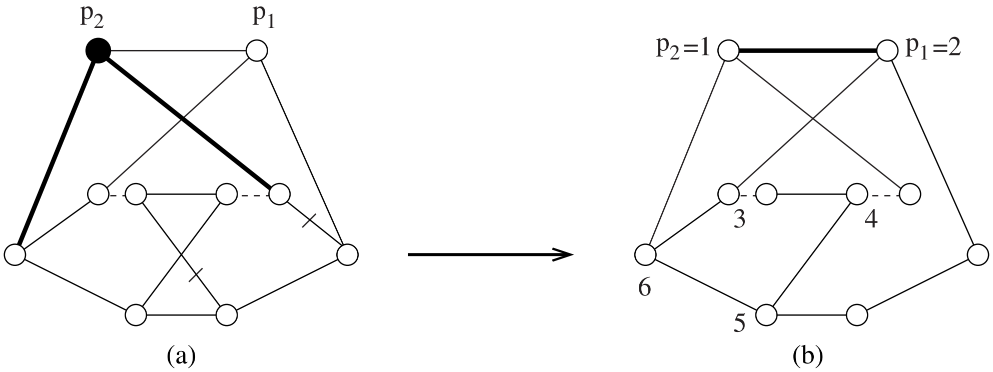


FIG. 11. *Bold edges are odd paths; thin edges are even paths; and dashed edges may have length zero. To obtain (b) from (a), resign on the black vertex and delete the "crossed" edges. The numbers $i = 1, \dots, 6$ indicate the vertices $r_i^{F'}$.*

Suppose this is false, and let $P_1 \in \mathcal{O}_{3,1}^F$ and $P_2 \in \mathcal{O}_{3,2}^F$ be disjoint outside F. Let p_1 be the end of P_1 in the core of F, and let p_2 be the end of P_2 in the core of F. Let $K \in \mathcal{K}_{p_1}(F)$. If $p_2 \neq p_1$, let P be a path in the cage of F that misses p_1 and connects p_2 to CAGE(K) (as CAGE(F) is 2-connected, such P exists); if $p_2 = p_1$, let P consist only of p_2. Then $P_2 \cup P \in \mathcal{O}_{3,2}^K \cup \mathcal{O}_{4,2}^K$ and $P_1 \in \mathcal{O}_{3,1}^K$. Moreover, these paths are disjoint. This contradicts (43) and (45). So (52) follows.

(53) *All links in $\mathcal{O}_3^F$ have the same end in the core of F; we call that end p.*

If not, then as $\mathcal{O}_{3,1}^F$ and $\mathcal{O}_{3,2}^F$ are nonempty, there would be a link in $\mathcal{O}_{3,1}^F$ and a link, in $\mathcal{O}_{3,2}^F$ that have different ends in the core of F. By (52) the union of two such links would contain a link of F that contradicts Lemma 7. So (53) follows.

Let $\mathcal{B}$ be the bridge of $\{r_1^F, r_2^F, p\}$ that contains CAGE(F).

(54) P_{12}^F *and all links in* $\mathcal{O}_3^F$ *lie outside* $\mathcal{B}$.

That P_{12}^F lies outside $\mathcal{B}$ follows as $\mathcal{L}_5^F = \emptyset$. Suppose $\mathcal{B}$ contains a link P in $\mathcal{O}_3^F$. Then as $\mathcal{B} - r_1^F - r_2^F - p$ is connected, it contains a path Q from $P - r_1^F - r_2^F - p$ to $F - r_1^F - r_2^F - p$. Now $P \cup Q$ contains a link of F with one end outside $\{r_1^F, r_2^F, p\}$. This contradicts Lemma 7. So (54) follows.

So $\{r_1^F, r_2^F, p\}$ is a 3-vertex cutset separating the core of F from the links in $\mathcal{O}_3^F$. As this is not an improper 3-vertex cutset, bridge $\mathcal{B}$ contains an odd circuit. Hence, by Lemma 8, $\mathcal{B}$ contains an odd path that connects p to one of r_1^F and r_2^F and that does not contain the other vertex in $\{r_1^F, r_2^F\}$. Clearly such a path contains an odd link of F with at most one end in $\{r_1^F, r_2^F\}$. As $\Lambda_1^F \cup \Lambda_2^F = \mathcal{O}_3^F$, this contradicts (54). So the lemma follows. □

LEMMA 10. *Let K be a $K_{3,3}^1$-subdivision in a clean signed graph, let Q_1 be an st-link in $\mathcal{O}_{2,1}^K$ with $s \in P_{14}^K$ and $t \in (P_{14}^K)_{sr_4^F}$, and let Q_2 be an $r_2^K p$-link of $K \cup Q_1$ with $p \in (Q_1 \cup (P_{14}^K)_{r_4^K s}) - s$. Then the unique $r_2^K r_4^K$-path P' in $(Q_1 \cup Q_2 \cup P_{14}^F) - s$ is even.*

Proof. Suppose P' is odd. If necessary resign on p such that $P' - Q_2$ is even, and contract $P' - Q_2$, $(P_{14}^K)_{r_4^K t}$ and $(P_{14}^F)_{r_1^F s}$. This yields a subdivision of the signed graph in Figure 9(a). As illustrated in Figure 9, that signed graph has a $K_{3,3}^{1,1}$-minor, a contradiction. □

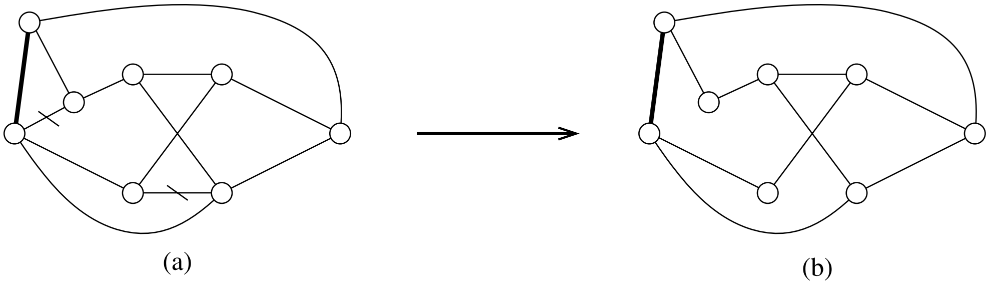


FIG. 12. *Bold edges are odd paths; thin edges are even paths. To obtain* (b) *from* (a)*, delete the "crossed" edges.*

LEMMA 11. *Let F be an extreme $K^1_{3,3}$-extension in a clean signed graph. Then $\mathcal{L}^F_{5,1} = \emptyset$ or $\mathcal{E}^F_{3,1} \cup \mathcal{E}^F_{4,1} = \emptyset$.*

Proof. Suppose this is not true. Then we may assume that there exists a p_1p_2-link $P \in \mathcal{L}^F_{5,1}$ and an r^F_1q-link $Q \in \mathcal{E}^F_{3,1} \cup \mathcal{E}^F_{4,1}$ with $p_2 \in P^F_{14}$ and that $q \in \text{CORE}(F) \cup T^F_2$. Choose $K \in \mathcal{K}_q(F)$. By 35-symmetry we may assume that $q \in P^K_{45} \cup P^K_{65} \cup P^K_{52}$. Let R be the intersection of P^K_{65} with the r^K_6q-subpath of $P^K_{45} \cup P^K_{65} \cup P^K_{52}$. By Lemma 7, P is even and disjoint with Q. Now deleting R and $(P^F_{14})_{r^F_1 p_2}$ from $K \cup P \cup Q$ yields a $K^1_{3,3}$-subdivision F' with $P^{F'}_{12} = (P^F_{12})_{r^F_1 p_1}$. As $P^{F'}_{12}$ is properly contained in P^F_{12}, this contradicts the extremeness of F. (See Figure 12 for the special case that $q = r^K_5$.) □

The results so far say that certain combinations of links cannot occur; here is a lemma that says that certain links force other ones.

LEMMA 12. *Let F be an extreme $K^1_{3,3}$-extension in a 3-connected clean signed graph with no blockvertex and no improper 3-vertex cutset. If $\mathcal{O}^F_{2,1} \cup \mathcal{L}^F_{5,1} \neq \emptyset$, then $\mathcal{L}^F_{3,1} \cup \mathcal{L}^F_{4,1} \neq \emptyset$.*

Proof. Let F be a counterexample. As $\mathcal{O}^F_{2,1} \cup \mathcal{L}^F_{5,1} \neq \emptyset$, it follows from Lemma 9 that $\mathcal{L}^F_{5,2} = \emptyset$. So, as also $\mathcal{L}^F_{3,1} \cup \mathcal{L}^F_{4,1} = \emptyset$, it follows from Lemma 7 that r^F_1 does not lie in the bridge $\mathcal{B}$ of $\{r^F_2, r^F_4, r^F_6\}$ that contains $\text{CAGE}(F) \cup T^F_2$. As $\{r^F_2, r^F_4, r^F_6\}$ is no improper 3-vertex cutset, $\mathcal{B}$ contains an odd circuit. Hence, by Lemma 8, $\mathcal{B}$ contains an odd path that has both ends in $\{r^F_2, r^F_4, r^F_6\}$ and that is disjoint from the third vertex in $\{r^F_2, r^F_4, r^F_6\}$. Such a path contains an odd link of F. By Lemma 7, that odd link is in $\mathcal{O}^F_{2,2} \cup \mathcal{O}^F_{3,2}$. As that contradicts Lemma 9, the lemma follows. □

LEMMA 13. *Let F be an extreme $K^1_{3,3}$-extension in a 3-connected clean signed graph that has no blockvertex and no improper 3-vertex cutset. If $Q \in \mathcal{O}^F_{2,1}$ with ends on P^F_{1j} with $j = 4, 6$ and $P \in \mathcal{O}^F_{4,2}$, then P intersects $Q \cup P^F_{1j}$.*

Proof. Let P and Q be as indicated. Assume P and $Q \cup P^F_{1j}$ do not intersect. By Lemma 9, $\mathcal{O}^F_{3,1} \cup \mathcal{O}^F_{4,1} = \emptyset$, and thus, by Lemma 12, $\mathcal{E}^F_{3,1} \cup \mathcal{E}^F_{4,1} \neq \emptyset$. Let $R \in \mathcal{E}^F_{3,1} \cup \mathcal{E}^F_{4,1}$. Then, by Lemma 7, R is internally vertex disjoint with P and Q. Hence, we have the signed graph in Figure 13(a) as a minor. As indicated in Figure 13 that signed graph has a $K^{1,2}_{3,3}$-minor, a contradiction. □

10. Handles. A *handle* of a $K^1_{3,3}$-extension F is a link in $\mathcal{O}^F_2$ with no end in $\{r^F_1, r^F_2\}$. The following lemma says that in a counterexample to Theorem 3(ii) each extreme $K^1_{3,3}$-extension has a handle.

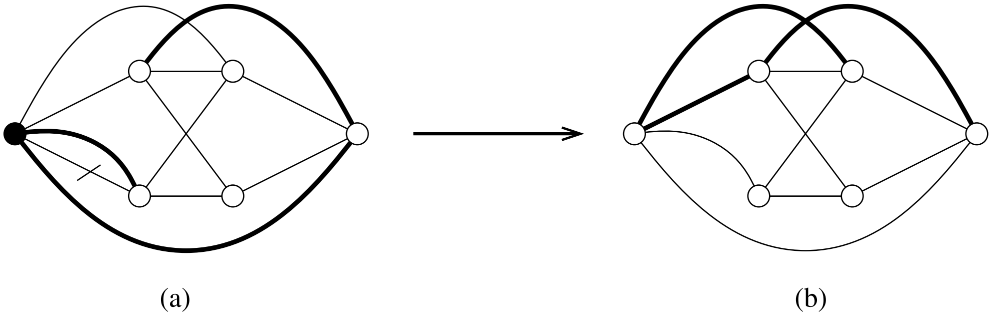


FIG. 13. *Bold edges are odd; thin edges are even. To obtain* (b) *from* (a), *resign on the black vertex and delete the "crossed" edge.*

LEMMA 14. *Each extreme $K^1_{3,3}$-extension F in a 3-connected clean signed graph (G,Σ) with no blockvertex and no improper 3-vertex cutset has a handle.*

Proof. Let (G,Σ) and F form a counterexample; thus, F has no handle. Let $B := T^F_1 \cup \text{CAGE}(F) \cup T^F_2$.

(55) $(G,\Sigma) - r^F_1 - r^F_2$ *contains an odd circuit, say, C.*

Suppose this is not true; then we may assume, by resigning, that all edges not incident with r^F_1 or r^F_2 are even. It is easy to see that this resigning can be done such that all edges in B are even. In other words $\Sigma \subseteq (\delta_G(r^F_1) \cup \delta_G(r^F_2)) - B$.

As (G,Σ) has no blockvertex, there exists an odd circuit disjoint from r^F_2. As $G - r^F_1 - r^F_2$ is connected, F has a link Q_1 that closes with $F - r^F_2$ an odd circuit. Moreover, as (55) is false, all such odd circuits go through r^F_1. So, as $\Sigma \subseteq (\delta_G(r^F_1) \cup \delta_G(r^F_2)) - B$, we have that $Q_1 \in \mathcal{L}^F_{1,1} \cup \mathcal{O}^F_{2,1} \cup \mathcal{O}^F_{3,1} \cup \mathcal{O}^F_{4,1} \cup \mathcal{L}^F_{5,1}$. By symmetry F also has a link $Q_2 \in \mathcal{L}^F_{1,2} \cup \mathcal{O}^F_{2,2} \cup \mathcal{O}^F_{3,2} \cup \mathcal{O}^F_{4,2} \cup \mathcal{L}^F_{5,2}$ that closes with $F - r^F_1$ an odd circuit.

First assume that P^F_{12} consists of a single edge. Then, $Q_1, Q_2 \notin \mathcal{L}^F_1 \cup \mathcal{L}^F_5$, so by Lemma 9 and by symmetry, we may assume that $Q_1 \in \mathcal{O}^F_{2,1}$ and $Q_2 \in \mathcal{O}^F_{4,2}$. We also may assume that Q_1 has its ends on P^F_{14}. By Lemma 13, Q_2 intersects $Q_1 \cap P^F_{14}$. From this and as $\Sigma \subseteq (\delta_G(r^F_1) \cup \delta_G(r^F_2)) - B$, one easily deduces a contradiction against Lemma 10.

So we may assume that P^F_{12} does not consist of a single edge. As G is 3-connected, $\mathcal{L}^F_5 \neq \emptyset$. So we may as well assume that $Q_1 \in \mathcal{L}^F_{5,1}$. By Lemmas 12 and 11 there exists a link $Q \in \mathcal{L}^F_{3,1} \cup \mathcal{L}^F_{4,1}$. Hence, $\mathcal{L}^F_{5,1}$ and $\mathcal{L}^F_{3,1} \cup \mathcal{L}^F_{4,1}$ are nonempty, so by Lemma 9, $\Lambda^F_2 = \emptyset$. This implies that $Q_2 \in \mathcal{L}^F_{1,2}$. By Lemma 7, Q is vertex disjoint with Q_1, and as $\mathcal{L}^F_{5,2} \subseteq \Lambda^F_2 = \emptyset$, Q is also disjoint with Q_2. Contract all edges in $P^F_{12} \cup Q_1 \cup Q_2$ that are not incident with $\{r^F_1, r^F_2\}$ and not incident with a vertex on P^F_{14}; they are all even. The resulting signed graph has the signed graph in Figure 14(a) as a minor. As illustrated in Figure 14, that signed graph has a $K^2_{3,3}$-minor. This contradiction proves (55).

We may assume that $\Lambda^F_2 = \mathcal{O}^F_{4,2}$. Indeed, by Lemma 9 and 12-symmetry we may assume that $\Lambda^F_2 = \emptyset$ or $\Lambda^F_2 = \mathcal{O}^F_{4,2}$. As by definition, $\mathcal{O}^F_{4,2}$ is contained in Λ^F_2, which means the sets are equal.

(56) *If B has an odd r^F_2p-link with $p \neq r^F_1$, then P^F_{12} is a single edge.*

Assume that P^F_{12} is not an edge. Then, as G is 3-connected, $\mathcal{L}^F_5 \neq \emptyset$. So as $\Lambda^F_2 = \mathcal{O}^F_{4,2}$, we have that $\mathcal{L}^F_{5,2} = \emptyset$, so $\mathcal{L}^F_{5,1} \neq \emptyset$. Hence, by Lemma 12, there exists a link

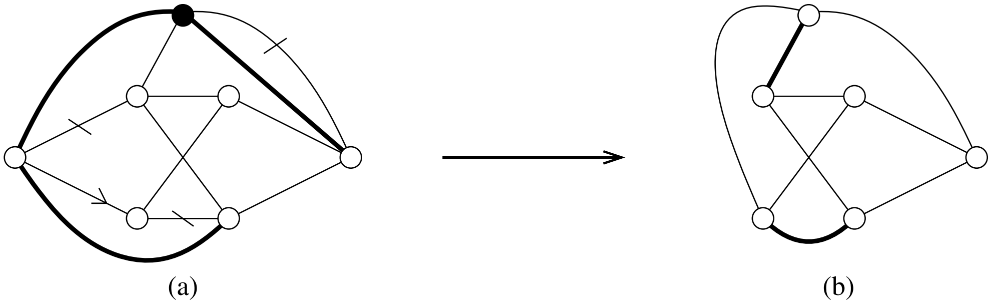


FIG. 14. *Bold edges are odd; thin edges are even. To obtain* (b) *from* (a)*, resign on the black vertex, delete the "crossed" edges, and contract the "directed" edge.*

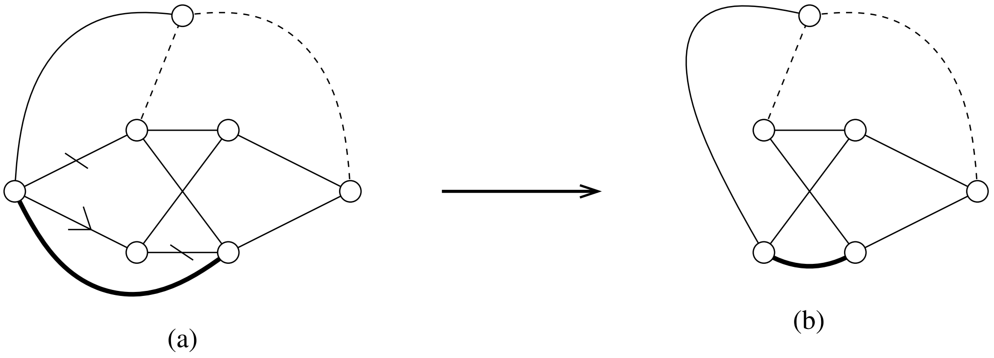


FIG. 15. *Bold edges are odd; thin edges are even; and both in* (a) *and in* (b) *exactly one of the dashed edges is odd. To obtain* (b) *from* (a)*, delete the "crossed" edges and contract the "directed" edge.*

$R \in \mathcal{L}^F_{3,1} \cup \mathcal{L}^F_{4,1}$. By Lemma 11, $R \in \mathcal{O}^F_{3,1} \cup \mathcal{O}^F_{4,1}$. Hence, by Lemma 9, $\mathcal{O}^F_{4,2} = \emptyset$, so $\Lambda^F_2 = \emptyset$.

Let P be an odd r^F_2p-link of B with $p \neq r^F_1$. As $\Lambda^F_2 = \emptyset$, path P intersects P_{12}. So P contains a link in $\mathcal{L}^F_5$; as this collection is equal to $\mathcal{L}^F_{5,1}$ we get that $p \in T^F_1$. Let Q be the shortest path on P_{12} from r^F_1 to P. As $\mathcal{L}^F_{5,2} = \emptyset$ and as P intersects P_{12}, the subgraphs R and $P \cup Q$ share no other vertex than r^F_1. Hence, (G, Σ) has a minor as in Figure 15(a), which has a $K^2_{3,3}$-minor. This contradiction proves (56).

(57) *There exists a vertex* $p \notin \{r^F_1, r^F_2\}$ *such that each path in* $G - r^F_1 - r^F_2$ *from* B *to* C *contains* p.

If not, then in $G - r^F_1 - r^F_2$ there exist two vertex disjoint paths from C to B. So B has an odd link J contained in $G - r^F_1 - r^F_2$. As F has no handle, it follows from Lemma 7 that J is not a link of F, so J intersects P_{12}. But this implies that P_{12} is not an edge and that its union with J contains an odd r^F_2p-link with $p \neq r^F_1$. As this contradicts (56), (57) follows.

Let $\mathcal{B}$ be the union of the bridges of $\{r^F_1, r^F_2, p\}$ that contain edges of B. Assume p is chosen such that $\mathcal{B}$ is as small as possible. Note that $\mathcal{B}$ is 2-connected and that $\mathcal{B} - r^F_1 - r^F_2$ is connected. Let P_1, P_2, and P_3 be three vertex disjoint paths from C to $\{r^F_1, r^F_2, p\}$. Take a path P' from p to $B - r^F_1 - r^F_2$ with no internal vertices in B;

let u be its end vertex in B.

(58) P_{12} *is a single edge.*

This follows from (56), as $C \cup P_1 \cup P_2 \cup P_3 \cup P'$ contains an odd $r^F_2 p$-link with $p \neq r^F_1$.

So each link of B, except P_{12}, is a link of F.

$C \cup P_1 \cup P_2 \cup P_3 \cup P'$ contains an odd $r^F_1 u$-link of F and an odd $r^F_2 u$-link of F. So as $\Lambda^F_2 = \mathcal{O}^F_{4,2}$, we have that $u \in T^F_1$ and thus that $\mathcal{O}^F_{2,1}$ and $\mathcal{O}^F_{4,2}$ are not empty. Hence, we have by Lemma 9 and (58) that $\Lambda^F_1 = \mathcal{O}^F_{2,1}$ and $\Lambda^F_2 = \mathcal{O}^F_{4,2}$.

(59) $\mathcal{B}$ *contains a link* P *in* $\mathcal{O}^F_{2,1} \cup \mathcal{O}^F_{4,2}$.

As $\{r^F_1, r^F_2, p\}$ is not an improper 3-vertex cutset, $\mathcal{B}$ contains as odd circuit. From this and as $\mathcal{B}$ is 2-connected, it follows that $\mathcal{B}$ contains an odd $r^F_1 r^F_2$-path, say, Q. As $\mathcal{B} - r^F_1 - r^F_2$ is connected, it contains a path R that connects $Q - r^F_1 - r^F_2$ with $B - r^F_1 - r^F_2$. The union of R and Q contains an odd link P of F that has at most one end in $\{r^F_1, r^F_2\}$. By (58), $P \in \Lambda^F_1 \cup \Lambda^F_2 = \mathcal{O}^F_{2,1} \cup \mathcal{O}^F_{4,2}$. So (59) follows.

Let q be the end of P not in $\{r^F_1, r^F_2\}$. By 46-symmetry, we may assume that $q \in P^F_{14} - r^F_1$. Take the subpath Q of P' from p to $q \in P \cup T^F_1$. Then as Q can be extended to an $r^F_1 p$-link as well as an $r^F_2 p$-link of $F \cup P$ of either parity, it is straightforward to argue from Lemma 13 that $q \in (Q \cup P^F_{14}) - r^F_1$ and from Lemma 10 that $q \in P^F_{16} - r^F_1$. This is absurd. □

11. Proof of Theorem 3(ii). We finally prove Theorem 3(ii). Assume that (G, Σ) is a 3-connected clean signed graph with no blockvertex and no improper 3-vertex cutset. Let F be an extreme $K^1_{3,3}$-extension in (G, Σ). By Lemma 14 and by 12-symmetry, we may assume that F has a handle in $\mathcal{O}^F_{2,1}$.

Let $\mathcal{F}$ be the set of all $K^1_{3,3}$-extensions F' with $P^{F'}_{12} = P^F_{12}$, $T^{F'}_2 = T^F_2$, $\textsc{cage}(F') = \textsc{cage}(F)$, and $\{r^{F'}_4, r^{F'}_6\} = \{r^F_4, r^F_6\}$; obviously each $F' \in \mathcal{F}$ is extreme.

(60) *Each* $F' \in \mathcal{F}$ *has a handle in* $\mathcal{O}^{F'}_{2,1}$.

If not, then by Lemma 14 some $F' \in \mathcal{F}$ has a handle P in $\mathcal{O}^{F'}_{2,2}$. As $\mathcal{O}^F_{2,1} \neq \emptyset$, it follows from Lemma 9 that $\mathcal{O}^F_{2,2} = \emptyset$. Hence, $P \notin \mathcal{O}^F_{2,2}$. Therefore this handle intersects $T^F_1 - r^F_1$, and thus it contains a link of F that contradicts Lemma 7. So (60) follows.

Hence, Lemma 9 implies

(61) $\Lambda^{F'}_2 = \mathcal{O}^{F'}_{4,2}$ *for each* $F' \in \mathcal{F}$.

The *tip* of a link in $\mathcal{O}^F_{2,1}$, so in particular of a handle, is the end that lies farthest from r^F_1 on T^F_1.

(62) *Let* P *be a handle of* F *with tip* s *on* P^F_{14}, *and let* $L \in \mathcal{L}^F_{2,1}$ *with ends* x *in* $(P^F_{14})_{r^F_1 s} - r^F_1 - s$ *and* y *in* $(P^F_{14})_{s r^F_4} - s$. *Then there exists a* $K^1_{3,3}$*-extension* F' *in* $\mathcal{F}$ *with* $P^{F'}_{16} = P^F_{16}$ *and* $(P^{F'}_{14})_{y r^F_4} = (P^F_{14})_{y r^F_4}$ *that has a handle with tip* y.

In proving this we clearly may assume that L consists of a path that is internally disjoint with P and possibly a part of P. If L is odd, then it is a handle of F with tip y. Hence, we may assume that L is even. We may also assume that the only odd

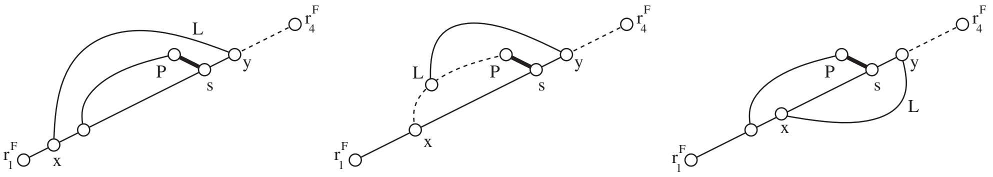


FIG. 16. *Bold edges are odd; thin edges are even; and dashed edges may have length zero.*

edge on $P \cup L$ is the edge of P incident with s. Figure 16 depicts the three possible arrangements of P and L along P^F_{14}. Let F' be the $K^1_{3,3}$-extension obtained from F by replacing $(P^F_{14})_{xy}$ with L. One easily checks in Figure 16 that F' satisfies all claims in (62).

A *single border of* F is any pair (r^F_1, s) where s is the tip of a handle. A pair (r, s) is a *linked border of* F if s is the tip of a handle and there exists an rr'-link in $\mathcal{L}^F_{6,1}$ with $r' \in (T^F_1)_{r^F_1 s} - s$; any such rr'-link is a *join for the linked border* (r, s). A pair (r, s) is a *double border of* F if r and s are both tips of a handle, one lying in P^F_{14} and the other in P^F_{16}, and there exists a link in $\mathcal{L}^F_{6,1}$ with both ends in $(T^F_1)_{rs} - r - s$; any such link is a *join for the double border* (r, s). A *border of* F is a single, linked, or double border of F. Note that if (r, s) is a border, then one among r and s lies on P^F_{14} and the other on P^F_{16}. Moreover, $s \neq r^F_1$ and $r = r^F_1$ exactly when (r, s) is a single border. Note that by Lemma 7, joins for borders are even.

If (r, s) is a border, let $B[r, s] = F - (T^F_1)_{rs}$, and let $\mathcal{L}[r, s]$ be the collection of links of F with one end in $B[r, s] - r^F_1 - r - s$ and the other end in $(T^F_1)_{rs} - r^F_1 - r - s$.

(63) *If* (r^F_1, s) *is a single border of* F *with* $\mathcal{L}[r^F_1, s] \not\subseteq \mathcal{L}^F_{2,1} \cup \mathcal{L}^F_{6,1}$, *then* $\mathcal{L}[r^F_1, s] \cap \mathcal{O}^F_{4,2} \neq \emptyset$ *and* $\Lambda^F_1 = \mathcal{O}^F_{2,1}$.

To prove this, let Q be a handle with end s, and let $P \in \mathcal{L}[r^F_1, s] \setminus (\mathcal{L}^F_{2,1} \cup \mathcal{L}^F_{6,1})$. Then, by Lemma 7, P has an end on $P^F_{12} - r^F_1$. Let P' be the shortest subpath of P from P^F_{12} to $Q \cup T^F_1$. Clearly, by changing P if necessary, we may assume that P consists of P' and possibly a subpath of Q. If P was even, (G, Σ) would have the signed graph in Figure 17(a) as a minor. As illustrated in Figure 17 that signed graph has a $K^{1,1}_{3,3}$-minor. So P is odd. As by Lemma 7, $\mathcal{O}^F_{5,1} = \emptyset$, this means that $P \in \mathcal{O}^F_{4,2}$. So $\mathcal{L}[r^F_1, s] \cap \mathcal{O}^F_{4,2} \neq \emptyset$ indeed. Moreover, as $\mathcal{O}^F_{4,2} \neq \emptyset$, it follows from Lemma 9 that $\Lambda^F_1 = \mathcal{O}^F_{2,1} \cup \mathcal{L}^F_{5,1}$. In other words, $\mathcal{O}^F_{3,1} \cup \mathcal{O}^F_{4,1} = \emptyset$. So, as $\mathcal{O}^F_{2,1} \neq \emptyset$ it follows from Lemma 12 that $\mathcal{E}^F_{3,1} \cup \mathcal{E}^F_{4,1} \neq \emptyset$. Hence, by Lemma 11, $\mathcal{L}^F_{5,1}$ is empty. Thus $\Lambda^F_1 = \mathcal{O}^F_{2,1}$ indeed, and (63) follows.

The *value for* F of a border (r, s) is defined as the number of edges in $B[r, s]$. Choose $F \in \mathcal{F}$ and a border (r, s) for F such that

(64) *the value for* F *of* (r, s) *is as small as possible.*

By 46-symmetry assume that s lies on P^F_{14} and that r lies on P^F_{16}. Then we have the following:

(65) $$\mathcal{L}[r, s] \cap \mathcal{L}^F_{2,1} = \emptyset.$$

Suppose this is not true; let $L \in \mathcal{L}[r, s] \cap \mathcal{L}^F_{2,1}$. Let x be the end of L in $(T^F_1)_{rs}$, and let y be the other end of L. If x and y lie on P^F_{14}, then by (62) there exists a

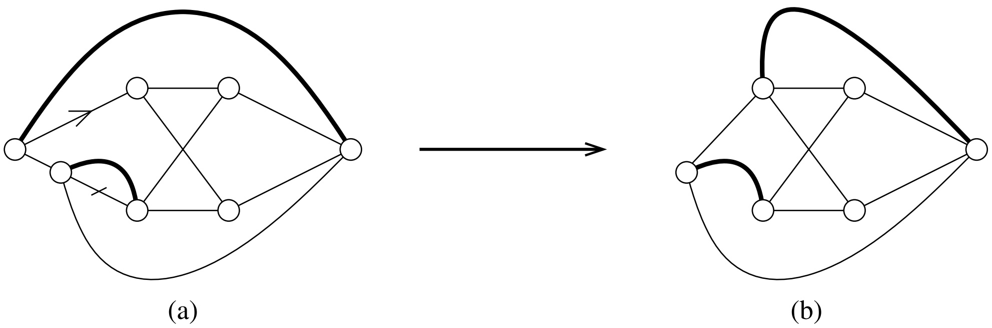


FIG. 17. *Bold edges are odd; thin edges are even. To obtain* (b) *from* (a)*, delete the "crossed" edge and contract the "directed" edge.*

$K^1_{3,3}$-extension F' such that (r, y) is a border of F'. The value for F' of (r, y) is clearly smaller than the value for F of (r, s). By (64) this is impossible, so x and y lie on P^F_{16}. In fact, by 46-symmetry and symmetry between r and s, this also means that (r, s) is not a double border. Hence, as $s \in P^F_{14}$, (r, s) is a linked border. Let P be a join for (r, s).

If L intersected P, it would do so internally and (y, s) would be a linked border for F (with a join in $L \cup P$). As the value of (y, s) is smaller than that of (r, s), it follows from (64) that this is impossible, so L and P are disjoint.

If L was odd, it would be a handle and (y, s) would be a double border, again contradicting (64). So L is even. Let F' be the $K^1_{3,3}$-extension obtained from F by replacing $(P^F_{16})_{xy}$ with L. Clearly, $F' \in \mathcal{F}$. Now (y, s) is a linked border of F'. The value for F' of (y, s) is clearly smaller than the value for F of (r, s). By (64) this is impossible, so (65) follows.

(66) $$\mathcal{L}[r, s] \cap \mathcal{L}^F_{6,1} = \emptyset.$$

Suppose this is not true; let $L \in \mathcal{L}[r, s] \cap \mathcal{L}^F_{6,1}$. Let y be the end of L in $B[r, s]$. If y lies on P^F_{16}, then (y, s) would be a linked border of F that has a smaller value than (r, s), contradicting (64) (L would be a join for that border). So, $y \in P^F_{14}$. By 46-symmetry and symmetry between r and s, this also implies that (r, s) is not a double border. Now, as $L \in \mathcal{L}^F_{6,1}$, (r, s) is a linked border; let R be a join for (r, s), and let Q be a handle with tip s. By (65), L and R are internally vertex disjoint, and by construction they do not share any end. By Lemma 7, L and R are both even. Moreover, both these paths are internally disjoint with Q; otherwise, we would have a link in $\mathcal{O}^F_{6,1}$. Now, let $K \in \mathcal{K}(F)$, and let K' be the $K^1_{3,3}$-subdivision obtained from K by replacing P^K_{45} and P^K_{63} with L and R. Then K' is extreme in $K' \cup Q$. As Q is a link of K' that violates Lemma 7 with respect to K', (66) follows.

(67) (r, s) *is a linked or double border of* F.

Suppose this is not true; then (r, s) is a single border and $r = r^F_1$. As G is 3-connected, $\{r^F_1, s\}$ is not a 2-vertex cutset, so $\mathcal{L}[r^F_1, s] \neq \emptyset$. By (65), (66), and (63), there exists an $L \in \mathcal{L}[r^F_1, s] \cap \mathcal{O}^F_{4,2}$, and $\Lambda^F_1 = \mathcal{O}^F_{2,1}$. In particular, $\mathcal{O}^F_{3,1} \cup \mathcal{O}^F_{4,1} = \emptyset$, so by Lemma 12, $\mathcal{E}^F_{3,1} \cup \mathcal{E}^F_{4,1} \neq \emptyset$. As also $\mathcal{L}^F_{5,1} = \emptyset$, it follows from (61) that $\mathcal{L}^F_5 = \emptyset$. So P^F_{12} is a single edge. From this, (65), and (66), it follows that the bridge, say, $\mathcal{B}$, of $\{r^F_1, s, r^F_2\}$ containing CAGE(F) is distinct from the bridge, say, $\mathcal{A}$, of $\{r^F_1, s, r^F_2\}$ containing

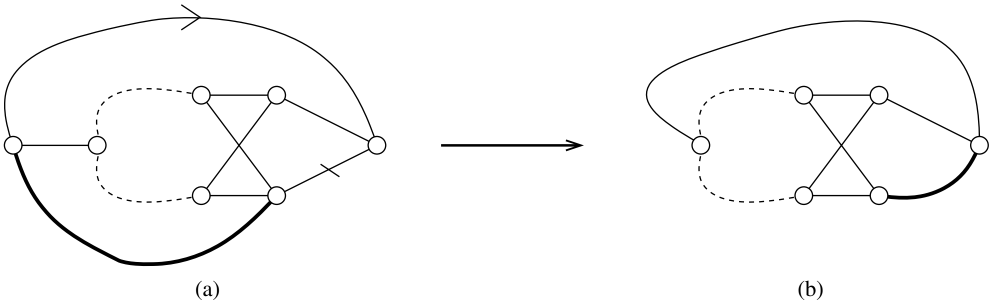


FIG. 18. *Bold edges are odd; thin edges are even; and both in* (a) *and in* (b) *exactly one of the dashed edges is odd. To obtain* (b) *from* (a)*, delete the "crossed" edge and contract the "directed" edge.*

$(T^F_{12})_{r^F_1 s}$. Hence, as (G, Σ) has no improper 3-vertex cutset, $\mathcal{B}$ is not bipartite. By Lemma 8, $\mathcal{B}$ contains an odd path P from s to one of r^F_1 and r^F_2 that misses the other vertex in $\{r^F_1, r^F_2\}$. As P^F_{12} is a single edge, P^F_{12} is not contained in $\mathcal{B}$. Therefore P contains a link $Q \in \mathcal{O}^F_{2,1} \cup \mathcal{O}^F_{4,2}$.

Let R be a handle with tip s. Then R lies in $\mathcal{A}$. As $Q \in \mathcal{B}$, links R and Q are internally disjoint. This means that if $Q \in \mathcal{O}^F_{4,2}$, then, by Lemma 13, Q has an end in P^F_{14}. However, then links Q and R contradict Lemma 10. So $Q \in \mathcal{O}^F_{2,1}$.

As L lies in $\mathcal{A}$ and Q lies in $\mathcal{B}$, these links are internally vertex disjoint. Since L has an end in P^F_{14}, it follows from Lemma 13 that Q has its ends in P^F_{14}. As Q lies in $\mathcal{B}$ its tip, say, y, lies in $(P^F_{14})_{sr^F_4} - s$. Hence, by (64), Q is not a handle. So the other end of Q is r^F_1. But then Q and L violate Lemma 10. This proves (67).

(68) *$B[r, s]$ has an odd rs-link T with the following three properties:*
T intersects $(P^F_{14})_{r^F_1 s}$ internally; r^F_1 does not lie on T; and if (r, s) is a
linked border, then T intersects P^F_{16} only in r.

Indeed such a path is contained in the union of a handle with tip s, a join for (r, s) and T^F_1.

(69) *No odd $r^F_2 w$-link of $B[r, s] \cup T$ with $w \in (T^F_2 \cup \textsc{core}(F)) - r^F_2$ contains r^F_1.*

Assume this is false; let P be an odd $r^F_2 w$-link of $B[r, s] \cup T$ with $w \in (T^F_2 \cup \textsc{core}(F)) - r^F_2$ that contains r^F_1. Let Y be the subpath of P^F_{14} from P to T. Note that by (68) Y has neither r^F_4 nor r^F_6 as one of its ends. By resigning on the vertices of Y, if necessary, we see that (G, Σ) has the signed graph in Figure 18(a) as a minor. As illustrated in Figure 18, that signed graph has $K^2_{3,3}$ as a minor. This contradiction proves (69).

(70) $$\mathcal{E}^F_{3,1} \cup \mathcal{E}^F_{4,1} = \emptyset.$$

Suppose this is false; let $P \in \mathcal{E}^F_{3,1} \cup \mathcal{E}^F_{4,1}$. Paths P and T are disjoint as otherwise F has a link that violates Lemma 7. This means that $P^F_{12} \cup P$ contradicts (69), so (70) follows.

Hence, as $\mathcal{O}^F_{2,1} \neq \emptyset$, it follows from Lemma 12 that $\mathcal{O}^F_{3,1} \cup \mathcal{O}^F_{4,1} \neq \emptyset$. Hence, by Lemma 9, $\Lambda^F_2 = \emptyset$.

(71) $$\mathcal{L}[r, s] = \emptyset.$$

Suppose this is false; let $L \in \mathcal{L}[r,s]$. By (65), (66), and Lemma 7, L has an end, say, y, on $P^F_{12} - r^F_1$. Let x be the other end of L. By the properties of T listed in (68) we may assume that if L meets T, then $x \in P^F_{14}$ (if not, we can replace L with another path in $T \cup L$ that does end in P^F_{14}). In any case, $L \in \mathcal{L}[r^F_1, t]$ for $t = s$ or $t = r$. As $\mathcal{O}^F_{3,1} \cup \mathcal{O}^F_{4,1} \neq \emptyset$, it follows from (63) that t is not the tip of a handle. So $L \in \mathcal{L}[r^F_1, r]$ and (r,s) is a linked border and, as $x \notin P^F_{14}$, the paths T and L are vertex disjoint. Moreover, as $\mathcal{O}^F_{4,2}$ and $\mathcal{O}^F_{5,1}$ are both empty, L is even. Hence, the concatenation of $(P^F_{12})_{r^F_2 y}$, L, $(P^F_{16})_{x r^F_1}$, and any link in $\mathcal{O}^F_{3,1} \cup \mathcal{O}^F_{4,1}$ violates (69). So (71) follows.

As $\{r, r^F_1, s\}$ is not an improper 3-vertex cutset, there exists a link Q of F that closes with $B[r,s]$ an odd circuit. As $\mathcal{E}^F_{3,1} \cup \mathcal{E}^F_{4,1} = \Lambda^F_2 = \mathcal{O}^F_{5,1} = \emptyset$, link $Q \in \mathcal{L}^F_{2,1} \cup \mathcal{L}^F_1$. By (64), Q cannot be in $\mathcal{O}^F_{2,1}$. If $Q \in \mathcal{L}^F_1$, then as Q closes with B an odd circuit, $L \cup P^F_{12}$ contains an even $r^F_2 r^F_1$-path, which together with any link in $\mathcal{O}^F_{3,1} \cup \mathcal{O}^F_{4,1}$ forms a link violating (69). So $Q \in \mathcal{E}^F_{2,1}$. As Q closes with $B[r,s]$ an odd circuit, r^F_1 is an end of Q. Let q be the other end of Q. Let u be the vertex among r and s that is farthest from q along T^F_1. Let F^* be the $K^1_{3,3}$-extension in $\mathcal{F}$ obtained from F by replacing $(T^F_1)_{r^F_1 q}$ with Q. Vertex u is not the tip of a handle of F, as otherwise (q,u) is a linked border of F^* that has a smaller value than (r,s) has. So u is r, border (r,s) is linked, and q lies on P^F_{14}. By (71), Q and T are disjoint. Hence, by the last property of T listed in (68), $T \cup (P^F_{14})_{sq} \in \mathcal{O}^{F^*}_{6,1}$. This contradicts Lemma 7, which completes the proof of Theorem 3(ii). □

Acknowledgment. We thank Romeo Rizzi for comments on an early draft of this manuscript.

REFERENCES

[1] F. Barahona, *Planar multicommodity flows, max cut and the Chinese postman problem*, in Polyhedral Combinatorics, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 1, W. Cook and P. D. Seymour, eds., American Mathematical Society, Providence, R.I., 1990, pp. 189–202.

[2] J. F. Geelen and B. Guenin, *Packing odd circuits in Eulerian graphs*, J. Combin. Theory Ser. B, 86 (2002), pp. 280–295.

[3] A. M. H. Gerards, *Graphs and Polyhedra—Binary Spaces and Cutting Planes*, CWI Tract 73, Centrum Wisk. Inform., Amsterdam, 1990.

[4] A. M. H. Gerards, *Multicommodity flows and polyhedra*, CWI Quart., 6 (1993), pp. 281–296.

[5] B. Guenin, *A characterization of weakly bipartite graphs*, J. Combin. Theory Ser. B, 83 (2001), pp. 112–168.

[6] B. Guenin, *personal communication*, 2004.

[7] D. W. Hall, *A note on primitive skew curves*, Bull. Amer. Math. Soc., 49 (1943), pp. 935–937.

[8] T. R. Jensen and B. Toft, *Graph Coloring Problems*, John Wiley and Sons, New York, 1995.

[9] B. Rothschild and A. Whinston, *On two commodity network flows*, Oper. Res., 14 (1966), pp. 377–387.

[10] P. D. Seymour, *The matroids with the max-flow min-cut property*, J. Combin. Theory Ser. B, 23 (1977), pp. 189–222.

[11] P. D. Seymour, *Decomposition of regular matroids*, J. Combin. Theory Ser. B, 28 (1980), pp. 305–359.

[12] P. D. Seymour, *Matroids and multicommodity flows*, European J. Combin., 2 (1981), pp. 257–290.

[13] K. Truemper, *Max-flow min-cut matroids: Polynomial testing and polynomial algorithms for maximum flow and shortest routes*, Math. Oper. Res., 12 (1987), pp. 72–96.

SIAM J. Discrete Math.
Vol. 21, No. 2, pp. 303–312

A MAJORIZATION BOUND FOR THE EIGENVALUES OF SOME GRAPH LAPLACIANS[*]

TAMON STEPHEN[†]

Abstract. Grone and Merris conjectured that the Laplacian spectrum of a graph is majorized by its conjugate vertex degree sequence. In this paper, we prove that this conjecture holds for a class of graphs, including trees. We also show that this conjecture and its generalization to graphs with Dirichlet boundary conditions are equivalent.

Key words. graph Laplacian, majorization, graph spectrum, degree sequence, Dirichlet Laplacian

AMS subject classifications. 05C50, 05C07

DOI. 10.1137/040619594

1. Introduction.

One way to extract information about the structure of a graph is to encode the graph in a matrix and study the invariants of that matrix, such as the spectrum. In this paper, we study the spectrum of the combinatorial Laplacian matrix of a graph.

The *combinatorial Laplacian* of a simple graph $G = (V, E)$ on the set of n vertices $V = \{v_1, \dots, v_n\}$ is the $n \times n$ matrix $L(G)$ defined by

$$L(G)_{ij} = \begin{cases} \deg(v_i) & \text{if } \ i = j, \\ -1 & \text{if } \ \{i, j\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Here $\deg(v)$ is the *degree* of v, that is, the number of edges on v. The matrix $L(G)$ is positive semidefinite, and so its eigenvalues are real and nonnegative. We list them in nonincreasing order and with multiplicity:

$$\lambda_1(L(G)) \geq \lambda_2(L(G)) \geq \cdots \geq \lambda_{n-1}(L(G)) \geq \lambda_n(L(G)) = 0.$$

When the context is clear, we can write $\lambda_i(G)$, or simply λ_i. We abbreviate the sequence of n eigenvalues as $\lambda(L(G))$.

We are interested in the Grone–Merris (GM) conjecture that the spectrum $\lambda(L(G))$ is majorized by the conjugate partition of the nonincreasing sequence of vertex degrees of G [5]. This question is currently being studied, see, for example, [4], but has yet to be resolved. We extend the class of graphs for which the conjecture is known to hold to include trees, among other graphs. We also show that if GM holds for graph Laplacians, it also holds for more general Dirichlet Laplacians (cf. [2]) as conjectured by Duval [3].

[*] Received by the editors November 24, 2004; accepted for publication (in revised forms) October 19, 2006; published electronically April 13, 2007. An extended abstract of the results in this paper appeared in *Proceedings of the European Conference on Combinatorics, Graph Theory, and Applications*, Berlin, Germany, 2005.
http://www.siam.org/journals/sidma/21-2/61959.html

[†] Department of Mathematics, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada (tamon@sfu.ca).

2. Background and definitions.

2.1. Graphs. Given a graph $G = (V, E)$ with $n = |V|$ vertices and $m = |E|$ edges, there are several ways to represent G as a matrix. There is the *edge-incidence matrix*, an $n \times m$ matrix that records in each column the two vertices incident on a given edge. For directed graphs, we can consider a signed edge-incidence matrix:

$$\partial(G)_{ve} = \begin{cases} 1 & \text{if } v \text{ is the head of edge } e, \\ -1 & \text{if } v \text{ is the tail of edge } e, \\ 0 & \text{otherwise.} \end{cases}$$

There is also an $n \times n$ matrix $A(G)$ called the *adjacency matrix*, which is defined by

$$A(G)_{ij} = \begin{cases} 1 & \text{if } \{i, j\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

The diagonal of $A(G)$ is zero.

We can encode the (vertex) degree sequence of G in nonincreasing order as a vector $d(G)$ of length n and in an $n \times n$ matrix $D(G)$ whose diagonal is $d(G)$ and whose off-diagonal elements are 0. Then the combinatorial Laplacian $L(G)$ that we study is simply $D(G) - A(G)$. It is easy to check that if we (arbitrarily) orient G and consider the matrix $\partial(G)$ above, we also have $L(G) = \partial(G)\partial(G)^t$.

The *complement* of a graph $G = (V, E)$ is the graph $\overline{G}$ on V whose edges are exactly those not included in G.

Remark 2.1. The Laplacian is sometimes defined with entries *normalized* by dividing by the square roots of the degrees. However, we do not do that here.

2.2. Majorization. We recall that a *partition* $p = p(i)$ is a nonincreasing sequence of natural numbers, and its *conjugate* is the sequence $p^T(j) := |\{i : p(i) \leq j\}|$. Then p^T has exactly $p(1)$ nonzero elements. When convenient, we can add or drop trailing zeros in a partition. For nonincreasing real sequences s and t of length n, we say that s is *majorized* by t (denoted $s \trianglelefteq t$) if, for all $k \leq n$,

$$\sum_{i=1}^{k} s_i \leq \sum_{i=1}^{k} t_i \tag{2.1}$$

and

$$\sum_{i=1}^{n} s_i = \sum_{i=1}^{n} t_i. \tag{2.2}$$

The concept of majorization extends to vectors by comparing the nonincreasing vectors produced by sorting the elements of the vector into nonincreasing order. Given a vector v, call the sorted vector v' which contains the elements of v sorted in nonincreasing order (with multiplicity) sort(v).

In the context of majorization of unsorted vectors, we will often want to refer to the *concatenation* of two vectors x and y (i.e., the vector which contains the elements of x followed the elements of y). This is denoted x, y, as, for example, in Lemma 2.3.

There is a rich theory of majorization inequalities which occur throughout mathematics; see, for example, [9]. Matrices are an important source of such inequalities. Notably, the relationship between the diagonal and spectrum of a Hermitian matrix is characterized by majorization; see, for example, [7].

We will use the following lemmas about majorization, which can be found in [9].

LEMMA 2.2. *If x and y are vectors and P is a doubly stochastic matrix and $x = Py$, then $x \trianglelefteq y$.*

This yields two simple corollaries.

LEMMA 2.3. *For any vectors $x \trianglelefteq y$ and any vector z, we have $x, z \trianglelefteq y, z$.*

LEMMA 2.4. *If x and y are nonincreasing sequences, and $x = y$ except that at indices $i < j$ we have $x_i = y_i - a$ and $x_j = y_j + a$, where $a \geq 0$, then $x \trianglelefteq y$.*

Lemma 2.4 says that for nonincreasing sequences, transferring units from lower to higher indices reduces the vector in the majorization partial order. In particular, if x, x', y, y' are all *nonincreasing* sequences, $x' \trianglelefteq x$ and $y' \trianglelefteq y$, then

$$x' + y' \trianglelefteq x' + y \trianglelefteq x + y. \tag{2.3}$$

LEMMA 2.5. *Let A and B be positive semidefinite (more generally, Hermitian) matrices. Then*

$$\lambda(A), \lambda(B) \trianglelefteq \lambda(A + B).$$

This is Theorem G.1.b in Chapter 9 of [9], although the majorization is reversed in the printing available to the author.

LEMMA 2.6. *For positive semidefinite (more generally, Hermitian) matrices A and B,*

$$\lambda(A + B) \trianglelefteq \lambda(A) + \lambda(B).$$

This is a theorem of Fan (Theorem G.1 in Chapter 9 of [9]).

LEMMA 2.7. *Let A be an $m \times n$ $0 - 1$ (or incidence) matrix with row sums $r_1, \ldots, r_m$ and columns sums $c_1, \ldots, c_n$ both indexed in nonincreasing order. Let r^T be the conjugate of the partition $(r_1, \ldots, r_m)$ and c be the partition $(c_1, \ldots, c_n)$. Then*

$$c \trianglelefteq r^T. \tag{2.4}$$

This is known as the Gale–Ryser theorem (Theorem C.1 in Chapter 9 of [9]).

2.3. The GM conjecture. In the notation of this section, the GM conjecture is

$$\lambda(G) \trianglelefteq d^T(G). \tag{2.5}$$

Note that

$$\sum_{i=1}^{n} d_i^T = \sum_{i=1}^{n} d_i = \mathrm{trace}(L(G)) = \sum_{i=1}^{n} \lambda_i.$$

If we ignore isolated vertices (which contribute only zero entries to λ and d), we will have $d_1^T = n$. Using this fact, it is possible to show that

$$\lambda_1 \leq d_1^T. \tag{2.6}$$

Three short proofs of this are given in [4]. The authors then continue to prove the second majorization inequality:

$$\lambda_1 + \lambda_2 \leq d_1^T + d_2^T. \tag{2.7}$$

However, their proof would be difficult to extend.

There are several other facts which fit well with the GM conjecture. One is that if the GM conjecture holds, then the instances where (2.5) holds with equality are well understood; these would be the threshold graphs of section 3.1. Also, since d and λ are, respectively, the diagonal and spectrum of $L(G)$, we have $d \trianglelefteq \lambda$. Combining this with GM gives $d \trianglelefteq d^T$, a fact that has been proved combinatorially. We refer the reader to [4] for further discussion.

3. GM on classes of graphs. In this section, we give further evidence for the GM conjecture by remarking that it holds for several classes of graphs, including threshold graphs, regular graphs, and trees.

3.1. Threshold graphs. The GM conjecture was originally formulated in the context of *threshold* graphs, which are a class of graphs with several extremal properties. An introduction to threshold graphs is [8]. Threshold graphs are the graphs that can be constructed recursively by adding isolated vertices and taking graph complements. It turns out that they are also characterized by degree sequences: the convex hull of possible (unordered) degree sequences of an n vertex graph defines a polytope. The extreme points of this polytope are the degree sequences that have a unique labelled realization, and these are exactly the threshold graphs.

Threshold graphs are interesting from the point of view of spectra. Both Hammer and Kelmans [6] and Grone and Merris [5] investigated the question of which graphs have integer spectra. They found that threshold graphs are one class of graphs that have integer spectra and showed for these graphs that $\lambda(G) = d^T(G)$. We could interpret the GM conjecture as saying that threshold graphs are extreme in terms of spectra and that these extreme spectra can be understood as conjugate degree sequences.

3.2. Complements. Threshold graphs are built using graph complements of existing graphs, and so it is not surprising that the GM conjecture behaves well under taking complements. Indeed, the relationship between $\lambda(G)$ and $\lambda(\overline{G})$ is the same as between $d_n^T(G)$ and $d_n^T(\overline{G})$. For a graph G with n vertices, the ith largest vertex of G is the $(n-i)$th largest vertex of $\overline{G}$, and we have $d_i(G) = n - 1 - d_{n-i}(\overline{G})$. Translating this to the conjugate partition d^T yields $d_i^T(G) = n - d_{n-1-i}^T(\overline{G})$ with $d_n^T(G) = d_n^T(\overline{G}) = 0$.

Now notice that $L(G) + L(\overline{G}) = nI_n - J_n$, where J_n is the $n \times n$ matrix of ones. The matrix $nI_n - J_n$ sends the special eigenvector e_n (n ones) to 0 and acts as the scalar n on $e_n^{\perp}$. Both $L(G)$ and $L(\overline{G})$ also send e_n to 0, giving us $\lambda_n(G) = \lambda_n(\overline{G}) = 0$. Since $L(G)$ and $L(\overline{G})$ sum to nI_n on $e_n^{\perp}$, they have the same set of eigenvectors on $e_n^{\perp}$, and for each eigenvector, the corresponding eigenvalues for $L(G)$ and $L(\overline{G})$ sum to n. Thus $\lambda_i(G) = n - \lambda_{n-1-i}(\overline{G})$. As a consequence, GM holds for G if and only if GM holds for $\overline{G}$.

3.3. Regular and nearly regular graphs. For some small classes of graphs, it can be easily shown that the GM conjecture holds. Consider a k-regular graph G on n vertices (in a *k-regular* graph, all vertices have degree k). Then the degree sequence $d(G)$ is k repeated n times, and its conjugate $d^T(G)$ is n repeated k times followed by $n-k$ zeros. Thus d^T majorizes every nonnegative sequence of sum kn whose largest term is at most n, and, in particular, $\lambda \trianglelefteq d^T$. Indeed, this proof shows that GM holds for what we might call *nearly regular* graphs, that is, graphs whose vertices have degree either k or $(k-1)$.

3.4. Graphs with low maximum degree. Using facts about the initial GM inequalities, we can prove that GM must hold for graphs with low maximal degree. For example, if a graph has maximum vertex degree 2, then $d_3^T = d_4^T = \cdots = d_n^T = 0$, and so, for $k = 2, 3, \ldots, n$,

$$\sum_{i=1}^{k} \lambda_i \le \sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} d_i^T = \sum_{i=1}^{k} d_i^T .$$

More generally, the GM inequalities for $k \ge \text{max\_deg}(G)$ hold trivially. Thus GM holds for graphs of maximum degree 2 by (2.6). Using Duval and Reiner's result (2.7), we get that GM holds for graphs of maximum degree 3.

3.5. Trees and more. It is tempting to try to prove GM inductively by breaking graphs into simpler components on which GM clearly holds. In this section, we show that if G is "almost" the union of two smaller graphs on which GM holds, then GM holds for G as well. We apply this construction to show that GM holds for trees.

Take two graphs $A = (V_A, E_A)$ and $B = (V_B, E_B)$ on disjoint vertex sets V_A and V_B. Define their *disjoint sum* to be $A+B = (V_A \cup V_B, E_A \cup E_B)$. Assuming V_A and V_B are not empty, this is a disconnected graph. Now take two graphs $G = (V, E_G)$ and $H = (V, E_H)$ on the same vertex set V. Define their *union* as $G \cup H = (V, E_G \cup E_H)$.

Given the spectra and conjugate degree sequences of A and B, the spectrum of $A+B$ is (up to ordering) $\lambda(A+B) = (\lambda(A), \lambda(B))$, while the conjugate degree sequence of $A+B$ is $d^T(A+B) = d^T(A) + d^T(B)$ (taking each vector to have length n). Then if $\lambda(A) \trianglelefteq d^T(A)$ and $\lambda(B) \trianglelefteq d^T(B)$, we see that

$$\lambda(A+B) \trianglelefteq \lambda(A) + \lambda(B) \trianglelefteq d^T(A) + d^T(B) = d^T(A+B),$$

where the first majorization follows from Lemma 2.6 and the second from (2.3).

In a typical situation, where neither A or B is very small, we would expect the majorization $\lambda(A+B) \trianglelefteq d^T(A+B)$ to hold with considerable slack. We can use this slack to show that if we add a few more edges to $A+B$, the majorization will still hold.

THEOREM 3.1. *Take graphs A or B on disjoint vertex sets V_A and V_B. Let $G = A + B$, and on $V = V_A \cup V_B$ let C be a graph of "new edges" between V_A and V_B. Assume that GM holds on A, B, and C, i.e., that $\lambda(A) \trianglelefteq d^T(A)$, $\lambda(B) \trianglelefteq d^T(B)$, and $\lambda(C) \trianglelefteq d^T(C)$. Additionally, assume that $d_i^T(C) \le d_i^T(A), d_i^T(B)$ for all i and that $d_1^T(B) \le d_m^T(A)$, where m is the largest nonzero index of $d^T(C)$ (equivalently, m is the maximum vertex degree in C). Let $H = C \cup G$. Then*

$$\lambda(H) \trianglelefteq d^T(H). \tag{3.1}$$

Proof. Our strategy is to understand $d^T(H)$ in terms of the conjugate degree sequences of its constituent graphs. In particular, we show that

$$\text{sort}(d^T(A), d^T(B)) + d^T(C) \trianglelefteq d^T(H). \tag{3.2}$$

Then we can apply the majorizations of λ by d^T for A, B, C to the above terms and apply (2.3) to get

$$\begin{gathered}\operatorname{sort}(\lambda(A), \lambda(B)) + \lambda(C) \ \trianglelefteq\ \operatorname{sort}(d^T(A), \lambda(B)) + \lambda(C)\\ \trianglelefteq\ \operatorname{sort}(d^T(A), d^T(B)) + \lambda(C) \ \trianglelefteq\ \operatorname{sort}(d^T(A), d^T(B)) + d^T(C) \trianglelefteq\ d^T(H).\end{gathered}$$

The two terms on the left-hand side of this equation are spectra of $L(G)$ and $L(C)$, respectively. Hence by Lemma 2.6 their sum majorizes the spectrum of $L(H) = L(G) + L(C)$:

$$\lambda(H) \trianglelefteq \lambda(G) + \lambda(C) \trianglelefteq d^T(H).$$

It remains to prove (3.2), which is a statement entirely about conjugate degree sequences. For convenience, we will use the terminology of Ferrer's diagrams to describe these nonincreasing nonnegative integer sequences. That is, if s is such a sequence, we will describe s as a diagram of rows and columns with row i (reading from top to bottom) of length $s(i)$ and hence column j (reading from left to right) of length $s^T(j)$.

We begin with the following.

Claim 3.2. *Let k be the larger of* $\text{max\_deg}(A)$ *and* $\text{max\_deg}(B)$*. We have*

$$(d_1^T(G), d_2^T(G), \ldots, d_k^T(G), d_1^T(C), \ldots, d_m^T(C)) \trianglelefteq d^T(H).$$

Proof of claim. The term on the right-hand side is the concatenation of two partitions, $d^T(G)$ and $d^T(C)$. The columns of $d^T(G)$ index the vertices of G, and the length of a column gives the degree of the corresponding vertex. Since this claim is purely about degree sequences, we introduce a series of intermediate "partial graphs" where edges are allowed to have only one end. Degree sequences and their conjugates are still well defined for such objects.

Consider two copies of V, calling them V^1 and V^2. Take $G_0 = G$ on V^1 and $C_0 = C$ on V^2. Let $l = 2\,\text{max\_deg}(C)$. For $i = 1, 2, \ldots, l$, define graphs G_i and C_i by moving one end of one edge from each nonisolated vertex of C_{i-1} on V^1 to V^2. That is, let G_i be G_{i-1} plus these additional ends of edges, and let C_i be C_{i-1} with these ends of edges removed. Then we will have $G_l = H$, and C_l will be the empty graph 0_{V^2} on V^2.

We can now prove the claim via a chain of l majorizations:

$$d^T(G), d^T(C) = d^T(G_0), d^T(C_0) \trianglelefteq \cdots \trianglelefteq d^T(G_l), d^T(C_l) = d^T(H), d^T(0_{V^2}) = d^T(H)$$

if we can show that, for each $i = 1, 2, \ldots, l$,

$$d^T(G_{i-1}), d^T(C_{i-1}) \trianglelefteq d^T(G_i), d^T(C_i). \tag{3.3}$$

Compare the partitions in (3.3): The first row of $d^T(C_{i-1})$ on the left-hand side is removed, and each element from that row is inserted into a separate column of $d^T(G_{i-1})$ (representing a distinct vertex) to get $d^T(G_i)$. Where there are columns of equal length in $d^T(G_{i-1})$, they should be ordered so that those acquiring new elements come first. To see that this operation increases (or leaves unchanged) the partition in the majorization partial order, observe that after ignoring the (unchanged) contents of $d^T(C_i)$ it is equivalent to sorting the new row into the partition $d^T(G_{i-1})$, using Lemma 2.4 to move its final (rightmost) element to the proper column and repeating as necessary.

This completes the proof of the Claim 3.2. We note that

$$\begin{aligned} d^T(G) &= d^T(A) + d^T(B) \\ &= (d_1^T(A) + d_1^T(B), d_2^T(A) + d_2^T(B), \ldots, d_k^T(A) + d_k^T(B), 0, \ldots, 0), \end{aligned}$$

and hence

$$(d_1^T(A) + d_1^T(B), d_2^T(A) + d_2^T(B), \ldots, d_k^T(A) + d_k^T(B), d_1^T(C), \ldots, d_m^T(C)) \trianglelefteq d^T(H).$$

If we sort the vector on the left into nonincreasing order, the first m terms will remain fixed by the assumptions that $d_m^T(A) \geq d_1^T(B) \geq d_1^T(C)$. Since we have assumed that $d_i^T(C) \leq d_i^T(B)$ for all i, we can apply Lemma 2.4 to the reordered sequence to get

$$\begin{aligned} (d_1^T(A) + d_1^T(C), &d_2^T(A) + d_2^T(C), \ldots, d_m^T(A) + d_m^T(C), \\ &d_{m+1}^T(A), \ldots, d_k^T(A), d_1^T(B), \ldots, d_k^T(B)) \trianglelefteq d^T(H). \end{aligned}$$

The right-hand term decomposes as

$$(d_1^T(A), \ldots, d_k^T(A), d_1^T(B), \ldots, d_k^T(B)) + (d_1^T(C), \ldots, d_m^T(C), 0, \ldots, 0).$$

Since we assume $d_m^T(A) \geq d_1^T(B)$, the first m entries of $(d^T(A), d^T(B))$ will remain unchanged if the vector is sorted. This gives (3.2) and completes the proof of Theorem 3.1. □

More generally, we could replace the conditions in the statement of Theorem 3.1 with the condition (3.2), which can be checked combinatorially. The conditions in the theorem statement and (3.2) are most likely to be satisfied if C is small relative to A and B.

A useful case is when C consists of k disjoint edges. Then $m = 1$ and $d_1^T(C) = 2k$. Without loss of generality, we can take $d_1(A) \geq d_1(B)$, and the only condition that we will need to check is that $d_1(A), d_1(B) \geq d_1(C)$; i.e., both A and B must have at least $2k$ nonisolated vertices.

The strategy for applying Theorem 3.1 to show that a given graph H satisfies GM is to find a "cut" C for it that contains few edges and divides H into relatively large components. For example, we have the following result.

Corollary 3.3. *The GM conjecture holds for trees.*

Proof. Proceed by induction on the diameter of the graph. If T has diameter 1 or 2, then there is a vertex v which is the neighbor of all the remaining vertices, and T is a threshold graph. So GM holds with equality for T.

Otherwise, we can find some edge e that does not have a leaf vertex. Since T is a tree, e is a cut edge and divides T into two nontrivial connected components, A and B. We apply induction to A and B and apply Theorem 3.1 to $H = (A + B) \cup C$, where C is the graph on the vertex set of T containing the single edge e. □

Remark 3.4 (small graphs). The facts in this section allow us to check that GM holds for some small graphs without directly computing eigenvalues. For example, since the GM condition is closed under complement (see section 3.2) for graphs on up to 5 vertices, it is enough to observe that either G or $\overline{G}$ has maximum degree ≤ 3. Out of 156 graphs on 6 vertices, 146 can be decomposed into smaller graphs $(A + B) \cup C$ using Theorem 3.1. Calculating the eigenvalues of the remaining 10 does not yield a counterexample.

For any particular larger graph G, we could attempt to verify that GM holds for G by breaking G (or $\overline{G}$) into smaller graphs across cuts that have relatively few edges and applying Theorem 3.1.

4. Simplices and pairs. The most recent work relating to the GM conjecture has been to study the spectra of more general structures than graphs, such as simplicial complexes and simplicial family pairs. In this section, we show that the generalization of GM to graphs with Dirichlet boundary conditions is equivalent to the original conjecture and may be useful in approaching GM.

4.1. Simplicial complexes. In [4], the authors look at *simplicial complexes*, which are higher-dimensional analogues of simple graphs (see, for example, [10]). A set of faces of a given dimension i is called an *i-family.* Given a simplicial complex Δ, we can denote the i-family of all faces in Δ of dimension i as $\Delta^{(i)}$. For example, a graph is a one-dimensional complex, and its edge set is the 1-family $\Delta^{(1)}$. Define the degree sequence d of an i-family to be the list of the numbers of i-faces from the family incident on each vertex and sorted into nonincreasing order. We can then define $d(\Delta, i)$ as the degree sequence of $\Delta^{(i)}$, which we can abbreviate to $d(\Delta)$ or d when the context is clear.

We define the *chain group* $C_i(\Delta)$ of formal linear combinations of elements of $\Delta^{(i)}$ and generalize the signed incidence matrix ∂ of section 2.1 to a signed boundary map $\partial_i : C_i(\Delta) \to C_{i-1}(\Delta)$. This allows us to define a *Laplacian* on $C_i(\Delta)$, namely $L_i(\Delta) = \partial_i \partial_i^T$, and study its corresponding spectrum $s(\Delta, i)$, sometimes abbreviated $s(\Delta)$ or s.

Duval and Reiner [4] looked at *shifted* simplicial complexes, which are a generalization of threshold graphs to complexes. They showed that for a shifted complex Δ and any i, we have $s(\Delta, i) = d^T(\Delta, i)$. They then conjectured that GM also holds for complexes, i.e., that for any complex and any i we have

$$s(\Delta, i) \trianglelefteq d^T(\Delta, i). \tag{4.1}$$

They also show that some related facts, such as (2.6), generalize to complexes.

4.2. Simplicial pairs. In [3], Duval continues by studying *relative (family) pairs* (K, K'), where the set $K = \Delta^{(i)}$ for some i is taken modulo a family of $(i-1)$-faces $K' \subseteq \Delta^{(i-1)}$. When $K' = \emptyset$, this reduces to the situation of the previous section.

Remark 4.1. In the case $i = 1$, this is the edge set of a graph (K) with a set of *deleted* boundary vertices K'. An edge attached to a deleted vertex will not be removed—it remains as part of the pair, but we now think of the edge as having a hole on one (or both) ends.

This type of graph with a boundary appears in conformal invariant theory. In this language, the relative Laplacian of an (edge, vertex) pair is sometimes referred to as a *Dirichlet Laplacian* and its eigenvalues as *Dirichlet eigenvalues*; see, for example, [2]. Recently [1] used the spectrum of the Dirichlet Laplacian in the analysis of "chip-firing games," which are processes on graphs that have an absorbing (Dirichlet) boundary at some vertices.

We can form chain groups $C_i(K)$ and $C_{i-1}(K, K')$ and use these to define a (signed) boundary operator on the pair $\partial(K, K') : C_i(K) \to C_{i-1}(K, K')$. Hence we get a Laplacian for family pairs $L(K, K') = \partial(K, K')\partial(K, K')^T$. Considered as a matrix, $L(K, K')$ will be the principal submatrix of $L(K)$ whose rows are indexed by the i-faces in $\Delta^{(i-1)} - K'$. Finally, we get a spectrum $s(K, K')$ for family pairs from the eigenvalues of $L(K, K')$.

Duval defines the degree $d_v(K, K')$ of vertex v (in the case of a graph, v is allowed to be in K') relative to the pair (K, K') as the number of faces in K that contain v such that $K - \{v\}$ is in $\Delta^{(i-1)} - K'$. This allows him to define the degree sequence

$d(K, K')$ for pairs and to conjecture that GM holds for relative pairs:

$$s(K, K') \trianglelefteq d^T(K, K'). \tag{4.2}$$

4.3. The GM conjecture for relative pairs. It turns out that at least in the case of (edge, vertex) pairs that (4.2) follows from the original GM conjecture for graphs.

THEOREM 4.2. *GM for graphs ⇒ GM for (edge, vertex) pairs.*

Proof. Let $G = (V, E)$ be a graph with $D \subseteq V$ a set of "deleted" vertices. Let $U = V - D$ be the remaining "undeleted" vertices. We will assume that GM holds only on the undeleted part of the graph, i.e., $G|_U$. So we have $s(G|_U) \trianglelefteq d^T(G|_U)$. We can ignore the edges in $G|_D$ completely, since they have no effect on either $s(E, D)$ or $d(E, D)$. The remaining edges connect vertices in D to vertices in U. Define G' to be the graph on V whose edge are exactly the edges of G between D and U. Let a be the degree sequence of the deleted vertices in G' and b be the degree sequence of the undeleted vertices in G'.

We can compute $d^T(E, D)$ in terms of the degree sequences and spectra of $G|_U$, G', and $G|_D$, since $d_i^T(E, D)$ is the number of vertices (deleted or not) attached to at least i nondeleted vertices. The number of such vertices in U will be $d_i^T(G|_U)$, and the number in D will be $d_i^T(G') = a^T$. Hence $d^T(E, D) = d_i^T(G|_U) + a^T$.

Now consider the Laplacian $L(E, D)$. This is the submatrix of $L(G)$ indexed by U. An edge (i, j) in $G|_U$ contributes to entries ii, ij, ji, jj in both $L(E, D)$ and $L(G)$. An edge in G', say from $i \in U$ to $j \in D$, contributes only to entry ii, and an edge in $G|_D$, does not affect $L(E, D)$. So we have $L(E, D) = L(G|_U) + \mathrm{Diag}(b)$, and by Lemma 2.6 we have

$$s(E, D) \trianglelefteq s(G|_U) + b. \tag{4.3}$$

We complete our equivalence by appealing to the Gale–Ryser theorem, (2.4), to claim that $b \trianglelefteq a^T$. This follows from the fact that a and b are row and column sums (in nonincreasing order) of the $|D| \times |U|$ bipartite incidence matrix for G'. Combining with the assumption that $s(G|_U) \trianglelefteq d^T(G|_U)$ and (4.3), we get

$$s(E, D) \trianglelefteq s(G|_U) + b \trianglelefteq d^T(G|_U) + a^T = d^T(E, D). \qquad \square$$

This proof relies on the bipartite structure of G', and so it is not immediately obvious how to extend it to higher-dimensional complexes. It would be interesting to do this.

Remark 4.3. Because the induction used to prove Theorem 4.2 requires only that the "undeleted" part of the graph satisfy GM, it is tempting to attack the original GM conjecture by showing that if GM holds for a pair $(G, \{v\})$, then GM holds for G.

Acknowledgments. I thank Vic Reiner for introducing me to this question and for comments throughout, and Art Duval and the anonymous referee for helpful comments. This work began when I was a postdoctoral fellow at the Institute for Mathematics and its Applications (IMA) at the University of Minnesota.

REFERENCES

[1] F. R. K. CHUNG AND R. B. ELLIS, *A chip-firing game and Dirichlet eigenvalues*, Discrete Math., 257 (2002), pp. 341–355.

[2] F. R. K. Chung and R. P. Langlands, *A combinatorial Laplacian with vertex weights*, J. Combin. Theory Ser. A, 75 (1996), pp. 316–327.

[3] A. M. Duval, *A common recursion for Laplacians of matroids and shifted simplicial complexes*, Doc. Math., 10 (2005), pp. 583–618.

[4] A. M. Duval and V. Reiner, *Shifted simplicial complexes are Laplacian integral*, Trans. Amer. Math. Soc., 354 (2002), pp. 4313–4344.

[5] R. Grone and R. Merris, *The Laplacian spectrum of a graph* II, SIAM J. Discrete Math., 7 (1994), pp. 221–229.

[6] P. L. Hammer and A. K. Kelmans, *Laplacian spectra and spanning trees of threshold graphs*, Discrete Appl. Math., 65 (1996), pp. 255–273.

[7] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1990. Corrected reprint of the 1985 original.

[8] N. V. R. Mahadev and U. N. Peled, *Threshold Graphs and Related Topics*, Ann. Discrete Math. 56, North–Holland, Amsterdam, 1995.

[9] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Math. Sci. Engrg. 143, Academic Press, New York, 1979.

[10] J. R. Munkres, *Elements of Algebraic Topology*, Addison–Wesley, Menlo Park, CA, 1984.

SIAM J. DISCRETE MATH.
Vol. 21, No. 2, pp. 313–328

A NEW APPROXIMATION ALGORITHM FOR THE NONPREEMPTIVE SCHEDULING OF INDEPENDENT JOBS ON IDENTICAL PARALLEL PROCESSORS[*]

GIUSEPPE PALETTA[†] AND PAOLAMARIA PIETRAMALA[‡]

Abstract. We consider the scheduling problem of n independent jobs on m identical parallel processors in order to minimize makespan, the completion time of the last job. We propose a new approximation algorithm that iteratively combines partial solutions to the problem. The worst-case performance ratio of the algorithm is $\frac{z+1}{z} - \frac{1}{mz}$, where z is the number of initial partial solutions that are obtained by partitioning the set of jobs into z families of subsets which satisfy suitable properties. The computational behavior of our worst-case performance ratio provided encouraging results on three families of instances taken from the literature.

Key words. optimization, scheduling, approximation algorithm, worst-case performance ratio

AMS subject classifications. 68W25, 90C59

DOI. 10.1137/050634657

1. Introduction.

In this paper, we consider the scheduling problem of n independent jobs on m parallel machines. Each job i must be processed without interruption by only one of the m machines (nonpreemptive environment); the machines are identical, and thus the processing time a_i of the job i is independent of the machine processing it (identical parallel processor environment). The objective is to minimize the makespan, i.e., the total time required to complete all the jobs. Using the standard three field classification scheme (Graham et al. (1979)), this problem is usually denoted as $P||C_{max}$.

The problem is well known to be NP-hard in strong sense for an arbitrary $m \geq 2$ (Garey and Johnson (1979) and Ullman (1976)). It is unlikely that there exists a polynomial-time algorithm for producing a minimal makespan, and so we consider heuristic algorithms in the hope of providing near-optimal results. If we look at the approximation algorithms for this problem, we can refer to the list scheduling family of Graham (1966, 1969), which includes the largest processing time (LPT), and to the multifit decreasing (MFD) scheduling algorithm of Coffman, Garey, and Johnson (1978). The LPT algorithm runs in $O(n\log(n) + nm)$-time and has the worst-case ratio equal to $\frac{4}{3} - \frac{1}{3m}$, whereas the MFD algorithm runs in $O(n\log(n) + knm)$-time and has a worst-case ratio equal to $\frac{13}{11} + \frac{1}{2^k}$ (Friesen (1984) and Yue (1990)), where k represents the number of times that a bin packing problem is solved by using the lowest fit decreasing algorithm (Coffman, Garey, and Johnson (1978)).

The literature of parallel machine scheduling problems, on the heuristic algorithms, has been extensively reviewed by Cheng and Sin (1990), Lawler et al. (1993), and Mokotoff (2001). An overview of existing results and of recent research areas is presented in the handbook edited by Leung (2004).

[*] Received by the editors June 28, 2005; accepted for publication (in revised form) November 15, 2006; published electronically April 27, 2007.

http://www.siam.org/journals/sidma/21-2/63465.html

[†] Dipartimento di Economia e Statistica, Università della Calabria, 87036 Arcavacata di Rende, Cosenza, Italy (g.paletta@unical.it).

[‡] Dipartimento di Matematica, Università della Calabria, 87036 Arcavacata di Rende, Cosenza, Italy (pietramala@mat.unical.it).

The algorithm proposed in this paper is based on the idea of combining iteratively partial solutions until a feasible solution for the scheduling problem is obtained. The initial partial solutions are obtained by partitioning the set of jobs into z families of subsets satisfying suitable properties that will be described below.

The algorithm runs in $O(n\log(n)+nm)$-time as the *LPT* algorithm and produces a solution with a worst-case performance ratio equal to $\frac{z+1}{z}-\frac{1}{mz}$ if $z>1$, where z is the number of initial partial solutions, whereas, if $z=1$, the algorithm produces an optimal solution. This bound of the algorithm is very poor whenever z is less than 3. However, as it will be described below, z is at least equal to $\lceil\frac{n}{m\rho}\rceil$ (where ρ is the ratio between $\max_{i=1,\dots,n}\{a_i\}$ and $\min_{i=1,\dots,n}\{a_i\}$ and $\lceil x\rceil$ denotes the smallest integer not less than x) so that when $n>6m\rho$, our algorithm works very well compared to *LPT* and *MFD* algorithms. Also, our bound is comparable with $\frac{t+1}{t}-\frac{1}{mt}$ (where t is the least number of jobs on any processor), that is, the worst-case performance bound given by Coffman and Sethi (1976) for the *LPT* approach.

The paper is organized as follows. Section 2 presents the definitions and the properties of the partitions that are used to design the algorithm. Section 3 contains the description of the algorithm. Section 4 includes the statements of the inductive assertions about the efficiency and performance of our algorithm. Finally, the computational results obtained from three families of instances taken from the literature are presented in section 5.

2. Definitions and preliminary results. Let $I=\{1,\dots,i,\dots,n\}$ be the set of n independent jobs, $M=\{1,\dots,j,\dots,m\}$ be the set of m identical parallel processors, and $A=\{a_1,\dots,a_i,\dots,a_n\}$ be the set of processing times of the jobs. Let us start with a precise definition of the partitions to be studied.

Let $\mathcal{I}=\{I_1^1,\dots,I_j^1,\dots,I_m^1,\dots,I_1^r,\dots,I_j^r,\dots,I_m^r,\dots,I_1^z,\dots,I_j^z,\dots,I_m^z\}$ be a partition of the set I. Let $a_j^r:=\sum_{i\in I_j^r}a_i$ be the sum of the processing times of the jobs belonging to I_j^r, $r=1,\dots,z$ and $j=1,\dots,m$.

DEFINITION 1. *A partition*

$$\mathcal{I}=\{I_1^1,\dots,I_j^1,\dots,I_m^1,\dots,I_1^r,\dots,I_j^r,\dots,I_m^r,\dots,I_1^z,\dots,I_j^z,\dots,I_m^z\}$$

*of the set I is called an α-*partition *if the following properties are satisfied:*

(a) $a_1^1\ge\dots\ge a_j^1\ge\dots\ge a_m^1\ge\dots\ge a_1^r\ge\dots\ge a_m^r\ge\dots\ge a_1^z\ge\dots\ge a_m^z$.

(b) $a_1^r\le 2a_m^r$, $r=1,\dots,z$.

(c) $a_m^r+a_m^z>a_1^r$, $r=1,\dots,z-1$.

(d) a_j^r, $j=1,\dots,m$ *and* $r=1,\dots,z-1$, *is equal to the sum of at least an* $a_i\in A$ *so that* $a_i\ge a_1^z$.

We associate the *α-partition* $\mathcal{I}$ of I with the family $\mathcal{P}=\{\mathcal{I}^1,\dots,\mathcal{I}^r,\dots,\mathcal{I}^z\}$ of z partial solutions, where each $\mathcal{I}^r=\{I_1^r,\dots,I_j^r,\dots,I_m^r\}$, $r=1,\dots,z$, represents the rth partial solution of the scheduling problem. In particular, with respect to the partial solution $\mathcal{I}^r$, $r=1,\dots,z$, each I_j^r, $j=1,\dots,m$, represents the set of jobs that must be performed by the machine j. Each $\mathcal{I}^r$, $r=1,\dots,z$, is associated with the m-set $G^r=\{a_1^r,\dots,a_j^r,\dots,a_m^r\}$ that is referred to the set of total processing times of the job-sets of $\mathcal{I}^r$.

In view of property (a), the m elements of each G^r, $r=1,\dots,z$, are sorted in nonincreasing order with respect to their size.

In what follows, let us denote by $\Delta^r:=a_1^r-a_m^r$, $r=1,\dots,z$, the gap between the maximum and the minimum element of the m-set G^r.

Example 1. Let us focus on the instance $I=\{1,2,\dots,25\}$, $M=\{1,2,3,4,5\}$, and $A=\{47,46,39,33,31,31,31,27,27,26,25,25,25,24,23,19,19,19,18,18,9,6,6,5,2\}$.

The partition $\mathcal{I} = \{I_1^1 = \{1\}, I_2^1 = \{2\}, I_3^1 = \{4, 22, 23\}, I_4^1 = \{3, 24\}, I_5^1 = \{5, 21, 25\}, I_1^2 = \{6\}, I_2^2 = \{7\}, I_3^2 = \{8\}, I_4^2 = \{9\}, I_5^2 = \{10\}, I_1^3 = \{11\}, I_2^3 = \{12\}, I_3^3 = \{13\}, I_4^3 = \{14\}, I_5^3 = \{15\}, I_1^4 = \{16\}, I_2^4 = \{17\}, I_3^4 = \{18\}, I_4^4 = \{19\}, I_5^4 = \{20\}\}$ is an α-*partition* of I because it satisfies all the properties of Definition 1. This α-*partition* is associated with the family $\mathcal{P} = \{\mathcal{I}^1, \mathcal{I}^2, \mathcal{I}^3, \mathcal{I}^4\}$ of 4 partial solutions, where $\mathcal{I}^1 = \{I_1^1, I_2^1, I_3^1, I_4^1, I_5^1\}$, $\mathcal{I}^2 = \{I_1^2, I_2^2, I_3^2, I_4^2, I_5^2\}$, $\mathcal{I}^3 = \{I_1^3, I_2^3, I_3^3, I_4^3, I_5^3\}$, and $\mathcal{I}^4 = \{I_1^4, I_2^4, I_3^4, I_4^4, I_5^4\}$. The 5-set $G^1 = \{47, 46, 33 + 6 + 6, 39 + 5, 31 + 9 + 2\}$ and the gap $\Delta^1 = 5$ are associated with the partial solution $\mathcal{I}^1$, $G^2 = \{31, 31, 27, 27, 26\}$ and the gap $\Delta^2 = 5$ are associated with $\mathcal{I}^2$, $G^3 = \{25, 25, 25, 24, 23\}$ and the gap $\Delta^3 = 2$ are associated with $\mathcal{I}^3$, and $G^4 = \{19, 19, 19, 18, 18\}$ and the gap $\Delta^4 = 1$ are associated with $\mathcal{I}^4$.

Let $\mathcal{I} = \{I_1^1, \dots, I_j^1, \dots, I_m^1, \dots, I_1^r, \dots, I_j^r, \dots, I_m^r, \dots, I_1^{z-1}, \dots, I_m^{z-1}, I_1^z, \dots, I_j^z, \dots, I_p^z\}$ be a partition of the set I with $p \le m$; let $a_j^r := \sum_{i \in I_j^r} a_i$ be the sum of the processing times of the jobs belonging to I_j^r, $r = 1, \dots, z-1$ and $j = 1, \dots, m$; let $u_j := \sum_{i \in I_j^z} a_i$ be the sum of the processing times of the jobs belonging to I_j^z, $j = 1, \dots, p$ and $u_j := 0$, $p < j \le m$.

Definition 2. *A partition*

$$\mathcal{I} = \{I_1^1, \dots, I_j^1, \dots, I_m^1, \dots, I_1^r, \dots, I_j^r, \dots, I_m^r, \dots, I_1^{z-1}, \dots, I_j^{z-1}, \dots,$$
$$I_m^{z-1}, I_1^z, \dots, I_j^z, \dots, I_p^z\}$$

*of the set I is called a β-*partition *if the following properties are satisfied:*

(a) $a_1^1 \ge \cdots \ge a_j^1 \ge \cdots \ge a_m^1 \ge \cdots \ge a_1^r \ge \cdots \ge a_m^r \ge \cdots \ge a_1^{z-1} \ge \cdots \ge a_m^{z-1} \ge u_1 \ge \cdots \ge u_p$.

(b) $a_1^r \le 2a_m^r$, $r = 1, \dots, z-1$, *and* $u_1 > 2u_m$.

(c) $a_m^r + u_p > a_1^r$, $r = 1, \dots, z-1$.

(d) a_j^r, $j = 1, \dots, m$ *and* $r = 1, \dots, z-1$, *is equal to the sum of at least an* $a_i \in A$ *so that* $a_i \ge u_1$.

We associate the β-*partition* $\mathcal{I}$ of I with the family $\mathcal{P} = \{\mathcal{I}^1, \dots, \mathcal{I}^r, \dots, \mathcal{I}^{z-1}, \mathbf{I}\}$ of z partial solutions, where each $\mathcal{I}^r = \{I_1^r, \dots, I_j^r, \dots, I_m^r\}$, $r = 1, \dots, z-1$, represents the rth partial solution and $\mathbf{I} = \{I_1^z, \dots, I_j^z, \dots, I_p^z, \emptyset_{p+1}, \dots, \emptyset_m\}$ represents the zth partial solution, where $m-p$ machines do not perform jobs. In particular, with respect to the partial solution $\mathbf{I}$, I_j^z, $j = 1, \dots, p$, represents the set of jobs that are performed by the machine j, and $\emptyset_j$, $j = p+1, \dots, m$, indicates that the machine j does not perform jobs. Each $\mathcal{I}^r$, $r = 1, \dots, z-1$, is associated with the m-set $G^r = \{a_1^r, \dots, a_j^r, \dots, a_m^r\}$, whereas $\mathbf{I}$ is associated with the m-set $U = \{u_1, \dots, u_p, 0, \dots, 0\}$ that is referred to the set of the total processing times of the job-sets of $\mathbf{I}$.

In view of property (a), the elements of each G^r, $r = 1, \dots, z-1$, and the elements of U are sorted in nonincreasing order with respect to their size.

As before, let us denote by $\Delta^r := a_1^r - a_m^r, r = 1, \dots, z-1$, the gap between the maximum and the minimum element of the m-set G^r and by $\Delta^U := u_1 - u_m$ the gap between the maximum and the minimum element of the m-set U.

Example 2. Let us focus on the instance $I = \{1, 2, \dots, 25\}$, $M = \{1, 2, 3, 4, 5\}$, and $A = \{50, 48, 44, 42, 39, 36, 35, 34, 32, 30, 29, 28, 28, 28, 28, 27, 26, 26, 23, 11, 10, 9, 9, 2, 1\}$. The partition $\mathcal{I} = \{I_1^1 = \{1\}, I_2^1 = \{5, 20\}, I_3^1 = \{2\}, I_4^1 = \{3\}, I_5^1 = \{4, 24\}, I_1^2 = \{6\}, I_2^2 = \{7\}, I_3^2 = \{8\}, I_4^2 = \{9\}, I_5^2 = \{10\}, I_1^3 = \{11\}, I_2^3 = \{12\}, I_3^3 = \{13\}, I_4^3 = \{14\}, I_5^3 = \{15\}, I_1^4 = \{16\}, I_2^4 = \{17\}, I_3^4 = \{18\}, I_4^4 = \{19\}, I_5^4 = \{21, 22, 25\}, I_1^5 = \{23\}\}$ is a β-*partition* of I because it satisfies all the properties of Definition 2. This β-*partition* is associated with the family $\mathcal{P} = \{\mathcal{I}^1, \mathcal{I}^2, \mathcal{I}^3, \mathcal{I}^4, \mathbf{I}\}$

of 5 partial solutions, where $\mathcal{I}^1 = \{I_1^1, I_2^1, I_3^1, I_4^1, I_5^1\}$, $\mathcal{I}^2 = \{I_1^2, I_2^2, I_3^2, I_4^2, I_5^2\}$, $\mathcal{I}^3 = \{I_1^3, I_2^3, I_3^3, I_4^3, I_5^3\}$, $\mathcal{I}^4 = \{I_1^4, I_2^4, I_3^4, I_4^4, I_5^4\}$, and $\mathbf{I} = \{I_1^5, \emptyset_2, \emptyset_3, \emptyset_4, \emptyset_5\}$. The 5-set $G^1 = \{50, 39 + 11, 48, 44, 42 + 2\}$ and the gap $\Delta^1 = 6$ are associated with the partial solution $\mathcal{I}^1$, $G^2 = \{36, 35, 34, 32, 30\}$ and the gap $\Delta^2 = 6$ are associated with $\mathcal{I}^2$, $G^3 = \{29, 28, 28, 28, 28\}$ and the gap $\Delta^3 = 1$ are associated with $\mathcal{I}^3$, $G^4 = \{27, 26, 26, 23, 10 + 9 + 1\}$ and the gap $\Delta^4 = 7$ are associated with $\mathcal{I}^4$, and $U = \{9, 0, 0, 0, 0\}$ and the gap $\Delta^U = 9$ are associated with $\mathbf{I}$.

Definition 3. *Let $\mathcal{I}^r$ and $\mathcal{I}^q$ be two partial solutions related to an* α-partition *(*β-partition*) of I. Let us define* "combination" *among $\mathcal{I}^r$ and $\mathcal{I}^q$ ($\mathcal{I}^r \uplus \mathcal{I}^q$) as the m-family*

$$\mathcal{I}^r \uplus \mathcal{I}^q = \{I_1^r \cup I_m^q, \ldots, I_j^r \cup I_{m-j+1}^q, \ldots, I_m^r \cup I_1^q\}.$$

This new family corresponds to a new partial solution on the jobs belonging to $\mathcal{I}^r$ and $\mathcal{I}^q$. In particular, the set $I_j^r \cup I_{m-j+1}^q$, for each $j = 1, \ldots, m$, represents the jobs performed by the machine j. The total processing time needed for the machines to perform all the jobs belonging to $\mathcal{I}^r \uplus \mathcal{I}^q$ is computed by using the following definition.

Definition 4. *Let G^r and G^q be the sets of processing times of the partial solutions $\mathcal{I}^r$ and $\mathcal{I}^q$ related to an* α-partition *(*β-partition*) of I. Let us define* "sum" *among G^r and G^q ($G^r \oplus G^q$) as the m-set (not necessarily ordered)*

$$G^r \oplus G^q = \{a_1^r + a_m^q, \ldots, a_j^r + a_{m-j+1}^q, \ldots, a_m^r + a_1^q\}.$$

Notice that $a_j^r + a_{m-j+1}^q$ represents the total processing time needed for machine j to perform all the jobs belonging to $I_j^r \cup I_{m-j+1}^q$, and $\mathcal{I}^r \uplus \mathcal{I}^q$ is a partial solution that is not related to an *α-partition* or to a *β-partition* of I because the elements of $G^r \oplus G^q$ are not sorted in decreasing order with respect to their size.

Example 3. Let G^1 and G^2 be the sets of processing times of the partial solutions $\mathcal{I}^1$ and $\mathcal{I}^2$ which are related to the *α-partition* of Example 1; then

$$\begin{aligned} G^1 \oplus G^2 &= \{a_1^1 + a_5^2, a_2^1 + a_4^2, a_3^1 + a_3^2, a_4^1 + a_2^2, a_5^1 + a_1^2\} \\ &= \{47 + 26, 46 + 27, 45 + 27, 44 + 31, 42 + 31\} = \{73, 73, 72, 75, 73\} \end{aligned}$$

and

$$\begin{aligned} \mathcal{I}^1 \uplus \mathcal{I}^2 &= \{I_1^1 \cup I_5^2, I_2^1 \cup I_4^2, I_3^1 \cup I_3^2, I_4^1 \cup I_2^2, I_5^1 \cup I_1^2\} \\ &= \{\{1, 10\}, \{2, 9\}, \{4, 22, 23, 8\}, \{3, 24, 7\}, \{5, 21, 25, 6\}\}, \end{aligned}$$

where, for example, $a_5^1 + a_1^2 = 73$ represents the total processing time needed for machine 5 to perform the jobs belonging to $I_5^1 \cup I_1^2 = \{5, 21, 25, 6\}$.

The "*sum*" operator satisfies the properties indicated in Lemmas 1 and 3.

Lemma 1. *Let G^r and G^q be the sets of processing times of the partial solutions $\mathcal{I}^r$ and $\mathcal{I}^q$ which are relative to an α-partition (β-partition) of I. Put $S = G^r \oplus G^q$ and $\Delta^S = \max\{S\} - \min\{S\}$. Then*

1. $\max\{S\} \leq 2\min\{S\}$;
2. $\Delta^S \leq \max\{\Delta^r, \Delta^q\}$;
3. $\Delta^S < a_m^z$ *($\Delta^S < u_p$).*

Proof.

Statement 1.

Let

$$\max\{S\} = a_k^r + a_{m-k+1}^q \text{ for some } k, \quad 1 \leq k \leq m,$$

and

$$\min\{S\} = a_l^r + a_{m-l+1}^q \text{ for some } l, \quad 1 \leq l \leq m.$$

Then

$$\max\{S\} = a_k^r + a_{m-k+1}^q \leq (\text{property (a)}) \leq a_1^r + a_1^q \leq (\text{property (b)})$$

$$\leq 2a_m^r + 2a_m^q \leq (\text{property (a)}) \leq 2(a_l^r + a_{m-l+1}^q) = 2\min\{S\}.$$

Statement 2.

Let

$$\Delta^S = (a_k^r + a_{m-k+1}^q) - (a_l^r + a_{m-l+1}^q) = (a_k^r - a_l^r) + (a_{m-k+1}^q - a_{m-l+1}^q).$$

It can happen that $a_k^r - a_l^r \geq 0$ or not. We examine both cases separately.

First case: $a_k^r - a_l^r \geq 0$.

An immediate consequence of the definition of S and of property (a) gives rise to

$$a_{m-k+1}^q - a_{m-l+1}^q \leq 0.$$

It follows that

$$\Delta^S = (a_k^r - a_l^r) + (a_{m-k+1}^q - a_{m-l+1}^q) \leq a_k^r - a_l^r \leq (\text{property (a)}) \leq a_1^r - a_m^r = \Delta^r.$$

Second case: $a_k^r - a_l^r \leq 0$.

An immediate consequence of the definition of S and of property (a) gives rise to

$$a_{m-k+1}^q - a_{m-l+1}^q \geq 0.$$

Then

$$\Delta^S = (a_k^r - a_l^r) + (a_{m-k+1}^q - a_{m-l+1}^q) \leq a_{m-k+1}^q - a_{m-l+1}^q$$

$$\leq (\text{property (a)}) \leq a_1^q - a_m^q = \Delta^q.$$

Consequently,

$$\Delta^S \leq \max\{\Delta^r, \Delta^q\}.$$

Statement 3.

Property (c) ensures that

$$\Delta^r < a_m^z \quad \text{and} \quad \Delta^q < a_m^z (\Delta^r < u_p \quad \text{and} \quad \Delta^q < u_p).$$

From Statement 2, we deduce that

$$\Delta^S < a_m^z (\Delta^S < u_p). \qquad \square$$

LEMMA 2. *Let $G^r = \{a_1^r, \ldots, a_j^r, \ldots, a_m^r\}$ and $U = \{u_1, \ldots, u_p, 0, \ldots, 0\}$, $p < m$, be the sets of processing times of the partial solutions $\mathcal{I}^r$ and $\mathbf{I}$ related to a β-partition of I. Set $S = U \oplus G^r$. Then*

1. $\min\{S\} = a^r_{m-p}$;
2. $\max\{S\} = u_k + a^r_{m-k+1}$ for some $k, 1 \leq k \leq p$.

Proof.

Statement 1.

Let

$$S = \{u_1 + a^r_m, \dots, u_p + a^r_{m-p+1}, \dots, a^r_{m-p}, \dots, a^r_1\}.$$

Property (a) yields $a^r_1 \geq \cdots \geq a^r_{m-p}$. Moreover, for each $j = 1, \dots, p$,

$$a^r_{m-p} \leq a^r_1 < (\text{property (c)}) < u_p + a^r_m \leq (\text{property (a)}) \leq u_j + a^r_{m-j+1},$$

and hence we deduce that $\min\{S\} = a^r_{m-p}$.

Statement 2.

It follows from property (c) that, for all $j = 1, \dots, p$,

$$u_j + a^r_{m-j+1} \geq a^r_1 \geq (\text{property (a)}) \geq a^r_2 \geq \cdots \geq a^r_{m-p},$$

from which we deduce that

$$\max\{S\} = \max_{j=1,\dots,p}\{u_j + a^r_{m-j+1}\}. \qquad \square$$

Lemma 3. *Let G^r and U be the sets of processing times of the partial solutions $\mathcal{I}^r$ and $\mathbf{I}$ related to a β-partition of I. Set $S = U \oplus G^r$ and $\Delta^S = \max\{S\} - \min\{S\}$. Then*

1. $\max\{S\} \leq 2\min\{S\}$;
2. $\Delta^S \leq \max\{\Delta^U, \Delta^r\} = \Delta^U$;
3. $\Delta^S \leq u_1$.

Proof.

Statement 1.

We distinguish two cases.

First case: $p < m$.

Lemma 2 states that $\min\{S\} = a^r_{m-p}$ and $\max\{S\} = u_k + a^r_{m-k+1}$ for some $k, 1 \leq k \leq p$. Then

$$\max\{S\} = u_k + a^r_{m-k+1} \leq (\text{property (a)}) \leq a^r_{m-p} + a^r_{m-p} = 2a^r_{m-p} = 2\min\{S\}.$$

Second case: $p = m$.

First,

$$a^r_1 < (\text{property (c)}) < u_m + a^r_m \leq (\text{property (a)}) \leq u_j + a^r_{m-j+1},\ j = 1, \dots, m.$$

So, we obtain

$$a^r_1 \leq \min_{j=1,\dots,m}\{u_j + a^r_{m-j+1}\}. \qquad \text{(i)}$$

Now

$$\max\{S\} = \max_{j=1,\dots,m}\{u_j + a^r_{m-j+1}\} \leq (\text{property (a)}) \leq u_1 + a^r_1$$

$$\leq (\text{property (a)}) \leq 2a^r_1 \leq \text{from (i)} \leq 2\min\{S\}.$$

Statement 2.

We distinguish two cases.

First case: $p < m$.

First, we show that $\max\{\Delta^U, \Delta^r\} = \Delta^U$. In fact, we have

$$\Delta^r = a_1^r - a_m^r < \text{(property (c))} < u_p \leq \text{(property (a))} \leq u_1 = \Delta^U, \quad r = 1, \ldots, z-1.$$

Moreover, Lemma 2 ensures that, for some k, $1 \leq k \leq p$, one has

$$\Delta^S = u_k + a_{m-k+1}^r - a_{m-p}^r \leq (\text{because } a_{m-k+1}^r - a_{m-p}^r \leq 0)$$

$$\leq u_k \leq \text{(property (a))} \leq u_1 = \Delta^U.$$

Second case: $p = m$.

First, we show that $\max\{\Delta^U, \Delta^r\} = \Delta^U$. In this case, we obtain

$$\Delta^r - \Delta^U = a_1^r - a_m^r - (u_1 - u_m) < \text{(property (c))}$$

$$< a_m^r + u_m - a_m^r - u_1 + u_m = 2u_m - u_1 < \text{(property (b))} < 0.$$

Now let $\max\{S\} = u_k + a_{m-k+1}^r$ for some $k, 1 \leq k \leq m$, and $\min\{S\} = u_l + a_{m-l+1}^r$ for some $l, 1 \leq l \leq m$. So,

$$\Delta^S = (u_k - u_l) + (a_{m-k+1}^r - a_{m-l+1}^r).$$

If $u_k - u_l \geq 0$ (hence $a_{m-k+1}^r - a_{m-l+1}^r \leq 0$), then

$$\Delta^S \leq u_k - u_l \leq \text{(property (a))} \leq u_1 - u_m = \Delta^U.$$

If $u_k - u_l < 0$ (hence $a_{m-k+1}^r - a_{m-l+1}^r \geq 0$), then

$$\Delta^S \leq a_{m-k+1}^r - a_{m-l+1}^r \leq \text{(property (a))} \leq a_1^r - a_m^r = \Delta^r \leq \Delta^U.$$

Summarizing, we derive

$$\Delta^S \leq \Delta^U.$$

Statement 3.

From Statement 2, we deduce that

$$\Delta^S \leq \Delta^U = u_1 - u_m \leq u_1. \qquad \square$$

DEFINITION 5. *Let G^r and G^q be the sets of processing times of the partial solutions $\mathcal{I}^r$ and $\mathcal{I}^q$ related to an α-*partition *(β-*partition*) of I. Let us define "*ordered sum*" among G^r and G^q as the ordered m-set* Ord*$(G^r \oplus G^q)$ whose elements are the elements of $G^r \oplus G^q$ sorted in nonincreasing order with respect to their size.*

DEFINITION 6. *Let $\mathcal{I}^r$ and $\mathcal{I}^q$ be two partial solutions related to an α-*partition *(β-*partition*) of I. Let us define "*ordered combination*" among $\mathcal{I}^r$ and $\mathcal{I}^q$ as the m-family* Ord*$(\mathcal{I}^r \uplus \mathcal{I}^q)$ whose sets are those of $\mathcal{I}^r \uplus \mathcal{I}^q$ sorted so that the jth element of* Ord*$(G^r \oplus G^q)$ represents the total processing time of the jth job-set of* Ord*$(\mathcal{I}^r \uplus \mathcal{I}^q)$.*

Thus, we have $Ord(G^r \oplus G^q) = Ord(G^q \oplus G^r)$ and $Ord(\mathcal{I}^r \uplus \mathcal{I}^q) = Ord(\mathcal{I}^q \uplus \mathcal{I}^r)$. In the following, the partial solution $Ord(\mathcal{I}^r \uplus \mathcal{I}^q)$ is called "*combined partial solution*" to distinguish it from the initial partial solutions obtained by the procedure described in the next section.

Example 4. With reference to Example 3, we have

$$Ord(G^1 \oplus G^2) = Ord(73, 73, 72, 75, 73) = \{75, 73, 73, 73, 72\}$$

and

$$\begin{aligned} Ord(\mathcal{I}^1 \uplus \mathcal{I}^2) &= Ord(\{1, 10\}, \{2, 9\}, \{4, 22, 23, 8\}, \{3, 24, 7\}, \{5, 21, 25, 6\}) \\ &= \{\{3, 24, 7\}, \{5, 21, 25, 6\}, \{1, 10\}, \{2, 9\}, \{4, 22, 23, 8\}\}. \end{aligned}$$

Lemma 4. *Let* $\mathcal{P} = \{\mathcal{I}^1, \ldots, \mathcal{I}^r, \ldots, \mathcal{I}^z\}$ *be a family of* z *partial solutions related to an* α*-partition of* I*. Then* $\{\mathrm{Ord}(\ldots(\mathrm{Ord}(\mathrm{Ord}(\mathcal{I}^1 \uplus \mathcal{I}^2) \uplus \mathcal{I}^3) \uplus \ldots) \uplus \mathcal{I}^r), \mathcal{I}^{r+1}, \ldots, \mathcal{I}^z\}$, $r = 2, \ldots, z$, *is a family of* $z - r + 1$ *partial solutions that is related to an* α*-partition of* I.

Proof. Property (a) is guaranteed because the "*ordered combination*" operator is iteratively performed among the first two partial solutions related to an α*-partition.* Properties (b) and (c) are guaranteed, respectively, by Statements 1 and 3 of Lemma 1. Property (d) is guaranteed because it was guaranteed by the partial solutions in $\mathcal{P}$. □

By using the same arguments in the proof of Lemma 4, we have the following lemma.

Lemma 5. *Let* $\mathcal{P} = \{\mathcal{I}^1, \ldots, \mathcal{I}^r, \ldots, \mathcal{I}^{z-1}, \mathbf{I}\}$ *be a family of* z *partial solutions related to a* β*-partition of* I*. Then* $\{\mathrm{Ord}(\ldots(\mathrm{Ord}(\mathrm{Ord}(\mathcal{I}^1 \uplus \mathcal{I}^2) \uplus \mathcal{I}^3) \uplus \ldots) \uplus \mathcal{I}^r), \mathcal{I}^{r+1}, \ldots, \mathcal{I}^{z-1}, \mathbf{I}\}$, $r = 2, \ldots, z-1$, *is a family of* $z - r + 1$ *partial solutions that is related to a* β*-partition of* I.

Example 5. With reference to Example 1, we have $\{Ord(\mathcal{I}^1 \uplus \mathcal{I}^2), \mathcal{I}^3, \mathcal{I}^4\}$, which is a family of three partial solutions, where $Ord(\mathcal{I}^1 \uplus \mathcal{I}^2)$ and $Ord(G^1 \oplus G^2)$ are reported in Example 4.

3. Algorithms.

The proposed algorithm, which uses the procedure named *IPS* (initial partial solutions) described later, partitions the jobs so as to obtain an α*-partition* or a β*-partition* of I, i.e., a family of initial partial solutions to the scheduling problem. Then, as indicated in Lemmas 4 and 5, it iteratively combines, in turn, the initial partial solutions with the current *combined partial solution* by utilizing the *ordered combination* operator. The iterative process is repeated until a solution of the scheduling problem is obtained. The algorithm, named *MPS* (multiprocessor scheduling), can be summarized as follows.

Algorithm *MPS*.

Initialization

- Use the procedure *IPS* to obtain the family $\mathcal{P}$ of z initial partial solutions that are related to an α*-partition* or to a β*-partition* of I. If *IPS* returns only a partial solution, then Stop (the solution is optimal);
- Set $C = \{C_1 = 0, \ldots, C_j = 0, \ldots, C_m = 0\}$ and $\mathcal{T} = \{\mathcal{T}_1 = \emptyset_1, \ldots, \mathcal{T}_j = \emptyset_j, \ldots, \mathcal{T}_m = \emptyset_m\}$, where $\mathcal{T}$ represents the current *combined partial solution* and C_j the processing time of the job-set $\mathcal{T}_j$.

Construction

- For $r = 1, \ldots, z-1$, compute $C = Ord(C \oplus G^r)$ and $\mathcal{T} = Ord(\mathcal{T} \uplus \mathcal{I}^r)$, where G^r and $\mathcal{I}^r$ have been provided by the *IPS* procedure;
- If *IPS* returns an α*-partition* of I, then compute $C = Ord(C \oplus G^z)$ and $\mathcal{T} = Ord(\mathcal{T} \uplus \mathcal{I}^z)$, where G^z and $\mathcal{I}^z$ have been provided by the *IPS* procedure;

```
    - If $IPS$ returns a $\beta$-partition of $I$, then compute $C = Ord(C \oplus U)$ and $\mathcal{T} = Ord(\mathcal{T} \uplus \mathbf{I})$, where $U$ and $\mathbf{I}$ have been provided by the $IPS$ procedure.
```

At the end, *MPS* returns $\mathcal{T} = \{\mathcal{T}_1, \dots, \mathcal{T}_j, \dots, \mathcal{T}_m\}$ and $C = \{C_1, \dots, C_j, \dots, C_m\}$, where $\mathcal{T}$ represents a solution of the problem and C the completion times of the machines. In particular, each $\mathcal{T}_j$, $j = 1, \dots, m$, represents the jobs assigned to machine j and C_j the processing time needed for each machine j to perform the jobs which are assigned to it.

3.1. Determining the initial partial solutions. The procedure *IPS*, which finds the initial partial solutions related to an *α-partition* or a *β-partition* of I, first orders the jobs so that $a_1 \geq \cdots \geq a_i \geq \cdots \geq a_n$. Then it builds an *$\alpha$-partition* or a *$\beta$-partition* of I by processing the jobs in turn, starting with job 1. Now we suppose that, when job i must be inserted, there are either $z-1$ families $\mathcal{I}^r$, $r = 1, \dots, z-1$, that satisfy $a_1^r \leq 2a_m^r$ and the family $\mathbf{I}$ that satisfies $u_1 > 2u_m$, or there are z families $\mathcal{I}^r$. Also, we suppose that all families are sorted in nonincreasing order with respect to their processing times. Let $\Pi = \{\pi(1), \pi(2), \dots\}$ be the permutation of the indexes of the current partial solutions of type $\mathcal{I}^r$ so that $\Delta^{\pi(1)} \geq \Delta^{\pi(2)} \geq \cdots$. Then *IPS* inserts job i as follows. First, it selects, among the ordered families of type $\mathcal{I}^r$, the family $\mathcal{I}^q = \mathcal{I}^{\pi(1)}$ with the biggest gap $\Delta^q = \Delta^{\pi(1)}$ between the processing times of the first and the last job-sets. Now if $a_i \leq \Delta^q$, then job i is inserted into the last job-set of $\mathcal{I}^q$, that is, into I_m^q; the job-sets of $\mathcal{I}^q$ are sorted in nonincreasing order with respect to their processing times, and Π is arranged in nonincreasing order with respect to the gaps. If $a_i > \Delta^q$ and all job-sets of $\mathbf{I}$ are not empty, then job i is inserted into the last set of $\mathbf{I}$, that is, into I_m^z, and the job-sets of $\mathbf{I}$ are sorted in nonincreasing order with respect to their processing times. If $a_i > \Delta^q$ and some job-sets of $\mathbf{I}$ are empty, then job i is inserted into the first job-set empty of $\mathbf{I}$. In this case, it is not necessary to sort the job-sets because they are already ordered. If $a_i > \Delta^q$ and all job-sets of $\mathbf{I}$ are empty, then the job i is inserted into the first set of $\mathbf{I}$, and z is increased by one. Also, if job i is inserted into $\mathbf{I}$ and $2u_m \geq u_1$, then $\mathbf{I}$ becomes $\mathcal{I}^z$, index z is inserted into Π, that is again arranged, and $\mathbf{I}$ is placed equal to m empty sets. The procedure can be formally described as follows.

```
    PROCEDURE IPS
    Initialization
        - Order the jobs so that $a_1 \geq \cdots \geq a_i \geq \cdots \geq a_n$. Set $z = 1$ ($z$ = number of initial partial solutions);
        - Consider $\mathbf{I} = \{I_1^z = \{1\}, I_2^z = \emptyset_2, \dots, I_m^z = \emptyset_m\}$, $U = \{u_1 = a_1, u_2 = 0, \dots, u_m = 0\}$;
        - Set $\Delta^U = a_1$, $p = 1$ ($p$ represents the number of elements of $U$ not equal to 0), and $\Pi = \emptyset$.
Construction
    For each $i = 2, \dots, n$
        - If $\Pi \neq \emptyset$, then set $q = \pi(1)$, $\Delta^{max} = \Delta^q$, and consider the $m$-set $G^q$ and the family $\mathcal{I}^q$, else set $\Delta^{max} = 0$;
        - If $a_i \leq \Delta^{max}$, then
            - $a_m^q = a_m^q + a_i$, $I_m^q = I_m^q \cup \{i\}$, sort the elements of the $m$-set $G^q$
              so that $a_1^q \geq \cdots \geq a_j^q \geq \cdots \geq a_m^q$, and arrange the family $\mathcal{I}^q$ so
              that $a_j^q$ is the total time required by the jobs belonging to $I_j^q$;
            - $\Delta^q = a_1^q - a_m^q$, and arrange the set $\Pi$ so that $\Delta^{\pi(1)} \geq \Delta^{\pi(2)} \geq \cdots$;
        End If $a_i \leq \Delta^{max}$;
        - If $a_i > \Delta^{max}$, then
```

```
            - If all job-sets of $\mathbf{I}$ are empty ($\Delta^U = \infty$), then $z = z + 1$,
              $\mathbf{I} = \{I_1^z = \{i\}, I_2^z = \emptyset_2, \dots, I_m^z = \emptyset_m\}$,
              $U = \{u_1 = a_i, u_2 = 0, \dots, u_m = 0\}$, and $p = 1$;
            - If all job-sets of $\mathbf{I}$ are not empty ($\Delta^U \neq \infty$ and $p = m$), then
              $u_m = u_m + a_i$, $I_m^z = I_m^z \cup \{i\}$, sort the elements of the set $U$ so that
              $u_1 \geq \cdots \geq u_m$, and arrange the family $\mathbf{I}$ so that $u_j$ is the total time
              required by the jobs belonging to $I_j^z$;
            - If some job-sets of $\mathbf{I}$ are empty ($\Delta^U \neq \infty$ and $1 \leq p < m$), then
              $p = p + 1$, $u_p = a_i$, $I_p^z = I_p^z \cup \{i\}$;
            - $\Delta^U = u_1 - u_m$;
            - If $p = m$ and $2u_m \geq u_1$, then $G^z = U$, $\mathcal{I}^z = \mathbf{I}$, $\Delta^z = \Delta^U$, insert $z$
              into $\Pi$, and arrange the set $\Pi$ so that $\Delta^{\pi(1)} \geq \Delta^{\pi(2)} \geq \cdots$,
              $\mathbf{I} = \{\emptyset_1, \dots, \emptyset_j, \dots, \emptyset_m\}$, $U = \{0, , 0, \dots, 0\}$, and $\Delta^U = \infty$;
        End If $a_i > \Delta^{max}$;
    End For $i$.
```

It is easy to show that the procedure *IPS* produces a partition so that the first job-set of each family is a singleton, and, consequently, the processing time of the last job-set of a family is greater than the processing time of the first job-set of the next family. Also, *IPS* produces an *α-partition* of I if $U = \{u_1 = 0, \dots, u_m = 0\}$ or a *β-partition* of I if $U \neq \{u_1 = 0, \dots, u_m = 0\}$.

4. Efficiency and performance of the algorithm. With regard to efficiency, it is easy to show that the *MPS* algorithm, as the *LPT* algorithm, runs in $O(n \log(n) + nm)$-time, that is, the running time of the procedure *IPS*.

Now we consider the performance of the *MPS* algorithm.

Let us use $C_{max} = \max_{j=1,\dots,m}\{C_j\}$ and C^*_{max} to denote the makespan of the *MPS* solution and the optimal makespan, respectively. Denote by $C_{min} = \min_{j=1,\dots,m}\{C_j\}$ and by $\Delta := C_{max} - C_{min}$.

To obtain the performance ratio of our algorithm, it is necessary to show the following lemmas.

LEMMA 6. *If the procedure* IPS *returns a family of z initial partial solutions relative to an α-partition of I, then $a_m^z \geq \Delta$.*

Proof. Property (c) states that, for $q = 1, \dots, z - 1$,

$$\Delta^q = a_1^q - a_m^q < a_m^z.$$

Moreover, we obtain from property (b) that

$$\Delta^z = a_1^z - a_m^z \leq 2a_m^z - a_m^z = a_m^z.$$

Then

$$a_m^z \geq \max_{q=1,\dots,z}\{\Delta^q\}.$$

Using Lemmas 1 and 4, we note that

$$\Delta \leq \max_{q=1,\dots,z}\{\Delta^q\}.$$

Consequently, $a_m^z \geq \Delta$. □

LEMMA 7. *If the procedure* IPS *returns a family of z initial partial solutions relative to a β-partition of I, then $u_1 \geq \Delta$.*

Proof. Using Lemmas 3 and 5, we note that

$$\Delta \leq \max\{\Delta^1, \ldots, \Delta^{z-1}, \Delta^U\} = \Delta^U. \qquad \text{(i)}$$

Since

$$\Delta^U = u_1 - u_m \leq u_1,$$

it follows from (i) that $u_1 \geq \Delta$. □

We are now able to find a lower bound on C^*_{max} with respect to the number z of initial partial solutions relative to an *α-partition* or a *β-partition* of I produced by *IPS*.

THEOREM 1. *If the procedure* IPS *returns a family* $\mathcal{P}$ *of* z *initial partial solutions relative to an* α*-partition or a* β*-partition of* I*, then*

$$C^*_{max} \geq z\Delta.$$

Proof. We need to distinguish two cases.

First case: $\mathcal{P}$ *is relative to an* α*-partition of* I*.*

Of course,

$$C^*_{max} \geq z a^z_m.$$

Moreover, Lemma 6 ensures that $a^z_m \geq \Delta$.

It follows that

$$C^*_{max} \geq z a^z_m \geq z\Delta.$$

Second case: $\mathcal{P}$ *is relative to a* β*-partition of* I*.*

In view of property (d), the number of elements $a_i \in A$ so that $a_i \geq u_1$ in $a^1_1, \ldots, a^1_m, \ldots, a^{z-1}_1, \ldots, a^{z-1}_m$ is at least $m(z-1)$. Focusing only on these $m(z-1)$ elements $a_i \geq u_1$, together with u_1, we get $m(z-1)+1$ jobs with processing times greater than or equal to u_1 (we can ignore the other jobs). Then

$$C^*_{max} \geq \left\lceil \frac{m(z-1)+1}{m} \right\rceil u_1 = \left\lceil (z-1) + \frac{1}{m} \right\rceil u_1 = z u_1.$$

Moreover, Lemma 7 ensures that $u_1 \geq \Delta$, and so

$$C^*_{max} \geq z\Delta. \qquad \square$$

We are now in a position to show results about the performance ratio of the proposed algorithm.

THEOREM 2. *If the procedure* IPS *returns a family* $\mathcal{P}$ *of* z *initial partial solutions relative to an* α*-partition (or a* β*-partition) of* I*, then*

$$\frac{C_{max}}{C^*_{max}} \leq \frac{z+1}{z} - \frac{1}{mz} \qquad \text{if } z > 1$$

and

$$C_{max} = C^*_{max} \qquad \text{if } z = 1.$$

Proof. If the procedure *IPS* returns only the m-set G^1 (or only the set U), we have an optimal solution. In fact, $a_1^1 = a_1 = \max\{a_i\}$ (respectively, $u_1 = a_1$) and $C^*_{max} \geq a_1$. If $z > 1$, let $C'_{min} = C_{min} + \frac{1}{m}\Delta$. Then

$$C_{max} = C_{min} + \Delta = C'_{min} - \frac{1}{m}\Delta + \Delta = C'_{min} + \left(1 - \frac{1}{m}\right)\Delta.$$

Since

$$C^*_{max} \geq C'_{min},$$

it follows that

$$C_{max} \leq C^*_{max} + \left(1 - \frac{1}{m}\right)\Delta.$$

Moreover, Theorem 1 states that $C^*_{max} \geq z\Delta$. We can conclude that

$$\frac{C_{max}}{C^*_{max}} \leq 1+\frac{1}{C^*_{max}}\left(1 - \frac{1}{m}\right)\Delta \leq 1+\frac{1}{z\Delta}\left(1 - \frac{1}{m}\right)\Delta = 1+\frac{m-1}{mz} = \frac{z+1}{z}-\frac{1}{mz}. \qquad \square$$

We exhibit an example that shows that the worst-case performance ratio of our algorithm cannot be improved.

Example 6. Let us focus on the instance $I = \{1, 2, 3, 4, 5\}, M = \{1, 2\}$, and $A = \{3, 3, 2, 2, 2\}$. The optimal solution is obtained when jobs 1 and 2 are performed by machine 1 and jobs 3, 4, and 5 are performed by machine 2. This solution is associated with the completion time of the machines $\{3 + 3, 2 + 2 + 2\} = \{6, 6\}$, and the related makespan is equal to 6.

The procedure *IPS* returns $\mathcal{I}^1 = \{I_1^1 = \{1\}, I_2^1 = \{2\}\}$, $\mathcal{I}^2 = \{I_1^2 = \{3\}, I_2^2 = \{4\}\}$, $\mathbf{I} = \{I_1^3 = \{5\}, \emptyset\}$, $G^1 = \{3, 3\}$, $G^2 = \{2, 2\}$, and $U = \{2, 0\}$. Altogether, *MPS* returns $\mathcal{T} = Ord(Ord(\mathcal{I}^1 \uplus \mathcal{I}^2) \uplus \mathbf{I}\} = \{\mathcal{T}_1 = \{1, 5, 4\}, \mathcal{T}_2 = \{2, 3\}\}$ and $C = Ord(Ord(G^1 \oplus G^2) \oplus U) = \{2 + 3 + 2, 0 + 3 + 2\} = \{7, 5\}$, and the related makespan is equal to 7. Then

$$\frac{C_{max}}{C^*_{max}} = \frac{7}{6} \quad \text{and} \quad \frac{z+1}{z} - \frac{1}{mz} = \frac{4}{3} - \frac{1}{6} = \frac{7}{6}.$$

4.1. Estimate of the worst-case performance ratio. We estimated the worst-case performance ratio of our algorithm. In the following, it is supposed that $a_1 \geq \cdots \geq a_i \geq \cdots \geq a_n$ and $\rho = \frac{a_1}{a_n}$.

PROPOSITION 1. *The procedure* IPS *returns z initial partial solutions with $z \geq \lceil \frac{n}{m\rho} \rceil \geq 1$.*

Proof. The procedure *IPS* returns z initial partial solutions where each I_1^r, $r = 1, \ldots, z$, is a singleton.

When *IPS* returns an *α-partition* of I, $a_1^r \leq a_1^1 = a_1$, $r = 1, \ldots, z$, it follows that $\sum_{j=1,\ldots,m} a_j^r \leq m a_1^r \leq m a_1$, $r = 1, \ldots, z$. Hence

$$\sum_{i=1,\ldots,n} a_i = \sum_{r=1,\ldots,z} \sum_{j=1,\ldots,m} a_j^r \leq \sum_{r=1,\ldots,z} m a_1 \leq z m a_1.$$

When *IPS* returns a *β-partition* of I, $a_1^r \leq a_1^1 = a_1$, $r = 1, \ldots, z - 1$, and $u_1 \leq a_1^1 = a_1$, it follows that $\sum_{j=1,\ldots,m} a_j^r \leq m a_1^r \leq m a_1$, $r = 1, \ldots, z - 1$, and $\sum_{j=1,\ldots,p} u_j \leq p a_1^r \leq m a_1^r \leq m a_1$. Hence

$$\sum_{i=1,\ldots,n} a_i = \sum_{r=1,\ldots,z-1} \sum_{j=1,\ldots,m} a_j^r + \sum_{j=1,\ldots,p} u_j \leq \sum_{r=1,\ldots,z-1} m a_1 + m a_1 \leq z m a_1.$$

Summarizing, we obtain

$$z \geq \left\lceil \frac{1}{ma_1} \sum_{i=1,\dots,n} a_i \right\rceil \geq \left\lceil \frac{na_n}{ma_1} \right\rceil = \left\lceil \frac{n}{m\rho} \right\rceil. \qquad \square$$

Through Proposition 1 and Theorem 2, we immediately obtain the following result.

Corollary 1. *The algorithm* MPS *returns a solution with*

$$\frac{C_{max}}{C^*_{max}} \leq 1 + \frac{1}{\left\lceil \frac{n}{\rho m} \right\rceil} - \frac{1}{m\left\lceil \frac{n}{\rho m} \right\rceil}.$$

This estimate is very poor when ρ is very large, whereas, when $\rho \leq \frac{n}{6m}$, the estimate is better than the worst-case ratio of the *LPT* and *MFD* algorithms.

When $a_1 < 2a_n$, we obtain the following results.

Proposition 2. *For all instances so that $a_1 < 2a_n$, the procedure* IPS *returns $z = \lceil \frac{n}{m} \rceil \geq 1$ initial partial solutions.*

Proof. When *IPS* returns an *α-partition* of I, each I^r_j, $r = 1, \dots, z$ and $j = 1, \dots, m$, is a singleton, while when *IPS* returns a *β-partition* of I, each I^r_j, $r = 1, \dots, z-1$ and $j = 1, \dots, m$, is a singleton, just like each I^z_j, $j = 1, \dots, p$, is also a singleton.

It follows that $z = \lceil \frac{n}{m} \rceil \geq 1$. $\square$

Corollary 2. *For all instances so that $a_1 < 2a_n$, the algorithm* MPS *returns a solution with*

$$\frac{C_{max}}{C^*_{max}} \leq 1 + \frac{1}{\left\lceil \frac{n}{m} \right\rceil} - \frac{1}{m\left\lceil \frac{n}{m} \right\rceil}.$$

This estimate is better than the worst-case ratio of the *LPT* and *MFD* algorithms when $\frac{n}{m} \geq \frac{11}{2}$.

5. Computational results. First, we analyze the computational behavior of the average worst-case ratio bound $\frac{z+1}{z} - \frac{1}{mz}$ of *MPS* and the average worst-case ratio bound $\frac{t+1}{t} - \frac{1}{mt}$ given by Coffman and Sethi (1976) for the *LPT* approach, and then we compare our bound with the *MFD* bound. The *MPS* and *LPT* algorithms have been coded in Fortran 77 and run on the three families of instances that have been used by Frangioni, Necciari, and Scutellà (2004).

In the first family of instances, denoted NONUNIFORM, the number of machines m were $5, 10$, and 25, the number of jobs n were $100, 500$, and 1000, and the intervals for the integer processing times were $[1, 100], [1, 1000]$, and $[1, 10000]$. Ten instances were randomly generated for each choice of m, n and of the processing time intervals, for a total of 270 instances. The generator, which was presented by Frangioni, Necciari, and Scutellà, when an interval $[a, b]$ of the processing times is given, produces instances where 98% of the processing times are uniformly distributed in the interval $[(b-a)0.9, b]$, while the remaining processing times fall within the interval $[a, (b-a)0.02]$. The generator is available from http://www.di.unipi.it/di/groups/optimize/Data/index.html.

The last two families of instances have been derived from difficult bin packing instances and are available at the OR-Library of J. E. Beasley from http://mscmga.ms.ic.ac.uk/jeb/orlib/binpackinfo.html.

Table 1
Behavior of the average worst-case ratio bounds on BINPACK instances.

		LPT		*MPS*	
m	n	$\frac{t+1}{t} - \frac{1}{mt}$	*sec.*	$\frac{z+1}{z} - \frac{1}{mz}$	*sec.*
[46,52]	120	1.32654	0.1	1.32654	0.1
[97,106]	250	1.33005	0.0	1.33005	0.1
[196,207]	500	1.33167	0.0	1.33167	0.3
[393,411]	1000	1.33250	0.1	1.33250	0.6

Table 2
Behavior of the average worst-case ratio bounds on TRIPLET instances.

		LPT		*MPS*	
m	n	$\frac{t+1}{t} - \frac{1}{mt}$	*sec.*	$\frac{z+1}{z} - \frac{1}{mz}$	*sec.*
20	60	1.31666	0.0	1.31666	0.0
40	120	1.32500	0.0	1.32500	0.0
83	249	1.32931	0.0	1.32931	0.1
167	501	1.33133	0.1	1.33133	0.1

In each instance of the last two families, the number of machines m is the number of bins in the best known solution of bin packing instances. In the second family, denoted BINPACK, the number of jobs n were $120, 250, 500, 1000$, and the processing times were uniformly distributed in $[20, 100]$. Twenty instances were generated for each choice of n, for a total of 80 instances. In the third family, denoted TRIPLET, the number of jobs n were $60, 120, 249, 501$, and the processing times were in $[25, 50]$. Twenty instances were generated for each choice of n, for a total of 80 instances.

Tables 1, 2, and 3 compare the behavior of the average worst-case performance bounds in BINPACK, TRIPLET, and NONUNIFORM instances. For each algorithm, the entries give the average worst-case performance bound and the average running time (expressed in seconds, on a Pentium III, 933 MHz, 256 MbRAM, and including the sorting time). The worst-case performance bounds and the running times were averaged for each group of 10 NONUNIFORM instances (Table 3), whereas they were averaged for each group of 20 BINPACK and TRIPLET instances (Tables 1 and 2, respectively).

All the instances of the three families (see Tables 1, 2, and 3) were solved very quickly by both algorithms. The running time of *MPS* is only slightly higher than the running time of *LPT*.

On NONUNIFORM instances (Table 3), the average worst-case ratio bound of *MPS* is slightly better, whereas, on BINPACK and TRIPLET instances (Tables 1 and 2), *MPS* and *LPT* have the same average worst-case ratio bound. Because all TRIPLET instances satisfy the property $a_1 < 2a_n$, the average worst-case performance bound of *MPS* can be computed by using Corollary 2.

The average worst-case ratio bounds of *MPS* and *LPT* decrease as the ratio $\frac{n}{m}$ increase (Table 3), whereas, when the ratio $\frac{n}{m}$ is constant, the average worst-case performance bounds of *MPS* and *LPT* decrease as m decreases (Tables 1 and 2).

As shown in Table 3, the average worst-case ratio bounds were independent from the intervals within which the processing times of the jobs were generated (see each row of the table) because, for each interval $[a, b]$, the generator produced instances where 98% of the processing times were distributed in the interval $[(b - a)0.9, b]$.

Table 3
Behavior of the average worst-case ratio bounds on NONUNIFORM instances.

		$\mathbf{a_j \in [1, 100]}$				$\mathbf{a_j \in [1, 1000]}$				$\mathbf{a_j \in [1, 10000]}$			
		LPT		***MPS_A***		***LPT***		***MPS_A***		***LPT***		***MPS_A***	
m	**n**	$\frac{t+1}{t} - \frac{1}{mt}$	**sec.**	$\frac{z+1}{z} - \frac{1}{mz}$	**sec.**	$\frac{t+1}{t} - \frac{1}{mt}$	**sec.**	$\frac{z+1}{z} - \frac{1}{mz}$	**sec.**	$\frac{t+1}{t} - \frac{1}{mt}$	**sec.**	$\frac{z+1}{z} - \frac{1}{mz}$	**sec.**
5	100	1.04000	0.0	1.04000	0.0	1.04000	0.0	1.04000	0.0	1.04000	0.0	1.04000	0.1
	500	1.00808	0.0	1.00802	0.1	1.00808	0.0	1.00807	0.0	1.00808	0.0	1.00807	0.1
	1000	1.00402	0.1	1.00402	0.1	1.00403	0.0	1.00402	0.0	1.00403	0.1	1.00403	0.0
10	100	1.09000	0.0	1.09000	0.0	1.09000	0.0	1.09000	0.0	1.09000	0.0	1.09000	0.0
	500	1.01800	0.0	1.01800	0.0	1.01800	0.0	1.01800	0.0	1.01800	0.0	1.01800	0.0
	1000	1.00909	0.1	1.00900	0.1	1.00909	0.1	1.00900	0.0	1.00909	0.0	1.00907	0.1
25	100	1.24000	0.0	1.24000	0.0	1.24000	0.0	1.24000	0.0	1.24000	0.0	1.24000	0.0
	500	1.04800	0.0	1.04800	0.1	1.04800	0.0	1.04800	0.2	1.04800	0.0	1.04800	0.3
	1000	1.02400	0.0	1.02400	0.0	1.02400	0.0	1.02400	0.0	1.02400	0.0	1.02400	0.0

The average worst-case ratio bounds $\frac{z+1}{z} - \frac{1}{mz}$ and $\frac{t+1}{t} - \frac{1}{mt}$ were always better than the worst-case ratio $\frac{13}{11} + \frac{1}{2^k} \approx 1.1818$ of the *MFD* algorithm, except in instances with $\frac{n}{m} \leq 4$.

The numerical results computed by using Corollary 1 on BINPACK and NONUNIFORM instances were not reported because, as expected ($n \ll 6m\rho$), they were very poor.

6. Conclusions. We have designed a new approximation algorithm, which runs in $O(n\log(n)+nm)$-time as the *LPT* algorithm of Graham, for the scheduling problem of independent jobs on identical parallel processors in order to minimize makespan. The worst-case performance ratio of the algorithm is $\frac{z+1}{z} - \frac{1}{mz}$, where z is the number of initial partial solutions that are obtained by partitioning the set of jobs into z families of subsets which satisfy suitable properties.

The computational results showed that our worst-case performance ratio outperforms the one of the *MFD* algorithm on a few instances taken from the literature, except in instances with $\frac{n}{m} \leq 4$. Also, our worst-case performance ratio bound is comparable to that given by Coffman et al. for the *LPT* algorithm.

An estimate of our bound is also given, and it was quite good when $n > 6m\rho$.

Acknowledgments. We are grateful both to the editor and to the referees for their suggestions, which helped us to improve the presentation of this paper.

REFERENCES

T. Cheng and C. Sin (1990), *A state of the art review of parallel machine scheduling research*, European J. Oper. Res., 47, pp. 271–292.

E. G. Coffman, Jr., M. R. Garey, and D. S. Johnson (1978), *An application of bin-packing to multiprocessor scheduling*, SIAM J. Comput., 7, pp. 1–17.

E. G. Coffman, Jr., and R. Sethi (1976), *A generalized bound on LPT sequencing*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Bleue, 10, pp. 17–25.

A. Frangioni, E. Necciari, and M. G. Scutellà (2004), *A multi-exchange neighborhood for minimum makespan machine scheduling problems*, J. Comb. Optim., 8, pp. 195–220.

D. K. Friesen (1984), *Tighter bounds for the Multifit processor scheduling algorithm*, SIAM J. Computing, 13, pp. 170–181.

M. R. Garey and D. S. Johnson (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, San Francisco.

R. L. Graham (1966), *Bounds for certain multiprocessing anomalies*, Bell System Tech. J., 45, pp. 1563–1581.

R. L. Graham (1969), *Bounds on multiprocessing timing anomalies*, SIAM J. Appl. Math., 17, pp. 416–429.

R. L. Graham, E. L. Lawler, J. K. Lenstra, and A. H. G. Rinnooy Kan (1979), *Optimization and approximation in deterministic sequencing and scheduling. A survey*, Ann. Discrete Math., 5, pp. 287–326.

E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, and D. B. Shmoys (1993), *Sequencing and scheduling: Algorithms and complexity in logistics of production and inventory*, in Handbooks in Operation Research and Management Science, Vol. 4, North–Holland, Amsterdam, pp. 445–522.

J. Leung, ed. (2004), *Handbook of Scheduling: Algorithms, Models, and Performance Analysis*, Chapman and Hall/CRC, Boca Raton, FL.

E. Mokotoff (2001), *Parallel machine scheduling problems: A survey*, Asia-Pacific J. Oper. Res., 18, pp. 193–243.

J. D. Ullman (1976), *Complexity of sequencing problems*, in Computer and Job Shop Scheduling Theory, E. G. Coffman, ed., Wiley, New York, pp. 139–164.

M. Yue (1990), *On the exact upper bound for the multifit processor scheduling algorithm*, Ann. Oper. Res., 24, pp. 233–259.

SIAM J. Discrete Math.
Vol. 21, No. 2, pp. 329–348

GRAPH POWERS, DELSARTE, HOFFMAN, RAMSEY, AND SHANNON*

NOGA ALON† AND EYAL LUBETZKY‡

Abstract. The kth p-power of a graph G is the graph on the vertex set $V(G)^k$, where two k-tuples are adjacent iff the number of their coordinates which are adjacent in G is not congruent to 0 modulo p. The clique number of powers of G is polylogarithmic in the number of vertices; thus graphs with small independence numbers in their p-powers do not contain large homogeneous subsets. We provide algebraic upper bounds for the asymptotic behavior of independence numbers of such powers, settling a conjecture of [N. Alon and E. Lubetzky, *Combinatorica*, 27 (2007), pp. 13–33] up to a factor of 2. For precise bounds on some graphs, we apply Delsarte's linear programming bound and Hoffman's eigenvalue bound. Finally, we show that for any nontrivial graph G, one can point out specific induced subgraphs of large p-powers of G with neither a large clique nor a large independent set. We prove that the larger the Shannon capacity of $\overline{G}$ is, the larger these subgraphs are, and if G is the complete graph, then some p-power of G matches the bounds of the Frankl–Wilson Ramsey construction, and is in fact a subgraph of a variant of that construction.

Key words. graph powers, Delsarte's linear programming bound, eigenvalue bounds, Ramsey theory, cliques and independent sets

AMS subject classifications. 05C69, 05D10, 05E35, 94B65

DOI. 10.1137/060657893

1. Introduction.

The kth Xor graph power of a graph G, $G^{\oplus k}$, is the graph whose vertex set is the Cartesian product $V(G)^k$, where two k-tuples are adjacent iff an odd number of their coordinates is adjacent in G. This product was used in [21] to construct edge colorings of the complete graph with two colors, containing a smaller number of monochromatic copies of K_4 than the expected number of such copies in a random coloring.

In [4], the authors studied the independence number, α, and the clique number, ω, of high Xor powers of a fixed graph G, motivated by problems in coding theory: cliques and independent sets in such powers correspond to maximal codes satisfying certain natural properties. It is shown in [4] that while the clique number of $G^{\oplus k}$ is linear in k, the independence number $\alpha(G^{\oplus k})$ grows exponentially: the limit $\alpha(G^{\oplus k})^{\frac{1}{k}}$ exists and is in the range $[\sqrt{|V(G)|}, |V(G)|]$. Denoting this limit by $x_\alpha(G)$, the problem of determining $x_\alpha(G)$ for a given graph G proves to be extremely difficult, even for simple families of graphs. Using spectral techniques, it is proved in [4] that $x_\alpha(K_n) = 2$ for $n \in \{2, 3, 4\}$, where K_n is the complete graph on n vertices, and it is conjectured that $x_\alpha(K_n) = \sqrt{n}$ for every $n \geq 4$. The best upper bound given in [4] on $x_\alpha(K_n)$ for $n \geq 4$ is $n/2$.

*Received by the editors April 21, 2006; accepted for publication (in revised form) November 20, 2006; published electronically April 27, 2007.
http://www.siam.org/journals/sidma/21-2/65789.html

†School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540, and Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, 69978, Israel (nogaa@tau.ac.il). This author's research was supported in part by the Israel Science Foundation, by a U.S.–Israeli BSF grant, by NSF grant CCR-0324906, by a Wolfensohn fund, and by the State of New Jersey.

‡School of Mathematics, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, 69978, Israel (lubetzky@tau.ac.il). This author's research was partially supported by a Charles Clore Foundation Fellowship.

The graph product we introduce in this work, which generalizes the Xor product, is motivated by Ramsey theory. In [9], Erdős proved the existence of graphs on n vertices without cliques or independent sets of size larger than $O(\log n)$ vertices, and that in fact almost every graph satisfies this property. Ever since, there have been many attempts to provide explicit constructions of such graphs. Throughout the paper, without being completely formal, we call a graph "Ramsey" if it has neither a "large" clique nor a "large" independent set. The famous Ramsey construction of Frankl and Wilson [10] provided a family of graphs on n vertices, FW_n, with a bound of $\exp\left(\sqrt{(2+o(1))\log n \log\log n}\right)$ on the independence and clique numbers, using results from extremal finite set theory. Thereafter, constructions with the same bound were produced in [3] using polynomial spaces and in [11] using low degree matrices. Recently, the old Frankl–Wilson record was broken in [6], where the authors provided, for any $\varepsilon > 0$, a polynomial-time algorithm for constructing a Ramsey graph on n vertices without cliques or independent sets on $\exp\left((\log n)^{\varepsilon}\right)$ vertices. The disadvantage of this latest revolutionary construction is that it involves a complicated algorithm, from which it is hard to tell the structure of the resulting graph.

Relating the above to graph products, the Xor product may be viewed as an operator, $\oplus_k$, which takes a fixed input graph G on n vertices and produces a graph on n^k vertices, $H = G^{\oplus k}$. The results of [4] imply that the output graph H satisfies $\omega(H) \leq nk = O(\log(|V(H)|))$, and that if G is a nontrivial d-regular graph, then H is d'-regular, with $d' \to \frac{1}{2}|V(H)|$ as k tends to infinity. Thus, $\oplus_k$ transforms any nontrivial d-regular graph into a random looking graph, in the sense that it has an edge density of roughly $\frac{1}{2}$ and a logarithmic clique number. However, the lower bound $\alpha(H) \geq \sqrt{|V(H)|}$, which holds for every even k, implies that $\oplus_k$ cannot be used to produce good Ramsey graphs.

In order to modify the Xor product into a method for constructing Ramsey graphs, one may try to reduce the high lower bound on the independence numbers of Xor graph powers. Therefore, we consider a generalization of the Xor graph product, which replaces the modulo 2 (adjacency of two k-tuples is determined by the parity of the number of adjacent coordinates) with some possibly larger modulo $p \in \mathbb{N}$. Indeed, we show that by selecting a larger p, the lower bound on the independence number, $\alpha(H)$, is reduced from $\sqrt{|V(H)|}$ to $|V(H)|^{1/p}$, at the cost of a polynomial increase in $\omega(H)$. The generalized product is defined as follows.

Definition 1.1. *Let $k, p \in \mathbb{N}$. The kth p-power of a graph G, denoted by $G^{k_{(p)}}$, is the graph whose vertex set is the Cartesian product $V(G)^k$, where two k-tuples are adjacent iff the number of their coordinates which are adjacent in G is not congruent to* 0 *modulo p, that is,*

$$(u_1, \ldots, u_k)\ (v_1, \ldots, v_k) \in E(G^k) \quad \textit{iff} \quad |\{i : u_i v_i \in E(G)\}| \not\equiv 0 \pmod{p}.$$

Throughout the paper, we use the abbreviation G^k for $G^{k_{(p)}}$ when there is no danger of confusion.

In section 2 we show that the limit $\alpha(G^k)^{\frac{1}{k}}$ exists and equals $\sup_k \alpha(G^k)^{\frac{1}{k}}$; denote this limit by $x_\alpha^{(p)}$. A simple lower bound on $x_\alpha^{(p)}$ is $|V(G)|^{1/p}$, and algebraic arguments show that this bound is nearly tight for the complete graph: $x_\alpha^{(p)}(K_n) = O(n^{1/p})$. In particular, we obtain that

$$\sqrt{n} \leq x_\alpha(K_n) = x_\alpha^{(2)}(K_n) \leq 2\sqrt{n-1},$$

improving the upper bound of $n/2$ for $n \geq 4$ given in [4], and determining that the behavior of x_α for complete graphs is as stated in Question 4.1 of [4] up to a factor of 2.

For the special case $G = K_n$, it is possible to apply coding theory techniques in order to bound $x_\alpha^{(p)}(G)$. The problem of determining $x_\alpha^{(p)}(K_n)$ can be translated into finding the asymptotic maximum size of a code over the alphabet $[n]$, in which the Hamming distance between any two codewords is divisible by p. The related problem for *linear* codes over a field has been well studied: see, e.g., [23] for a survey on this subject. However, as we later note in section 2, the general nonlinear case proves to be quite different, and the upper bounds on linear divisible codes do not hold for $x_\alpha^{(p)}(K_n)$. Yet, other methods for bounding sizes of codes are applicable. In section 3 we demonstrate the use of Delsarte's linear programming bound in order to obtain precise values of $\alpha(K_3^{k_{(3)}})$. We show that $\alpha(K_3^{k_{(3)}}) = 3^{k/2}$ whenever $k \equiv 0 \pmod 4$, while $\alpha(K_3^{k_{(3)}}) < \frac{1}{2}3^{k/2}$ for $k \equiv 2 \pmod 4$; hence the series $\alpha(K_3^{k+1_{(3)}})/\alpha(K_3^{k_{(3)}})$ does not converge to a limit.

Section 4 gives a general bound on $x_\alpha^{(p)}$ for d-regular graphs in terms of their eigenvalues, using Hoffman's eigenvalue bound. The eigenvalues of p-powers of G are calculated using tensor products of matrices over $\mathbb{C}$, in a way somewhat similar to performing a Fourier transform on the adjacency matrix of G. This method may also be used to derive tight results on $\alpha(G^{k_{(p)}})$, and we demonstrate this on the above-mentioned case of $p = 3$ and the graph K_3, where we compare the results with those obtained in section 3 by the Delsarte bound.

Section 5 shows, using tools from linear algebra, that indeed the clique number of $G^{k_{(p)}}$ is polylogarithmic in k, and thus p-powers of graphs attaining the lower bound of $x_\alpha^{(p)}$ are Ramsey. We proceed to show a relation between the Shannon capacity of the complement of G, $c(\overline{G})$, and the Ramsey properties of p-powers of G. Indeed, for any nontrivial graph G, we can point out a large Ramsey-induced subgraph of some p-power of G. The larger $c(\overline{G})$ is, the larger these Ramsey subgraphs are. When $G = K_p$ for some prime p, we obtain that $H = K_p^{p^2_{(p)}}$ is a Ramsey graph matching the bound of Frankl and Wilson, and in fact, H contains an induced subgraph which is a modified variant of FW_{N_1} for some N_1 and is contained in another variant of FW_{N_2} for some N_2. The method of proving these bounds on $G^{k_{(p)}}$ provides yet another (simple) proof for the Frankl–Wilson result.

2. Algebraic lower and upper bounds on $x_\alpha^{(p)}$.

In this section, we define the parameter $x_\alpha^{(p)}$ and provide lower and upper bounds for it. The upper bounds follow from algebraic arguments, using graph representation by polynomials.

2.1. The limit of independence numbers of p-powers.

The following lemma establishes that $x_\alpha^{(p)}$ exists and gives simple lower and upper bounds on its range for graphs on n vertices.

LEMMA 2.1. *Let G be a graph on n vertices, and let $p \geq 2$. The limit of $\alpha(G^{k_{(p)}})^{\frac{1}{k}}$ as $k \to \infty$ exists, and, denoting it by $x_\alpha^{(p)}(G)$, it satisfies*

$$n^{1/p} \leq x_\alpha^{(p)}(G) = \sup_k \alpha(G^{k_{(p)}})^{\frac{1}{k}} \leq n.$$

Proof. Observe that if I and J are independent sets of G^k and G^l, respectively, then the set $I \times J$ is an independent set of G^{k+l}, as the number of adjacent coordinates between any two k-tuples of I and between any two l-tuples of J is $0 \pmod p$. Therefore, the function $g(k) = \alpha(G^k)$ is supermultiplicative and strictly positive, and we may apply Fekete's lemma (cf., e.g., [15, p. 85]) to obtain that the limit of $\alpha(G^k)^{\frac{1}{k}}$ as $k \to \infty$ exists, and satisfies

$$\lim_{k\to\infty} \alpha(G^k)^{\frac{1}{k}} = \sup_k \alpha(G^k)^{\frac{1}{k}}. \tag{2.1}$$

Clearly, $\alpha(G^k) \leq n^k$, and it remains to show the lower bound on $x_\alpha^{(p)}$. Notice that the following set is an independent set of G^p:

$$I = \{\ (u, \ldots, u)\ :\ u \in V(G)\} \subset G^p,$$

since for all $u, v \in V(G)$ there are either 0 or p adjacent coordinates between the two corresponding p-tuples in I. By (2.1), we obtain that $x_\alpha^{(p)}(G) \geq |I|^{1/p} = n^{1/p}$. □

2.2. Bounds on $x_\alpha^{(p)}$ of complete graphs. While the upper bound $|V(G)|$ on $x_\alpha^{(p)}(G)$ is clearly attained by an edgeless graph, proving that a family of graphs attains the lower bound requires some effort. The next theorem states that complete graphs achieve the lower bound of Lemma 2.1 up to a constant factor.

THEOREM 2.2. *The following holds for all integers $n, p \geq 2$:*

$$x_\alpha^{(p)}(K_n) \leq 2^{H(1/p)}(n-1)^{1/p}, \tag{2.2}$$

where $H(x) = -x\log_2(x) - (1-x)\log_2(1-x)$ is the binary entropy function. In particular, $x_\alpha^{(p)}(K_n) = \Theta(n^{1/p})$. In the special case where $n = p = q^r$ for some prime q and $r \geq 1$, the lower bound roughly matches the upper bound:

$$p^{\frac{2}{p+1}} \leq x_\alpha^{(p)}(K_p) \leq \left(\mathrm{e}p^2\right)^{1/p}.$$

Taking $p = 2$ and noting that $H(\frac{1}{2}) = 1$, we immediately obtain the following corollary for Xor graph products, which determines the asymptotic behavior of x_α for complete graphs.

COROLLARY 2.3. *For all $n \geq 2$, the complete graph on n vertices satisfies*

$$\sqrt{n} \leq x_\alpha(K_n) \leq 2\sqrt{n-1}.$$

Proof of Theorem 2.2. The upper bound will follow from an argument on polynomial representations, an approach which was used in [3] to bound the Shannon capacity of certain graphs. Take $k \geq 1$, and consider the graph $H = K_n^k$. For every vertex of H, $u = (u_1, \ldots, u_k)$, we define the following polynomial in $\mathbb{R}[x_{i,j}]$, where $i \in [k]$, $j \in [n]$:

$$f_u(x_{1,1}, \ldots, x_{k,n}) = \prod_{t=1}^{\lfloor k/p \rfloor} \left(k - tp - \sum_{i=1}^{k} x_{i,u_i}\right). \tag{2.3}$$

Next, give the following assignment of values for $\{x_{i,j}\}$, x_v, to each $v = (v_1, \ldots, v_k) \in V(H)$:

$$x_{i,j} = \delta_{v_i,j}, \tag{2.4}$$

where δ is the Kronecker delta. Definitions (2.3) and (2.4) imply that for every two such vertices $u = (u_1, \ldots, u_k)$ and $v = (v_1, \ldots, v_k)$ in $V(H)$,

$$f_u(x_v) = \prod_{t=1}^{\lfloor k/p \rfloor} \left(k - tp - \sum_{i=1}^{k} \delta_{u_i,v_i}\right) = \prod_{t=1}^{\lfloor k/p \rfloor} \left(|\{i\ :\ u_i \neq v_i\}| - tp\right). \tag{2.5}$$

Notice that, by the last equation, $f_u(x_u) \neq 0$ for all $u \in V(H)$, and consider two distinct nonadjacent vertices $u, v \in V(H)$. The Hamming distance between u and v (considered as vectors in $\mathbb{Z}_n^k$) is by definition 0 (mod p) (and is not zero). Thus, (2.5) implies that $f_u(x_v) = 0$.

Recall that for all u, the assignment x_u gives values $x_{i,j} \in \{0,1\}$ for all i, j, and additionally $\sum_{j=1}^{n} x_{i,j} = 1$ for all i. Therefore, it is possible to replace all occurrences of $x_{i,n}$ by $1 - \sum_{j=1}^{n-1} x_{i,j}$ in each f_u, and then proceed and reduce the obtained result modulo the polynomials,

$$\bigcup_{i\in[k]} \left(\{x_{i,j}^2 - x_{i,j} : j \in [n]\} \;\cup\; \{x_{i,j}x_{i,l} : j, l \in [n], j \neq l\}\right),$$

without affecting the value of the polynomials on the above-defined substitutions. In other words, after replacing $x_{i,n}$ by $1 - \sum_{j<n} x_{i,j}$, we repeatedly replace $x_{i,j}^2$ by $x_{i,j}$, and let all the monomials containing $x_{i,j}x_{i,l}$ for $j \neq l$ vanish. This gives a set of multilinear polynomials $\{\tilde{f}_u\}$ satisfying

$$\begin{cases} \tilde{f}_u(x_u) \neq 0 & \text{for all } u \in V(H), \\ \tilde{f}_u(x_v) = 0 & \text{for } u \neq v,\ uv \notin E(H), \end{cases}$$

where the monomials of $\tilde{f}_u$ are of the form $\prod_{t=1}^{r} x_{i_t,j_t}$ for some $0 \leq r \leq \lfloor \frac{k}{p} \rfloor$, a set of pairwise distinct indices $\{i_t\} \subset [k]$, and indices $\{j_t\} \subset [n-1]$.

Let $\mathcal{F} = \mathrm{Span}(\{\tilde{f}_u : u \in V(H)\})$, and let I denote a maximum independent set of H. A standard argument shows that $F = \{\tilde{f}_u : u \in I\}$ is linearly independent in $\mathcal{F}$. Indeed, suppose that $\sum_{u\in I} a_u \tilde{f}_u(x) = 0$; then substituting $x = x_v$ for some $v \in I$ gives $a_v = 0$. It follows that $\alpha(H) \leq \dim \mathcal{F}$, and thus

$$\alpha(H) \leq \sum_{r=0}^{\lfloor k/p \rfloor} \binom{k}{r} (n-1)^r \leq \left(2^{H(1/p)}(n-1)^{1/p}\right)^k, \tag{2.6}$$

where in the last inequality we used the fact that $\sum_{i \leq \lambda n} \binom{n}{i} \leq 2^{nH(\lambda)}$ (cf., e.g., the remark following Corollary 4.2 in [2], and also [5, p. 242]). Taking the kth root and letting k grow to ∞, we obtain

$$x_\alpha^{(p)}(K_n) \leq 2^{H(1/p)}(n-1)^{1/p},$$

as required.

In the special case of K_p (that is, $n = p$), note that $2^{H(\frac{1}{p})} = p^{\frac{1}{p}}(\frac{p}{p-1})^{\frac{p-1}{p}} \leq (\mathrm{e}p)^{\frac{1}{p}}$ and hence in this case $x_\alpha^{(p)}(K_p) \leq (\mathrm{e}p^2)^{1/p}$. If $p = q^r$ is a prime-power, we can provide a nearly matching lower bound for $x_\alpha^{(p)}(K_p)$ using a construction of [4], which we shortly describe for the sake of completeness.

Let $\mathcal{L}$ denote the set of all lines with finite slopes in the affine plane $GF(p)$, and write down the following vector w_ℓ for each $\ell \in \mathcal{L}$, $\ell = ax + b$ for some $a, b \in GF(p)$:

$$w_\ell = (a, ax_1 + b, ax_2 + b, \ldots, ax_p + b),$$

where $x_1, \ldots, x_p$ denote the elements of $GF(p)$. For every two distinct lines ℓ, ℓ', if $\ell \| \ell'$, then $w_\ell, w_{\ell'}$ has a single common coordinate (the slope a). Otherwise, $w_\ell, w_{\ell'}$ has a single common coordinate, which is the unique intersection of ℓ, ℓ'. In any case, we obtain that the Hamming distance of w_ℓ and $w_{\ell'}$ is p; hence $W = \{w_\ell : \ell \in \mathcal{L}\}$ is an independent set in K_p^{p+1}. By (2.1), we deduce that

$$x_\alpha^{(p)}(K_p) \geq p^{\frac{2}{p+1}},$$

completing the proof. □

Remark 2.4. The proof of Theorem 2.2 used representation of the vertices of K_n^k by polynomials of kn variables over $\mathbb{R}$. It is possible to prove a similar upper bound on $x_\alpha^{(p)}(K_n)$ using a representation by polynomials of k variables over $\mathbb{R}$. To see this, use the natural assignment of $x_i = v_i$ for $v = (v_1, \ldots, v_k)$, denoting it by x_v, and assign the following polynomial to $u = (u_1, \ldots, u_k)$:

$$f_u(x_1, \ldots, x_k) = \prod_{t=1}^{\lfloor k/p \rfloor} \left(k - tp - \sum_{i=1}^{k} \prod_{\substack{j=1 \\ j \neq u_i}}^{n} \frac{x_i - j}{u_i - j} \right). \tag{2.7}$$

The expression $\prod_{j \neq u_i} \frac{x_i - j}{u_i - j}$ is the monomial of the Lagrange interpolation polynomial and is equal to δ_{x_i, u_i}. Hence, we obtain that $f_u(x_u) \neq 0$ for any vertex u, whereas $f_u(x_v) = 0$ for any two distinct nonadjacent vertices u, v. As the Lagrange monomials yield values in $\{0, 1\}$, we can convert each f_u to a multilinear combination of these polynomials, $\tilde{f}_u$, while retaining the above properties. Note that there are n possibilities for the Lagrange monomials (determined by the value of u_i), and it is possible to express one as a linear combination of the rest. From this point, a calculation similar to that in Theorem 2.2 for the dimension of $\mathrm{Span}(\{\tilde{f}_u : u \in V\})$ gives the upper bound (2.2).

Remark 2.5. The value of $\alpha(K_n^{k_{(p)}})$ corresponds to a maximum size of a code C of k-letter words over $\mathbb{Z}_n$, where the Hamming distance between any two codewords is divisible by p. The case of *linear* such codes when $\mathbb{Z}_n$ is a field, that is, we add the restriction that C is a linear subspace of $\mathbb{Z}_n^k$, has been thoroughly studied; it is equivalent to finding a linear subspace of $\mathbb{Z}_n^k$ of maximal dimension, such that the Hamming weight of each element is divisible by p. It is known for this case that if p and n are relatively prime, then the dimension of C is at most k/p (see [22]), and hence the size of C is at most $n^{k/p}$. However, this bound does not hold for the nonlinear case (notice that this bound corresponds to the lower bound of Lemma 2.1). We give two examples of this:

1. Take $p = 3$ and $n = 4$. The divisible code bound implies an upper bound of $4^{1/3} \approx 1.587$, and yet $x_\alpha^{(3)}(K_4) \geq \sqrt{3} \approx 1.732$. This follows from the geometric construction of Theorem 2.2, which provides an independent set of size 9 in $K_3^{4_{(3)}} \subset K_4^{4_{(3)}}$, using only the coordinates $\{0, 1, 2\}$ (this result can be slightly improved by adding an all-3 vector to the above construction in the 12th power).
2. Take $p = 3$ and $n = 2$. The linear code bound is $2^{1/3} \approx 1.26$, whereas the following construction shows that $\alpha(K_2^{12_{(3)}}) \geq 24$, implying that $x_\alpha^{(3)}(K_2) \geq 24^{1/12} \approx 1.30$. Let $\{v_1, \ldots, v_{12}\}$ denote the rows of a binary Hadamard matrix of order 12 (such a matrix exists by Paley's theorem; cf., e.g., [12]). For all $i \neq j$, v_i and v_j have precisely 6 common coordinates, and hence the set $I = \{v_i\} \cup \{\overline{v}_i\}$ (where $\overline{v}_i$ denotes the complement of v_i modulo 2) is an independent set of size 24 in $K_2^{12_{(3)}}$. In fact, I is a maximum independent set

of $K_2^{12(3)}$, as Delsarte's linear programming bound (described in section 3) implies that $\alpha(K_2^{12(3)}) \leq 24$.

2.3. The value of $x_{\alpha}^{(3)}(K_3)$. While the upper bound of Theorem 2.2 on $x_\alpha^{(p)}(K_n)$ is tight up to a constant factor, the effect of this constant on the independence numbers is exponential in the graph power, and we must resort to other techniques in order to obtain more accurate bounds. For instance, Theorem 2.2 implies that

$$1.732 \approx \sqrt{3} \leq x_\alpha^{(3)}(K_3) \leq 2^{H(\frac{1}{3})} 2^{\frac{1}{3}} = \frac{3}{2^{1/3}} \approx 2.381.$$

In sections 3 and 4, we demonstrate the use of Delsarte's linear programming bound and Hoffman's eigenvalue bound for the above problem, and in both cases obtain the exact value of $\alpha(K_3^{k(3)})$ under certain divisibility conditions. However, if we are merely interested in the value of $x_\alpha^{(3)}(K_3)$, a simpler consideration improves the bounds of Theorem 2.2 and shows that $x_\alpha^{(3)}(K_3) = \sqrt{3}$.

LEMMA 2.6. *For any $k \geq 1$, $\alpha(K_3^{k(3)}) \leq 3 \cdot \sqrt{3}^k$, and in particular $x_\alpha^{(3)}(K_3) = \sqrt{3}$.*

Proof. Treating vertices of K_3^k as vectors of $\mathbb{Z}_3^k$, notice that every two vertices $x = (x_1, \ldots, x_k)$ and $y = (y_1, \ldots, y_k)$ satisfy

$$\sum_{i=1}^{k} (x_i - y_i)^2 \equiv |\{i : x_i \neq y_i\}| \pmod 3,$$

and hence if I is an independent set in K_3^k, then

$$\sum_i (x_i - y_i)^2 \equiv 0 \pmod 3 \text{ for all } x, y \in I.$$

Let I denote a maximum independent set of K_3^k, and let $I_c = \{x \in I : \sum_i x_i^2 \equiv c \pmod 3\}$ for $c \in \{0, 1, 2\}$. For every $c \in \{0, 1, 2\}$ we have

$$\sum_i (x_i - y_i)^2 = 2c - 2x \cdot y \equiv 0 \pmod 3 \text{ for all } x, y \in I_c,$$

and hence $x \cdot y = c$ for all $x, y \in I_c$. Choose c for which $|I_c| \geq |I|/3$, and subtract an arbitrary element $z \in I_c$ from all the elements of I_c. This gives a set J of size at least $|I|/3$, which satisfies

$$x \cdot y = 0 \text{ for all } x, y \in J.$$

Since Span(J) is a self-orthogonal subspace of $\mathbb{Z}_3^k$, its dimension is at most $k/2$, and hence $|J| \leq 3^{k/2}$. Altogether, $\alpha(K_3^k) \leq 3 \cdot \sqrt{3}^k$, as required. □

3. Delsarte's linear programming bound for complete graphs. In this section, we demonstrate how Delsarte's linear programming bound may be used to derive precise values of independence numbers in p-powers of complete graphs. As this method was primarily used on binary codes, we include a short proof of the bound for a general alphabet.

3.1. Delsarte's linear programming bound. The linear programming bound follows from the relation between the distance distribution of codes and the Krawtchouk polynomials, defined as follows.

Definition 3.1. *Let $n \in \mathbb{N}$ and take $q \geq 2$. The Krawtchouk polynomials $\mathcal{K}_k^{n;q}(x)$ for $k = 0, \ldots, n$ are defined by*

$$\mathcal{K}_k^{n;q}(x) = \sum_{j=0}^{k} \binom{x}{j} \binom{n-x}{k-j} (-1)^j (q-1)^{k-j}. \tag{3.1}$$

Definition 3.2. *Let C be an n-letter code over the alphabet $\{1, \ldots, q\}$. The distance distribution of C, $B_0, B_1, \ldots, B_n$, is defined by*

$$B_k = \frac{1}{|C|} |\{(w_1, w_2) \in C^2 : \delta(w_1, w_2) = k\}| \quad (k = 0, \ldots, n),$$

where δ denotes the Hamming distance.

The Krawtchouk polynomials $\{\mathcal{K}_k^{n;q}(x)\}$ are sometimes defined with a normalizing factor of q^{-k}. Also, it is sometimes customary to define the distance distribution with a different normalizing factor, letting $A_k = \frac{B_k}{|C|}$, in which case A_k is the probability that a random pair of codewords has a Hamming distance k.

The Krawtchouk polynomials $\{\mathcal{K}_k^{n;q} : k = 0, \ldots, n\}$ form a system of orthogonal polynomials with respect to the weight function $w(x) = \frac{n!}{\Gamma(1+x)\Gamma(n+1-x)} (q-1)^x$, where Γ is the gamma function. For further information on these polynomials see, e.g., [20].

Delsarte [7] (see also [18]) presented a remarkable method for bounding the maximal size of a code with a given set of restrictions on its distance distribution. This relation is given in the next proposition, for which we include a short proof.

Proposition 3.3. *Let C be a code of n-letter words over the alphabet $[q]$, whose distance distribution is $B_0, \ldots, B_n$. The following holds:*

$$\sum_{i=0}^{n} B_i \mathcal{K}_k^{n;q}(i) \geq 0 \quad \textit{for all } k = 0, \ldots, n. \tag{3.2}$$

Proof. Let $G = \mathbb{Z}_q^n$, and for every two functions $f, g : G \to \mathbb{C}$, define (as usual) their inner product $\langle f, g \rangle$ and their delta-convolution, $f * g$, as

$$\langle f, g \rangle = \int_G f(x) \overline{g(x)} dx = \frac{1}{|G|} \sum_{T \in G} f(T) \overline{g(T)},$$

$$(f * g)(s) = \int_G f(x) \overline{g(x-s)} dx.$$

Denoting the Fourier expansion of f by $f = \sum_{S \in G} \widehat{f}(S) \chi_S$, where $\chi_S(x) = \omega^{S \cdot x}$ and ω is the qth root of unity, it follows that for any $k = 0, \ldots, n$,

$$\sum_{S \in G : |S| = k} \widehat{f}(S) = \frac{1}{|G|} \sum_{i=0}^{n} \mathcal{K}_k^{n;q}(i) \sum_{T \in G : |T| = i} f(T), \tag{3.3}$$

where $|S|$ and $|T|$ denote the Hamming weights of $S, T \in G$. Since the delta-convolution satisfies

$$\widehat{f * g}(S) = \widehat{f}(S) \overline{\widehat{g}(S)},$$

every f satisfies

$$\widehat{f*f}(S) = |\widehat{f}(S)|^2 \geq 0. \tag{3.4}$$

Let f denote the characteristic function of the code C, $f(x) = \mathbf{1}_{\{x\in C\}}$, and notice that

$$(f*f)(S) = \frac{1}{|G|}\sum_{T\in G} f(T)\overline{f(T-S)} = \frac{1}{|G|}|\{T : T, T-S \in C\}|,$$

and thus

$$B_i = \frac{|G|}{|C|}\sum_{T:|T|=i}(f*f)(T). \tag{3.5}$$

Putting together (3.3), (3.4), and (3.5), we obtain

$$0 \leq \sum_{S:|S|=k}\widehat{f*f}(S) = \frac{1}{|G|}\sum_{i=0}^{n}\mathcal{K}_k^{n;q}(i)\sum_{T:|T|=i}(f*f)(T) = \frac{|C|}{|G|^2}\sum_{i=0}^{n}\mathcal{K}_k^{n;q}(i)B_i,$$

as required. □

Let $F \subset [n]$ be a set of forbidden distances between distinct codewords. Since $|C| = \sum_i B_i$, the following linear program provides an upper bound on the size of any code with no pairwise distances specified by F:

$$\text{maximize} \sum_i B_i \text{ subject to the constraints}$$

$$\left\{\begin{array}{rl} B_0 = 1, & \\ B_i \geq 0 & \text{for all } i, \\ B_i = 0 & \text{for all } i \in F, \\ \sum_{i=0}^{n} B_i\mathcal{K}_k^{n;q}(i) \geq 0 & \text{for all } k = 0,\ldots,n. \end{array}\right.$$

By examining the dual program, it is possible to formulate this bound as a minimization problem. The following proposition has been proved in various special cases (cf., e.g., [8], [16]). For the sake of completeness, we include a short proof of it.

Proposition 3.4. *Let C be a code of n-letter words over the alphabet $[q]$, whose distance distribution is $B_0,\ldots,B_n$. Let $P(x) = \sum_{k=0}^{n}\alpha_k\mathcal{K}_k^{n;q}(x)$ denote an n-degree polynomial over $\mathbb{R}$. If $P(x)$ has the two properties*

$$\alpha_0 > 0 \quad \text{and } \alpha_i \geq 0 \quad \text{for all } i = 1,\ldots,n, \tag{3.6}$$

$$P(d) \leq 0 \quad \text{whenever } B_d > 0 \text{ for } d = 1,\ldots,n, \tag{3.7}$$

then $|C| \leq P(0)/\alpha_0$.

Proof. The MacWilliams transform of the vector $(B_0,\ldots,B_n)$ is defined as follows:

$$B'_k = \frac{1}{|C|}\sum_{i=0}^{n}\mathcal{K}_k^{n;q}(i)B_i. \tag{3.8}$$

By the Delsarte inequalities (stated in Proposition 3.3), $B'_k \geq 0$, and furthermore

$$B'_0 = \frac{1}{|C|}\sum_{i=0}^{n}\mathcal{K}_0^{n;q}(i)B_i = \frac{1}{|C|}\sum_i B_i = 1.$$

Therefore, as (3.6) guarantees that $\alpha_i \geq 0$ for $i > 0$, we get

$$\sum_{k=0}^{n} \alpha_k B'_k \geq \alpha_0. \tag{3.9}$$

On the other hand, $B_0 = 1$, and by (3.7), whenever $B_i > 0$ for some $i > 0$ we have $P(i) \leq 0$, and thus

$$\sum_{i=0}^{n} B_i P(i) \leq P(0). \tag{3.10}$$

Combining (3.9) and (3.10) with (3.8) gives

$$\alpha_0 \leq \sum_{k=0}^{n} \alpha_k B'_k = \frac{1}{|C|} \sum_{i=0}^{n} B_i \sum_{k=0}^{n} \alpha_k \mathcal{K}_k^{n;q}(i) = \frac{1}{|C|} \sum_{i=0}^{n} B_i P(i) \leq \frac{P(0)}{|C|},$$

and the result follows. □

We proceed with an application of the last proposition in order to bound the independence numbers of p-powers of complete graphs. In this case, the distance distribution is supported by $\{i : i \equiv 0 \pmod{p}\}$, and in section 3.2 we present polynomials which satisfy the properties of Proposition 3.4 and provide tight bounds on $\alpha(K_3^{k_{(3)}})$.

3.2. Improved estimations of $\boldsymbol{\alpha(K_3^{k_{(3)}})}$. Recall that the geometric construction of Theorem 2.2 describes an independent set of size p^2 in $K_p^{p+1_{(p)}}$ for every p, which is a prime-power. In particular, this gives an independent set of size $3^{k/2}$ in $K_3^{k_{(3)}}$ for every $k \equiv 0 \pmod{4}$. Using Proposition 3.4 we are able to deduce that indeed $\alpha(K_3^k) = 3^{k/2}$ whenever $k \equiv 0 \pmod{4}$, whereas for $k \equiv 2 \pmod{4}$ we prove that $\alpha(K_3^k) < \frac{1}{2}3^{k/2}$.

Theorem 3.5. *The following holds for any even integer k:*

$$\begin{cases} \alpha(K_3^k) = 3^{k/2}, & k \equiv 0 \pmod{4}, \\ \frac{1}{3}3^{k/2} \leq \alpha(K_3^k) < \frac{1}{2}3^{k/2}, & k \equiv 2 \pmod{4}. \end{cases}$$

Proof. Let k be an even integer, and define the following polynomials:

$$P(x) = \frac{2}{3}3^{k/2} + \sum_{\substack{t=1 \\ t \not\equiv 0 (\mathrm{mod}\ 3)}}^{k} \mathcal{K}_t^{k;3}(x), \tag{3.11}$$

$$Q(x) = \frac{2}{3}3^{k/2} + \sum_{\substack{t=0 \\ t \equiv 0 (\mathrm{mod}\ 3)}}^{k} \mathcal{K}_t^{k;3}(x). \tag{3.12}$$

Clearly, both P and Q satisfy (3.6), as $\mathcal{K}_0^{n;q} = 1$ for all n, q. It remains to show that P, Q satisfy (3.7) and to calculate $P(0), Q(0)$. As the following calculation will prove useful later on, we perform it for a general alphabet q and a general modulo p.

Denoting the qth root of unity by $\omega = \mathrm{e}^{2\pi i/q}$, we have

$$\begin{aligned}
\sum_{\substack{t=0\\ t\equiv 0 (\mathrm{mod}\ p)}}^{k} \mathcal{K}_t^{k;q}(s) &= \sum_{\substack{t=0\\ t\equiv 0 (\mathrm{mod}\ p)}}^{k} \sum_{j=0}^{t} \binom{s}{j}\binom{k-s}{t-j}(-1)^j (q-1)^{t-j} \\
&= \sum_{j=0}^{s}\binom{s}{j}(-1)^j \sum_{\substack{l=0\\ j+l\equiv 0 (\mathrm{mod}\ p)}}^{k-s} \binom{k-s}{l}(q-1)^l \\
&= \sum_{j=0}^{s}\binom{s}{j}(-1)^j \sum_{l=0}^{k-s}\binom{k-s}{l}(q-1)^l \frac{1}{q}\sum_{t=0}^{q-1}\omega^{(j+l)t} \\
&= \delta_{s,0}\cdot q^{k-1} + \frac{1}{q}\sum_{t=1}^{q-1}(1+(q-1)\omega^t)^{k-s}(1-\omega^t)^s, \qquad (3.13)
\end{aligned}$$

where the last equality is by the fact that $\sum_{j=0}^{s}\binom{s}{j}(-1)^j = \delta_{s,0}$, and therefore the summand for $t=0$ vanishes if $s \neq 0$ and is equal to q^{k-1} if $s=0$. Repeating the above calculation for $t \not\equiv 0 \pmod p$ gives

$$\begin{aligned}
\sum_{\substack{t=0\\ t\not\equiv 0 (\mathrm{mod}\ p)}}^{k} \mathcal{K}_t^{k;q}(s) &= \sum_{j=0}^{s}\binom{s}{j}(-1)^j \sum_{l=0}^{k-s}\binom{k-s}{l}(q-1)^l \left(1-\frac{1}{q}\sum_{t=0}^{q-1}\omega^{(j+l)t}\right) \\
&= \delta_{s,0}\cdot(q^k - q^{k-1}) - \frac{1}{q}\sum_{t=1}^{q-1}(1+(q-1)\omega^t)^{k-s}(1-\omega^t)^s. \qquad (3.14)
\end{aligned}$$

Define

$$\xi_s = \frac{1}{q}\sum_{t=1}^{q-1}(1+(q-1)\omega^t)^{k-s}(1-\omega^t)^s,$$

and consider the special case $p=q=3$. The fact that $\omega^2 = \overline{\omega}$ implies that

$$\xi_s = \frac{2}{3}\mathrm{Re}\left((1+2\omega)^{k-s}(1-\omega)^s\right) = \frac{2}{3}\mathrm{Re}\left((\sqrt{3}i)^{k-s}(\sqrt{3}\mathrm{e}^{-\frac{\pi}{6}i})^s\right) = \frac{2}{3}\sqrt{3}^k\cos\left(\frac{\pi k}{2}-\frac{2\pi s}{3}\right), \tag{3.15}$$

and for even values of k and $s \equiv 0 \pmod 3$ we deduce that

$$\xi_s = \frac{2}{3}3^{k/2}(-1)^{k/2}. \tag{3.16}$$

Therefore, $\xi_s = \frac{2}{3}3^{k/2}$ whenever $s \equiv 0 \pmod 3$ and $k \equiv 0 \pmod 4$, and (3.14) gives the following for any $k \equiv 0 \pmod 4$:

$$\begin{aligned}
P(0) &= \frac{2}{3}3^{k/2} + \frac{2}{3}3^k - \xi_0 = \frac{2}{3}3^k, \\
P(s) &= \frac{2}{3}3^{k/2} - \xi_s = 0 \ \text{ for any } 0 \neq s \equiv 0 \pmod 3.
\end{aligned}$$

Hence, $P(x)$ satisfies the requirements of Proposition 3.4 and we deduce that for any $k \equiv 0 \pmod 4$,

$$\alpha(K_3^k) \le \frac{P(0)}{\frac{2}{3}3^{k/2}} = 3^{k/2}.$$

As mentioned before, the construction used for the lower bound on $x_\alpha^{(p)}(K_3)$ implies that this bound is indeed tight whenever $4 \mid k$.

For $k \equiv 2 \pmod 4$ and $s \equiv 0 \pmod 3$ we get $\xi_s = -\frac{2}{3}3^{k/2}$, and by (3.13) we get

$$\begin{aligned} Q(0) &= \frac{2}{3}3^{k/2} + 3^{k-1} + \xi_0 = 3^{k-1}, \\ Q(s) &= \frac{2}{3}3^{k/2} + \xi_s = 0 \quad \text{for any } 0 \neq s \equiv 0 \pmod 3. \end{aligned}$$

Again, $Q(x)$ satisfies the requirements of Proposition 3.4 and we obtain the following bound for $k \equiv 2 \pmod 4$:

$$\alpha(K_3^k) \leq \frac{Q(0)}{\frac{2}{3}3^{k/2}+1} = \frac{3^k}{2 \cdot 3^{k/2} + 3} < \frac{1}{2}3^{k/2}.$$

To conclude the proof, take a maximum independent set of size $\sqrt{3}^l$ in K_3^l, where $l = k-2$, for a lower bound of $\frac{1}{3}3^{k/2}$. □

4. Hoffman's bound on independence numbers of p-powers.

In this section we apply spectral analysis in order to bound the independence numbers of p-powers of d-regular graphs. The next theorem generalizes Theorem 2.9 of [4] by considering tensor powers of adjacency matrices whose values are pth roots of unity.

THEOREM 4.1. *Let G be a nontrivial d-regular graph on n vertices, whose eigenvalues are $d = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and let $\lambda = \max\{\lambda_2, |\lambda_n|\}$. The following holds for any $p \geq 2$:*

(4.1)
$$x_\alpha^{(p)}(G) \leq \max\left\{\sqrt{n^2 - 2\left(1-\cos\left(\frac{2\pi}{p}\right)\right)d(n-d)}, \lambda\sqrt{2-2\cos\left(\frac{2\pi}{p}\left\lfloor\frac{p}{2}\right\rfloor\right)}\right\}.$$

Proof. Let $A = A_G$ denote the adjacency matrix of G, and define the matrices B_t for $t \in \mathbb{Z}_p$ as follows:

$$B_t = J_n + (\omega^t - 1)A, \tag{4.2}$$

where $\omega = e^{2\pi i/p}$ is the pth root of unity, and J_n is the all-ones matrix of order n. In other words,

$$(B_t)_{uv} = \omega^{tA_{uv}} = \begin{cases} \omega^t & \text{if } uv \in E(G), \\ 1 & \text{if } uv \notin E(G). \end{cases}$$

By the definition of the matrix tensor product $\otimes$, it follows that for all $u = (u_1, \ldots, u_k)$ and $v = (v_1, \ldots, v_k)$ in G^k,

$$(B_t^{\otimes k})_{u,v} = \omega^{t|\{i \,:\, u_i v_i \in E(G)\}|},$$

and

$$\sum_{t=0}^{p-1}(B_t^{\otimes k})_{u,v} = \begin{cases} p & \text{if } |\{i : u_i v_i \in E(G)\}| \equiv 0 \pmod p, \\ 0 & \text{otherwise.} \end{cases}$$

Recalling that $uv \in E(G^k)$ iff $|\{i : u_i v_i \in E(G)\}| \not\equiv 0 \pmod p$, we get

$$A_{G^k} = J_{n^k} - \frac{1}{p}\sum_{t=0}^{p-1} B_t^{\otimes k} = \frac{p-1}{p}J_{n^k} - \frac{1}{p}\sum_{t=1}^{p-1} B_t^{\otimes k}. \tag{4.3}$$

The above relation enables us to obtain expressions for the eigenvalues of G^k and then apply the following bound, proved by Hoffman (see [13], [17]): every regular nontrivial graph H on N vertices, whose eigenvalues are $\mu_1 \geq \cdots \geq \mu_N$, satisfies

$$\alpha(H) \leq \frac{-N\mu_N}{\mu_1 - \mu_N}. \tag{4.4}$$

Recall that J_n has a single nonzero eigenvalue of n corresponding to the all-ones vector $\underline{1}$. Hence, (4.2) implies that $\underline{1}$ is an eigenvector of B_t with an eigenvalue of $n + (\omega^t - 1)d$, and the remaining eigenvalues of B_t are $\{(\omega^t - 1)\lambda_i : i > 1\}$. By well-known properties of tensor products, we obtain that the largest eigenvalue of $H = G^k$ (which is its degree of regularity) is

$$\begin{aligned}\mu_1 &= n^k - \frac{1}{p}\sum_{t=0}^{p-1}(n + (\omega^t - 1)d)^k = n^k - \frac{1}{p}\sum_{j=0}^{k}\binom{k}{j}(n-d)^{k-j}d^j\sum_{t=0}^{p-1}\omega^{jt} \\ &= n^k - \sum_{\substack{j=0 \\ j\equiv 0 (\text{mod } p)}}^{k}\binom{k}{j}(n-d)^{k-j}d^j, \end{aligned} \tag{4.5}$$

and the remaining eigenvalues are of the form

$$\mu(\lambda_{i_1}, \ldots, \lambda_{i_s}) = -\frac{1}{p}\sum_{t=1}^{p-1}(n + (\omega^t - 1)d)^{k-s}\prod_{j=1}^{s}(\omega^t - 1)\lambda_{i_j}, \tag{4.6}$$

where $0 < s \leq k$ and $1 < i_j \leq n$ for all j (corresponding to an eigenvector which is a tensor-product of the eigenvectors of λ_{i_j} for $j = 1, \ldots, s$ and $\underline{1}^{\otimes k-s}$). The following holds for all such choices of s and $\{\lambda_{i_j}\}$:

$$\begin{aligned}|\mu(\lambda_{i_1}, \ldots, \lambda_{i_s})| &\leq \max_{1\leq t\leq p-1}\left|(n + (\omega^t - 1)d)^{k-s}\prod_{i=1}^{s}(\omega^t - 1)\lambda_{i_j}\right| \\ &\leq \max_{1\leq t\leq p-1}|n + (\omega^t - 1)d|^{k-s}(|\omega^t - 1|\lambda)^s \\ &\leq \max_{1\leq t\leq p-1}\left(\max\{|n + (\omega^t - 1)d|, \lambda|\omega^t - 1|\}\right)^k.\end{aligned}$$

Since for any $1 \leq t \leq p - 1$ we have

$$\begin{aligned}|n + (\omega^t - 1)d|^2 &= n^2 - 2\left(1 - \cos\left(\frac{2\pi t}{p}\right)\right)d(n-d) \leq n^2 - 2\left(1 - \cos\left(\frac{2\pi}{p}\right)\right)d(n-d), \\ |\omega^t - 1|^2 &= 2 - 2\cos\left(\frac{2\pi t}{p}\right) \leq 2 - 2\cos\left(\frac{2\pi}{p}\left\lfloor\frac{p}{2}\right\rfloor\right),\end{aligned}$$

it follows that

$$|\mu(\lambda_{i_1}, \ldots, \lambda_{i_s})| \leq (\max\{\rho_1, \rho_2\})^k,$$

where

$$\begin{aligned}\rho_1 &= \sqrt{n^2 - 2\left(1 - \cos\left(\frac{2\pi}{p}\right)\right)d(n-d)}, \\ \rho_2 &= \lambda\sqrt{2 - 2\cos\left(\frac{2\pi}{p}\left\lfloor\frac{p}{2}\right\rfloor\right)}.\end{aligned}$$

By the same argument, (4.5) gives

$$|\mu_1| \geq n^k - \rho_1^k,$$

and applying Hoffman's bound (4.4), we get

$$\alpha(G^k) \leq \frac{-n^k \mu_{n^k}}{\mu_1 - \mu_{n^k}} \leq \frac{(\max\{\rho_1, \rho_2\})^k}{1 - (\frac{\rho_1}{n})^k + (\frac{\max\{\rho_1,\rho_2\}}{n})^k}. \tag{4.7}$$

To complete the proof, we claim that $\max\{\rho_1, \rho_2\} \leq n$, and hence the denominator in the expression above is $\Theta(1)$ as $k \to \infty$. Clearly, $\rho_1 \leq n$, and a simple argument shows that $\lambda \leq n/2$ and hence $\rho_2 \leq n$ as well. To see this, consider the matrix A^2 whose diagonal entries are d; we have

$$nd = \operatorname{tr} A^2 = \sum_i \lambda_i^2 \geq d^2 + \lambda^2,$$

implying that $\lambda \leq \sqrt{d(n-d)} \leq \frac{n}{2}$. Altogether, taking the kth root and letting k tend to ∞ in (4.7), we obtain that $x_\alpha^{(p)}(G) \leq \max\{\rho_1, \rho_2\}$, as required. □

Examples. For $p = 2, 3$ the above theorem gives

$$x_\alpha^{(2)}(G) \leq \max\{|n - 2d|, 2\lambda\},$$
$$x_\alpha^{(3)}(G) \leq \max\{\sqrt{n^2 - 3d(n-d)}, \sqrt{3}\lambda\}.$$

Since the eigenvalues of K_3 are $\{2, -1, -1\}$, this immediately provides another proof for the fact that $x_\alpha^{(3)}(K_3) \leq \sqrt{3}$. Note that, in general, the upper bounds derived in this method for $x_\alpha^{(p)}(K_n)$ are useful only for small values of n, and tend to n as $n \to \infty$, whereas by the results of section 2 we know that $x_\alpha^{(p)}(K_n) = \Theta(n^{1/p})$.

Consider $d = d(n) = \frac{n}{2} + O(\sqrt{n}\,)$, and let $G \sim G_{n,d}$ denote a random d-regular graph on n vertices. By the results of [14], $\lambda = \max\{\lambda_2, |\lambda_n|\} = O(n^{3/4})$, and thus Theorem 4.1 implies that $x_\alpha^{(2)}(G) = O(n^{3/4})$, and $x_\alpha^{(3)}(G) \leq (1+o(1))\frac{n}{2}$. We note that one cannot hope for better bounds on $x_\alpha^{(3)}$ in this method, as ρ_1 attains its minimum at $d = \frac{n}{2}$.

Remark 4.2. The upper bound (4.1) becomes weaker as p increases. However, if p is divisible by some $q \geq 2$, then clearly any independent set of $G^{k_{(p)}}$ is also an independent set of $G^{k_{(q)}}$, and in particular $x_\alpha^{(p)}(G) \leq x_\alpha^{(q)}(G)$. Therefore, when applying Theorem 4.1 on some graph G, we can replace p by the minimal $q \geq 2$, which divides p. For instance, $x_\alpha^{(4)}(G) \leq x_\alpha^{(2)}(G) \leq \max\{|n-2d|, 2\lambda\}$, whereas substituting $p = 4$ in (4.1) gives the slightly weaker bound $x_\alpha^{(4)}(G) \leq \{\sqrt{(n-d)^2 + d^2}, 2\lambda\}$.

Remark 4.3. In the special case $G = K_n$, the eigenvalues of G are $\{n-1, -1, \ldots, -1\}$, and the general expression for the eigenvalues of G^k in (4.6) takes the form (note that $\lambda_{i_j} = -1$ for all $1 \leq j \leq s$)

$$\mu(s) = -\frac{1}{p} \sum_{t=1}^{p-1} (1 + (n-1)\omega^t)^{k-s} (1 - \omega^t)^s,$$

and as $s > 0$, we obtain the following from (3.14):

$$\mu(s) = \sum_{\substack{t=0 \\ t \not\equiv 0 (\text{mod } p)}}^{k} \mathcal{K}_t^{k;q}(s).$$

Similarly, comparing (4.5) to (3.14) gives

$$\mu_1 = \sum_{\substack{t=0 \\ t\not\equiv 0 (\bmod\ p)}}^{k} \mathcal{K}_t^{k;q}(0).$$

It is possible to deduce this result directly, as K_n^k is a Cayley graph over $\mathbb{Z}_n^k$ with the generator set $S = \{x : |x| \not\equiv 0 \pmod p\}$, where $|x|$ denotes the Hamming weight of x. It is well known that the eigenvalues of a Cayley graph are equal to the character sums of the corresponding group elements. Since for any $k = 0, \ldots, n$ and any $x \in \mathbb{Z}_n^k$ the Krawtchouk polynomial $\mathcal{K}_k^{n;q}$ satisfies

$$\mathcal{K}_k^{n;q}(|x|) = \sum_{y \in \mathbb{Z}_n^k : |y| = k} \chi_y(x),$$

the eigenvalue corresponding to $y \in \mathbb{Z}_n^k$ is

$$\mu(y) = \sum_{x \in S} \chi_x(y) = \sum_{\substack{t=0 \\ t\not\equiv 0 \pmod p}}^{k} \sum_{x : |x| = t} \chi_x(y) = \sum_{\substack{t=0 \\ t\not\equiv 0 \pmod p}}^{k} \mathcal{K}_t^{k;q}(|y|).$$

Remark 4.4. The upper bound on $x_\alpha^{(p)}$ was derived from an asymptotic analysis of the smallest eigenvalue μ_{n^k} of G^k. Tight results on $\alpha(G^k)$ may be obtained by a careful analysis of the expression in (4.6). To illustrate this, we consider the case $G = K_3$ and $p = 3$. Combining the previous remark with (3.14) and (3.15), we obtain that the eigenvalues of $K_3^{k(3)}$ are

$$\begin{aligned} \mu_1 &= \frac{2}{3}3^k - \frac{2}{3}\sqrt{3}^k \cos\left(\frac{\pi k}{2}\right), \\ \mu(s) &= -\frac{2}{3}\sqrt{3}^k \cos\left(\frac{\pi k}{2} - \frac{2\pi s}{3}\right) \quad \text{for } 0 < s \leq k. \end{aligned}$$

Noticing that $\mu(s)$ depends only on the values of $s \pmod 3$ and $k \pmod 4$, we can determine the minimal eigenvalue of G^k for each given power k and deduce that

$$\begin{aligned} &\alpha(G^k) \leq 3^{k/2} && \text{if } k \equiv 0 \pmod 4, \\ &\alpha(G^k) \leq \frac{3^{k+1}}{3 + 2 \cdot 3^{(k+1)/2}} < \frac{1}{2}3^{(k+1)/2} && \text{if } k \equiv 1 \pmod 2, \\ &\alpha(G^k) \leq \frac{3^k}{3 + 2 \cdot 3^{k/2}} < \frac{1}{2}3^{k/2} && \text{if } k \equiv 2 \pmod 4, \end{aligned}$$

matching the results obtained by the Delsarte linear programming bound.

5. Ramsey subgraphs in large p-powers of any graph.

In order to prove a polylogarithmic upper bound on the clique sizes of p-powers of a graph G, we use an algebraic argument, similar to the method of representation by polynomials described in the section 2. We note that the same approach provides an upper bound on the size of independent sets. However, for this latter bound, we require another property, which relates the problem to strong graph products and to the Shannon capacity of a graph.

The kth *strong* power of a graph G (also known as the *and* power), denoted by $G^{\wedge k}$, is the graph whose vertex set is $V(G)^k$, where two distinct k-tuples $u \neq v$ are adjacent iff each of their coordinates is either equal or adjacent in G:

$$(u_1,\dots,u_k)(v_1,\dots,v_k) \in E(G^{\wedge k}) \text{ iff for all } i = 1,\dots,k:\ u_i = v_i \text{ or } u_iv_i \in E(G).$$

In 1956, Shannon [19] related the independence numbers of strong powers of a fixed graph G to the effective alphabet size in a zero-error transmission over a noisy channel. Shannon showed that the limit of $\alpha(G^{\wedge k})^{\frac{1}{k}}$ as $k \to \infty$ exists and equals $\sup_k \alpha(G^{\wedge k})^{\frac{1}{k}}$, by supermultiplicativity; this limit is denoted by $c(G)$, the Shannon capacity of G. It follows that $c(G) \geq \alpha(G)$, and in fact equality holds for all perfect graphs. However, for nonperfect graphs, $c(G)$ may exceed $\alpha(G)$, and the smallest (and most famous) example of such a graph is C_5, the cycle on 5 vertices, where $\alpha(C_5) = 2$ and yet $c(C_5) \geq \alpha(C_5^{\wedge 2})^{\frac{1}{2}} = \sqrt{5}$. The seemingly simple question of determining the value of $c(C_5)$ was solved only in 1979 by Lovász [17], who introduced the ϑ-function to show that $c(C_5) = \sqrt{5}$.

The next theorem states the bound on the clique numbers of $G^{k_{(p)}}$ and relates the Shannon capacity of $\overline{G}$, the complement of G, to bounds on independent sets of $G^{k_{(p)}}$.

Theorem 5.1. *Let G denote a graph on n vertices and let $p \geq 2$ be a prime. The clique number of $G^{k_{(p)}}$ satisfies*

$$\omega(G^{k_{(p)}}) \leq \binom{kn+p-1}{p-1}, \tag{5.1}$$

and if I is an independent set of both $G^{k_{(p)}}$ and $\overline{G}^{\wedge k}$, then

$$|I| \leq \binom{kn+\lfloor\frac{k}{p}\rfloor}{\lfloor\frac{k}{p}\rfloor}. \tag{5.2}$$

Moreover, if in addition G is regular, then

$$\omega(G^{k_{(p)}}) \leq \binom{k(n-1)+p}{p-1},\quad |I| \leq \binom{k(n-1)+\lfloor\frac{k}{p}\rfloor+1}{\lfloor\frac{k}{p}\rfloor}. \tag{5.3}$$

The above theorem implies that if S is an independent set of $\overline{G}^{\wedge k}$, then any independent set I of $G^{k_{(p)}}[S]$, the induced subgraph of $G^{k_{(p)}}$ on S, satisfies inequality (5.2). For large values of k, by definition there exists such a set S of size roughly $c(\overline{G})^k$. Hence, there are induced subgraphs of $G^{k_{(p)}}$ of size tending to $c(\overline{G})^k$ whose clique number and independence number are bounded by the expressions in (5.1) and (5.2), respectively.

In the special case $G = K_n$, the graph $\overline{G}^{\wedge k}$ is an edgeless graph for any k, and hence

$$\alpha(K_n^{k_{(p)}}) \leq \binom{k(n-1)+\lfloor\frac{k}{p}\rfloor+1}{\lfloor\frac{k}{p}\rfloor} \leq (\mathrm{e}p(n-1)+\mathrm{e}+o(1))^{k/p},$$

where the $o(1)$-term tends to 0 as $k \to \infty$. This implies an upper bound on $x_\alpha^{(p)}(K_n)$ which nearly matches the upper bound of Theorem 2.2 for large values of p.

Proof. Let $g_1 : V(G) \to \mathbb{Z}_p^m$ and $g_2 : V(G) \to \mathbb{C}^m$, for some integer m, denote two representations of G by m-dimensional vectors satisfying the following for any (not necessarily distinct) $u, v \in V(G)$:

$$\left\{ \begin{array}{ll} g_i(u) \cdot g_i(v) = 0 & \text{if } uv \in E(G) \\ g_i(u) \cdot g_i(v) = 1 & \text{otherwise} \end{array} \right. \quad (i = 1, 2). \tag{5.4}$$

It is not difficult to see that such representations exist for any graph G. For instance, the standard basis of n-dimensional vectors is such a representation for $G = K_n$. In the general case, it is possible to construct such vectors inductively, in a way similar to a Gram–Schmidt orthogonalization process. To see this, define the lower diagonal $|V(G)| \times |V(G)|$ matrix M as follows:

$$M_{k,i} = \left\{ \begin{array}{cl} -\sum_{j=1}^{i-1} M_{k,j} M_{i,j}, & i < k, \ v_i v_k \in E(G), \\ 1 - \sum_{j=1}^{i-1} M_{k,j} M_{i,j}, & i < k, \ v_i v_k \notin E(G), \\ 1, & i = k, \\ 0, & i > k. \end{array} \right.$$

The rows of M satisfy (5.4) for any distinct $u, v \in V(G)$, and it remains to modify the inner product of any vector with itself into 1 without changing the inner products of distinct vectors. This is clearly possible over $\mathbb{Z}_p$ and $\mathbb{C}$ using additional coordinates.

Consider $G^{k_{(p)}}$, and define the vectors $w_u = g_1(u_1) \circ \cdots \circ g_1(u_k)$ for $u = (u_1, \ldots, u_k) \in V(G^k)$, where $\circ$ denotes vector concatenation. By definition

$$w_u \cdot w_v \equiv k - |\{i : u_i v_i \in E(G)\}| \pmod p$$

for any $u, v \in V(G^k)$, and hence if S is a maximum clique of G^k, then $w_u \cdot w_v \not\equiv k \pmod p$ for any $u, v \in S$. It follows that if B is the matrix whose columns are w_u for $u \in S$, then $C = B^t B$ has values which are $k \pmod p$ on its diagonal and entries which are not congruent to k modulo p anywhere else. Clearly, $\text{rank}(C) \leq \text{rank}(B)$, and we claim that $\text{rank}(B) \leq kn$, and that, furthermore, if G is regular, then $\text{rank}(B) \leq k(n-1)+1$. To see this, notice that as the dimension of $\text{Span}(\{g_1(u) : u \in V\})$ is at most n, the dimension of the span of $\{w_u : u \in G^k\}$ is at most kn. If in addition G is regular, define $z = \sum_{u \in V} g_1(u)$ (assuming without loss of generality that $z \neq 0$), and observe that by (5.4), each of the vectors w_u is orthogonal to the following $k-1$ linearly independent vectors:

$$\{z \circ (-z) \circ \underline{0}^{\circ(k-2)}, \ \underline{0} \circ z \circ (-z) \circ \underline{0}^{\circ(k-3)}, \ldots, \ \underline{0}^{\circ(k-2)} \circ z \circ (-z)\}. \tag{5.5}$$

Similarly, the vectors $w'_u = g_2(u_1) \circ \cdots \circ g_2(u_k)$ satisfy the following for any $u, v \in V(G^k)$:

$$w'_u \cdot w'_v = k - |\{i : u_i v_i \in E(G)\}|.$$

Let I denote an independent set of $G^{k_{(p)}}$, which is also an independent set of $\overline{G}^{\wedge k}$. By the definition of $\overline{G}^{\wedge k}$, every $u, v \in I$ shares a coordinate i such that $u_i v_i \in E(G)$, and combining this with the definition of $G^{k_{(p)}}$, we obtain

$$0 < |\{i : u_i v_i \in E(G)\}| \equiv 0 \pmod p \ \text{ for any } u, v \in I.$$

Therefore, for any $u \neq v \in I$,

$$w'_u \cdot w'_v = k - tp \ \text{ for some } \ t \in \left\{1, \ldots, \left\lfloor \frac{k}{p} \right\rfloor \right\},$$

and if B' is the matrix whose columns are w'_u for $u \in I$, then $C' = B'^t B'$ has the entries k on its diagonal and entries of the form $k - tp$, $0 < t \le \lfloor \frac{k}{p} \rfloor$, anywhere else. Again, the definition of g_2 implies that $\operatorname{rank}(C') \le kn$, and in case G is regular, $\operatorname{rank}(C') \le k(n-1)+1$ (each w'_u is orthogonal to the vectors of (5.5) for $z = \sum_{u \in V} g_2(u)$).

Define the following polynomials:

$$f_1(x) = \prod_{\substack{j \in \mathbb{Z}_p \\ j \not\equiv k (\text{mod } p)}} (j - x), \qquad f_2(x) = \prod_{t=1}^{\lfloor \frac{k}{p} \rfloor} (k - tp - x). \tag{5.6}$$

By the discussion above, the matrices D, D' obtained by applying f_1, f_2 on each element of C, C', respectively, are nonzero on the diagonal and zero anywhere else, and, in particular, they are of full rank: $\operatorname{rank}(D) = |S|$ and $\operatorname{rank}(D') = |I|$. Recalling that the ranks of C and C' are at most kn, and at most $k(n-1)+1$ if G is regular, the proof is completed by the following simple lemma of [1].

Lemma 5.2 (see [1]). *Let $B = (b_{i,j})$ be an $n \times n$ matrix of rank d, and let $P(x)$ be an arbitrary polynomial of degree k. Then the rank of the $n \times n$ matrix $(P(b_{i,j}))$ is at most $\binom{k+d}{k}$. Moreover, if $P(x) = x^k$, then the rank of $(P(b_{i,j}))$ is at most $\binom{k+d-1}{k}$.* □

For large values of k, the upper bounds provided by the above theorem are

$$\omega(H) \le \binom{(1+o(1))kn}{p},$$
$$\alpha(H) \le \binom{(1+o(1))kn}{k/p}.$$

This gives the following immediate corollary, which states that large p-powers of any nontrivial graph G contain a large induced subgraph without large homogeneous sets.

Corollary 5.3. *Let G be some fixed nontrivial graph and fix a prime p.*

1. *Let S denote a maximum clique of G, and set $\lambda = \log \omega(G) = \log \alpha(\overline{G})$. For any k, the induced subgraph of $G^{k_{(p)}}$ on S^k, $H = G^{k_{(p)}}[S^k]$, is a graph on $N = \exp(k\lambda)$ vertices which satisfies*

$$\omega(H) = O(\log^p N), \qquad \alpha(H) \le N^{(1+o(1))\frac{\log(np)+1}{p\lambda}}.$$

2. *The above formula holds when taking $\lambda = \frac{\log \alpha(\overline{G}^{\wedge \ell})}{\ell}$ for some $\ell \ge 1$ dividing k, S a maximum clique of $\overline{G}^{\wedge \ell}$, and $H = G^{k_{(p)}}[S^{k/\ell}]$. In particular, for sufficiently large values of k, $G^{k_{(p)}}$ has an induced subgraph H on $N = \exp\big((1-o(1))k \log c(\overline{G})\big)$ vertices satisfying*

$$\omega(H) = O(\log^p N), \qquad \alpha(H) \le N^{(1+o(1))\frac{\log(np)+1}{p \log c(\overline{G})}}.$$

Remark 5.4. In the special case $G = K_n$, where n, p are large and $k > p$, the bound on $\omega(K_n^k)$ is $\binom{(1+o(1))kn}{p}$, whereas the bound on $\alpha(K_n^k)$ is $\binom{(1+o(1))kn}{k/p}$. Hence, the optimal mutual bound on these parameters is obtained at $k = p^2$. Writing $H = K_n^k$, $N = n^k = n^{p^2}$, and $p = n^c$ for some $c > 0$, we get

$$p = \sqrt{\frac{(2c + o(1)) \log N}{\log \log N}}$$

and

$$\max\{\omega(H),\alpha(H)\} \leq ((1+o(1))\mathrm{e}pn)^p = \exp\left(\left(\frac{1+c}{\sqrt{2c}}+o(1)\right)\sqrt{\log N \log\log N}\right).$$

The last expression is minimized for $c = 1$, and thus the best Ramsey construction in p-powers of K_n is obtained at $p = n$ and $k = p^2$, giving a graph H on N vertices with no independence set or clique larger than $\exp\left((1+o(1))\sqrt{2\log N \log\log N}\right)$ vertices. This special case matches the bound of the Frankl–Wilson Ramsey construction, and is in fact closely related to that construction, as we next describe.

The graph FW_N, where $N = \binom{p^3}{p^2-1}$ for some prime p, is defined as follows: its vertices are the N possible choices of (p^2-1)-element sets of $[p^3]$, and two vertices are adjacent iff the intersection of their corresponding sets is congruent to -1 modulo p. Observe that the vertices of the graph $K_n^{k(p)}$ for $n = p$ and $k = p^2$, as described above, can be viewed as k-element subsets of $[kn]$, where the choice of elements is restricted to precisely one element from each of the k subsets $\{(j-1)n+1,\ldots,jn\}$, $j \in [k]$ (the jth subset corresponds to the jth coordinate of the k-tuple). In this formulation, the intersection of two sets corresponds to the number of common coordinates between the corresponding k-tuples. As $k = p^2 \equiv 0 \pmod{p}$, it follows that two vertices in $K_p^{p^2(p)}$ are adjacent iff the intersection of their corresponding sets is not congruent to 0 modulo p. Altogether, we obtain that $K_p^{p^2(p)}$ is an induced subgraph of a slight variant of FW_N, where the differences are in the cardinality of the sets and the criteria for adjacency.

Another relation between the two constructions is the following: one can identify the vertices of $K_2^{p^3(p)}$ with all possible subsets of $[p^3]$, where two vertices are adjacent iff the intersection of their corresponding sets is not congruent to 0 modulo p. In particular, $K_2^{p^3(p)}$ contains all the (p^2-1)-element subsets of $[p^3]$, a variant of FW_N for the above value of N (the difference lies in the criteria for adjacency).

We note that the method of proving Theorem 5.1 can be applied to the graph FW_N, giving yet another simple proof for the properties of this well-known construction.

Acknowledgment. The authors would like to thank Simon Litsyn and Benny Sudakov for useful discussions.

REFERENCES

[1] N. Alon, *Problems and results in extremal combinatorics.* I, Discrete Math., 273 (2003), pp. 31–53.

[2] N. Alon, *Probabilistic methods in extremal finite set theory*, in Extremal Problems for Finite Sets, Bolyai Soc. Math. Stud. 3, P. Frankl, Z. Füredi, G. O. H. Katona, and D. Miklós, eds., Visegrád, Hungary, 1991, pp. 39–57.

[3] N. Alon, *The Shannon capacity of a union*, Combinatorica, 18 (1998), pp. 301–310.

[4] N. Alon and E. Lubetzky, *Codes and Xor graph products*, Combinatorica, 27 (2007), pp. 13–33.

[5] N. Alon and J. H. Spencer, *The Probabilistic Method*, 2nd ed., Wiley, New York, 2000.

[6] B. Barak, A. Rao, R. Shaltiel, and A. Wigderson, 2*-source dispersers for sub-polynomial entropy and Ramsey graphs beating the Frankl-Wilson construction*, in Proceedings of the 38th ACM Symposium on Theory of Computing, 2006, pp. 671–680.

[7] P. Delsarte, *Bounds for unrestricted codes by linear programming*, Philips Res. Rep., 27 (1972), pp. 272–289.

[8] P. Delsarte, *An algebraic approach to the association schemes of coding theory*, Philips Res. Rep. Suppl., 10 (1973), pp. 1–97.

[9] P. Erdős, *Some remarks on the theory of graphs*, Bull. Amer. Math. Soc., 53 (1947), pp. 292–294.

[10] P. Frankl and R. Wilson, *Intersection theorems with geometric consequences*, Combinatorica, 1 (1981), pp. 357–368.

[11] V. Grolmusz, *Low rank co-diagonal matrices and Ramsey graphs*, Electron. J. Combin., 7 (2000).

[12] M. Hall, *Combinatorial Theory*, 2nd ed., Wiley, New York, 1986.

[13] A. J. Hoffman, *On eigenvalues and colorings of graphs*, in Graph Theory and Its Applications, B. Harris, ed., Academic Press, New York, London, 1970, pp. 79–91.

[14] M. Krivelevich, B. Sudakov, V. Vu, and N. Wormald, *Random regular graphs of high degree*, Random Structures Algorithms, 18 (2001), pp. 346–363.

[15] J. H. van Lint and R. M. Wilson, *A Course in Combinatorics*, 2nd ed., Cambridge University Press, Cambridge, UK, 2001.

[16] S. Litsyn, *New upper bounds on error exponents*, IEEE Trans. Inform. Theory, 45 (1999), pp. 385–398.

[17] L. Lovász, *On the Shannon capacity of a graph*, IEEE Trans. Inform. Theory, 25 (1979), pp. 1–7.

[18] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North–Holland, Amsterdam, 1977.

[19] C. E. Shannon, *The zero-error capacity of a noisy channel*, IRE Trans. Inform. Theory, 2 (3) (1956), pp. 8–19.

[20] G. Szegő, *Orthogonal Polynomials*, 4th ed., Amer. Math. Soc. Colloq. Publ. 23, AMS, Providence, RI, 1975.

[21] A. Thomason, *Graph products and monochromatic multiplicities*, Combinatorica, 17 (1997), pp. 125–134.

[22] H. N. Ward, *Divisible codes*, Arch. Math. (Basel), 36 (1981), pp. 485–494.

[23] H. N. Ward, *Divisible codes: A survey*, Serdica Math. J., 27 (2001), pp. 263–278.

SIAM J. Discrete Math.
Vol. 21, No. 2, pp. 349–360

ON MAXIMUM COST $K_{T,T}$-FREE T-MATCHINGS OF BIPARTITE GRAPHS*

MÁRTON MAKAI†

Abstract. Frank examined the maximum $K_{t,t}$-free t-matching problem of simple bipartite graphs. As the C_6-free 2-matching problem is NP-hard (Geelen), this is a promising generalization of restricted 2-matchings. Given an arbitrary family $\mathcal{T}$ of $K_{t,t}$-subgraphs of the underlying graph, a $\mathcal{T}$-free t-matching is a subgraph of maximum degree at most t that contains no member of $\mathcal{T}$. We show that the maximum size $\mathcal{T}$-free t-matching problem also admits a nice min-max formula. Given an integer cost function on the edge-set which is vertex-induced on any member of $\mathcal{T}$, we also show an integer min-max formula for the maximum cost of $\mathcal{T}$-free t-matchings. As the maximum cost C_4-free 2-matching problem is NP-hard (Király), we cannot expect a nice characterization in general.

Key words. submodularity, restricted matchings, integer polyhedra

AMS subject classifications. 05C07, 05C70, 05C65, 90C10, 90C27, 90C46, 90C47, 90C57

DOI. 10.1137/060652282

1. Introduction.

Throughout the paper, we work with the finite simple bipartite graph $G = (V = A \cup B, E)$, and $t \geq 2$ will be an integer. For a graph H, $V(H)$ and $E(H)$ denote, respectively, its set of vertices and edges. If H is a subgraph of G, then its color classes will be denoted by $A(H) = A \cap V(H)$ and $B(H) = B \cap V(H)$. (The same notation will be used also in some other situations.) If H is a graph (hypergraph), $X \subseteq V(H)$, then $H[X]$ denotes the subgraph (subhypergraph) of H induced by X, and $|E(H[X])|$ is denoted by $i_H(X)$. For $X, Y \subseteq V(H)$, $X \cap Y = \emptyset$, $\delta_H(X, Y) = \delta_H(\{X, Y\})$ denotes the set of edges of H going between X and Y, $\delta_H(X)$ stands for $\delta_H(X, V - X)$, and for the singleton $\{v\}$, we use $\delta_H(v)$ rather than $\delta_H(\{v\})$. If δ is replaced by d, then it denotes the cardinality of the corresponding set. In these notations, the graph is sometimes replaced by its edge set, or if the graph (edge set) is clear from the context, its notation is omitted. For a function $g : V \to \mathbb{Z}$ and $X \subseteq V$, we use $g(X) = \sum_{v \in X} g(v)$; we do not distinguish subsets of V and their characteristic functions, nor do we distinguish vectors and functions on the same ground set.

For $f, g : V \to \mathbb{Z}$, $f \leq g$, an *(f, g)-factor* of G is a subgraph H of G such that (s.t.) $f(v) \leq d_H(v) \leq g(v)$ for every $v \in V$. The $(0, g)$-factors are called *g-matching*s, and the (g, g)-factors are called *g-factor*s. In the literature, a matching may consist of multiple copies of an edge; our notion of matching, consisting of subgraphs, is referred to as simple. Since we deal only with simple graphs and simple matchings, the adjective "simple" is omitted. The problem of searching for a matching with a maximum number of edges is known as the *maximum matching problem.* Similarly, given a cost function $c : E \to \mathbb{Z}$, the *maximum cost matching problem* is to search for a matching H maximizing its cost $c(E(H))$. It is known from bipartite matching theory

[*] Received by the editors February 17, 2006; accepted for publication (in revised form) November 6, 2006; published electronically May 7, 2007. This research was supported by the Hungarian National Foundation for Scientific Research, OTKA grants T037547 and TS 049788, by European MCRTN Adonet, contract grant 504438, and by the Egerváry Research Group of the Hungarian Academy of Sciences.

http://www.siam.org/journals/sidma/21-2/65228.html

[†] Department of Operations Research, Eötvös University, Pázmány Péter sétány 1/C, H-1117 Budapest, Hungary, and Communication Networks Laboratory, Pázmány Péter sétány 1/A, H-1117 Budapest, Hungary (marci@cs.elte.hu).

(see, e.g., [10]) that the maximum number of edges of a g-matching in a bipartite graph G is

$$\min_{Z \subseteq V} g(Z) + i_G(V - Z).$$

Similarly, a simple formula is known for the maximum cost of g-matchings (g-factors), although it is better to formulate this in polyhedral terms.

Generalizing this, Cunningham and Geelen proposed investigating the *maximum C_4-free 2-matching problem*, i.e., the problem of finding a maximum 2-matching that does not contain a cycle of length four. Hartvigsen obtained a min-max formula by a combinatorial algorithm and introduced the linear program

$$\begin{aligned}
&(1.1) \qquad && \max x(E), \\
&(1.2) && x \in \mathbb{R}^E,\ 0 \leq x \leq 1, \\
&(1.3) && x(\delta(v)) \leq 2 \quad \text{for every } v \in V, \\
&(1.4) && x(E(C)) \leq 3 \quad \text{for every subgraph } C \text{ of } G \text{ isomorphic to } C_4.
\end{aligned}$$

Clearly, the integer solutions of (1.2)–(1.4) are exactly the C_4-free 2-matchings. Hartvigsen proved the following integrality result.

THEOREM 1 (Hartvigsen [5, 6]). *The optimum of the linear program* (1.1)–(1.4) *is attained on an integer vector, and the optimum of its dual is attained on a half-integer vector.*

Király sharpened this by stating that the dual has, in fact, integer optimal solutions, and he obtained the following theorem by a relatively simple inductive proof.

THEOREM 2 (Király [9]). *For $g : V \to \{0, 1, 2\}$, the maximum number of edges of a C_4-free g-matching of the bipartite graph G is*

$$\min_{Z \subseteq V} g(Z) + i_G(V - Z) - c_2(G[V - Z]),$$

where $c_2(G[V - Z])$ denotes the number of C_4-components of $G[V - Z]$.

The crucial observation in the area was made by Frank. As Geelen proved that the *maximum C_6-free 2-matching problem* is NP-hard [4], Frank proposed generalizing the C_4-free 2-matching problem by *forbidding $K_{t,t}$ subgraphs in t-matchings.* His approach is based on the general set-pair covering theorem of Frank and Jordán [3], thus not leading to a combinatorial algorithm. Later, Király also was able to extend his proof for this case [8].

THEOREM 3 (Frank [2] with $f \equiv t$, Király [8]). *For $g : V \to \{0, 1, 2, \ldots, t\}$, the maximum number of edges of a $K_{t,t}$-free g-matching of the bipartite graph G is*

$$\min_{Z \subseteq V} g(Z) + i_G(V - Z) - c_t(G[V - Z]),$$

where $c_t(G[V - Z])$ denotes the number of $K_{t,t}$-components of $G[V - Z]$.

We emphasize that neither of these two approaches is algorithmic. The proof based on the Frank–Jordán theorem provides a polynomial time algorithm via the ellipsoid method [1, 3], but a purely combinatorial algorithm is not known. It may be possible to extend Hartvigsen's maximum C_4-free 2-matching algorithm for the above case, but it would be much more interesting to see an algorithmic approach via the Frank–Jordán theorem.

Frank's technique yielded another generalization of the problem. A complete bipartite graph $K_{k,l}$ with $k + l > t + 1$ and $k, l \geq 1$ is said to be a large biclique.

Frank [2] proved a min-max formula for the maximum number of edges of a subgraph H of G which contains no large biclique. Notice that, for $t = 1$, this is the maximum matching problem and that, for $t = 2$, this is the $K_{2,2}$-free 2-matching problem.

The natural question of maximum cost restricted matchings also arises. Király noticed [7] that Geelen's proof [4] can be modified to show that the maximum cost C_4-free 2-matching problem is NP-hard. Hence, the more general $K_{t,t}$-free t-matching problem is also NP-hard. On the other hand, Frank's approach enables us to handle *vertex-induced* cost functions, i.e., any cost function $c : E \to \mathbb{Z}$ for which there exists $c' : V \to \mathbb{R}$ s.t. $c(uv) = c'(u) + c'(v)$ for every $uv \in E$ [2].

The main purpose of this paper is to approximate better the borderline of tractability in the maximum cost $K_{t,t}$-free t-matching problem by a polyhedral study. We give min-max formulae for the maximum cost problem for a class of cost functions which is more general than the class of vertex-induced cost functions.

Suppose that we are given a function $g : V \to \{0, 1, 2, \dots, t\}$ and an arbitrary family $\mathcal{T}$ of the $K_{t,t}$-subgraphs of G, which is called the set of forbidden $K_{t,t}$'s. A *$\mathcal{T}$-free g-matching (g-factor, (f, g)-factor)* is a g-matching (g-factor, (f, g)-factor) that contains no member of $\mathcal{T}$. A cost function $c : E \to \mathbb{Z}$ is said to be *$\mathcal{T}$-induced* if, for every $T \in \mathcal{T}$, there exists $c_T : V(T) \to \mathbb{R}$ s.t. $c(uv) = c_T(u) + c_T(v)$ for every $uv \in E(T)$. In other words, $\mathcal{T}$-induced cost functions are vertex-induced on forbidden $K_{t,t}$'s. Formulated in polyhedral terms, our main result is the following.

THEOREM 4. *Let G be a bipartite graph, $f, g : V \to \{0, 1, 2, \dots, t\}$, $f \le g$, and let $\mathcal{T}$ be an arbitrary family of $K_{t,t}$-subgraphs of G. If $c : E \to \mathbb{Z}$ is a $\mathcal{T}$-induced cost function, then the optimum of the linear program*

$$\max cx \tag{1.5}$$

$$x \in \mathbb{R}^E,\ 0 \le x \le 1, \tag{1.6}$$

$$f(v) \le x(\delta(v)) \le g(v) \quad \textit{for every } v \in V, \tag{1.7}$$

$$x(E(T)) \le t^2 - 1 \quad \textit{for every } T \in \mathcal{T} \tag{1.8}$$

and the optimum of its dual are attained on integer vectors.

If there is no forbidden $K_{t,t}$, then all the integer cost functions are $\mathcal{T}$-induced, while if $\mathcal{T}$ is very dense in G, then the set of $\mathcal{T}$-induced cost functions coincides with the set of vertex-induced cost functions. It is not hard to see that Theorem 4 implies the following for maximum $\mathcal{T}$-free g-matchings.

THEOREM 5. *Let G be a bipartite graph, $g : V \to \{0, 1, 2, \dots, t\}$, and let $\mathcal{T}$ be an arbitrary family of $K_{t,t}$-subgraphs of G. Then the maximum number of edges of a $\mathcal{T}$-free g-matching is*

$$\min_{Z \subseteq V} g(Z) + i_G(V - Z) - c_{\mathcal{T}}(G[V - Z]),$$

where $c_{\mathcal{T}}(G[V - Z])$ denotes the number of $\mathcal{T}$-components of $G[V - Z]$.

The proof of Theorem 4 is based on the primal-dual method and on the following theorem, which characterizes the existence of $\mathcal{T}$-free (l, u)-factors.

THEOREM 6. *For $l, u : V \to \mathbb{Z}$, $0 \le l \le u \le t$, G has a $\mathcal{T}$-free (l, u)-factor if and only if, for each $X \subseteq A$ and $Y \subseteq B$,*

$$l(X) \le u(Y) + i_G(X \cup B - Y) - c_{\mathcal{T}}(G[X \cup B - Y]) \tag{1.9}$$

and

$$l(Y) \le u(X) + i_G(Y \cup A - X) - c_{\mathcal{T}}(G[Y \cup A - X]) \tag{1.10}$$

hold, where $c_{\mathcal{T}}(G)$ denotes the number of components of G that are members of $\mathcal{T}$.

In the rest of the paper, the proofs of Theorems 4 and 6 are presented. The proof of the latter result is based on a slightly modified version of the Frank–Jordán theorem. The goal of this generalization is twofold. The first aim is to end up with a proof of Theorem 6, while the second one is to show that there are possible generalizations of the Frank–Jordán theorem where the uncrossing operation is not so apparent.

2. Proof of Theorem 4.

Proof of Theorem 4. Let c be a $\mathcal{T}$-induced cost function, and let $(y, \pi, \lambda, z) \geq 0$, $y : E \to \mathbb{Z}$, $\pi : V \to \mathbb{Z}$, $\lambda : V \to \mathbb{Z}$, $z : \mathcal{T} \to \mathbb{Z}$, be a (not necessarily optimal) integer dual solution, where π_v is associated with the constraint $x(\delta(v)) \leq g(v)$, and λ_v is associated with $-x(\delta(v)) \leq -f(v)$. (Note that there always exists a dual solution, say $y = c$, $\pi = 0$, $\lambda = 0$, $z = 0$.)

An edge $uv \in E$ is said to be *tight* if the dual inequality

$$y_{uv} + \pi_u + \pi_v - \lambda_u - \lambda_v + \sum_{T \in \mathcal{T} : uv \in E(T)} z_T \geq c(uv) \tag{2.1}$$

holds with equality. For $F \subseteq E$ and $\mathcal{S} \subseteq \mathcal{T}$, we introduce the notations $F_{\text{tight}} = \{e \in F : e \text{ is tight}\}$, $F_0 = \{e \in F : y_e = 0\}$, $F_+ = F - F_0$, $\mathcal{S}_0 = \{T \in \mathcal{S} : z_T = 0\}$, and $\mathcal{S}_+ = \mathcal{S} - \mathcal{S}_0$. Therefore, the set of tight edges e with $y_e = 0$ is denoted by $E_{\text{tight},0}$. Moreover, let us choose (y, π, λ, z) so that the vector (w_1, w_2, w_3, w_4) defined by $w_1 = \sum_{e\in E} y_e + \sum_{v \in V}(\pi_v g(v) - \lambda_v f(v)) + \sum_{T \in \mathcal{T}}(t^2-1)z_T$, $w_2 = \sum_{T\in\mathcal{T}}(t^2-1)z_T$, $w_3 = \sum_{e\in E} y_e$, $w_4 = \sum_{v\in V}(\pi_v + \lambda_v)$ is lexicographically as small as possible.

In what follows, we either construct a primal solution which satisfies the complementary slackness conditions with respect to (y, π, λ, z), or we construct a dual solution (y', π', λ', z') s.t.

$$\sum_{e\in E} y'_e + \sum_{v\in V}(\pi'_v g(v) - \lambda'_v f(v)) + \sum_{T\in\mathcal{T}}(t^2-1)z'_T$$
$$< \sum_{e\in E} y_e + \sum_{v\in V}(\pi_v g(v) - \lambda_v f(v)) + \sum_{T\in\mathcal{T}}(t^2-1)z_T.$$

First, we need some technical observations.

Lemma 7.

(i) $E_+ \subseteq E_{\text{tight}}$.

(ii) *The members of $\mathcal{T}_+$ are disjoint.*

(iii) *For any $v \in V$, either $g(v) = 0$ and $\delta(v)_+ = \emptyset$, or $g(v) > 0$ and $|\delta(v)_+| < g(v)$.*

(iv) *If $\lambda_v > 0$, then $|\delta(v)_+| < f(v)$.*

(v) *If $T \in \mathcal{T}_+$ and $e \in E(T)$, then $e \in E_{\text{tight},0}$.*

(vi) *If $v \in \bigcup\{V(T) : T \in \mathcal{T}_+\}$, then $g(v) = t$.*

(vii) *If $v \in \bigcup\{V(T) : T \in \mathcal{T}_+\}$ and $\lambda_v > 0$, then $f(v) = t$.*

(viii) *If $T \in \mathcal{T}_+$, $uv \in E$, $u \in V(T)$, and $v \notin V(T)$, then $y_{uv} = 0$.*

(ix) *At least one of $\pi_v = 0$ or $\lambda_v = 0$ holds for every $v \in V$.*

Proof. If any of the above statements does not hold, then we show that (y, π, λ, z) can be replaced by (y', π', λ', z') so that the corresponding (w'_1, w'_2, w'_3, w'_4) is strictly smaller than (w_1, w_2, w_3, w_4). We define (y', π', λ', z') only on the coordinates where it changes compared to (y, π, λ, z).

(i) If $e \in E_+$ is not tight, then let $y'_e = y_e - 1$, and therefore $w'_1 < w_1$.

(ii) Let $T, S \in \mathcal{T}_+$ s.t. $V(T) \cap V(S) \neq \emptyset$. Then let $z'_T = z_T - 1$ and $z'_S = z_S - 1$, $\pi'_v = \pi_v + 1$ for every $v \in V(T) \cap V(S)$, and $y'_e = y_e + 1$ for every $e \in E[V(T) - V(S)] \cup E[V(S) - V(T)]$. Setting $a = |A(T) \cap A(S)|$ and $b = |B(T) \cap B(S)|$, we have $g(A(T) \cap A(S)) \leq t(a+b)$, $w'_1 \leq w_1 - 2(t^2 - 1) + g(A(T) \cap A(S)) + 2(t-a)(t-b) \leq 2 - a(t-b) - b(t-a) \leq w_1$, and $w'_2 < w_2$.

(iii) If $g(v) = 0$ and $\delta(v)_+ \neq \emptyset$ for some $v \in V$, then let $\pi'_v = \pi_v + 1$ and $y'_{uv} = y_{uv} - 1$ for each $uv \in \delta(v)_+$, and therefore $w'_1 < w_1$. If $|\delta(v)_+| \geq g(v) > 0$, then we do the same operation, but, in this case, $w'_1 \leq w_1$, $w'_2 = w_2$, and $w'_3 < w_3$.

(iv) Otherwise, let $\lambda'_v = \lambda_v - 1$ and $y'_{uv} = y_{uv} - 1$ for every $uv \in \delta(v)_+$, and thus $w'_1 \leq w_1$, $w'_2 = w_2$, $w'_3 \leq w_3$, and $w'_4 < w_4$.

(v) If $e \in E(T)$ is not tight, then let $z'_T = z_T - 1$ and $y_h = y_h + 1$ for every $h \in E(T) - \{e\}$, and thus $w_1 \leq w'_1$ and $w'_2 < w_2$. Let C be the set of components of $(V, E_{\mathrm{tight},0})[V(T)]$. By (iii), $|\delta(v)_+| \leq t - 1$, and so C does not contain singletons. Using that c is $\mathcal{T}$-induced, it can be seen that if $I \in C$ and $ij, jk, kl \in E(I)$, then $il \in E(I)$; i.e., C contains only complete bipartite graphs. If $|C| = 1$, then we are done. Otherwise, T has a $K_{2,2}$-subgraph $(\{a_1, a_2, b_1, b_2\}, \{a_1b_1, a_1b_2, a_2b_1, a_2b_2\})$ s.t. $a_1b_1, a_2b_2 \in E_{\mathrm{tight},0}$ and $a_1b_2, a_2b_1 \in E_{\mathrm{tight},+}$, which contradicts that c is $\mathcal{T}$-induced.

(vi) By (ii), there is a unique $T \in \mathcal{T}_+$ s.t. $v \in T$, and we may assume $v \in A(T)$. If $g(v) \leq t - 1$, then let $z'_T = z_T - 1$ and $\pi'_a = \pi_a$ for every $a \in A(T)$. Then $w'_1 \leq w_1$ and $w'_2 < w_2$.

(vii) Suppose that the statement does not hold for some $v \in A(T)$. Then let $\lambda'_v = \lambda_v - 1$, $z'_T = z_T - 1$, and $y'_{ab} = y_{ab} + 1$ for every $a \in A(T) - \{v\}$, $b \in B(T)$. Then $w'_1 \leq w_1$ and $w'_2 < w_2$.

(viii) If $uv \in E_+$, $u \in A(T)$, and $v \in B - B(T)$, then let $z'_T = z_T - 1$, $y'_{uv} = y_{uv} - 1$, and $\pi'_a = \pi_a + 1$ for every $a \in A(T)$. Then $w'_1 \leq w_1$ and $w'_2 < w_2$.

(ix) If $\pi_v > 0$ and $\lambda_v > 0$ for some $v \in V$, then let $\pi'_v = \pi_v - 1$ and $\lambda'_v = \lambda_v - 1$. Then $w'_1 \leq w_1$, $w'_2 \leq w_2$, $w'_3 \leq w_3$, and $w'_4 < w_4$. □

We define a graph G' from the graph $(V, E_{\mathrm{tight},0})$ by shrinking $A(T)$ and $B(T)$ to new vertices T_A and T_B for each $T \in \mathcal{T}_+$; we delete $\bigcup\{E(T) : T \in \mathcal{T}_+\}$ from the edge set; and finally, to obtain a simple graph, we delete the parallel copies of edges. Thus, the set of old and new vertices in G' is $V_{\mathrm{old}} = V - \bigcup\{V(T) : T \in \mathcal{T}_+\}$ and $V_{\mathrm{new}} = \{T_A, T_B : T \in \mathcal{T}_+\}$. The sets A_{old}, B_{old}, A_{new}, and B_{new} are defined similarly.

In order to construct a $\mathcal{T}$-free (f, g)-factor of G, we try to construct a $\mathcal{T}'$-free (l, u)-factor of G' with $u : V(G') \to \mathbb{Z}$, $l : V(G') \to \mathbb{Z}$, and $\mathcal{T}'$. First, let

$$u(v) = \begin{cases} g(v) - |\delta(v)_+| & \text{if } v \in V_{\mathrm{old}} \text{ and } \lambda_v = 0, \\ f(v) - |\delta(v)_+| & \text{if } v \in V_{\mathrm{old}} \text{ and } \lambda_v > 0, \\ 1 & \text{if } v \in V_{\mathrm{new}}, \end{cases}$$

$$l(v) = \begin{cases} g(v) - |\delta(v)_+| & \text{if } v \in V_{\mathrm{old}} \text{ and } \pi_v > 0, \\ \max(f(v) - |\delta(v)_+|, 0) & \text{if } v \in V_{\mathrm{old}} \text{ and } \pi_v = 0, \\ 1 & \text{if } v \in V_{\mathrm{new}}, v = T_X, \\ & \text{and } \pi_u + \lambda_u > 0 \text{ for each } u \in X \cap V(T), \\ 0 & \text{if } v \in V_{\mathrm{new}}, v = T_X, \\ & \text{and } \pi_u + \lambda_v = 0 \text{ for some } u \in X \cap V(T). \end{cases}$$

Next, let $\mathcal{T}'$ be a family of subgraphs of G' containing each $T \in \mathcal{T}$ s.t. $V(T) \cap \bigcup\{V(S) : S \in \mathcal{T}_+\} = \emptyset$ and $G'[V(T)]$ is isomorphic to $K_{t,t}$.

Case 1. G' has a $\mathcal{T}'$-free (l, u)-factor H'. For satisfying the complementary slackness conditions with respect to (y, π, λ, z), we have to define H s.t.

$$\begin{gathered} E(H) \subseteq E_{\text{tight}}, \\ y_e > 0 \Rightarrow e \in E(H), \\ \pi_v > 0 \Rightarrow |\delta_H(v)| = g(v), \\ \lambda_v > 0 \Rightarrow |\delta_H(v)| = f(v), \\ z_T > 0 \Rightarrow |E(H[T])| = t^2 - 1. \end{gathered}$$

First, each edge of H' has an inverse image at the shrinking operation. Thus, for each $e \in E(H')$, we put exactly one of these edges into H. Next, for $T \in \mathcal{T}_+$, $A(T)$ is incident with at most one edge of (the already defined) H. If there is such an edge, then let a_T be its end vertex in $A(T)$. Otherwise, $\pi_u + \lambda_u = 0$ for some $u \in A(T)$, and thus we let $a_T = u$. We choose b_T similarly. For each $T \in \mathcal{T}_+$, we put $E(T) - \{a_T b_T\}$ into H. Last, we put E_+ into H.

The construction shows that H is an (f, g)-factor of G and that it meets the complementary slackness conditions. We have to prove only that H is $\mathcal{T}$-free. Clearly, for $T \in \mathcal{T}_+$, H does not contain T as a subgraph. Suppose now that H contains T for some $T \in \mathcal{T}_0$ and T shares some vertices with an $S \in \mathcal{T}_+$. Now $d_H(A(S), B - V(S)) \leq 1$ and $d_H(B(S), A - V(S)) \leq 1$, and thus $|B(T) - B(S)| \leq 1$ and $|B(T) - B(S)| \leq 1$. As T and S are different, we may assume $|A(T) - A(S)| = 1$. Thus, either $|B(T) - B(S)| = 0$, or $|B(T) - B(S)| = 1$. It is not hard to see that both lead to contradiction. So consider $T \in \mathcal{T}$ with $V(T) \subseteq V - \bigcup\{V(S) : S \in \mathcal{T}_+\}$, and suppose that T is a subgraph of H. Let C be the set of components of the graph $(V, E_{\text{tight},0})[V(T)]$. Using that c is $\mathcal{T}$-induced, it can be seen that if $I \in C$ and $ij, jk, kl \in E(I)$, then $il \in E(I)$; i.e., C contains only complete bipartite graphs. If $|C| = 1$, then $E(T) \subseteq E_{\text{tight},0}$, T is forbidden in H', and H' cannot contain T, which is a contradiction. As $E(T) \subseteq E_{\text{tight},0} \cup E_{\text{tight},+}$, (iii) implies that C does not contain singletons. Hence, $|C| \geq 2$, and T has a $K_{2,2}$-subgraph $(\{a_1, a_2, b_1, b_2\}, \{a_1 b_1, a_1 b_2, a_2 b_1, a_2 b_2\})$ s.t. $a_1 b_1, a_2 b_2 \in E_{\text{tight},0}$ and $a_1 b_2, a_2 b_1 \in E_{\text{tight},+}$, which contradicts that c is $\mathcal{T}$-induced.

Case 2. G' has no $\mathcal{T}'$-free (l, u)-factor. In this case, we construct the dual solution (y', π', λ', z') so that

$$\begin{aligned} &\sum_{e \in E} y'_e + \sum_{v \in V} (\pi'_v g(v) - \lambda'_v f(v)) + \sum_{T \in \mathcal{T}} (t^2 - 1) z'_T \\ &\qquad\qquad < \sum_{e \in E} y_e + \sum_{v \in V} (\pi_v g(v) - \lambda_v f(v)) + \sum_{T \in \mathcal{T}} (t^2 - 1) z_T. \end{aligned}$$

By Theorem 6, there exists $X' \subseteq A'(G')$, $Y' \subseteq B'(G')$ satisfying

$$l(X') > u(Y') + i_{G'}(X' \cup B'(G') - Y') - c_{\mathcal{T}'}(G'[X' \cup B'(G') - Y']) \tag{2.2}$$

or

$$l(Y') > u(X') + i_{G'}(Y' \cup A'(G') - X') - c_{\mathcal{T}'}(G'[Y' \cup A'(G') - X']). \tag{2.3}$$

By symmetry, we may assume that (2.2) holds. Moreover, we choose X' and Y' so that $X' \cup B'(G') - Y'$ is minimal. Let $\mathcal{C}'$ be the set of $\mathcal{T}'$-components of $G'[X' \cup B'(G') - Y']$. Let I' be the set of edges of $E(G'[X' \cup B'(G') - Y'])$ which are not in $\mathcal{T}'$-components, and let $I'_{\mathcal{T}'}$ be the set of edges of $E(G'[X' \cup B'(G') - Y'])$ in $\mathcal{T}'$-components.

LEMMA 8. *Let $x \in X'$. If $x \in V(T)$ for some $\mathcal{T}'$-component T of $G'[X' \cup B'(G') - Y']$, then $l(x) = t$, and $|I' \cap \delta_{G'}(x)| < l(x)$ otherwise.*

Let $y \in B'(G') - Y'$. If $y \in V(T)$ for some $\mathcal{T}'$-component T of $G'[X' \cup B'(G') - Y']$, then $u(y) = t$, and $|I' \cap \delta_{G'}(y)| < u(y)$ otherwise.

Proof. Otherwise, for the first case, we reset X' to $X' - x$, and we reset Y' to $Y' + y$ for the second one. □

If $x \in X'$, $y \in B'(G') - Y'$, and $xy \in I'$, then, by Lemma 8, $l(x) \geq 2$ and $u(y) \geq 2$, and hence x and y cannot be shrunk vertices. This implies that each edge of I' has a unique inverse image at the shrinking operation. Let X and Y be the inverse images of X' and Y' at the shrinking operation.

The dual solution (y, π, λ, z) changes as follows. Let

$$y'_{ab} = \begin{cases} y_{ab} - 1 & \text{if } a \in A - X \text{ and } b \in Y \text{ and } y_{ab} > 0, \\ y_{ab} + 1 & \text{if } a \in X \text{ and } b \in B - Y \text{ and } y_{ab} > 0, \\ y_{ab} + 1 = 1 & \text{if } ab \in I', \\ y_{ab} & \text{otherwise,} \end{cases}$$

$$\pi'_v = \begin{cases} \pi_v - 1 & \text{if } v \in X \text{ and } \pi_v > 0, \\ \pi_v + 1 & \text{if } v \in Y \text{ and } \lambda_v = 0, \\ \pi_v & \text{otherwise,} \end{cases}$$

$$\lambda'_v = \begin{cases} \lambda_v + 1 & \text{if } v \in X \text{ and } \pi_v = 0, \\ \lambda_v - 1 & \text{if } v \in Y \text{ and } \lambda_v > 0, \\ \lambda_v & \text{otherwise,} \end{cases}$$

$$z'_T = \begin{cases} z_T - 1 & \text{if } T_A \in A'(G') - X',\ T_B \in Y', \text{ and } z_T > 0, \\ z_T + 1 & \text{if } T_A \in X',\ T_B \in B'(G') - Y', \text{ and } z_T > 0, \\ z_T + 1 = 1 & \text{if } T \in I'_{\mathcal{T}'}, \\ z_T & \text{otherwise.} \end{cases}$$

First, it easily follows from the construction that $(y', \pi', \lambda', z') \geq 0$ and the dual inequality (2.1) remains true for (y', π', λ', z').

Next, we have to compute the change of the dual objective function. If $x \in X'$ is a shrunk vertex, then let $x_1, x_2, \ldots, x_t$ be the inverse vertices. Then either $\pi_{x_i} > 0$ and $g(x_i) = t$, or, if $\pi_{x_i} = 0$, then $l(x) > 0$ implies $l(x) = 1$, $\pi_{x_i} + \lambda_{x_i} > 0$, and therefore $f(x_i) = t$ by (vii). Also, $\delta(x_i)_+ = \emptyset$. Next, suppose that $x \in X'$ is not shrunk, but x is in some $\mathcal{T}'$-component of $G'[X' \cup (B'(G') - Y')]$. If $\pi_x > 0$, then $t = l(x) = g(x) - |\delta(x)_+|$ by Lemma 8, and if $\pi_x = 0$, then $t = l(x) = f(x) - |\delta(x)_+|$ also by Lemma 8. Last, suppose that $x \in X'$ is not shrunk, but x is not in some $\mathcal{T}'$-component of $G'[X' \cup (B'(G') - Y')]$. If $\pi_x > 0$, then $l(x) = g(x) - |\delta(x)_+|$, and if $\pi_x = 0$, then $l(x) = f(x) - |\delta(x)_+|$. These together imply

$$\sum_{x \in X} d_+(v) - \sum_{v \in X, \pi_v > 0} g(v) - \sum_{v \in X, \pi_v = 0} f(v) + (t^2 - 1)|\mathcal{T}_+[X \cup B]| = -l(X').$$

Similarly, let $y \in Y'$. If y is shrunk and $y_1, y_2, \ldots, y_t$ are the inverse vertices, then either $\lambda_{y_i} = 0$ and $g(y_i) = t$, or, if $\lambda_{y_i} > 0$, then $f(y_i) = t$ by (vii). If y is not shrunk, then either $\lambda_y = 0$ and $u(y) = g(y) - |\delta(y)_+|$, or $\lambda_y = 0$ and $u(y) = f(y) - |\delta(y)_+|$. Thus,

$$-\sum_{x \in Y} d_+(v) + \sum_{v \in Y, \lambda_v = 0} g(v) + \sum_{v \in Y, \lambda_v > 0} f(v) - (t^2 - 1)|\mathcal{T}_+[A \cup Y]| = u(Y').$$

The change of the dual objective function is

$$\begin{aligned}\sum_{e\in E}(y'_e-y_e)+g(v)\sum_{v\in V}(\pi'_v-\pi_v)+f(v)\sum_{v\in V}(\lambda'_v-\lambda_v)+(t^2-1)\sum_{T\in\mathcal{T}}(z'_T-z_T)&\\ =|E_+[X\cup(B-Y)]|-|E_-[(A-X)\cup Y]|+|I'|&\\ -\sum_{v\in X,\pi_v>0}g(v)-\sum_{v\in X,\pi_v=0}f(v)+\sum_{v\in Y,\lambda_v=0}g(v)+\sum_{v\in Y,\lambda_v>0}f(v)&\\ +\,(t^2-1)\left(|\mathcal{T}_+[X\cup(B-Y)]|-|\mathcal{T}_+[(A-X)\cup Y]|+c_{\mathcal{T}'}(G'[X'\cup(B'(G')-Y')])\right)&\\ =\sum_{v\in X}d_+(v)-\sum_{v\in X,\pi_v>0}g(v)-\sum_{v\in X,\pi_v=0}f(v)+(t^2-1)|\mathcal{T}_+[X\cup B]|&\\ -\sum_{v\in Y}d_+(v)+\sum_{v\in Y,\lambda_v=0}g(v)+\sum_{v\in Y,\lambda_v>0}f(v)-(t^2-1)|\mathcal{T}_+[A\cup Y]|&\\ +|I'|+(t^2-1)c_{\mathcal{T}'}(G'[X'\cup(B'(G')-Y')])&\\ =-l(X')+u(Y')+|I'|+(t^2-1)c_{\mathcal{T}'}(G'[X'\cup(B'(G')-Y')])<0.&\end{aligned}$$

This is a contradiction, finishing the proof. □

3. Proof of Theorem 6. The proof of Theorem 6 follows the structure of Frank's proof for Theorem 3 [2]. However, it can easily be seen that if an arbitrary family of $K_{t,t}$'s is forbidden, then the argument described there does not work. To address this problem, a slight extension of the Frank–Jordán theorem is used.

Consider now a bipartite graph on the vertex set $V = A \cup B$ with edge set $\mathcal{E} = A \times B$, and let $\mathcal{P} = \{(X, Y) : \emptyset \subsetneq X \subseteq A, \emptyset \subsetneq Y \subseteq B\}$ be called the *set of pairs.* For $U \subseteq V$, $\mathcal{P}[U] = \{(X_1, X_2) \in \mathcal{P} : X_1 \cup X_2 \subseteq U\}$. A pair (X_1, X_2) is said to be trivial if at least one of X_1 and X_2 is singleton.

In this section, when we use the word *collection*, this means a multiset of pairs; i.e., a pair belongs to the collection with *multiplicity.* Thus, it is rather convenient to consider a collection of pairs as a nonnegative function mapping $\mathcal{P}$ into $\mathbb{Z}$. The sum of two collections is defined by the sum of these functions. Other algebraic operations are handled similarly.

Two pairs $X, Y \in \mathcal{P}$ are *independent* if $\delta_{\mathcal{E}}(X) \cap \delta_{\mathcal{E}}(Y) = \emptyset$, while a collection of pairs is called independent if its members are pairwise independent. More generally, a collection $\mathcal{F}$ of pairs satisfying $\sum_{U\in\mathcal{F}} \delta_{\mathcal{E}}(U) \le h$ is called *h-independent.* (Remember that sets and their characteristic functions are not distinguished.) We define the partial order $\preceq$ on $\mathcal{P}$ as $(X_1, X_2) \preceq (Y_1, Y_2)$ if and only if $X_1 \subseteq Y_1$ and $X_2 \supseteq Y_2$. Two pairs $X, Y \in \mathcal{P}$ are *comparable* if $X \preceq Y$ or $Y \preceq X$. Two pairs are *crossing* if they are neither comparable nor independent. A collection of pairs is called *cross-free* if it contains no two crossing pairs.

If we are given a function $p : \mathcal{P} \to \mathbb{Z}$, we say that the pair X is positive if $p(X) > 0$. The nonnegative function $p : \mathcal{P} \to \mathbb{Z}$ is said to be *skew-bisupermodular* if, for every two positive crossing pairs $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$, there exists a cross-free collection of positive pairs $\mathcal{G}_{X,Y}$ satisfying

$$(3.1)\qquad \delta_{\mathcal{E}}(X)+\delta_{\mathcal{E}}(Y)\ge\sum_{U\in\mathcal{G}_{X,Y}}\delta_{\mathcal{E}}(U)\quad\text{and}\quad p(X)+p(Y)\le\sum_{U\in\mathcal{G}_{X,Y}}p(U)$$

and, for any collection $\mathcal{H}$ of positive pairs and any sequence of collections

$$(3.2)\qquad \mathcal{H}=\mathcal{H}_0,\mathcal{H}_1,\mathcal{H}_2,\ldots,$$

where $\mathcal{H}_{i+1}$ is obtained from $\mathcal{H}_i$ by decreasing the multiplicities of two crossing pairs $X, Y \in \mathcal{H}_i$ by 1 and increasing the multiplicities of the members of $\mathcal{G}_{X,Y}$ by 1, resulting in a finite sequence. Hence, the last member $\mathcal{U}_{\mathcal{H}}$ is a cross-free collection.

The nonnegative function $z : \mathcal{E} \to \mathbb{Z}$ is said to be a *cover* of p if $p(X_1, X_2) \leq z(\delta_{\mathcal{E}}(X_1, X_2))$ for every $(X_1, X_2) \in \mathcal{P}$. The cardinality of a minimum cover of p is $\tau_p = \min \sum_{e \in \mathcal{E}} z(e)$, where the minimum ranges over all covers z of p. Similarly, $\max \sum_{U \in \mathcal{F}} p(U)$ is denoted by ν_p, where the maximum is taken over all independent collections $\mathcal{F}$.

THEOREM 9. *For any skew-bisupermodular function p, $\nu_p = \tau_p$.*

Proof. $\nu_p \leq \tau_p$ can be seen easily. If $\nu_p = 0$, then $z = 0$ is a cover, and $\mathcal{F} = \emptyset$ is an independent collection, which together give equality. Thus, $\nu_p > 0$ can be assumed. Similarly, we assume $|A||B| \geq 2$. We can observe that, for any $e \in \mathcal{E}$, the function $p_e : \mathcal{P} \to \mathbb{Z}$ defined by $p_e(X) = \max\{p(X) - d_{\{e\}}(X), 0\}$ is a skew-bisupermodular function. If there exists an edge $e \in \mathcal{E}$ s.t. $\nu_{p_e} \leq \nu_p - 1$, then $\tau_p \leq \tau_{p_e} + 1 = \nu_{p_e} + 1 \leq \nu_p$ by using $\tau_p \leq \tau_{p_e} + 1$ by induction, and we are done. Thus, $\nu_{p_e} = \nu_p$ for every $e \in \mathcal{E}$. For each $e \in \mathcal{E}$, let us consider the independent collection $\mathcal{H}_e$ s.t. $p(\mathcal{H}_e) = \nu_{p_e}$, and let $\mathcal{H} = \sum_{e \in \mathcal{E}} \mathcal{H}_e$. By the construction, $p(\mathcal{H}) = \nu_p |A||B|$, and $\mathcal{H}$ is $|A||B| - 1$-independent. According to the definition of skew-bisupermodularity, there exists a cross-free collection $\mathcal{U}_{\mathcal{H}}$ which is $|A||B| - 1$-independent, and $p(\mathcal{U}_{\mathcal{H}}) \geq \nu_p |A||B|$. By applying Dilworth' theorem to the partial order $\preceq$ restricted to $\mathcal{U}_{\mathcal{H}}$, $\mathcal{U}_{\mathcal{H}}$ decomposes into at most $|A||B| - 1$ antichains s.t. each member of $\mathcal{U}_{\mathcal{H}}$ is contained in as many antichains as its multiplicity. But then there is an antichain $\mathcal{A}$ with $p(\mathcal{A}) \geq \frac{\nu_p |A||B|}{|A||B| - 1} > \nu_p$, contradicting the definition of ν_p. □

Let us define $p_\nu^A : 2^A \to \mathbb{Z}$ by

$$p_\nu^A(Z) = \max\{p(\mathcal{G}) : \mathcal{G} \text{ is an independent subcollection of } \mathcal{P}[Z \cup B]\}$$

and $p_\nu^B : 2^B \to \mathbb{Z}$ by

$$p_\nu^B(Z) = \max\{p(\mathcal{G}) : \mathcal{G} \text{ is an independent subcollection of } \mathcal{P}[A \cup Z]\}.$$

The proofs of the following four theorems are the same as the proofs of the analogous theorems of Frank and Jordán [3].

THEOREM 10. *Let $m : V \to \mathbb{Z}$ be a nonnegative function with $m(A) = m(B)$, and let p be a skew-bisupermodular function on $\mathcal{P}$. Then there exists a cover z of p s.t. $z(\delta_{\mathcal{E}}(v)) = m(v)$ for every $v \in V$ if and only if $m(Z) \geq p_\nu^A(Z)$ for every $Z \subseteq A$ and $m(Z) \geq p_\nu^B(Z)$ for every $Z \subseteq B$.*

Proof. The necessity can be easily seen. For the sufficiency, we define $p' : \mathcal{P} \to \mathbb{Z}$, $p' \geq 0$ by $p'(a, B) = m(a)$, $p'(A, b) = m(b)$ for $a \in A$, $b \in B$, and $p'(X_1, X_2) = p(X_1, X_2)$ for the other pairs. $p(a, B) \leq p_\nu^A(a) \leq m(a)$, and similarly $p(A, b) \leq m(b)$. Therefore, $p \geq p'$, and p' is skew-bisupermodular, since it is obtained from p by increasing the value on the trivial pairs (a, B) and (A, b), and these pairs cross no other. Let z be a minimum cover of p'. Clearly, $z(\delta_{\mathcal{E}}(v)) \geq m(v)$ for each $v \in V$. If we have equality for each v, then we are done. Thus, there exists an independent collection $\mathcal{F}$ s.t. $m(A) = m(B) < z(\mathcal{E}) = p'(\mathcal{F})$. $\mathcal{F}$ cannot contain trivial pairs of form both (a, B) and (A, b), and we may assume that it contains only pairs of the first type. Then let $Z = \{a \in A : (a, B) \in \mathcal{F}\}$ and $\mathcal{F}' = \mathcal{F} - \{(a, B) : a \in A\}$. Then $m(A) < p'(\mathcal{F})$ implies $m(A - Z) < p'(\mathcal{F}') = p(\mathcal{F}') \leq p_\nu^A(A - Z)$, which is a contradiction. □

The following statement can be proved similarly.

Theorem 11. *Let $m : A \to \mathbb{Z}$ be a nonnegative function, and let p be a skew-bisupermodular function on $\mathcal{P}$. Then there exists a cover z of p s.t. $z(\delta_{\mathcal{E}}(v)) = m(v)$ for every $v \in A$ if and only if $m(Z) \geq p_\nu^A(Z)$ for every $Z \subseteq A$.*

We call a function $q : V \to \mathbb{Z}$ *supermodular* if $q(X) + q(Y) \leq q(X \cap Y) + q(X \cup Y)$ holds for every $X, Y \subseteq V$. If, in addition, $q(\emptyset) = 0$ and q is monotone increasing (i.e., $q(X) \leq q(Y)$ whenever $X \subseteq Y$), then q is said to be a *contra-polymatroid function.* The polyhedron $C(q) = \{x \in \mathbb{R}^V : x(U) \geq q(U)\ \forall U \subseteq V\}$ is said to be the *contra-polymatroid* defined by q. We will use that contra-polymatroids are integer polyhedra.

Theorem 12. *p_ν^A and p_ν^B are contra-polymatroid functions.*

Proof. By symmetry, it is enough to prove the statement for p_ν^A. It is clear that p_ν^A is nonnegative, monotone increasing and $p_\nu^A(\emptyset) = 0$. Thus, we have to prove supermodularity, i.e., the inequality $p_\nu^A(X) + p_\nu^A(Y) \leq p_\nu^A(X \cap Y) + p_\nu^A(X \cup Y)$ for every $X, Y \subseteq A$. Let $\mathcal{G}_X$ and $\mathcal{G}_Y$ be collections which give the maximum in the definition of $p_\nu^A(X)$ and $p_\nu^A(Y)$. Then we can apply the uncrossing procedure to $\mathcal{G} = \mathcal{G}_X + \mathcal{G}_Y$ which results in a cross-free family $\mathcal{U}_{\mathcal{G}}$. Clearly, for each $ab \in \mathcal{E}$, $d_{\{ab\}}(\mathcal{U}_{\mathcal{G}}) \leq 2$ if $a \in X \cap Y$, $d_{\{ab\}}(\mathcal{U}_{\mathcal{G}}) \leq 1$ if $a \in X \cup Y - X \cap Y$, and $d_{\{ab\}}(\mathcal{U}_{\mathcal{G}}) = 0$ if $a \in A - X \cup Y$. Let $\mathcal{U}_{\mathcal{G}}^{\min}$ consist of the minimal elements $\mathcal{U}_{\mathcal{G}}$ with respect to the partial order $\preceq$. (If a minimal element has multiplicity 2, then it is taken only once.) Then $\mathcal{U}_{\mathcal{G}}^{\min}$ is an independent collection of $\mathcal{P}[(X \cap Y) \cup B]$, and $\mathcal{U}_{\mathcal{G}} - \mathcal{U}_{\mathcal{G}}^{\min}$ is an independent collection of $\mathcal{P}[X \cup Y \cup B]$. Thus, $p_\nu^A(X) + p_\nu^A(Y) \leq p(\mathcal{U}_{\mathcal{G}}^{\min}) + p(\mathcal{U}_{\mathcal{G}} - \mathcal{U}_{\mathcal{G}}^{\min}) \leq p_\nu^A(X \cap Y) + p_\nu^A(X \cup Y)$. □

Theorem 13. *Let $g : V \to \mathbb{Z}$ be a nonnegative function, and let p be a skew-bisupermodular function on $\mathcal{P}$. Then there exists a cover z of p s.t. $z(\delta_{\mathcal{E}}(v)) \leq g(v)$ for every $v \in V$ if and only if $g(Z) \geq p_\nu^A(Z)$ for every $Z \subseteq A$ and $g(Z) \geq p_\nu^B(Z)$ for every $Z \subseteq B$*

Proof. The restriction of g to A is in $C(p_\nu^A)$, and hence there is a minimal member $m : A \to \mathbb{Z}$ of $C(p_\nu^A)$ s.t. $m(a) \leq g(a)$ for each $a \in A$. Similarly, we can consider a minimal member $m : B \to \mathbb{Z}$ of $C(p_\nu^B)$ s.t. $m(b) \leq g(b)$ for each $b \in B$. The integer members of these two contra-polymatroids are the degree sequences of covers, and hence $m(A) = m(B)$. Then there exists a cover with degree function m, which completes the proof. □

Proof of Theorem 6. If there exists a $\mathcal{T}$-free (l, u)-factor, then for $X \subseteq A$ and $Y \subseteq B$, (1.9) and (1.10) clearly hold. We prove now the opposite direction, thus supposing that (1.9) and (1.10) hold for every $X \subseteq A$ and $Y \subseteq B$.

We define a skew-bisupermodular function $p : \mathcal{P} \to \mathbb{Z}$. For every $T \in \mathcal{T}$, let $p(A(T), B(T)) = 1$. If $a \in A$, $\emptyset \neq Z \subseteq B$, and $G[a \cup Z]$ is a complete bipartite graph, then let $p(a, Z) = \max\{|Z| - u(a), 0\}$. Similarly, if $b \in B$, $\emptyset \neq Z \subseteq A$, and $G[Z \cup b]$ is a complete bipartite graph, then let $p(Z, b) = \max\{|Z| - u(b), 0\}$. On other pairs, p is defined to be 0.

Lemma 14. *p is skew-bisupermodular.*

Proof. First, p is nonnegative. Second, suppose that X and Y are positive crossing pairs. If $X = (a, B_1)$ and $Y = (a, B_2)$ are trivial pairs, then let $\mathcal{G}_{X,Y} = \{(a, B_1 \cap B_2), (a, B_1 \cup B_2)\}$. If $X = (a, B_1)$ is trivial and $Y = (T_1, T_2)$ is a forbidden $K_{t,t}$, then let $\mathcal{G}_{X,Y} = \{(a, B_1 \cup T_2)\}$. If $X = (T_1, T_2)$ and $Y = (S_1, S_2)$ are forbidden $K_{t,t}$'s, then let $\mathcal{G}_{X,Y} = \{(a, T_2 \cup S_2) : a \in T_1 \cap S_1\} \cup \{(T_1 \cup S_1, b) : b \in T_2 \cap S_2\}$. It can easily be checked that (3.1) is satisfied.

Thus, we have to prove the existence of the sequence (3.2) for every collection $\mathcal{H}$. Suppose that $\mathcal{H}_0 = \mathcal{H}, \mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_i$ have already been defined. If $\mathcal{H}_i$ is cross-free, then we are done with $\mathcal{U}_{\mathcal{H}} = \mathcal{H}_i$. Otherwise, $\mathcal{H}_i$ contains two crossing pairs X and Y.

Then $\mathcal{H}_{i+1}$ is obtained from $\mathcal{H}_i$ by decreasing the multiplicities of X and Y and by increasing the multiplicities of the members of $\mathcal{G}_{X,Y}$ by 1. The uncrossing operation for a trivial pair and a $K_{t,t}$ or for two $K_{t,t}$'s decreases the sum of multiplicities of $K_{t,t}$'s, and these multiplicities cannot increase in other operations. Hence, such an uncrossing can be applied finitely many times. When two crossing trivial pairs are uncrossed, then $\sum_{U\in\mathcal{H}_j}(|U_1|-|U_2|)^2$ increases, which is upper bounded by $|\mathcal{H}_j|\max\{|A|,|B|\}^2$. Hence, the number of $K_{t,t}$ pairs decreases finitely many times, and between each two such operations, there are finitely many other operations. □

Next, $g : V \to \mathbb{Z}$ is defined by $g(v) = d_G(v) - l(v)$ for every $v \in V$. By applying (1.9) to $X = \{v\}$ and $Y = \emptyset$, we get that $g(v) \geq 0$ for every $v \in A$. Similarly, $g(v) \geq 0$ for every $v \in B$ by (1.10). Now we ask whether a cover z of p exists s.t. $z(\delta_{\mathcal{E}}(v)) \leq g(v)$ for every $v \in V$.

Case 1. There exists such a cover. Then let z be a minimal cover satisfying $z(\delta_{\mathcal{E}}(v)) \leq g(v)$ for every $v \in V$. (Minimal means that $\sum_{e\in\mathcal{E}} z(e)$ is as small as possible.)

LEMMA 15. *If $z(ab) > 0$, then $ab \in E$. Moreover, z is $0-1$ valued, and $\{ab \in E : z(ab) = 0\}$ is a $\mathcal{T}$-free (l,u)-factor.*

Proof. If $z(ab) > 0$, then there exists a positive pair (X_1, X_2) s.t. $a \in X_1$, $b \in X_2$, and $z(\delta_{\mathcal{E}}(X_1, X_2)) = p(X_1, X_2)$. But this implies $ab \in E$. If $z(ab) \geq 2$, then there is a trivial pair $X = (X_1, X_2)$, $a \in X_1$, $b \in X_2$ s.t. $z(\delta_{\mathcal{E}}(X)) = p(X)$. Suppose $|X_1| = 1$. Then z does not cover $(X_1, X_2 - \{b\})$, which is a contradiction.

The fact that $\{ab \in E : z(ab) = 0\}$ is a $\mathcal{T}$-free (l,u)-factor easily follows from the definition of p and g. □

Then, by Lemma 15, we are done.

Case 2. There does not exist such a cover. Then, by Theorem 13, there exists a set $Z \subseteq A$ s.t. $g(Z) < p_\nu^A(Z)$ or $Z \subseteq B$ s.t. $g(Z) < p_\nu^B(Z)$. By symmetry, we may assume the first, and let us choose Z to be minimal among these sets. Let $\mathcal{G}$ be a family which gives the maximum in the definition of $p_\nu^A(Z)$. Suppose, moreover, that the number of $K_{t,t}$ pairs in $\mathcal{G}$ is as small as possible, and, subject to this, the number of its trivial pairs is minimal.

LEMMA 16. *For every $a \in A$, $\mathcal{G}$ contains at most one trivial pair of form (a, B_1). Similarly, for every $b \in B$, $\mathcal{G}$ contains at most one trivial pair of form (A_1, b).*

Proof. If (a, B_1) are (a, B_2) trivial pairs in $\mathcal{G}$, then $\mathcal{G}$ could be replaced by $\mathcal{G} - \{(a, B_1), (a, B_2)\} + (a, B_1 \cup B_2)$. This operation does not decrease $p(\mathcal{G})$ and does not change the number of $K_{t,t}$ pairs but decreases the number of trivial pairs, which is a contradiction. □

LEMMA 17. *If (T_1, T_2) is a $K_{t,t}$ member of $\mathcal{G}$, $a \in T_1$, then $\mathcal{G}$ has no trivial member of form (a, B_1). Similarly, for $b \in T_2$, $\mathcal{G}$ has no trivial member of form (A_1, b).*

Proof. Otherwise, $\mathcal{G}$ could be replaced by $\mathcal{G}-\{(a, B_1), (T_1, T_2)\}+(a, B_1\cup T_2)$. This operation does not decrease $p(\mathcal{G})$ but decreases the number of $K_{t,t}$ pairs in $\mathcal{G}$. □

LEMMA 18. *If (X_1, X_2) and (Y_1, Y_2) are two $K_{t,t}$'s of $\mathcal{G}$, then $(X_1 \cup X_2) \cap (Y_1 \cup Y_2) = \emptyset$.*

Proof. If $(X_1\cup X_2)\cap(Y_1\cup Y_2) \neq \emptyset$, then, by symmetry, we may suppose $X_1\cap Y_1 \neq \emptyset$ and $X_2 \cap Y_2 = \emptyset$. Then we could remove (X_1, X_2) and (Y_1, Y_2) from $\mathcal{G}$ and insert $(a, X_2 \cup Y_2)$ into $\mathcal{G}$ for every $a \in X_1 \cap Y_1$. This operation does not increase $p(\mathcal{G})$ but decreases the number of $K_{t,t}$ pairs in $\mathcal{G}$. □

LEMMA 19. *$\mathcal{G}$ contains no trivial pair of form (a, B_1).*

Proof. Suppose $(a, B_1) \in \mathcal{G}$. By the above lemmas, there is no pair $(X_1, X_2) \in \mathcal{G}$

s.t. $a \in X_1$. Thus, let $Z' = Z - \{a\}$ and $\mathcal{G}' = \mathcal{G} - \{(a, B_1)\}$. By $d_G(a) \geq |B_1|$ and $l(a) \leq u(a)$, $d_G(a) - l(a) \geq |B_1| - u(a)$. And, finally, $g(Z') = g(Z) - (d_G(a) - l(a)) < p(\mathcal{G}) - (|B_1| - u(a)) = p(\mathcal{G}')$, which contradicts the minimal choice of Z. □

Thus, $\mathcal{G}$ is composed of the trivial pairs $(A_1, b_1), (A_2, b_2), \dots, (A_r, b_r)$ and of $K_{t,t}$ pairs $(X_1, Y_2), (X_2, Y_2), \dots, (X_s, Y_s)$ s.t. the sets X_i, Y_j and the singletons b_k are pairwise disjoint. Let $Y = \{b_1, b_2, \dots, b_r\}$ and $X = Z$. We will show that X and Y contradict (1.9).

LEMMA 20. *Let $(T_1, T_2) \in \mathcal{G}$ be a $K_{t,t}$ pair, and let $ab \in E$. If $a \in T_1$ and $b \in B - T_2$, or $a \in Z - T_1$ and $b \in T_2$, then there is a pair $(X_1, X_2) \in \mathcal{G}$ s.t. $a \in X_1$ and $b \in X_2$.*

Proof. Let us prove the $a \in T_1$, $b \in B - T_2$ case. If the statement does not hold, then we can replace $\mathcal{G}$ by $\mathcal{G} - \{(T_1, T_2)\} + \{(a, \{b\} \cup T_2)\}$, which would decrease the number of $K_{t,t}$ pairs of $\mathcal{G}$. □

These lemmas together imply that if $(T_1, T_2) \in \mathcal{G}$ is a $K_{t,t}$ pair, then there is no edge $ab \in E$ s.t. $a \in Z - T_1$ and $b \in T_2$. The definition of $\mathcal{G}$ implies that $A_k = Z$ for every $k = 1, 2, \dots, r$.

Then $(X_1, Y_2), (X_2, Y_2), \dots, (X_s, Y_s)$ define some of the $K_{t,t}$ components of $G[Z \cup B - Y]$. Hence,

$$i_G(Z \cup Y) - u(Y) + s = p(\mathcal{G}) > g(Z) = \sum_{v \in Z} (d_G(v) - l(v)) = i_G(Z \cup B) - l(Z),$$

which completes the proof. □

Acknowledgments. The author wishes to thank András Frank, Tamás Király, Zoltán Király, and Zsuzsanna Vaik for useful discussions on the topic.

REFERENCES

[1] T. Fleiner, *Uncrossing a family of set-pairs*, Combinatorica, 21 (2001), pp. 145–150.

[2] A. Frank, *Restricted t-matchings in bipartite graphs*, Discrete Appl. Math., 131 (2003), pp. 337–346.

[3] A. Frank and T. Jordán, *Minimal edge-coverings of pairs of sets*, J. Combin. Theory Ser. B, 65 (1995), pp. 73–110.

[4] J. F. Geelen, *The C_6-free 2-Factor Problem in Bipartite Graphs is NP-Complete*, manuscript, 1999.

[5] D. Hartvigsen, *The square-free 2-factor problem in bipartite graphs*, in Integer Programming and Combinatorial Optimization (Graz, 1999), Lecture Notes in Comput. Sci. 1610, Springer, Berlin, 1999, pp. 234–241.

[6] D. Hartvigsen, *Finding maximum square-free 2-matchings in bipartite graphs*, J. Combin. Theory Ser. B, 96 (2006), pp. 693–705.

[7] Z. Király, *The minimum cost square-free 2-factor problem in bipartite graphs is NP-complete, private communication.*

[8] Z. Király, *$K_{t,t}$-Free t-Matchings in Bipartite Graphs*, manuscript, 2000.

[9] Z. Király, *C_4-Free 2-Factors in Bipartite Graphs*, Tech. report TR-2001-13, MTA-ELTE EGRES, Egerváry Research Group on Combinatorial Optimization, Budapest, Hungary, 2001.

[10] W. T. Tutte, *A short proof of the factor theorem for finite graphs*, Canad. J. Math., 6 (1954), pp. 347–352.

SIAM J. DISCRETE MATH.
Vol. 21, No. 2, pp. 361–373

AN IMPROVED APPROXIMATION OF THE ACHROMATIC NUMBER ON BIPARTITE GRAPHS[*]

GUY KORTSARZ[†] AND SUNIL SHENDE[†]

Abstract. The achromatic number of a graph $G = (V, E)$ with $|V| = n$ vertices is the largest number k with the following property: the vertices of G can be partitioned into k independent subsets $\{V_i\}_{1 \leq i \leq k}$ such that for every distinct pair of subsets V_i, V_j in the partition, there is at least one edge in E that connects these subsets. We describe a greedy algorithm that computes the achromatic number of a bipartite graph within a factor of $O(n^{4/5})$ of the optimal. Prior to our work, the best known approximation factor for this problem was $n \log \log n / \log n$ as shown by Kortsarz and Krauthgamer [*SIAM J. Discrete Math.*, 14 (2001), pp. 408–422].

Key words. approximation algorithms, coloring of graphs and hypergraphs, graph algorithms

AMS subject classifications. 68W25, 05C15, 05C85

DOI. 10.1137/S0895480104442947

1. Introduction.

Consider a connected, undirected graph $G = (V, E)$ with $|V| = n$ vertices and $|E| = m$ edges. An *achromatic* coloring is an assignment of colors to the vertices of G such that adjacent vertices receive distinct colors and, furthermore, for every pair of distinct colors, there is at least one edge in the graph whose endpoints are assigned those colors. Equivalently, if we *contract* all the vertices with the same color into a single vertex and merge parallel edges, then the resulting graph is a clique. The *achromatic number* of G, denoted as $\psi(G)$, is the *largest* number k, $1 \leq k \leq n$, such that G admits an achromatic coloring with k colors.

The achromatic number problem is to determine $\psi(G)$ for any given graph G. This problem has been studied extensively; for instance, see the surveys of Edwards [4] and of Hughes and MacGillivray [10]. We focus on the algorithmic aspects of the problem. Yannakakis and Gavril [16] proved that the achromatic number problem is NP-hard. Farber et al. [5] showed that the problem remains NP-hard for bipartite graphs. Bodlaender [1] established that the problem is NP-hard on graphs that are simultaneously cographs and interval graphs. Cairnie and Edwards [2] showed that the problem is NP-hard even on trees.

Since an exact solution to the problem appears to be intractable, there has been an interest in approximating the achromatic number. An *approximation algorithm with ratio* $\alpha \geq 1$ for the achromatic number problem takes as input a graph G and returns, in time polynomial in the input size, a number $p \geq \psi(G)/\alpha$ such that G admits an achromatic coloring with p colors.

1.1. Previous work.

In any achromatic coloring, every set of monochromatic vertices in the graph (called a *color class*) is clearly an independent set; to maximize the number of colors, it seems natural to look for small independent sets. Hence, one might use the following *greedy* approach for finding an achromatic coloring with a

[*] Received by the editors April 12, 2004; accepted for publication (in revised form) December 3, 2006; published electronically May 7, 2007. A preliminary version of the paper appeared in *Proceedings of the* 11*th Annual European Symposium on Algorithms* (ESA), Springer, Berlin, 2003, pp. 385–396.

http://www.siam.org/journals/sidma/21-2/44294.html

[†] Department of Computer Science, Rutgers University, 227 Penn Street, Camden, NJ 08102 (guyk@camden.rutgers.edu, shende@camden.rutgers.edu).

large number of colors: iteratively remove from the graph maximal independent sets of small size. However, the problem of finding a *minimum maximal independent set* cannot be approximated within a ratio of $n^{1-\epsilon}$ for any $\epsilon > 0$, unless P = NP [8].

On the other hand, using a semigreedy approach to extracting small independent sets, Chaudhary and Vishwanathan [3] gave the first sublinear approximation algorithm for the achromatic number problem with an approximation ratio of $O(n/\sqrt{\log n})$ for any graph with n vertices. They conjectured that the achromatic number can be approximated within a ratio of $O(\sqrt{\psi(G)})$ for any graph G. In support of their conjecture, they gave an algorithm that returns an $O(\sqrt{\psi(G)}) = O(n^{7/20})$ ratio approximation for graphs G with *girth* (i.e., length of the shortest simple cycle) at least 7. For graphs G with girth at least 6, Krysta and Loryś [14] described an algorithm with approximation ratio $O(\sqrt{\psi(G)}) = O(n^{3/8})$; this ratio was improved slightly to $O(n \log\log n/\log n)$ by Kortsarz and Krauthgamer [11]. This latter paper also showed that the Chaudhary–Vishwanathan conjecture holds for graphs of girth 5 and demonstrated an algorithm for such graphs that approximates the achromatic number within a ratio of $O(n^{1/3})$.

To summarize the upper bounds on approximating the achromatic number for general or bipartite graphs, the best known approximation ratio guarantees are just barely sublinear in the number of vertices. We do know that graphs with large girth (at least 5) admit algorithms with relatively low approximation ratio for the achromatic number. This conclusion hinges on the result that $\psi(G) = \theta(m/n)$ for graphs G with n vertices, m edges, and girth at least 5 [11]. But, considering the complete bipartite graph, we encounter a graph with girth 4 and achromatic number equal to 2 that satisfies $2 \ll m/n = \Omega(n)$.

As for lower bounds, the first hardness of approximation result for general graphs was given by Kortsarz and Krauthgamer [11]. They showed that unless P = NP, the problem cannot be approximated within a ratio of $2 - \epsilon$ for any $\epsilon \geq 0$. In the preliminary conference version of the present paper [13], we stated (without a complete proof) the first nonconstant lower bound for the problem. The result was that unless NP admits a randomized quasi-polynomial-time algorithm, it is impossible to approximate the achromatic number on n-vertex bipartite graphs within a ratio of $(\ln n)^{1/4-\epsilon}$. The methods used for proving the hardness result are built upon a combination of one-round two-prover techniques and zero-knowledge techniques as suggested in Feige et al. [6]. In Kortsarz, Radhakrishnan, and Sivasubramanian [12], the lower bound on the approximation ratio is improved to $\sqrt{\log n}$, with details to appear in the forthcoming journal version of that paper.

1.2. Our contribution.

For graphs with n vertices, all previous results for the achromatic number problem had been unable to obtain approximations better than a factor of $\tilde{O}(n)$, where $\tilde{O}(n)$ is the class of functions that are essentially $O(n)$ ignoring logarithmic factors, i.e., functions of the form $O(n \log^k n)$ for some constant k. In this paper, we give a combinatorial algorithm for the problem when restricted to bipartite graphs. Our algorithm lowers the $\tilde{O}(n)$ barrier on the approximation ratio; specifically, it achieves a ratio of $O(n^{4/5})$ for approximating the achromatic number of every bipartite graph.

2. Preliminaries.

Consider a graph $G = (V, E)$. Following standard terminology, we use $d_G(u)$ and $N_G(u)$ to denote, respectively, the *degree* and the set of *adjacent neighbors* of any vertex u in the graph. Wherever possible, we will simplify notation by omitting G from subscripts when the graph G is clear from the context. For any subset $U \subseteq V$ of the vertices, the subgraph of G induced by U is denoted as $G[U]$. If $G[U]$ has no induced edges, then U is said to be an *independent set* in G.

Given disjoint subsets of vertices U, W in the graph, we say that they are *adjacent* if there exist adjacent vertices $u \in U$ and $v \in W$. The set U *covers* W if every vertex in W is adjacent to some vertex in U.

A *proper k-coloring* of the graph is a mapping that assigns to every vertex a corresponding color in the range $[1, k]$ such that adjacent vertices receive distinct colors. Thus, any proper k-coloring of a graph partitions its vertex set into k independent sets—one per color—called its *color classes.* An *achromatic k-coloring* is a proper coloring where all distinct pairs of color classes are adjacent. The partition formed by the color classes is called an *achromatic partition*; henceforth, we will use the terms achromatic coloring and achromatic partition interchangeably.

The *achromatic number problem* is to determine for any given graph G the *largest* number k such that G has an achromatic k-coloring. Note that, in contrast, the *chromatic number problem* is to determine for graph G the *smallest* number k such that G has a proper k-coloring (which, by minimality of k, is also an achromatic coloring).

The chromatic and achromatic numbers of a graph G are denoted by $\chi(G)$ and $\psi(G)$, respectively. Clearly, $\psi(G) \geq \chi(G)$ and, indeed, the problem of finding the achromatic number, being a maximization problem, is fundamentally different from that of finding the chromatic number, a minimization problem. For instance, when $\psi(G) = O(1)$, an achromatic coloring of G with $\psi(G)$ colors can be found in polynomial time by guessing $\binom{\psi(G)}{2}$ *critical* edges with distinct color combinations on their endpoints (see [5] for a more efficient algorithm). In contrast, even when $\chi(G) = 3$, it can be NP-hard to find a 4-coloring of G [7]. However, the general cases for both problems are known to be NP-hard, as is the bipartite case for the achromatic number problem.

3. Achromatic partitions and matchings.

The following lemmas are well known [15, 4, 3, 14] and are stated here without proof for completeness; we will use these lemmas extensively in the development and the analysis of our algorithm.

Lemma 1. *Let U be a subset of vertices of a graph G. Then any achromatic k-coloring of the subgraph, $G[U]$, can be extended greedily to an achromatic k'-coloring of G with $k' \geq k$ colors.*

Lemma 2. *Consider v, an arbitrary vertex in a graph G, and let $G \setminus v$ denote the graph that results when v and all its incident edges are deleted from G. Then $\psi(G) - 1 \leq \psi(G \setminus v) \leq \psi(G)$.*

Note that in the above lemma, if v is an isolated vertex in G, then its removal does not affect the achromatic number, i.e., $\psi(G \setminus v) = \psi(G)$. Hence, the lemma can be restated more generally as follows. Let U be any *ordered* subset of vertices of G. Suppose that we remove vertices in U from the graph one by one in the order prescribed for U. Let $U_c \subseteq U$ be the subset of vertices v such that v is *not isolated* in the subgraph of G that exists just prior to v's removal. Then $\psi(G \setminus U)$ is bounded above by $\psi(G)$ and below by $\psi(G) - |U_c|$.

A subset of edges of the graph G is called a *matching* if no two distinct edges in the subset share a common endpoint. Let $M = \{(u_1, v_1), \ldots, (u_k, v_k)\}$ be a matching with the sets of endpoints $X = \{u_1, \ldots, u_k\}$ and $Y = \{v_1, \ldots, v_k\}$. Then

- M is said to be *independent* if M is the induced subgraph, $G[X \cup Y]$;
- M is said to be *semi-independent* if X and Y are independent sets, and the edges in M, *ordered as above*, respect the following additional property: for

all $j > i \geq 1$, it holds that u_i is not adjacent to v_j.

Note that in a semi-independent matching, u_i may well be adjacent to v_k for $1 \leq k < i$. Hence, not every semi-independent matching is independent (although the converse is trivially true). A semi-independent matching can be used to obtain an achromatic coloring of the induced subgraph of its vertices as stated in the lemma below; a weaker version of this result, based on using an independent matching, is used in [3].

Lemma 3 (see [15]). *Let M be a semi-independent matching of size $\binom{t}{2}$ in G, and let $V(M)$ be the set of vertices in M. Then an achromatic t-coloring of the subgraph $G[V(M)]$ can be computed efficiently.*

We now focus exclusively on bipartite graphs for the remainder of the paper. Given independent sets of vertices U and V, we denote by $G(U, V, E)$ the bipartite graph G with bipartition (U, V) and edge set $E \subseteq U \times V$. For subsets $U' \subseteq U$ and $V' \subseteq V$, we use the alternative notation $G[U', V']$ for the induced subgraph $G[U' \cup V']$ to make explicit the subsets of the original bipartition that induce the subgraph.

For any vertex $v \in V$ in the bipartite graph $G(U, V, E)$, the induced subgraph, $G[N_G(v), \{v\}]$, consisting of v and its neighbors is called the *star centered at* v in G. Suppose that U does not contain any isolated vertices. A simple iterative procedure that we will call the *star removal algorithm* can be used to compute an achromatic coloring of G as follows. In iteration $i \geq 1$ of the algorithm, we choose an arbitrary surviving vertex $v_i \in V$ of nonzero degree in the current graph. The star centered at v_i in the current graph is removed in the iteration along with all the other edges incident on the star's vertices. The resulting graph is used for the next iteration. Note that the surviving portion of U, in this resulting graph, contains isolated vertices if and only if there are no further edges left. When all the edges of G have been eliminated, we process the sequence of stars removed in successive iterations. If an arbitrary edge (u_i, v_i) is chosen from the ith star, it is not difficult to see that the the resulting sequence of edges, $M = \{(u_1, v_1), \ldots, (u_k, v_k)\}$, forms a semi-independent matching.

Letting $\Delta_G(V)$ denote the largest degree of any vertex in bipartition V of G, it follows that k, the size of the semi-independent matching M, must be at least $|U|/\Delta_G(V)$. In conjunction with Lemmas 1 and 3, we get the following result.

Lemma 4. *Let $G(U, V, E)$ be a bipartite graph with no isolated vertices in U. Then the star removal algorithm produces an achromatic partition of size at least $\Omega(\sqrt{|U|/\Delta_G(V)})$.*

4. Achromatic partitions and reducing congruences.

Hell and Miller [9] define a very natural equivalence relation on the vertex set of any graph G. The relation, also called the *reducing congruence* of G [4, 10], is defined as follows: any pair of vertices of G are *equivalent* if and only if they have exactly the same set of neighbors in the graph.

We denote by $S_G(v)$ the equivalence class of vertex v under the reducing congruence for G; we will drop the subscript in the notation whenever G is clear from the context. Let q be the number of distinct equivalence classes under the reducing congruence for G. Assume that the vertices of G are indexed so that $S(v_1), \ldots, S(v_q)$ denote the distinct equivalence classes. Then v_i is the representative of its equivalence class, $S(v_i)$, and we refer to every member of the class as being a *copy* of v_i. Note that, by definition, two equivalent vertices cannot be adjacent to each other in G; hence $S(v_i)$ is an independent set in G. The following result can be shown.

Theorem 1 (see [11]). *Let G be a bipartite graph whose reducing congruence has q equivalence classes. Then there is an efficient algorithm to compute an achromatic* *coloring of G with at least*

$$\min\{\psi(G)/q,\ \sqrt{\psi(G)}\}$$

colors. Hence, the achromatic number of a bipartite graph can be approximated to within a ratio of $O(\max\{q, \sqrt{\psi(G)}\})$.

Given the graph G and the collection of its equivalence classes under the reducing congruence, the *reduced degree*, $d^*_G(v)$, of any vertex v is the maximum number of pairwise nonequivalent neighbors of v in G. Equivalently, if $\{v_i : 1 \leq i \leq q\}$ is the collection of distinct representatives of the equivalence classes, then the reduced degree of v in G is the degree of its representative in the induced subgraph, $G[\{v_i : 1 \leq i \leq q\}]$.

LEMMA 5. *Let u, w be a pair of vertices of G such that $S(u) \neq S(w)$ and $d^*(u) \leq d^*(w)$. Then there is a vertex z of G_k that is adjacent to w but not to u.*

Proof. Otherwise, if every neighbor of w is also a neighbor of u, we would have $N(w) \subseteq N(u)$. But $d^*(u) \leq d^*(w)$, which implies that $N(u) = N(w)$ and hence that $S(u) = S(w)$; this clearly contradicts our initial assumption. □

5. Intuitive description of the algorithm. Our goal is to show that for any bipartite graph $G(U, V, E)$ with n vertices, we can find an achromatic partition of size at least $\psi(G)/(Kn^{4/5})$ for some constant $K > 0$. Let ψ^* be an estimate of the true value of $\psi(G)$. Our approximation algorithm uses the parameter ψ^* to obtain an achromatic partition of an induced subgraph of G, with the guarantee that it will produce a large number of color classes in the partition *when the value of ψ^* equals $\psi(G)$*. Hence, it suffices to run the algorithm for *all* possible values of ψ^*, viz. $\psi^* = 1, 2, \ldots, n$, and use the best solution from among all the runs.

To explain the key ideas underlying the algorithm, it is convenient to use terms like *small* and *large* in an informal sense to qualify the relative sizes of various sets. We postpone more precise characterizations of these terms, but merely observe here that by *small* we mean of size roughly $O(n^{4/5})$ or n^δ for some $0 < \delta \leq 4/5$, and by *large* we mean of size roughly $\omega(n^{4/5})$.

We may assume that G has no isolated vertices because such vertices have no effect on the achromatic number of G. Also, $\psi(G)$ may be assumed to be large, for otherwise even the achromatic coloring induced by the initial bipartition $\{U, V\}$ will achieve a small ratio of approximation.

Next, consider the reducing congruence on G. Since G has no isolated vertices, its equivalence classes under the reducing congruence can be cleanly partitioned into those that are subsets of U (the *U-equivalence classes*) and those that are subsets of V (the *V-equivalence classes*). Let q_U (respectively, q_V) be the number of U-equivalence (respectively, V-equivalence) classes under the reducing congruence on G, and let $q = q_U + q_V$ denote the total number of equivalence classes. If q were small, then Theorem 1 (via the algorithm described in [11]) would guarantee a good approximation ratio for $\psi(G)$. Hence, we can assume that q and $\psi(G)$ are both large, i.e., have magnitude $\omega(n^{4/5})$.

Since q is large, the average size of an equivalence class under the reducing congruence is roughly $O(n^{1/5})$. We call such classes the *light* equivalence classes. By the Markov inequality, there will be only a few equivalence classes that are not light. The effect of such classes on our algorithm is negligible; for the sake of a simplified description, the maximum size of an equivalence class that is not light is not pertinent.

The heart of the approximation algorithm is a subroutine, Ach-Bip, that takes as input a bipartite graph $G[U_0, V_0] \subseteq G$ and a guessed value ψ^* of the achromatic number to iteratively compute a sequence, $A_1, A_2, \ldots, A_k$, of color classes. These classes form an achromatic partition of $G[\cup_{1\leq i\leq k} A_i]$. Broadly speaking, in iteration i, with $i \geq 1$, Ach-Bip works as follows:

- It starts with a subgraph $G_{i-1} = G[U_{i-1}, V_{i-1}]$.
- If G_{i-1} has no light U_{i-1}-equivalence classes, then the subroutine call exits. Otherwise, a set, A_i, of independent vertices in G_{i-1} is computed with the following properties: A_i is small in size and *covers* a relatively large set $U_i \subseteq U_{i-1} \setminus A_i$. The latter property ensures that color classes $A_{i+1}, A_{i+2}, \ldots, A_k$ computed in future iterations are adjacent to the class A_i.
- If the removal of A_i and some related vertices from G_{i-1} does not reduce the guessed achromatic number significantly, then the next iteration is initiated on a subgraph $G_i = G[U_i, V_i] \subset G_{i-1}$.

The subroutine Ach-Bip is first executed on the graph $G(U, V)$. Recall that q, the number of equivalence classes of G under the reducing congruence, is large. Since $q = q_U + q_V$, either q_U or q_V or both must be large. If q_U is large, then this first subroutine call may produce a large enough collection of color classes. However, if the call exits because there are no light U_{i-1}-equivalence classes at the beginning of some iteration i (with i being a relatively small number), then we still have the unexplored possibility that q_V, the number of V-equivalence classes, is large.

To exploit this case, the algorithm calls Ach-Bip again. In this second call, the input to the subroutine is the subgraph $G_{i-1} = G[V_{i-1}, U_{i-1}]$ that remains at the conclusion of the first call. Note in particular that the roles of the bipartitions are *interchanged*, viz. that V_{i-1} is treated as the first bipartition and U_{i-1} as the second one. There can be two possible outcomes when the second call to Ach-Bip halts. It may halt after finding a large enough achromatic partition of an induced subgraph of G_{i-1}. On the other hand, it is possible that the second call also halts within a small number of iterations—small enough that the ratio, of the actual achromatic number divided by the size of the larger of the achromatic partitions classes found in the calls, is not a good enough approximation guarantee.

In this latter event, our algorithm still manages to ensure—by design—that the graph, remaining *after* the two calls to Ach-Bip, still has a large achromatic number. In particular, the achromatic number is still at most $\psi^*/2$ less than $\psi(G)$, the achromatic number of the original graph G. Provided that our guess, ψ^*, is close to the optimal value, Theorem 1 allows us to compute a large achromatic partition of the remaining graph. This coloring can be extended (via Lemma 1) to obtain a guaranteed ratio of approximating the achromatic number of G. This completes the informal overview of the algorithm.

6. Formal description of the algorithm.

The approximation algorithm, Approx-Bip, is described below. As mentioned earlier, we execute the algorithm once for each possible value of the guessed achromatic number, ψ^*. The best solution from all the runs is used. The overall runtime is still polynomial in the input size and may be improved slightly (by a logarithmic factor) by deploying binary search over the possible range of values of ψ^*.

We introduce a few notational abbreviations that simplify the formal description and analysis of procedure Ach-Bip.

- We call a set *heavy* if it contains at least $n^{1/5}$ vertices. Otherwise, it is called *light*. In any bipartite graph $G(U, V, E)$, a vertex $v \in V$ is said to be U*-heavy* if the reduced degree of v (under the reducing congruence of G) is at least $n^{1/5}$.
- An assignment of values to several variables in a sequential manner is abbre-

```
   Input: A bipartite graph $G_0(U_0, V_0)$, and a guessed achromatic number $\psi^*$
   Output: An achromatic partition $\{A_1, A_2, \ldots\}$ of the induced graph $G_0[\cup_i A_i]$
1  if $\psi^* < 8n^{4/5}$ then
2      return $\mathcal{A} = \{U_0, V_0\}$
3  end
4  $\mathcal{A} = \emptyset$ ;
5  for $i = 1, 2, \ldots$ do
6      if there are no light $U_{i-1}$-equivalence classes in $G_{i-1}$ then
       /* Condition 1 */
7          return $\mathcal{A}$
8      end
9      $u \leftarrow$ a vertex in $U_{i-1}$ with minimum reduced degree in $G_{i-1}$ ;
10     $U',\ V',\ G' \leftarrow\ U_{i-1} \setminus S_{G_{i-1}}(u),\ V_{i-1} \setminus N_{G_{i-1}}(u),\ G[U', V']$ ;
11     $C_i \leftarrow \emptyset$
12     while $(U' \neq \emptyset)$ and $\exists$ a $U'$-heavy vertex in $V'$ do
13         $v \leftarrow$ a $U'$-heavy vertex in $V'$ with maximum reduced degree in $G'$ ;
14         Add $v$ to $C_i$ ;
15         $U',\ V',\ G' \leftarrow\ U' \setminus N_{G'}(v),\ V' \setminus \{v\},\ G[U', V']$
16     end
17     $q' \leftarrow$ the number of $U'$-equivalence classes in $G'$ ;
18     if $q' > n^{3/5}$ then                                  /* Condition 2 */
19         return the partition obtained by applying the star removal algorithm
           to $G'$
20     end
21     $D_i \leftarrow \{w \in U' \mid S_{G'}(w)$ is a light equivalence class$\}$ ;
22     for every heavy $U'$-equivalence class $S_{G'}(w)$ do
23         add an arbitrary neighbor of $S_{G'}(w)$ to $C_i$
24     end
25     $A_i \leftarrow S_{G_{i-1}}(u) \cup C_i$ ;
26     $L_i \leftarrow$ the set of isolated vertices in the graph
       $G[U_{i-1} \setminus (A_i \cup D_i), V_{i-1} \setminus (A_i \cup D_i)]$ ;
27     if it is not $\psi^*$-safe for $G_{i-1}$ to delete $(A_i \cup D_i \cup L_i)$ then
       /* Condition 3 */
28         return $\mathcal{A}$
29     end
30     add $A_i$ to $\mathcal{A}$ ;
31     $U_i,\ V_i,\ G_i \leftarrow\ U_{i-1} \setminus (A_i \cup D_i \cup L_i),\ V_{i-1} \setminus (A_i \cup D_i \cup L_i),\ G[U_i, V_i]$
32 end
```

PROCEDURE Ach-Bip.

viated on a single line, e.g., on line 10 of procedure Ach-Bip, the statement

$$U',\ V',\ G' \leftarrow\ U_{i-1} \setminus S_{G_{i-1}}(u),\ V_{i-1} \setminus N_{G_{i-1}}(u),\ G[U', V']$$

simply means that U' is assigned the value $U_{i-1} \setminus S_{G_{i-1}}(u)$, then V' is assigned the value $V_{i-1} \setminus N_{G_{i-1}}(u)$, and lastly G' is set to be the induced graph $G[U', V']$.

The following definition is of critical importance to the analysis of subroutine Ach-Bip.

```
Input: $G(U, V, E)$, a bipartite graph; $\psi^*$, a positive integer
Output: An achromatic partition of $G$
$\mathcal{A}_1 \leftarrow$ the achromatic partition returned by the call Ach-Bip($G(U, V)$, $\psi^*$).
Let $G^{[1]} = G[U^{[1]}, V^{[1]}]$ be the induced subgraph that remains when the
procedure call halts ;
$\mathcal{A}_2 \leftarrow$ the achromatic partition returned by the call Ach-Bip($G[V^{[1]}, U^{[1]}]$, $\psi^*$).
Note that the roles of the bipartitions are interchanged. Let
$G^{[2]} = G[U^{[2]}, V^{[2]}]$ be the induced subgraph that remains when this second
application of the procedure halts ;
If either of the achromatic partitions $\mathcal{A}_1$ or $\mathcal{A}_2$ is of size at least $\psi^*/(16n^{4/5})$,
then the corresponding achromatic coloring is extended to an achromatic
coloring of $G$ which is returned as final output ;
Otherwise, apply the algorithm of Theorem 1 on the subgraph $G^{[2]}$. The
achromatic coloring thereby obtained can be extended to an achromatic
coloring of $G$ which is returned as final output.
```

Algorithm 2: Approx-Bip.

Definition 1. *Starting with a subgraph G_0 of the graph G, let $G_0 \supset G_1 \supset \cdots \supset G_j$ be a sequence of induced subgraphs of G obtained by successively removing vertices (and their adjacent edges). Let ψ^* be a positive integer. Then the deletion of some ordered set of vertices S_j from G_j is said to be* ψ^*-safe *for G_j if the total number of nonisolated vertices (including those in S_j) removed from the initial subgraph G_0 is at most $\psi^*/4$.*

7. Analyzing the approximation ratio.

Our goal is to show that the approximation ratio obtained by Approx-Bip is $O(n^{4/5})$. The analysis is conducted under the assumption that $\psi(G) \geq 8n^{4/5}$. Otherwise, returning an arbitrary achromatic partition, e.g., the original bipartition of size 2, as done in line 2 of procedure Ach-Bip, trivially gives an $O(n^{4/5})$ ratio.

Consider a run of the algorithm when presented with a bipartite graph $G(U, V, E)$ with n vertices, and with the parameter ψ^*, which is an estimate of $\psi(G)$. Since the algorithm makes two separate calls on the procedure Ach-Bip, we first analyze the procedure itself in isolation and derive some useful properties.

We start by observing that the execution of the main loop (lines 5–32) in procedure Ach-Bip could be halted in one of *three mutually exclusive ways* during some iteration $(k + 1) \geq 1$.

Condition 1: At the beginning of the iteration, there are no light U_k-equivalence classes in G_k.

Condition 2: The star removal algorithm can be applied during the iteration.

Condition 3: Just prior to the end of the iteration, it is found that the current deletion of $(A_{k+1} \cup D_{k+1} \cup L_{k+1})$ in left-to-right order is not ψ^*-safe for G_k.

Note that the induced subgraphs $\{G_i\}_{i \geq 1}$ form a monotone decreasing chain with respect to graph size. If the star removal algorithm (Condition 2) cannot be applied during any iteration, then eventually one of the other two conditions will become true since the graphs keep getting smaller with each iteration. This guarantees that procedure Ach-Bip *will* eventually terminate.

The schematic shown in Figure 1 depicts the various sets computed during iteration $i \geq 1$ of procedure Ach-Bip. We say that iteration $i \geq 1$ is *successful* if none

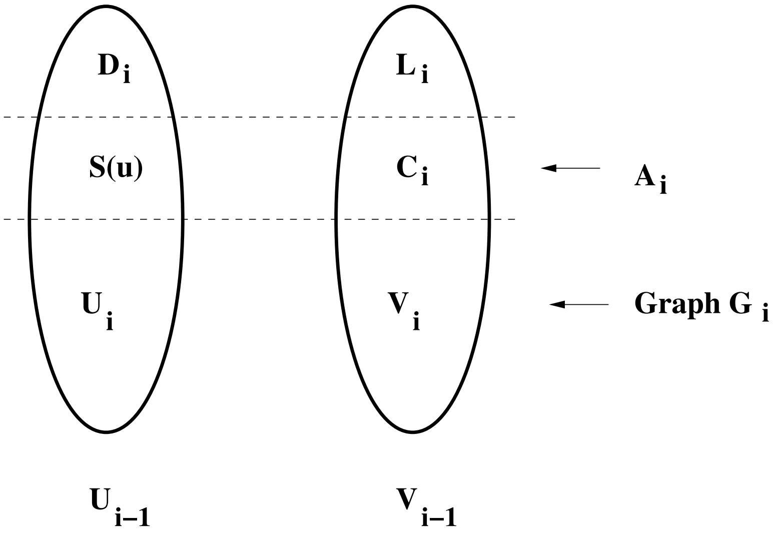


FIG. 1. *The graph G_{i-1}.*

of the conditions is triggered during the iteration, i.e., the procedure continues to the next iteration with the surviving subgraph G_i. Suppose that the first k iterations are successful and let $(k+1) \geq 1$ be the first unsuccessful iteration of the procedure.

LEMMA 6. *If procedure* Ach-Bip *halts during iteration $(k+1)$ due to Condition* 2*, then the achromatic partition that is returned has size at least $n^{1/5}$.*

Proof. During iteration $(k+1)$, consider the graph $G' = G[U', V']$ to which the star removal procedure is applied (lines 18–20). Since U' has at least $n^{3/5}$ equivalence classes, $|U'| > n^{3/5}$ must hold.

Consider any vertex $w \in U'$. Let u be the vertex chosen during the iteration on line 9 of the procedure. By construction, we observe that

$$U' = (U_k \setminus S_{G_k}(u)) \setminus \left(\bigcup_{v \in C_{k+1}} N_{G_k}(v) \right),$$
$$V' = (V_k \setminus N_{G_k}(u)) \setminus C_{k+1}.$$

Clearly, $w \notin S_{G_k}(u)$, and furthermore the choice of u on line 9 ensures that $d^*_{G_k}(u) \leq d^*_{G_k}(w)$. By Lemma 5, we conclude that there is a vertex $z \in (V_k \setminus N_{G_k}(u))$ that is adjacent to w. In fact, w must be adjacent to some vertex in V' since w is not adjacent to any vertex in C_{k+1}. It follows that U' does not have any isolated vertices.

Note that the inner loop condition (line 12) guarantees that every vertex in V' is adjacent to at most $n^{1/5}$ U'-equivalence classes in G'. From the discussion preceding Lemma 4, it is easy to see that the star removal algorithm will produce a collection of at least

$$\sqrt{\frac{n^{3/5}}{n^{1/5}}} = n^{1/5}$$

stars, and hence an achromatic partition of size at least $n^{1/5}$ can be returned, as claimed. □

Turning now to Condition 3, we need to show that if the procedure halts during iteration $(k+1)$ because a ψ^*-unsafe deletion for G_k is flagged, then the number, k, of classes in the partition $\mathcal{A}$ computed thus far must already be large enough. To this end, we establish a sequence of claims.

Claim 1. For $1 \leq i \leq k$, the set A_i is an independent set and is adjacent to A_j for every $j \in [i+1, k]$. In other words, $\mathcal{A}$ is indeed an achromatic partition of the subgraph $G[\cup_{1\leq i\leq k} A_i]$.

Proof. We first verify that at the end of a successful iteration i, the set of vertices A_i is an independent set. By construction, $A_i = S_{G_{i-1}}(u) \cup C_i$, where u is the vertex chosen on line 9. The vertices in $S_{G_{i-1}}(u)$ are mutually nonadjacent by definition. Moreover, $C_i \subseteq V_{i-1} \setminus N_{G_{i-1}}(u)$ and hence C_i is an independent set that is not adjacent to $S_{G_{i-1}}(u)$. Thus, A_i is independent as well.

Now, by construction, the vertices retained in the set U_i at the end of the iteration are exactly those that are *covered* by some vertex in $C_i \subset A_i$. The set A_j, for $i < j \leq k$, contains at least one vertex in $U_{j-1} \subseteq U_i$. Hence there is always an edge between A_i and A_j for $i < j \leq k$. □

Claim 2. For $1 \leq i \leq k$, the size of the set $(A_i \cup D_i)$, just prior to executing the safety check on line 27, is bounded by $4n^{4/5}$.

Proof. By construction, $A_i = S_{G_{i-1}}(u) \cup C_i$ prior to executing line 27. We know that $S_{G_{i-1}}(u)$ is a light equivalence class, and hence $|S_{G_{i-1}}(u)| < n^{1/5}$. A vertex $v \in V_{i-1}$ is added to C_i either during the inner loop (line 14) or later, if it happens to be adjacent to a heavy U'-equivalence class (line 23).

In the former case, just prior to the vertex v being added to C_i, it must have been adjacent to at least $n^{1/5}$ pairwise nonequivalent vertices in U'. These vertices (along with their copies) are removed from U' after v is added to C_i and before the next iteration of the inner loop commences. In other words, each vertex of U' eliminated in the inner loop corresponds to exactly one vertex in C_i that causes its elimination. Since the initial size of U' is bounded by n, it follows that no more than $n/n^{1/5} = n^{4/5}$ vertices could have been added to C_i during the execution of the inner loop.

The number of vertices, added to C_i because they are witness to being adjacent to some heavy U'-equivalence class (see lines 22–24), is at most the number of heavy U'-equivalence classes. This latter quantity is bounded above by the total number of U'-equivalence classes. Since U' has fewer than $n^{3/5}$ classes (otherwise, the star removal algorithm would have been used), at most $n^{3/5}$ vertices are added to C_i in the loop on lines 22–24. Thus, prior to executing the safety check, there are at most

$$n^{1/5} + n^{4/5} + n^{3/5} \leq 3 \cdot n^{4/5}$$

vertices in A_i.

U' has at most $n^{3/5}$ light equivalence classes when control reaches line 21. Subsequently, the set D_i is formed by collecting together all the vertices in these light equivalence classes. The size of D_i, just prior to executing the safety check, is therefore at most $n^{1/5} \cdot n^{3/5} = n^{4/5}$. Summing up, we see that $(A_i \cup D_i)$ contains no more than $4n^{4/5}$ vertices when it is tested for ψ^*-safety on line 27. □

Claim 3. If the first k iterations are successful, then the difference, $\psi(G_0) - \psi(G_k)$, is at most $4k \cdot n^{4/5}$.

Proof. Consider the graph G_i at the end of the ith successful iteration. For clarity, we can view the construction of G_i from G_{i-1} as taking place in two stages. First, the set of vertices $(A_i \cup D_i)$ is removed from G_{i-1}, giving us an intermediate graph G_i^-. Then L_i, the set of all isolated vertices in G_i^- (see line 26 in procedure Ach-Bip), are deleted from G_i^-, yielding G_i.

By Lemma 2 and Claim 2, we obtain the inequality $\psi(G_{i-1}) - \psi(G_i^-) \leq 4n^{4/5}$. Since L_i is an isolated set of vertices in G_i^-, their removal from G_i^- has no effect on the achromatic number; the number of vertices in L_i does not matter. Therefore, it holds that $\psi(G_i^-) = \psi(G_i)$ and hence

$$\psi(G_{i-1}) - \psi(G_i) \leq 4n^{4/5} \quad \text{for all } 1 \leq i \leq k.$$

The telescoping sum of the above k inequalities, one per successful iteration, yields

$$\psi(G_0) - \psi(G_k) \leq 4kn^{4/5},$$

as claimed. □

Lemma 7. *If procedure* Ach-Bip *halts during iteration* $(k+1)$ *under Condition* 3, *then the achromatic partition returned has size at least* $\lfloor \psi^*/16n^{4/5} \rfloor$.

Proof. For each successful iteration $i \in [1, k]$, it is ψ^*-safe for G_{i-1} to delete the corresponding set of vertices $(A_i \cup D_i \cup L_i)$. However, it is ψ^*-unsafe for G_k to delete $(A_{k+1} \cup D_{k+1} \cup L_{k+1})$, and by Definition 1 and Claim 2 this can happen only if

$$\begin{aligned} 4(k+1)n^{4/5} &\geq \sum_{i=1}^{k+1} |A_i \cup D_i| \\ &> \psi^*/4. \end{aligned}$$

We conclude that $\mathcal{A} = \{A_1, A_2, \ldots, A_k\}$ forms an achromatic partition of the subgraph $G[\cup_{1 \leq i \leq k} A_i]$ by Claim 1. It has size $k \geq \lfloor \psi^*/(16n^{4/5}) \rfloor$, as claimed. □

We now address Condition 1 in procedure Ach-Bip. If the procedure halts on this condition at the beginning of iteration $(k+1)$, then we have two possibilities. If $k \geq \lfloor \psi^*/(16n^{4/5}) \rfloor$, then the same conclusion as that of Lemma 7 holds. Otherwise, $k < \lfloor \psi^*/(16n^{4/5}) \rfloor$, and we may not necessarily have a good guarantee of an approximation ratio.

However, note that G_k, the graph at the beginning of iteration $(k+1)$, has no light U_k-equivalence classes, which is what triggers the condition. Since U_k has only heavy classes, it has no more than $n^{4/5}$ equivalence classes in total (because each heavy class has at least $n^{1/5}$ vertices and $|U_k| \leq n$).

Claim 4. Assume that both applications of procedure Ach-Bip on lines 1 and 2 of algorithm Approx-Bip halt on Condition 1 of procedure Ach-Bip. Let q_1 (respectively, q_2) be the number of $U^{[1]}$-equivalence classes in $G^{[1]}$ (respectively, the number of $U^{[2]}$-equivalence classes in $G^{[2]}$). Then, the graph $G^{[2]}$ has achromatic number at least $\psi(G) - \psi^*/2$ and has at most a total of $(q_1 + q_2) \leq 2n^{4/5}$ equivalence classes.

Proof. Observe that the removal of vertices (along with all their incident edges) from a graph cannot increase the number of equivalence classes: two vertices that were equivalent before the removal remain equivalent afterward. Hence, the number of $V^{[2]}$ equivalence classes is at most q_1 (note that the partitions are interchanged before the second application of procedure Ach-Bip on line 2). Thus $G^{[2]}$ has at most a total of $(q_1 + q_2)$ equivalence classes. The discussion preceding the statement of the claim shows that $(q_1 + q_2)$ is bounded above by $2n^{4/5}$.

Since neither application of procedure Ach-Bip halts on Condition 3, the vertices deleted during both applications are ψ^*-safe for deletion. Hence, by Definition 1, the net decrease in the achromatic number is at most $2\psi^*/4 = \psi^*/2$. □

Theorem 2. *For at least one value of* ψ^*, *viz. when* $\psi^* = \psi(G)$, *algorithm* Approx-Bip *achieves an approximation ratio of* $O(n^{4/5})$.

Proof. Lemma 6 shows that if either of the two applications of procedure Ach-Bip halts on Condition 2, then we are guaranteed an approximation ratio of $O(n^{4/5})$ regardless of the relationship between $\psi(G)$ and ψ^*.

Now, consider the situation when ψ^* happens to be equal to $\psi(G)$, but neither application of procedure Ach-Bip halts on Condition 2. There are two possibilities. If either one of the two procedure calls halts on Condition 3, then by Lemma 7, the corresponding returned partition (either $\mathcal{A}_1$ or $\mathcal{A}_2$) can be extended to an achromatic partition of G whose size provides the desired $O(n^{4/5})$-ratio approximation of $\psi(G)$.

Otherwise, it must be the case that both applications of procedure Ach-Bip halt on Condition 1. From Claim 4, we see that the residual graph $G^{[2]}$ has at most $2n^{4/5}$ equivalence classes in its reducing congruence and has an achromatic number that is at least $\psi(G)/2$ (since $\psi^* = \psi(G)$). Using the algorithm underlying Theorem 1 on graph $G^{[2]}$ provides an $O(\max\{n^{4/5}, \sqrt{\psi(G)}\}) = O(n^{4/5})$ approximation ratio for the achromatic number of $G^{[2]}$. The achromatic coloring of $G^{[2]}$ can be extended to G with the same approximation ratio guarantee. □

Acknowledgments. The first author would like to thank Robert Krauthgamer and Magnus M. Halldórsson for useful discussions. The authors are also grateful to the anonymous referees for their careful scrutiny of the manuscript and their various suggestions for improving the presentation.

REFERENCES

[1] H. L. Bodlaender, *Achromatic number is NP-complete for cographs and interval graphs*, Inform. Process. Lett., 31 (1989), pp. 135–138.

[2] N. Cairnie and K. Edwards, *Some results on the achromatic number*, J. Graph Theory, 26 (1997), pp. 129–136.

[3] A. Chaudhary and S. Vishwanathan, *Approximation algorithms for the achromatic number*, J. Algorithms, 41 (2001), pp. 404–416.

[4] K. Edwards, *The harmonious chromatic number and the achromatic number*, in Surveys in Combinatorics, 1997 (Invited Papers for 16th British Combinatorial Conference), R. A. Bailey, ed., London Math. Soc. Lecture Note Ser. 241, Cambridge University Press, Cambridge, UK, 1997, pp. 13–47.

[5] M. Farber, G. Hahn, P. Hell, and D. Miller, *Concerning the achromatic number of graphs*, J. Combin. Theory Ser. B, 40 (1986), pp. 21–39.

[6] U. Feige, M. M. Halldórsson, G. Kortsarz, and A. Srinivasan, *Approximating the domatic number*, SIAM J. Comput., 32 (2002), pp. 172–195.

[7] V. Guruswami and S. Khanna, *On the hardness of* 4*-coloring a* 3*-colorable graph*, SIAM J. Discrete Math., 18 (2004), pp. 30–40.

[8] M. M. Halldórsson, *Approximating the minimum maximal independence number*, Inform. Process. Lett., 46 (1993), pp. 169–172.

[9] P. Hell and D. J. Miller, *On forbidden quotients and the achromatic number*, in Proceedings of the 5th British Combinatorial Conference (1975), C. St. J. A. Nash-Williams and J. Sheehan, eds., Congr. Numer. 15, Utilitas Mathematica, Winnipeg, Canada, 1976, pp. 283–292.

[10] F. Hughes and G. MacGillivray, *The achromatic number of graphs: A survey and some new results*, Bull. Inst. Combin. Appl., 19 (1997), pp. 27–56.

[11] G. Kortsarz and R. Krauthgamer, *On approximating the achromatic number*, SIAM J. Discrete Math., 14 (2001), pp. 408–422.

[12] G. Kortsarz, J. Radhakrishnan, and S. Sivasubramanian, *Complete partitions of graphs*, in Proceedings of the 16th Annual Symposium on Discrete Algorithms (SODA), SIAM, Philadelphia, 2005, pp. 860–869.

[13] G. Kortsarz and S. Shende, *Approximating the achromatic number problem on bipartite graphs*, in Proceedings of the 11th Annual European Symposium on Algorithms (ESA), Lecture Notes in Comput. Sci. 2832, Springer, Berlin, 2003, pp. 385–396.

[14] P. Krysta and K. Loryś, *Efficient approximation algorithms for the achromatic number*, in Proceedings of the 7th Annual European Symposium on Algorithms (ESA), Lecture Notes in Comput. Sci. 1643, Springer, Berlin, 1999, pp. 402–413.

[15] A. Máté, *A lower estimate for the achromatic number of irreducible graphs*, Discrete Math., 33 (1981), pp. 171–183.

[16] M. Yannakakis and F. Gavril, *Edge dominating sets in graphs*, SIAM J. Appl. Math., 38 (1980), pp. 364–372.

SIAM J. Discrete Math.
Vol. 21, No. 2, pp. 374–384

SMALL DIAMETERS OF DUALS*

JAROSLAV NEŠETŘIL† AND IDA ŠVEJDAROVÁ‡

Abstract. We prove that dual graphs and relational structures are connected. Moreover we give efficient bounds for their diameter: a linear bound in the case of oriented graphs (and this is best up to a constant) and a polynomial bound in the case of relational structures.

Key words. homomorphisms duality, graphs, relational structures, duals

AMS subject classifications. 05C15, 68R05, 03C13

DOI. 10.1137/050629707

1. Introduction.

How do local properties of graphs influence their global properties? The local-global phenomena were studied extensively and, in general, this is an area of negative results. See the seminal work of Erdős on high chromatic sparse graphs [1] (extended in this setting in [8]). There are, however, positive aspects of this local-global paradigm. For example, for proper minor closed classes we can characterize the maximum number of colors which one can demand on a subgraph (this leads to the notion of treedepth; see [3]), and for oriented graphs (and more generally for relational structures) one obtains a rich spectrum of global properties which are defined locally. The present paper is devoted to one such area: *homomorphism dualities*.

Our simplest model involves oriented graphs. Recall that for oriented graphs $G = (V, E)$ and $G' = (V', E')$, a homomorphism $f : G \to G'$ is any mapping $f : V \to V'$ satisfying $(x, y) \in E \Rightarrow (f(x), f(y)) \in E'$. (See [2] for an introduction to graphs and their homomorphisms.) Let $G \to G'$ denote the existence of a homomorphism. A *homomorphism duality* (see [4], [2]) is any statement of the following type:

$$\text{for every graph } G \text{ the following holds: } F \not\to G \text{ iff } G \to H \tag{1}$$

(thus G is H-colorable iff G doesn't contain a homomorphical image of F). The pair (F, H) is called a *dual pair* and H is the *dual* of F. This will be denoted by $H = D_F$. (The dual is uniquely determined up to homomorphism equivalence.) The following is a consequence of the main result of [4].

Theorem 1. *The dual D_F exists iff F is homomorphically equivalent to an oriented tree.*

The original construction of duals was an indirect one; however, an explicit and easy construction of the dual D_F was introduced in [5]. Besides having the useful property (1), this construction (of size $2^{n \log n}$ for a tree with n vertices) is an interesting combinatorial structure in itself. However, due to its exponential size, not much

*Received by the editors April 21, 2005; accepted for publication (in revised form) December 11, 2006; published electronically May 7, 2007.

http://www.siam.org/journals/sidma/21-2/62970.html

†Department of Applied Mathematics and Institute for Theoretical Computer Science (ITI), Charles University, Malostranské nám. 25, 11800 Praha 1, Czech Republic (nesetril@kam.ms.mff.cuni.cz). The work of this author was partially supported by project 1M0021620808 of the Czech Ministry of Education.

‡Department of Applied Mathematics, Charles University, Malostranské nám. 25, 11800 Praha 1, Czech Republic, and Department of Mathematics, University of Illinois at Urbana-Champaign, 1409 W. Green St., Urbana, IL 61801 (isvejda2@math.uiuc.edu).

is known about its properties. The construction is reviewed and analyzed in section 2, where we prove the following.

Theorem 2. *After removing isolated vertices, D_F is a connected graph of diameter at most $|V(F)| + 3$ for every tree F.*

Although we can prove indirectly that the core of the dual is always a connected graph (see Theorem 5), for Theorem 2 we need a careful analysis of the explicit construction of D_F. For a fixed tree F, the vertices of D_F are (neighborly) mappings $V(F) \to V(F)$ with arcs defined by means of "switching." This result should be compared with a connectivity and diameter result for trees and their rotations (see [9]). Proof of Theorem 2 is given in section 2.

In the context of applications in constraint satisfaction problems (see [2]), it is important that statements similar to Theorems 1 and 2 hold for all finite relational structures. Let $\Delta = (\delta_i; i \in I)$ be a finite sequence of positive integers. A *relational structure A of type Δ* (*Δ-structure*, for short) is a pair $(X, (R_i; i \in I))$ where $R_i \subseteq X^{\delta_i}$ for all $i \in I$, i.e., R_i is some δ_i-ary relation. The base set X of A is sometimes denoted by $\underline{A}$. We use $R_i(A)$ instead of R_i when necessary and call its elements *edges*. A homomorphism $A \to A'$ of Δ-structures is defined as a mapping of vertices which preserves the relations R_i for all $i \in I$. We have the following.

Theorem 3. *Δ-structure admits a dual iff it is homomorphically equivalent to a Δ-tree.*

See [4] or section 3 for the definition of Δ-tree. We prove the connectivity even in the following case.

Theorem 4. *For any Δ-tree F, its dual D_F is connected after removing isolated vertices.*

By isolated vertices we mean vertices which do not belong to any edge or those that belong only to edges in $R_i(A)$ with $\delta_i = 1$. Theorem 4 will be proved in section 3. Section 4 contains some remarks and open problems.

2. Oriented graphs. Let T be an arbitrary oriented tree. Although we can construct many dual graphs (graphs D_T such that $T \not\to G \Leftrightarrow G \to D_T$ holds for every G), any two duals D and D' are homomorphically equivalent, meaning that we have $D \to D'$ and $D' \to D$. Thus, up to isomorphism, only one of the duals is a core (it has no proper retracts). This is why we often speak about *the* dual. In this section, D_T will denote the dual obtained by construction described in Definition 1, whereas Core(D_T) will be the dual that is a core. As a warm-up we prove that Core(D_T) is a connected graph.

Theorem 5. *Core(D_T) is a connected graph.*

Proof. For contradiction, suppose that there exist two graphs, D_1 and D_2, such that Core$(D_T) = D_1 + D_2$ (there is no edge uv for $u \in V(D_1)$ and $v \in V(D_2)$). Each of the two graphs contains at least one edge, otherwise Core(D_T) would not be a core. Choose arbitrarily $u_1u_2 \in E(D_1)$ and $v_1v_2 \in E(D_2)$ and pick some odd k such that $k > |V(T)|$. Next, build a new graph D' from Core(D_T) by inserting new vertices $w_1, \dots, w_k$ and edges w_jw_{j-1} and w_jw_{j+1} for all even j as well as u_1w_1 and v_1w_k. This D', contrary to Core(D_T), contains a path with alternating directions of edges with endpoints in D_1 and D_2. Clearly $D_1 \not\to D_2$ and $D_2 \not\to D_1$ (as $D_1 + D_2$ is a core). It follows that $D' \not\to D_1 + D_2$. On the other hand $T \not\to D'$ since if $\phi : T \to D'$ is a homomorphism, then $\phi[V(T)] \cap V(D_i) = \emptyset$ for some $i = 1, 2$, because T is connected and the length of the path which connects D_1 and D_2 is greater than $|V(T)|$. Without loss of generality $i = 2$. The subgraph induced by vertices $\phi[V(T)]$ consists of some vertices of D_1 and some vertices that belong to the path between D_1 and D_2. Formally, $\phi[V(T)] \subseteq (V(D_1) \cup \{w_1, \ldots, w_k\})$. However, the subgraph induced by vertices $V(D_1) \cup \{w_1, \ldots, w_k\}$ is homomorphically equivalent to D_1: consider homomorphism ψ such that $\psi \restriction V(D_1)$ is the identity, $\psi(w_j) = u_1$ for j even, and $\psi(w_j) = u_2$ for j odd. Then $\psi\phi$ is a homomorphism mapping T to D_1, which is a contradiction of $T \not\to D_1 + D_2$. Thus we indeed have $T \not\to D'$, which together with $D' \not\to D_1 + D_2$ contradicts the assumption that $D_1 + D_2$ is a dual of T. □

In [5], Nešetřil and Tardif introduced the following explicit construction of D_T.

Definition 1. *D_T is the graph defined as follows: its vertices are all mappings from $V(T)$ to $V(T)$ such that for every $u \in V(T)$ either $(u, f(u)) \in E(T)$ or $(f(u), u) \in E(T)$. Two mappings f and g form an edge (f, g) of D_T if for all $(u, v) \in E(T)$ we have $f(u) \neq v$ or $g(v) \neq u$.*

Theorem 6. *D_T defined above is a dual of T.*

We have shown (Theorem 5) that $\mathrm{Core}(D_T)$ is a connected graph, i.e., for every $u, v \in V(\mathrm{Core}(D_T))$ there exists an oriented path starting with u and ending with v. But the above proof of Theorem 5 does not construct such a path and does not provide any information about its length. In particular, we would like to estimate the diameter of D_T. In Theorem 7, we will prove a stronger statement: not only the core of D_T is connected, but D_T itself is connected after removing isolated vertices. Moreover, its diameter is linear in the number of vertices of T, which is perhaps surprising considering that the number of vertices of D_T can be exponential in $|V(T)|$ (see [7]).

To prove this, we first characterize the isolated vertices of D_T.

Definition 2. *A vertex $u \in V(T)$ is a* source *if its indegree is zero. It is a* problematic source *for $f \in V(D_T)$ if it is a source and, moreover, for all its neighbors w we have $f(w) = u$. Similarly, u is a* sink *if its outdegree is zero and it is a* problematic sink *for $f \in V(D_T)$ if it is a sink and $f(w) = u$ for all vertices w adjacent to u.*

The proof of the next lemma follows directly from Definitions 1 and 2.

Lemma 1 (characterization of isolated vertices of the dual). *Outdegree of f in D_T is zero iff there exists a problematic sink for f in T. Indegree of f in D_T is zero iff there exists a problematic source for f in T.*

Let Z be the set of sources in T. Let V^* be the set of mappings in $V(D_T)$ that go against the directions of edges of T whenever possible, i.e., $V^* = \{f \in V(D_T) |$ if $f(x) = y$ for an edge (x, y) of T, then $x \in Z\}$.

Lemma 2. *If $f \in V^*$ and $h \in V(D_T)$, then (f, h) is an edge of D_T iff for any $x \in Z$, $h(f(x)) \neq x$.*

Proof. The pair (f, h) is an edge iff we have $h(y) \neq x$ whenever (x, y) is an edge of T such that $f(x) = y$. But since $f \in V^*$, this is equivalent to the requirement that $h(f(x)) \neq x$ for every $x \in Z$. □

Corollary 1. *If $f \in V^*$ and $f' \in V(D_T)$, and $f(x) = f'(x)$ for all $x \in Z$, then each outneighbor of f' is an outneighbor of f.*

Lemma 3. *Suppose $f, g \in V^*$ and each of f, g has outdegree greater than zero. If y is a source in T and $f(x) = g(x)$ for all $x \in Z \setminus \{y\}$, then $d_{D_T}(f, g) \leq 2$.*

Proof. Let $z_1 = f(y)$ and $z_2 = g(y)$. Pick h_f and h_g such that (f, h_f) and (g, h_g) are edges of D_T. Define a mapping h in the following way: Let $h(z_2) = h_g(z_2)$ and $h(u) = h_f(u)$ for all $u \neq z_2$. Let x be a source in T. If $f(x) \neq z_2$, then $h(f(x)) = h_f(f(x)) \neq x$ by Lemma 2 since (f, h_f) is an edge. If $f(x) = z_2$, then we also have $g(x) = z_2$ since f and g coincide on all sources except for y. Thus

$h(f(x)) = h(z_2) = h_g(z_2) = h_g(g(x)) \neq x$ since (g, h_g) is an edge. So $h(f(x)) \neq x$ whenever $x \in Z$ and by another application of Lemma 2, (f, h) is an edge. Similarly, for $x \in Z$, either $g(x) = z_2$ and $h(g(x)) = h(z_2) = h_g(g(x)) \neq x$, or $g(x) \neq z_2$, so $x \neq y$ and $h(g(x)) = h_f(g(x)) = h_f(f(x)) \neq x$. In any case $h(g(x)) \neq x$, so (g, h) is an edge. □

LEMMA 4. *Suppose $f, g \in V^*$ and each of f, g has outdegree greater than zero. Then either $f(x) = g(x)$ for all $x \in Z$, or there exists an $x^* \in Z$ such that $f(x^*) \neq g(x^*)$ and there exists $g' \in V^*$ such that g' has outdegree greater than zero, and $g'(x^*) = f(x^*)$ and $g'(x) = g(x)$ for all $x \in Z \setminus \{x^*\}$.*

Proof. Pick a vertex $y_1 \in Z$ such that $f(y_1) \neq g(y_1)$. Define a mapping h_1: $h_1(y_1) = f(y_1)$ and $h_1(u) = g(u)$ for $u \neq y_1$. If the outdegree of h_1 is greater than zero, then let $g' = h_1$ and $x^* = y_1$ and we are done.

Otherwise T has a problematic sink for h_1. Since $h_1(u) = g(u)$ for $u \neq y_1$ and there are no problematic sinks for g, the only problematic sink for h_1 is $f(y_1)$. In particular, all neighbors of $f(y_1)$ are sources and g maps all of them except y_1 to $f(y_1)$. But f has no problematic sink, so $f(y_1)$ has some neighbor $y_2 \in Z$ such that $f(y_2) \neq g(y_2) = f(y_1)$. Define h_2: $h_2(y_2) = f(y_2)$ and $h_2(u) = g(u)$ for $u \neq y_2$. If h_2 has outdegree greater than zero, then let $g' = h_2$ and $x^* = y_2$. If not, find y_3 in a similar manner.

Consider the sequence $y_1, y_2, y_3, \ldots$. The way we defined it together with the fact that T is a tree guarantee that the elements in the sequence never repeat. But then it is finite and if y_k is its last element, then h_k defined as $h_k(y_k) = f(y_k)$ and $h_k(u) = g(u)$ for $u \neq y_k$ has outdegree greater than zero. Let $g' = h_k$ and $x^* = y_k$. □

COROLLARY 2. *If $f, g \in V^*$, each of f, g has outdegree greater than zero, $f(x) = g(x)$ whenever $x \notin Z$, and m is the number of vertices x such that $f(x) \neq g(x)$, then $d_{D_T}(f, g) \leq 2m$.*

Proof. By induction of m. If $m = 0$, then $f = g$ and the statement is trivial. For $m = 1$, the statement is the essence of Lemma 3. If $m > 1$, find g' as in Lemma 4. There is only one vertex x such that $g'(x) \neq g(x)$ and $m - 1$ vertices x such that $f(x) \neq g'(x)$. By induction hypothesis $d_{D_T}(g', g) \leq 2$ and $d_{D_T}(f, g') \leq 2(m - 1)$ and the claim follows. □

THEOREM 7. *Let T be an oriented tree with n vertices and D_T its dual constructed in Definition 1. Let f and g be two vertices of D_T which are not isolated. Then there exists an oriented path between f and g of length at most $n + 3$.*

Proof. If f has nonzero outdegree, then let f^* be the mapping such that $f^*(x) = f(x)$ whenever $x \in Z$, and $f^*(x) = z$ for an (arbitrary) edge (z, x) whenever $x \notin Z$. Such f^* belongs to V^* and it has a common neighbor with f by Corollary 1. If f has no outneighbor, then it has an inneighbor h. Let f^* be a mapping that agrees with h on the set of sources and maps the rest of the vertices against the directions of some incident edges. Then again $f^* \in V^*$ and, by Corollary 1, (f^*, f) is an edge.

Let $g^*(x) = f^*(x)$ whenever $x \notin Z$. If g has an outneighbor, then let $g^*(x) = g(x)$ for $x \in Z$. If not, pick an inneighbor h and let $g^*(x) = h(x)$ for $x \in Z$. This g^* belongs to V^* and has distance at most 2 from g.

Let $Y \subseteq Z$ be the set of all vertices x such that $f^*(x) \neq g^*(x)$ and suppose $Y = \{y_1, \ldots, y_m\}$. By Corollary 2, f^* and g^* are at most $2m$ apart. Consider the subgraph R of T with edges $(y_i, f^*(y_i))$ and $(y_i, g^*(y_i))$ for $i \leq m$. R has $2m$ edges and since it is either a tree or a forest, we get

$$n = |V(T)| \geq |V(R)| \geq |E(R)| + 1 = 2m + 1. \tag{2}$$

Therefore $2m \leq n-1$ and the distance of f^* and g^* is at most $n-1$. The mappings f^* and g^* were chosen so that $d(f^*, f) \leq 2$ and $d(g^*, g) \leq 2$, so $d(f, g) \leq n - 1 + 2 + 2 = n + 3$. □

This bound can be improved to $n + 2$. We do not include the proof because it is merely a tedious case analysis based on the ideas presented in the above proof. Daphne Liu kindly informed us that the above proof can yield an upper bound n (and even $n - 1$ if n is odd). For duals constructed in Definition 1 this is optimal. Let us remark that [6] contains example of trees T for which $\mathrm{Core}(D_T)$ has diameter at least $\lfloor \frac{n-1}{2} \rfloor$.

3. Relational structures. Let A be a relational structure of type $\Delta = (\delta_i; i \in I)$. We write $u \in \mathbf{a}$ for a vertex u and an edge $\mathbf{a} = (a_1, \ldots, a_{\delta_i})$ if there exists an index k such that $u = a_k$.

The *incidence graph* $\mathrm{Inc}(A)$ of the structure A is the bipartite graph with parts $\underline{A}$ and $\mathrm{Block}(A) = \{(i, \mathbf{a}) | i \in I, \mathbf{a} \in R_i(A)\}$. The edges are all pairs $(u, (i, \mathbf{a}))$ such that $u \in \mathbf{a}$. A is called a Δ-*tree* when $\mathrm{Inc}(A)$ is a tree.

Notice that if $\delta_i = \delta_{i'}$ for some $i \neq i'$, then A can have an edge $\mathbf{x}$ that belongs to both $R_i(A)$ and $R_{i'}(A)$. However, if we have also $\delta_i > 1$, then such A is not a Δ-tree.

As we mentioned in section 1, A admits a dual iff it is homomorphically equivalent to a Δ-tree. A simple construction of duals for relational structures similar to the one in Definition 1 appeared in [7].

Definition 3. *Let A be a Δ-tree. Let D_A be the relational structure with the base set $\underline{D_A} = \{f : \underline{A} \to \mathrm{Block}(A) | (u, f(u)) \in E(\mathrm{Inc}(A))$ for all $u \in \underline{A}\}$. The δ_i-tuple $(f_1, \ldots, f_{\delta_i})$ belongs to $R_i(D_A)$ iff for every $(x_1, \ldots, x_{\delta_i}) \in R_i(A)$ there exists $j \in \{1, \ldots, \delta_i\}$ such that $f_j(x_j) \neq (i, (x_1, \ldots, x_{\delta_i}))$.*

Theorem 8 (see [7]). *Let A be a Δ-tree. The structure D_A defined above is a dual of A.*

Analogously to the proof of Theorem 5 we can prove easily that the core of D_A is a connected Δ-structure. Again we will prove a stronger statement: even the structure D_A constructed in Definition 3 is connected after deleting all isolated vertices.

Sinks and sources together with the sets of their neighbors played a crucial role in characterizing the isolated vertices of duals of graphs. The classes of the equivalence defined below play a similar role in characterizing the isolated vertices of duals of relational structures.

Let c_l denote the lth vertex of the edge $\mathbf{c}$.

Definition 4. *For every $i \in I$ and $k \in \{1, \ldots, \delta_i\}$ we will define equivalence $\approx_{(i,k)}$ on $R_i(A)$: $\mathbf{x} \approx_{(i,k)} \mathbf{y}$ if there exist an integer $m \geq 1$ and a sequence of edges $\mathbf{x} = \mathbf{c}^1, \mathbf{c}^2, \ldots, \mathbf{c}^m = \mathbf{y}$ with $\mathbf{c}^1, \ldots, \mathbf{c}^m \in R_i(A)$ which satisfy the following: for every $j = 1, \ldots, m-1$ there is an index $l_j \neq k$ such that the edges $\mathbf{c}^j$ and $\mathbf{c}^{j+1}$ share a vertex v, and v occupies the l_jth position in both edges (that is, $c^j_{l_j} = c^{j+1}_{l_j}$).*

The relation $\approx_{(i,k)}$ is clearly an equivalence. $[\mathbf{x}]_{\approx_{(i,k)}}$ will denote the class of the equivalence $\approx_{(i,k)}$ containing the edge $\mathbf{x}$.

If $\mathbf{x}$ and $\mathbf{y}$ are edges of a Δ-tree A that belong to the same equivalence class and they share a vertex v, then v is in the same position in both edges, and moreover this position is different from k. This is because $(i, \mathbf{x}), v, (i, \mathbf{y})$ is the unique path from $(i, \mathbf{x})$ to $(i, \mathbf{y})$ in $\mathrm{Inc}(A)$, by the definition of Δ-tree. In particular, v is the only vertex that $\mathbf{x}$ and $\mathbf{y}$ share. We necessarily have $x = \mathbf{c}^1$ and $\mathbf{c}^2 = \mathbf{y}$ and v is in the same position in both, and that is different from k.

For $\Delta = (2)$, a Δ-tree is just an orientation of an ordinary tree. A class of equivalence $\approx_{(1,1)}$ is a set of edges with the same second coordinate, say $\{(u_r, v); r \leq s\}$

for some vertex v and some index s. Saying that T has a problematic sink for f is equivalent to saying that there exists a set of edges $\{(u_r, v); r \leq s\}$ such that $f(u_r) = v$ for every $r \leq s$, and moreover, v has no neighbors other than the vertices u_r. Definition 5 generalizes this idea for a general Δ.

DEFINITION 5. *Let f be a vertex of D_A, $i \in I$, and $k \in \{1, \dots, \delta_i\}$, and let C be a class of equivalence $\approx_{(i,k)}$. We call C a* problematic class *for f and k if every edge $\mathbf{a}$ in C satisfies the following two conditions:*

(1) *If $\mathbf{a}$ shares a vertex v with some edge $\mathbf{y}$ not in C, then v is the kth component of $\mathbf{a}$.*

(2) *$f(a_k) = (i, \mathbf{a})$.*

If C is a class of equivalence $\approx_{(i,k)}$ and $\mathbf{a}$ is an edge in C that violates at least one of the two conditions above, then we call $\mathbf{a}$ a good edge *for C.*

LEMMA 5. *Let f be a vertex of D_A, $i \in I$, and $k \in \{1, \dots, \delta_i\}$. If every class of equivalence $\approx_{(i,k)}$ contains a good edge, then f is the kth component in some edge of D_A.*

Proof. We need to find $f_1, \dots, f_{k-1}, f_{k+1}, \dots, f_{\delta_i}$ such that $(f_1, \dots, f_{\delta_i})$ is an edge in D_A for $f_k = f$. For every class C of equivalence $\approx_{(i,k)}$, pick a good edge $\mathbf{a}$ in C and label the edges of C according to their distance from $\mathbf{a}$. That is, $\mathbf{a}$ gets the label 0; any edge $\mathbf{b}$ that has a common vertex with $\mathbf{a}$ gets the label 1; any edge that has a common vertex with such $\mathbf{b}$ (but not with $\mathbf{a}$) is labeled 2, and so on. Every edge in C gets a label, and such labeling is unique, as a consequence of the definition of $\approx_{(i,k)}$ and of A being a Δ-tree. In addition, if the good edge $\mathbf{a}$ violates the first condition, label $\mathbf{y}$ with -1. Now if $\mathbf{c} \in R_i(A)$ and $l \neq k$, let $f_l(c_l)$ be the edge with the smallest label of all edges incident to c_l. Since A is a Δ-tree, such an edge is determined uniquely. Define f_l arbitrarily on the rest of the vertices of A, under the condition that the image of every vertex is an incident edge, and let $f_k = f$.

To show that $(f_1, \dots, f_{\delta_i})$ is an edge of D_A, we need to show that for every $\mathbf{c} \in R_i(A)$ there exists an index l such that $f_l(c_l) \neq (i, \mathbf{c})$. If $\mathbf{c}$ is incident to another edge with a smaller label, then this is obvious. If not, then $\mathbf{c}$ has label 0 and it is a good edge violating the second condition. But then $f(c_k) = f_k(c_k) \neq (i, \mathbf{c})$. □

LEMMA 6. *Let $i \in I$ and $k \in \{1, \dots, \delta_i\}$. If there exists a problematic class for f and k, then f is not the kth component of any edge in $R_i(D_A)$.*

Proof. Suppose for contradiction that $f = f_k$ for some $(f_1, \dots, f_{\delta_i}) \in R_i(D_A)$. Let C be the problematic class for f and k and pick an edge $\mathbf{x} \in C$. Since $(f_1, \dots, f_{\delta_i})$ is an edge of D_A, there exists some vertex x_j in $\mathbf{x}$ such that f_j maps x_j to an edge $\mathbf{y}$ different from $\mathbf{x}$. The edge $\mathbf{y}$ shares the vertex x_j with $\mathbf{x}$, and if $\mathbf{y} \notin C$, then by condition (1) in Definition 5 we have $j = k$. Thus $f_k(x_k) \neq (i, \mathbf{x})$, which contradicts condition (2). So we have $\mathbf{y} \in C$. We will denote $\mathbf{x}^1 = \mathbf{x}$, $\mathbf{x}^2 = \mathbf{y}$ and continue constructing the sequence $\{\mathbf{x}^m | m \in \mathbb{N}\}$. Suppose $\mathbf{x}^m \in C$ is already known for some m. Find $j \in \{1, \dots, \delta_i\}$ such that f_j maps the vertex x_j^m to an edge $\mathbf{z}$ different from $\mathbf{x}^m$. Such j exists for all $\mathbf{x}^m \in R_i(A)$ because $(f_1, \dots, f_{\delta_i})$ is an element of $R_i(D_A)$. Denote $\mathbf{x}^{m+1} = \mathbf{z}$. Since $\mathbf{x}^m \in C$ and the two conditions hold, $\mathbf{x}^{m+1}$ belongs to C for reasons analogous to those above. This way we obtain an infinite sequence of edges in C such that every two consecutive edges are different and have a vertex in common. Can some elements repeat in this sequence? Suppose that $\mathbf{x}^m = \mathbf{x}^{m+l}$ for some $m, l \in \mathbb{N}$. If this is true for more than one pair m,l, choose the pair with the smallest value of l. If $l \geq 3$, we get a cycle in $\mathrm{Inc}(A)$, which contradicts the definition of Δ-tree. Two subsequent edges are different, so $l = 2$. Since $\mathbf{x}^{m+1}$ follows after $\mathbf{x}^m$ in our sequence and they belong to C, the common vertex is the jth in both for some j and we have $f_j(x_j^m) = (i, \mathbf{x}^{m+1})$. Also

$\mathbf{x}^{m+2}$ follows after $\mathbf{x}^{m+1}$, so the common vertex is the j'th in both for some j' and $f_{j'}(x_{j'}^{m+1}) = (i, \mathbf{x}^{m+2}) = (i, \mathbf{x}^{m})$. But since A is a Δ-tree, $\mathbf{x}^{m}$ and $\mathbf{x}^{m+1}$ share only one vertex and thus $j = j'$. Hence, $(i, \mathbf{x}^{m+1}) = f_j(x_j^m) = f_{j'}(x_{j'}^{m+1}) = (i, \mathbf{x}^{m})$, which contradicts our assumption. Thus the elements of our sequence never repeat and we obtained an infinite branch in a finite Δ-system, which is a contradiction. □

As a direct consequence of Lemmas 5 and 6 we get the following.

Theorem 9 (characterization of isolated vertices of duals). *A vertex f is not the kth component of any edge in $R_i(D_A)$ iff A has a problematic class for f and k.*

Throughout this section, we will use the following notation. Let $\widetilde{A}$ be a Δ-tree, $\widetilde{D_A}$ its dual, and let $\widetilde{D_A} = \{\tilde{f}_1, \dots, \tilde{f}_j\}$. Then A will denote the Δ-tree that we obtain by adding a new edge to $\widetilde{A}$. That is, we select some index $i \in I$ and add $\delta_i - 1$ new vertices and one new edge $\mathbf{b} \in R_i(A)$ which contains all the new vertices and one old vertex $v \in \widetilde{A}$. Notice that no vertex of $\mathbf{b}$ other than v can belong to $\widetilde{A}$, for otherwise either A or $\widetilde{A}$ is not a Δ-tree. Let D_A be the dual of A. It is not hard to see that D_A has vertices $\{f_1, \dots, f_j, f'_1, \dots, f'_j\}$, such that the mappings f_t and f'_t for $t = 1, \dots, j$ are both derived from $\tilde{f}_t$ and they differ only on the vertex v. More precisely, f_t and f'_t coincide with $\tilde{f}_t$ on vertices of $\widetilde{A}$ except for v; they are defined in the only possible way on vertices of $\mathbf{b}$ different from v (that is, $f_t(u) = f'_t(u) = (i, \mathbf{b})$ for $u \in \mathbf{b}$, $u \neq v$); and $f_t(v) = \tilde{f}_t(v)$, while $f'_t(v) = (i, \mathbf{b})$. Notice that the elements $f'_1, \dots, f'_j$ are not necessarily distinct: if v belongs to more edges of $\widetilde{A}$, then there are some indices k and l such that the mappings $\tilde{f}_k$ and $\tilde{f}_l$ differ only on the vertex v, and f'_k is equal to f'_l.

This notation allows us to state the following corollary of Theorem 9.

Corollary 3. *Let g be a vertex of D_A. If g is the kth component of some edge $(g_1, \dots, g_{\delta_i}) \in R_i(D_A)$ but $\tilde{g}$ is isolated in $\widetilde{D_A}$, then there exists a single problematic class $[\mathbf{x}]_{\approx_{(i,k)}}$ for $\tilde{g}$ and k in $\widetilde{D_A}$, and this class is no longer problematic (for g and k) after inserting the edge $\mathbf{b}$.*

The next lemma reveals a close relationship between D_A and $\widetilde{D_A}$: if we delete the vertices $f'_1, \dots, f'_j$ in D_A and all edges containing them, we get exactly a copy of $\widetilde{D_A}$. The proof follows immediately from Definition 3.

Lemma 7. *Let $i \in I$. $(\tilde{f}_1, \dots, \tilde{f}_{\delta_i})$ is an edge of $\widetilde{D_A}$ iff $(f_1, \dots, f_{\delta_i})$ is an edge of D_A.*

The structure D_A may contain vertices that do not belong to any edge (isolated vertices) and also vertices that are only in unary edges (i.e., edges with $\delta_i = 1$). To simplify notation, we will extend the definition of isolated vertices so that it includes also the latter kind of vertices: from now on, a vertex u will be *isolated* iff for every $i \in I$ with $\delta_i > 1$ and for every $\mathbf{x} \in R_i(A)$ we have $u \notin \mathbf{x}$. We will prove that after removing such isolated vertices, D_A is a connected Δ-system.

The proof of connectedness of the duals for graphs (Theorem 7) can be generalized for relational structures. However, we chose a different approach, one that shows how the dual changes when we modify the original tree, and thus provides additional insight into its structure. This proof gives an alternative proof of Theorem 7 (without the bound on diameter).

We will need the following lemma.

Lemma 8. *If B is a component of D_A which contains more than one vertex, then some of the vertices $f_1, \dots, f_j$ belong to B (that is, B does not contain only $f'_1, \dots, f'_j$).*

Proof. Suppose that the new edge $\mathbf{b}$ belongs to $R_{i'}(A)$ and has a common vertex $a_{j_1} = b_{j_2}$ with some $\mathbf{a} \in R_i(A)$. Let $\delta_l > 1$ and let $(f'_1, \dots, f'_{\delta_l})$ be an edge in B

that does not contain any of the vertices $f_1, \ldots, f_j$. One can easily see that this can happen only if $l \neq i'$. If $\mathbf{a} \notin R_l(A)$, then we can define $g(b_{j_2}) = (i, \mathbf{a})$, $g(u) = f'_1(u)$ for $u \neq b_{j_2}$. The δ_l-tuple $(g, f'_2, \ldots, f'_{\delta_l})$ is an edge: fix an edge $\mathbf{c} \in R_l(A)$ and find an index j such that $f'_j(c_j) \neq (l, \mathbf{c})$ (it exists since $(f'_1, \ldots, f'_{\delta_l})$ is an edge). Then this j works for the new δ_l-tuple of mappings as well. If $\mathbf{a} \in R_l(A)$, then we will choose $m \neq j_1$ (we can do this because $\delta_l > 1$) and define $g(b_{j_2}) = (i, \mathbf{a})$, $g(u) = f'_m(u)$ for $u \neq b_{j_2}$. Then $(f'_1, \ldots, f'_{m-1}, g, f'_{m+1}, \ldots, f'_{\delta_l})$ is also an edge in $R_l(D_A)$. Moreover, in both cases $g \in \{f_1, \ldots, f_j\}$. □

The so-called zigzag paths, i.e., paths with alternating directions of edges, play an important role in the proof of Theorem 7. The equivalence classes defined below are analogues of zigzag paths for relational structures.

Definition 6. For every i we will define an equivalence $\sim_i$ on $R_i(A)$: $\mathbf{x} \sim_i \mathbf{y}$ if there exists a sequence $\mathbf{x} = \mathbf{c^1}, \mathbf{c}^2, \ldots, \mathbf{c}^m = \mathbf{y}$ of edges from $R_i(A)$ such that for every j there exists some index $l_j \in \{1, \ldots, \delta_i\}$ such that $c^j_{l_j} = c^{j+1}_{l_j}$.

Contrary to the definition of $\approx_{(i,k)}$, now the index l_j can be arbitrary. For every $\mathbf{x}$ and k we have $[\mathbf{x}]_{\approx_{(i,k)}} \subseteq [\mathbf{x}]_{\sim_i}$. In the following proofs, $\mathbf{a}$ is again a fixed edge in $R_i(\widetilde{A})$ which has a common vertex with the newly added edge $\mathbf{b}$.

Lemma 9. *Suppose that $\widetilde{A}$ has an edge $\mathbf{y}$ that does not belong to the class $[\mathbf{a}]_{\sim_i}$ and that $\delta_i > 1$. Let $g \in \{f_1 \ldots, f_j\}$ be a vertex such that $\tilde{g}$ is isolated in $\widetilde{D_A}$ but $g = z^1_k$ for some k and some edge $(z^1_1, \ldots, z^1_{\delta_i}) \in R_i(D_A)$. Then there exists an index r and edges $\mathbf{z}^1, \ldots, \mathbf{z}^r$ in $R_i(D_A)$ such that each consecutive pair shares a vertex, $\mathbf{z}^1 = (z^1_1, \ldots, z^1_{\delta_i})$ and $\mathbf{z}^r = (z^r_1, \ldots, z^r_{\delta_i})$ with $(\widetilde{z^r_1}, \ldots, \widetilde{z^r_{\delta_i}}) \in R_i(\widetilde{D_A})$ and $z^r_1, \ldots, z^r_{\delta_i} \in \{f_1, \ldots, f_j\}$.*

Proof. Since $\widetilde{A}$ is a connected Δ-structure, we may without loss of generality suppose that $\mathbf{y}$ has a common vertex with some edge in $[\mathbf{a}]_{\sim_i}$. Let $\mathbf{y} = \mathbf{c}^1, \mathbf{c}^2, \ldots, \mathbf{c}^m = \mathbf{b}$ be the shortest sequence of edges in A such that each consecutive pair shares a vertex. Since A is a Δ-tree, all these edges except $\mathbf{c}^1$ and possibly $\mathbf{c}^m$ belong to $[\mathbf{a}]_{\sim_i}$. Let $\mathbf{z}^1 = (z^1_1, \ldots, z^1_{\delta_i})$ and define $\mathbf{z}^2, \ldots, \mathbf{z}^m \in R_i(D_A)$ inductively in the following way. Suppose that $\mathbf{z}^{t-1}$ is already known. At least one of $\mathbf{c}^{t-1}$ and $\mathbf{c}^t$ belongs to $[\mathbf{a}]_{\sim_i}$; suppose that the shared vertex w is the sth in this edge. Then let $\mathbf{z}^t$ be a δ_i-tuple which differs from $\mathbf{z}^{t-1}$ only in its sth component, z^t_s, and the image of the vertex w under the mapping z^t_s is the edge $\mathbf{c}^{t-1}$, whereas $z^t_s(u) = z^{t-1}_s(u)$ for all $u \neq w$.

Considering that $\mathbf{z}^{t-1}$ is an edge of D_A and that the mappings in $\mathbf{z}^t$ are mostly the same as those in $\mathbf{z}^{t-1}$, the edge $\mathbf{c}^{t-1}$ is the only one that can prevent the δ_i-tuple $\mathbf{z}^t$ from being an edge of D_A. If $t \geq 3$, then $\mathbf{c}^{t-1}$ shares its s'th vertex with $\mathbf{c}^{t-2}$ and $s \neq s'$ because the path from $\mathbf{y}$ to $\mathbf{b}$ was the shortest possible. But then $z^t_{s'} = z^{t-1}_{s'}$ maps $c^{t-1}_{s'}$ to the edge $\mathbf{c}^{t-2}$ and $\mathbf{c}^{t-1}$ does not violate the condition for $\mathbf{z}^t$ being an edge. If $t = 2$, then $\mathbf{c}^1 = \mathbf{y}$ can violate the condition only if $\mathbf{y} \in R_i(A)$. But in that case, since $\mathbf{y} \notin [\mathbf{a}]_{\sim_i}$, the common vertex of $\mathbf{y}$ and $\mathbf{c}^2$ is the sth in $\mathbf{c}^2$ and the s'th in $\mathbf{y}$ for some $s \neq s'$ and the condition holds.

Thus we have constructed edges $\mathbf{z}^1, \ldots, \mathbf{z}^m$ in $R_i(D_A)$. Every two subsequent edges in this sequence differ only in one component and since $\delta_i > 1$, they share at least one vertex. For $l = 1, \ldots, \delta_i$, define the mappings z^{m+1}_l: $z^{m+1}_l(v) = (i, \mathbf{a})$ for the vertex v shared by $\mathbf{a}$ and $\mathbf{b}$ and $z^{m+1}_l(u) = z^m_l(u)$ for $u \neq v$. For all l, $z^{m+1}_l \in \{f_1, \ldots, f_j\}$ and trivially $(z^{m+1}_1, \ldots, z^{m+1}_{\delta_i}) \in R_i(A)$, so by Lemma 7, $(\widetilde{z^{m+1}_1}, \ldots, \widetilde{z^{m+1}_{\delta_i}})$ is an edge of $\widetilde{D_A}$. Moreover, if v is the rth vertex in $\mathbf{a}$, then $(z^{m+1}_1, \ldots, z^{m+1}_{\delta_i})$ shares the vertex $z^{m+1}_r = z^m_r$ with $\mathbf{z}^m$. □

Lemma 10. *Suppose that every edge of $\widetilde{A}$ belongs to the class $[\mathbf{a}]_{\sim_i}$. Then D_A is connected after removing isolated vertices.*

Proof. If $\mathbf{b} \in R_i(A)$ and $a_s = b_s$ holds for some s (that is, the common vertex of a and b occupies the same position in both edges), then $R_i(D_A) = \emptyset$. This is because A is homomorphically equivalent to the Δ-system B with $\underline{B} = \{b_1, \dots, b_{\delta_i}\}$ and $R_i(B) = R(B) = \{\mathbf{b}\}$ and clearly $R_i(D_B) = \emptyset$. If there exists some $i' \neq i$ such that $\delta_{i'} > 1$, then for any $\delta_{i'}$-tuple $f_1, \dots, f_{\delta'_i} \in \underline{D_A}$ we have $(f_1, \dots, f_{\delta'_i}) \in R_{i'}(D_A)$ (since $R_{i'}(A) = \emptyset$, nothing can prevent the existence of such edge), otherwise all vertices of D_A are isolated.

If $\mathbf{b} \in R_i(A)$, but $b_s = a_{s'}$ for some $s \neq s'$ (the common vertex occupies different positions in the two edges) or $\mathbf{b} \notin R_i(A)$, then describing $R_i(D_A)$ is also relatively easy. First, let us label the edges of A recursively according to their distance from $\mathbf{b}$ and let $c(\mathbf{x})$ denote the label given to the edge $\mathbf{x}$. More precisely, $c(\mathbf{b}) = 0$, $c(\mathbf{a}) = 1$, $c(\mathbf{y}) = 2$ for edges that have a common vertex with $\mathbf{a}$, etc. Since A is a Δ-tree, such labeling exists and is unique. Now define sets $H_r \subseteq \underline{D_A}$ for $r = 1, \dots, \delta_i$: f will belong to H_r iff for every edge $\mathbf{x}$ other than $\mathbf{b}$, $f(x_r)$ has the smallest label of all edges incident with x_r. If $\mathbf{b} \in R_i(A)$, we impose an additional requirement on f in H_s: $f(b_s) \neq (i, \mathbf{b})$. We will prove that $R_i(D_A) = H_1 \times H_2 \times \cdots \times H_{\delta_i}$ (cartesian product of sets). If $(g_1, \dots, g_{\delta_i}) \in H_1 \times H_2 \times \cdots \times H_{\delta_i}$, then for every edge $\mathbf{x} \in R_i(A)$ there is some r such that $g_r(x_r) \neq (i, \mathbf{x})$ (here we need the extra condition on g_s if $\mathbf{b} \in R_i(A)$), so $(g_1, \dots, g_{\delta_i}) \in R_i(A)$. Let's prove the other inclusion. For contradiction suppose that $(g_1, \dots, g_{\delta_i}) \in R_i(D_A)$ but $g_r \notin H_r$ for some r. The extra condition for the case $\mathbf{b} \in R_i(A)$ and $r = s$ clearly holds, if applicable, because otherwise $\mathbf{b}$ would be an edge violating the condition for existence of the edge $(g_1, \dots, g_{\delta_i})$. So there is some vertex u in the rth position in some edge $\mathbf{x} \in R_i(A) \setminus \{\mathbf{b}\}$ such that g_r maps u to some edge $\mathbf{x}^0$, but u belongs to some edge $\mathbf{y}$ which is closer to $\mathbf{b}$ than $\mathbf{x}^0$. Since $R(\widetilde{A}) = [\mathbf{a}]_{\sim_i}$, we can without loss of generality suppose that $\mathbf{x} = \mathbf{x}^0$ and $u = x^0_k$. Since $(g_1, \dots, g_{\delta_i})$ is an edge, there is some $l_0 \neq r$ for which $g_{l_0}(x^0_{l_0}) = (i, \mathbf{x}^1)$ for $\mathbf{x}^1 \neq \mathbf{x}^0$. Since A is a Δ-tree, $\mathbf{y}$ is the only edge incident to $\mathbf{x}^0$ such that $c(\mathbf{y}) < c(\mathbf{x}^0)$, and therefore $c(\mathbf{x}^1) > c(\mathbf{x}^0)$. Analogously there exists $l_1 \neq l_0$ for which $g_{l_1}(x^1_{l_1}) = (i, \mathbf{x}^2)$ for $\mathbf{x}^2 \neq \mathbf{x}^1$. Again $c(\mathbf{x}^2) > c(\mathbf{x}^1)$. The sequence $\mathbf{x}^1, \mathbf{x}^2, \dots$ can finish only if it reaches $\mathbf{b}$ at some point. However, considering that $c(\mathbf{x}^1) < c(\mathbf{x}^2) < \cdots$ and $c(\mathbf{b}) = 0$, this will never happen. The system A is finite, and thus we obtain a contradiction.

If there exists an $i' \neq i$ with $\delta_{i'} > 1$, then again all vertices of D_A belong to a single nontrivial connected component. This is because if $\mathbf{b} \in R_{i'}(A)$, then $(f_1, \dots, f_{\delta_{i'}})$ is an edge whenever $f_s(b_s) \neq (i', \mathbf{b})$, and if $\mathbf{b} \notin R_{i'}(A)$, then $R_{i'}(A) = \emptyset$, so all the $\delta_{i'}$-tuples are edges. If there is no such i', then the nontrivial connected component contains exactly the elements of $R_i(D_A) = H_1 \times H_2 \times \cdots \times H_{\delta_i}$, which has only one connected component. □

THEOREM 10. *If the Δ-system obtained from $\widetilde{D_A}$ by removing isolated vertices is connected, then D_A is also connected after removing isolated vertices.*

Proof. By deleting the vertices $f'_1, \dots, f'_j$ from D_A (and all edges incident with them) we get a copy of $\widetilde{D_A}$ (Lemma 7). By assumption, this copy has at most one nontrivial connected component (i.e., connected component with more than one vertex), say C. For contradiction suppose that there exists some nontrivial component C' in D_A such that $C' \cap C = \emptyset$. Necessarily, some of the vertices $f'_1, \dots, f'_j$ belong to C'. But C' also contains some $g \in \{f_1, \dots, f_j\}$ (Lemma 8), and since $g \notin C$, $\tilde{g}$ was isolated in $\widetilde{D_A}$. Thus, to prove the theorem, it suffices to find a path in $\mathrm{Inc}(D_A)$ beginning with g and ending in C for every $g \in \{f_1, \dots, f_j\}$ such that $\tilde{g}$ is isolated in $\widetilde{D_A}$ but $g = g_k$ for some k, some $i \in I$ with $\delta_i > 1$, and some $(g_1, \dots, g_{\delta_i}) \in R_i(D_A)$. This will contradict the existence of C'.

Let g be such a vertex. By Corollary 3 there was only one problematic class $[\mathbf{x}]_{\approx_{(i,k)}}$ for $\tilde{g}$ in $\tilde{A}$, and after adding the edge $\mathbf{b}$ this class is no longer problematic. This could happen only if $\mathbf{b}$ has a common vertex with an edge $\mathbf{a} \in [\mathbf{x}]_{\approx_{(i,k)}}$; in particular we have $\mathbf{a} \in R_i(A)$ for this edge. In this situation we distinguish two cases. If $\tilde{A}$ has an edge that does not belong to $[\mathbf{a}]_{\sim_i}$, then use Lemma 9 to find a sequence of edges connecting g with an element of C. If all edges of $\widetilde{A}$ belong to $[\mathbf{a}]_{\sim_i}$, Lemma 10 proves the connectivity directly. □

Proof of Theorem 4. Every Δ-tree A has an edge $\mathbf{x}$ that shares only one vertex with the rest of A (analogous to a leaf in a tree). If we remove $\mathbf{x}$, the resulting Δ-structure is again a Δ-tree. Thus any Δ-tree can be built in a finite number of steps from the empty Δ-tree (i.e., a Δ-tree B with $\underline{B} = \emptyset$) by inserting edges in such a way that the Δ-systems obtained in each step are Δ-trees. Therefore we can proceed by induction, with the inductive step being the essence of the previous theorem. □.

Distance $d(u, v)$ of vertices u and v in a relational structure is defined as the smallest k for which there exists a sequence $u = u_0, \dots, u_k = v$ such that u_i and u_{i+1} belong to an edge. (It is the distance of u and v in the Gaifman graph of A.) A closer look at the proof of Theorem 10 gives a polynomial upper bound on the diameter of D_A.

Lemma 11. *If A is a Δ-tree with n vertices, then the diameter of D_A, after removing isolated vertices, is at most $3n^2 + n - 4 + 4|\Delta|n$.*

Proof. First, let us determine how adding a single edge influences the diameter of the dual. If A is constructed by adding an edge to a Δ-tree $\widetilde{A}$, then the situation is either as in Lemma 10, and the diameter of D_A is at most 2, or the situation is as in Lemma 9. In this case, D_A contains two groups of vertices, $\{f_1, \dots, f_j\}$ and $\{f'_1, \dots, f'_j\}$, and by the proof of Lemma 8, any nonisolated vertex in the second group is at most distance two apart from some member of the first group. The first group induces a copy of $\widetilde{D_A}$ and contains a connected component C. Let $g \in \{f_1, \dots, f_j\}$ be a nonisolated vertex of D_A that does not belong to C. In the proof of Lemma 9 we constructed a sequence of edges $\mathbf{z}^1, \dots, \mathbf{z}^m$ such that $\mathbf{z}^1$ contains g and $\mathbf{z}^m$ contains a vertex in C. The distance of every such g from C is therefore at most m, and the distance of a nonisolated $f' \in \{f'_1, \dots, f'_j\}$ from C is at most $m + 2$. If f and h are nonisolated vertices of D_A and $\operatorname{diam}(C)$ is the diameter of C, then

$$d(f, h) \leq d(f, C) + \operatorname{diam}(C) + d(h, C) \leq \operatorname{diam}(C) + 2m + 4. \tag{3}$$

Since the sequence $\mathbf{c}^1, \dots, \mathbf{c}^m$ in the proof of Lemma 9 was the shortest possible, $\mathbf{c}^2, \dots, \mathbf{c}^{m-1}$ are nonunary edges, and $m - 2$ is bounded by the number of nonunary edges of $\widetilde{A}$. A Δ-tree $\widetilde{A}$ with t vertices can have at most $t - 1$ nonunary edges, so using (3),

$$\operatorname{diam}(D_A) \leq \operatorname{diam}(\widetilde{D_A}) + 2(t + 1) + 4. \tag{4}$$

Let A be a Δ-tree with n vertices. Build A by adding edges one by one (as in the proof of Theorem 4), but moreover in such a way that unary edges are inserted only after all others are in place. We can insert the nonunary edges in at most $n - 1$ steps and, using (4) repeatedly, the resulting structure has diameter at most $\sum_{t=0}^{n-2}(2(t + 1) + 4) = n^2 + 3n - 4$.

Two situations can occur when we add a unary edge. If there is a vertex in $\{f_1, \dots, f_j\}$ that was isolated before, but is not isolated anymore after we add the new edge, then by the arguments above, the diameter can increase by $2(n + 1) + 4$.

But this can happen only if adding this edge transformed a problematic class into a good class. Let k be the number of times this situation happened. If every vertex is in a unary edge, there are surely no problematic classes left, so $k \leq n$. If there is no such vertex, the diameter can increase by at most 4, since every nonisolated vertex in $\{f'_1, \dots, f'_j\}$ is a distance of at most 2 from the vertices $\{f_1, \dots, f_j\}$. There are no more than $|\Delta| - 1$ unary relations (provided that there exists a nonunary relation) and each vertex of A is incident to at most that many unary edges, so this situation can happen no more than $k(|\Delta| - 2) + (n - k)(|\Delta| - 1)$ times. Altogether, adding the unary edges can increase the diameter by at most $k(2(n+1)+4) + 4k(|\Delta| - 2) + 4(n-k)(|\Delta| - 1)$, which is bounded by $2n(n+1) + 4n(|\Delta| - 1)$.

Combining these estimates, D_A has diameter at most

$$n^2 + 3n - 4 + 2n(n+1) + 4n(|\Delta| - 1) = 3n^2 + n - 4 + 4|\Delta|n. \qquad \Box \tag{5}$$

4. Concluding remarks. The linearity of diameter suggests the existence of fast algorithms for D_T. For graphs, the proof of the connectivity of D_T (Theorem 7) yields an algorithm which finds a path from f to g in at most $2dn^2$ steps, where d is the maximum degree in T. Is there a linear algorithm?

Knowing that D_T is connected, one might also try to determine its connectivity. Are there always vertices of small degree in D_T? How does the minimal degree in D_T depend on the height of T?

REFERENCES

[1] P. Erdős, *Graph theory and probability*, Canad. J. Math., 11 (1959), pp. 34–38.

[2] P. Hell and J. Nešetřil, *Graphs and Homomorphisms*, Oxford University Press, Oxford, 2004.

[3] J. Nešetřil and P. Ossona de Mendez, *Colouring and homomorphisms of minor-closed classes*, Discrete Comput. Geom., 25 (2003), pp. 651–664.

[4] J. Nešetřil and C. Tardif, *Duality theorems for finite structures (characterizing gaps and good characterizations)*, J. Combin. Theory Ser. B, 80 (2000), pp. 80–97.

[5] J. Nešetřil and C. Tardif, *A dualistic approach to bounding the chromatic number of a graph*, ITI Series, 2001-036 (2001).

[6] J. Nešetřil and C. Tardif, *On maximal finite antichains in the homomorphism order of directed graphs*, Discuss. Math. Graph Theory, 23 (2003), pp. 325–332.

[7] J. Nešetřil and C. Tardif, *Short answers to exponentially long questions: Extremal aspects of homomorphism duality*, SIAM J. Discrete Math., 19 (2005), pp. 914–920.

[8] J. Nešetřil and X. Zhu, *On sparse graphs with given colorings and homomorphisms*, J. Combin. Theory Ser. B, 80 (2004), pp. 161–172.

[9] D. Slater, R. Tarjan, and W. Thurston, *Rotation distance, triangulations, and hyperbolic geometry*, J. Amer. Math. Soc., 1 (1988), pp. 647–681.

[10] I. Švejdarová, *Coloring of Graphs and Dual Structures*, B.Sc. thesis, Charles University, Prague, 2003 (in Czech).

SIAM J. Discrete Math.
Vol. 21, No. 2, pp. 385–395

FORBIDDEN K-SETS IN THE PLANE*

MICHA A. PERLES[†] AND ROM PINCHASI[‡]

Abstract. Let A be a set of nonnegative integers. We say that A is *skippable* if there are arbitrary large finite sets of points in the plane, not contained in a line, that determine no k-edge for any $k \in A$. In this paper we show, by construction, that there are arbitrary large skippable sets. We also characterize precisely the skippable sets with at most two elements.

Key words. k-sets, skippable sets, allowable sequences

AMS subject classifications. 05, 51M04

DOI. 10.1137/050640229

1. Introduction.

Let G be a finite set of points in the plane. We say that a line l is *determined* by G if l passes through at least two points of G. A line l, determined by G, is called a k-edge, if, in one of the two open half-planes bounded by l, there are precisely k points of G. We say that G *skips* k if G has no k-edge.

Definition 1.1. *Let A be a set of nonnegative integers. We say that A is* skippable *if there are arbitrary large finite sets in the plane, which are not collinear, that skip k for every $k \in A$.*

This notion of a skippable set was defined in [PP1]. It is shown there that for any $k \geq 0$ the set $\{k, k+2\}$ is not skippable. This means that if a noncollinear set of points G is large enough, then it has either a k-edge or a $(k+2)$-edge. In this paper we show that skippable sets do exist. In particular, in Theorem 3.1 we show that $\{k\}$ is skippable if and only if $k \geq 2$. Moreover, in Theorem 3.2 we show that one can find arbitrary large such sets of positive integers.

We also complete the picture from [PP1] and characterize precisely the skippable sets that consist of two elements (Theorem 4.5).

In what follows, by referring to a *set of points* we mean a finite set of points in the two-dimensional Euclidean plane.

If G is a set of points in a general position, namely no three points of G lie on one line, then clearly G has a k-edge for every $0 \leq k \leq |G| - 2$. Therefore in our study of skippable sets we will be concerned mainly with sets that are not in a general position. Figure 1 shows an example of a set that skips $k = 2$.

In fact, by adding an arbitrary large number of points very close to the center of the shape in Figure 1, we will remain with a set that skips $k = 2$, thus showing that $\{2\}$ is a skippable set. Similarly, the example in Figure 5 illustrates that $\{4, 5\}$ is a skippable set. The study of skippable sets was initiated by Kupitz and Perles (see [K93, K94]). In [KP], Kupitz and Perles construct arbitrary large sets G, not contained in a line, that skip every k for $k = |G|, |G| - 1, \ldots, |G| - \log_2 |G|$. The question of whether one can fix some skippable values and find arbitrary large noncollinear sets that skip each of them was suggested by Perles. In this paper we give an affirmative answer to this question.

*Received by the editors September 13, 2005; accepted for publication (in revised form) January 16, 2007; published electronically May 7, 2007.
http://www.siam.org/journals/sidma/21-2/64022.html

[†] Hebrew University of Jerusalem, Giveat Ram, Jerusalem, Israel (perles@math.huji.ac.il).
[‡] Israel Institute of Technology, Technion, Haifa, Israel (room@math.technion.ac.il). This research was conducted when this author was an M.Sc. student in the Hebrew University of Jerusalem.

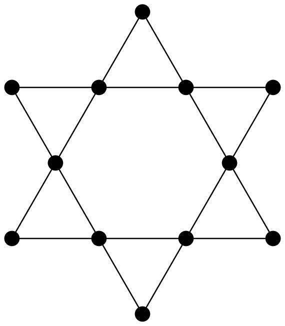


FIG. 1. *A set of points that skips $k = 2$.*

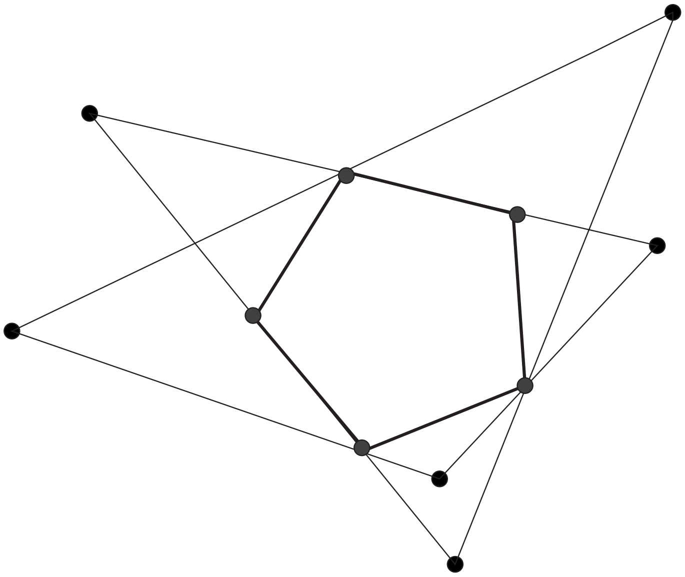


FIG. 2. *A tangency path to a pentagon.*

2. Tangency paths. In this section we define the notion of a *tangency path* and learn its properties in connection with skippable sets. The notion of a tangency path will be crucial for most of the constructions in this paper.

DEFINITION 2.1. *Let P be a convex polygon in the plane. A tangency path for P is a closed polygonal path with vertices $x_0, x_1, \ldots, x_{m-1}, x_m = x_0$ (m can be arbitrary) with the property that if l is the directed line $\overrightarrow{x_i x_{i+1}}$, then l is a tangent of P, $l \cap P$ is contained in the interior of the edge $[x_i, x_{i+1}]$, and the polygon P is in the closed half-plane to the left of l. In addition we require that the vertices of the path (namely, $x_0, x_1, \ldots, x_{m-1}$) are pairwise different. (See Figure 2 for an example.)*

NOTATION 2.2. *Let G be a set of points in the plane, and let l be any directed line. We denote by $A_G(l)$ the number of points of G that are inside the open half-plane to the right of l. We denote by $B_G(l)$ the number of points of G on l. When there is no ambiguity and the set G is known and fixed, we simply write $A(l)$ for $A_G(l)$ and $B(l)$ for $B_G(l)$.*

The following very simple lemma is the key observation regarding tangency paths. We recall that if γ is a closed oriented path in the plane, then the *index* of γ with

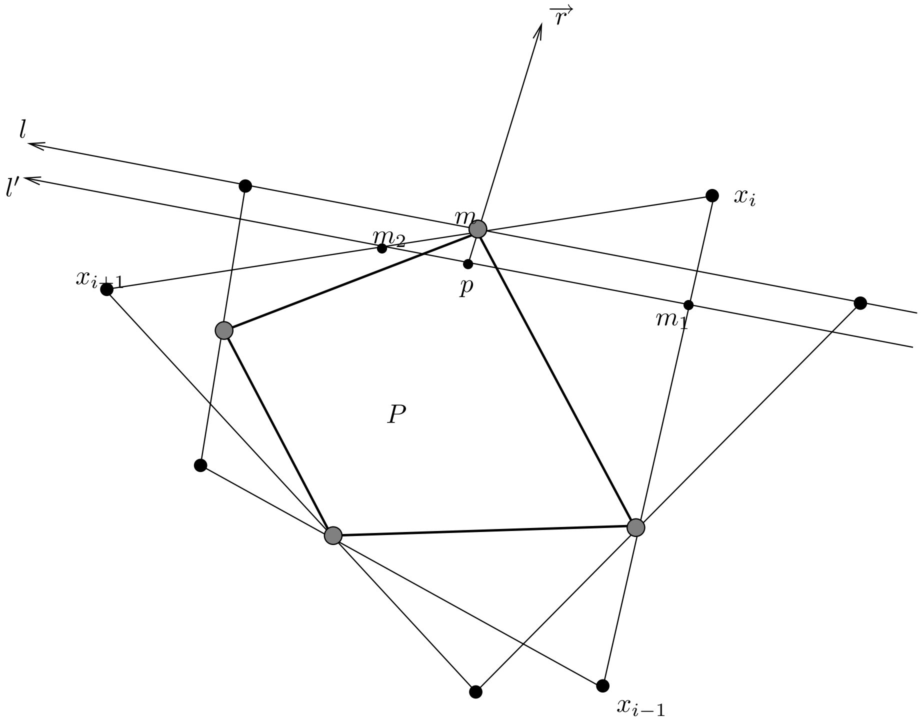


FIG. 3. *Lemma* 2.3.

respect to a point Q, not on γ, is the (counterclockwise) winding number of the path γ around Q. It is well known that if $\overrightarrow{r}$ is any ray emanating from Q, then the number of times γ crosses $\overrightarrow{r}$ from right to left minus the number of times γ crosses $\overrightarrow{r}$ from left to right equals the index of γ with respect to Q.

LEMMA 2.3. *Let P be a convex polygon, let $\mathcal{C}$ be a tangency path for P, and let G denote the set of vertices of $\mathcal{C}$. Then, for every directed line l that touches P so that P is contained in the closed half-plane to the left of l, we have*

$$A_G(l) + B_G(l)/2 = I(\mathcal{C}),$$

where $I(\mathcal{C})$ is the index of the closed path $\mathcal{C}$ with respect to any point in the interior of P.

Proof of Lemma 2.3. Let $x_0, x_1, \ldots, x_{n-1}$ be the set of vertices of $\mathcal{C}$ cyclicly ordered as they appear along the path $\mathcal{C}$. Let m be the midpoint of $l \cap P$. Let l' be a directed line parallel to l that is strictly to the left of l, intersects the interior of P, but still to the right of all vertices of $\mathcal{C}$ which are to the left of l. Let $p \in l' \cap P$ be such that the ray $\overrightarrow{r}$ emanating from p and passing through m does not include any vertex of $\mathcal{C}$ (see Figure 3).

Observe that if $x_i \in l$, then either the edge $[x_i, x_{x+1}]$ or the edge $[x_{i-1}, x_i]$ of $\mathcal{C}$ is included in l. Indeed, this follows because l is a tangent of P. Therefore, if k denotes the number of edges of $\mathcal{C}$ that are included in l, then $B(l) = 2k$. Clearly, every edge of $\mathcal{C}$ that is included in l is crossed by $\overrightarrow{r}$. Indeed, let $[x_i, x_{i+1}]$ be such an edge. By the definition of a tangency path, $[x_i, x_{i+1}] \supset l \cap P$. Therefore, $\overrightarrow{r}$ and $[x_i, x_{i+1}]$ meet at m. It follows that there are precisely $I(\mathcal{C}) - k$ edges of $\mathcal{C}$ that cross $\overrightarrow{r}$ and are not included in l.

If $[x_i, x_{i+1}]$ is an edge of $\mathcal{C}$ that crosses $\overrightarrow{r}$ and is not included in l, then precisely one of x_i and x_{i+1} is to the right of l. Indeed, if both x_i and x_{i+1} are to the right of l, then $[x_i, x_{i+1}] \cap P = \emptyset$. If none of x_i and x_{i+1} is to the right of l, then $[x_i, x_{i+1}] \cap \overrightarrow{r} = \emptyset$. In both cases we reach a contradiction.

On the other hand, we claim that if x_i is in the half-plane to the right of l, then precisely one of the edges $[x_i, x_{i+1}]$ and $[x_{i-1}, x_i]$ crosses $\overrightarrow{r}$. Indeed, observe that both $[x_i, x_{i+1}]$ and $[x_{i-1}, x_i]$ must cross the line l, for otherwise their intersection with P is empty. By the choice of l', both $[x_i, x_{i+1}]$ and $[x_{i-1}, x_i]$ must also cross l'. Let m_1 denote the intersection point of $[x_{i-1}, x_i]$ with l', and let m_2 denote the intersection point of $[x_i, x_{i+1}]$ with l'. Now p must be strictly between m_1 and m_2 on l', because P and therefore also p is to the left of both directed lines $\overrightarrow{x_i x_{i+1}}$ and $\overrightarrow{x_{i-1} x_i}$. It now follows that $\overrightarrow{r}$ crosses precisely one of $[x_i, x_{i+1}]$ and $[x_{i-1}, x_i]$ as required. This is because $\overrightarrow{r}$ meets precisely two edges of the triangle whose vertices are m_1, x_i, and m_2. One edge met by $\overrightarrow{r}$ is $[m_1 m_2]$; the other is either $[m_1, x_i]$, or $[m_2, x_i]$.

We can therefore conclude that $I(\mathcal{C}) - k = A(l)$. Combining this with $B(l) = 2k$ we obtain the desired result, namely, $A(l) + B(l)/2 = I(\mathcal{C})$. □

The following corollary is thus an immediate consequence of Lemma 2.3.

Corollary 2.4. *Let P be a convex polygon, and let $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_t$ be vertex disjoint tangency paths for P. Let G denote the set of vertices of all paths together. For every directed line l that touches P, so that P is contained in the open half-plane to the left of l, we have*

$$A_G(l) + B_G(l)/2 = \sum_{i=1}^{t} I(\mathcal{C}_i).$$

The next lemma relates between tangency paths and the property of skipping k for a certain value of k.

Lemma 2.5. *Let P be a convex polygon, and let $J \subset P$ be a finite set of points. Let $\mathcal{C}_1, \ldots, \mathcal{C}_t$ be a collection of vertex disjoint tangency paths for P, with the property that every edge of P is contained in at least one edge of some tangency path. Let G denote the set of vertices of all paths together. Then the set $S = G \cup vert(P) \cup J$ skips $k = I(\mathcal{C}_1) + \cdots + I(\mathcal{C}_t)$.*

Proof. Let l be a directed line determined by S. It is enough to show that the number of points of S in the open half-plane to the right of l is different from k.

Let l' be the directed line with the same direction as that of l such that l' touches P and P is contained in the closed half-plane to the left of l' (see Figure 4).

Case 1. $l = l'$. Then by Corollary 2.4, $A_G(l) + B_G(l)/2 = k$. Observe that $B_G(l) > 0$. Indeed, this is true if l contains an edge of P, because then there is an edge of some tangency path $\mathcal{C}_i$ that lies on l. If l touches P at a vertex, then it must pass through at least one more point of S which therefore belongs to G. It follows that $A_G(l) < k$. However, $A_G(l)$ is exactly the number of points of S in the open half-plane to the right of l.

Case 2. l is to the left of l'. Then the number of points of G in the open half-plane to the right of l is at least $A_G(l') + B_G(l') \geq A_G(l') + B_G(l')/2 = k$. Moreover, since l' passes through at least one vertex of P, we obtain that the number of points of G in the open half-plane to the right of l is at least $k + 1$.

Case 3. l is to the right of l'. Then the number of points of S that are on or to the right of l is at most $A_G(l') \leq k$. However, l passes through at least two points of S, and therefore the number of points of G in the open half-plane to the right of l is at most $k - 2$. □

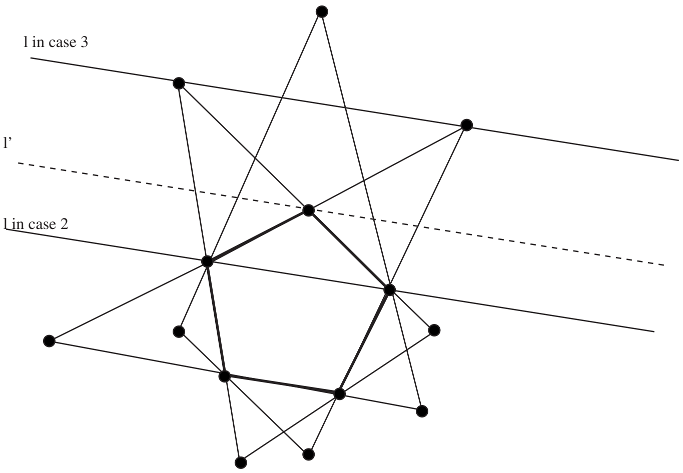


FIG. 4. *Lemma* 2.5.

We can impose another simple condition on the tangency paths $\mathcal{C}_1, \dots \mathcal{C}_t$ in Lemma 2.5, so that the resulting set S skips two consecutive values.

LEMMA 2.6. *Let P be a convex polygon, and let $J \subset P$ be a finite set of points. Let $\mathcal{C}_1, \dots, \mathcal{C}_t$ be a collection of vertex disjoint tangency paths for P, with the following two properties:*

- *Every edge of P is contained in at least two edges of some tangency paths.*
- *Every edge of a tangency path is collinear with at least one more edge of a (possibly other) tangency path.*

Let G denote the set of vertices of all paths together. Then $S = G \cup vert(P) \cup J$ skips k and $k-1$, where $k = I(\mathcal{C}_1) + \cdots + I(\mathcal{C}_t)$.

Proof. The proof goes exactly along the same lines as the proof of Lemma 2.5, except that now whenever $B(l') > 0$ we may conclude that $B(l') \geq 4$.

The example in Figure 5 shows such a case where the polygon P is the inner regular 7-gon. □

3. Construction of large skippable sets.

In this section we will use Lemma 2.5 to construct arbitrary large skippable sets. We will also show that $\{k\}$ is skippable for every $k \geq 2$, by constructing suitable arbitrary large sets of points that do not have a k-edge for a fixed value of $k \geq 2$. This is the goal of our next theorem.

We will need the following terminology in what follows. If P is a convex polygon, then a *diagonal* of P is a segment connecting two vertices of P. We say that the *order* of that diagonal is k if one of the open half-planes bounded by the line that contains this diagonal includes precisely $k-1$ vertices of P. Thus, for example, an edge of P is a diagonal of order 1.

For a convex polygon P and a point x outside P, the angle at which x *sees* P is the angle between the two tangents to P that pass through x.

It easy to see (and also follows immediately from Claim 4.1 in [PP1]) that $\{0\}$ and $\{1\}$ are not skippable. In other words, every noncollinear set of points S that is large enough must have a 0-edge and a 1-edge (in fact, one can drop the "large enough" condition here). In view of this we can now characterize precisely the skippable sets that consist of one element only.

Theorem 3.1. *The set $\{k\}$ is skippable if and only if $k \geq 2$.*

Proof. We already observed the "only if" part. For the "if" part fix some $k \geq 2$. Let Q be a regular $(2k+1)$-gon in the plane. Denote the vertices of Q in a cyclic counterclockwise order by $x_0, x_1, \ldots, x_{2k}$. Start from x_0, and draw a segment to x_k and from there to x_{2k} and from there to x_{3k} and so forth, where the indices are taken modulo $2k+1$. Since k and $2k+1$ are relatively prime, we will obtain a closed path $\mathcal{C}$ of length $2k+1$. This path is in fact combined from the diagonals of Q of order k. The intersections of all half-planes bounded by those diagonals and containing the center of Q form a smaller copy of a $(2k+1)$-gon that we denote by P. Let J be an arbitrary large set of points inside P.

P, J, and $\mathcal{C}$ satisfy the conditions of Lemma 2.5. Therefore $S = vert(P) \cup J \cup vert(Q)$ skips $I(\mathcal{C})$. This index is easy to calculate. Every vertex of Q sees P at an angle of $\pi/(2k+1)$. Therefore, the index of the path $\mathcal{C}$ with respect to any point inside P is $I(\mathcal{C}) = \frac{(2k+1)(\pi-\pi/(2k+1))}{2\pi} = k$. □

Remark. The construction in the proof of Theorem 3.1 was in fact suggested much earlier by Perles. The present notion of a tangency path gives us a convenient environment for presenting an elegant proof for the validity of the construction.

Using Lemma 2.5 as our main tool, we can also show very easily that there are arbitrary large skippable sets. This is the content of the next theorem. We will omit the very specific details of the construction but include enough for the reader to be able to complete the proof.

Theorem 3.2. *There are arbitrary large skippable sets.*

Proof. We start with a set G_1 that skips just one value k_1 and is constructed just like in the proof of Theorem 3.1. The construction consists of a polygon P_1 together with some points outside. We may add any number of points inside P_1 to get a larger set that also skips k_1.

Our construction for the proof of Theorem 3.2 is recursive. Assume that we have already constructed a set G_n that skips the values $k_1, \ldots, k_n$. Assume that G_n includes the set of vertices of a regular polygon P_n such that, no matter how many points we add to G_n inside P_n, the resulting set will still skip $k_1, \ldots, k_n$.

To construct G_{n+1} we will add points to G_n but only inside the polygon P_n. We add a very small copy of a set that skips, say, $k = 5$ and is constructed just like in the proof of Theorem 3.1. Namely, we add the set of vertices of an 11-gon plus the intersection of every two of its consecutive diagonals of order 5. These intersection points are the vertices of another (smaller) 11-gon that we denote by P_{n+1}. Then we add a set S of additional points outside (but very close to) P_{n+1}, so that they are still inside P_n, and such that $G_n \cup S$ may be regarded as a vertex disjoint union of tangency paths for P_{n+1}. One can easily be convinced that this can be done by adding at most, say, 10 extra points for each point of G_n. We thus get a resulting set G_{n+1} that can be regarded as G_n together with some extra points inside P_n, and therefore it still skips $k_1, \ldots, k_n$. However, it can also be regarded as a union of tangency paths for P_{n+1} and thus skips another value that we denote by k_{n+1}. This value must be greater than k_n, since the sum of the indices of all tangency paths to P_{n+1} is clearly greater than that of the tangency paths for P_n. Moreover, we can add any number of points inside P_{n+1}, and the resulting set of points will still skip $k_1, \ldots, k_{n+1}$. This shows that $\{k_1, \ldots, k_{n+1}\}$ is skippable and also concludes the induction step. □

Observe that the construction brought in the proof of Theorem 3.2 is exponential in the number of values that are skipped. This is clear from the proof. We can thus in general construct arbitrary large sets of points G that skip $\Omega(\log |G|)$ values of k. It is interesting to note that the totally different construction by Kupitz and Perles of arbitrary large sets G that skip the values $|G|, |G|-1, \dots, |G|-\log_2|G|$ is yet another example of a different nature for a set of n elements that skips roughly $\log n$ values between 1 and n. It is not known what the maximum number of values is (between 1 and n) that a noncollinear set of points of cardinality n can skip. A nontrivial (with constant multiplier less than 1) linear upper bound follows from the result in [PP1], which says that a noncollinear set of points cannot skip both k and $k+2$, provided that its cardinality is at least $2k+4$.

We thus leave this question open with no conjecture.

Problem. What is the maximum number of values that a noncollinear set of n points in the plane can skip?

Of course, in terms of the order of magnitude, the answer can be anything between $\log n$ and n.

4. Skippable sets of two elements. In this section we will characterize all skippable sets of two elements. It shown in [PP1] that $\{k, k+2\}$ is not skippable for any k. It follows from Theorem 3.1 that any set $\{k, l\}$ that contains either 0 or 1 is not skippable. We will show that, apart from another two sets of two elements that are not skippable, all others are skippable.

THEOREM 4.1. *For every $k \geq 4$, the set $\{k, k+1\}$ is skippable.*

Proof. We will use a construction that satisfies the conditions of Lemma 2.6. Fix $k \geq 4$. Let P be a regular $(2k-1)$-gon, and let $l_0, \dots, l_{2k-2}$ denote the lines containing the edges of P in a cyclic order. Let S_1 denote the intersection points $l_j \cap l_{j+2}$ for $j = 0, \dots, 2k-1$ (where the indices are taken modulo $2k-1$). S_1 is in fact the set of vertices of a regular $(2k-1)$-gon that contains P. Let S_2 denote the intersection points $l_j \cap l_{j+k-1}$ for $j = 0, \dots, 2k-1$. S_2 is the set of vertices of a regular $(2k-1)$-gon that contains S_1 inside. Observe that we need $k \geq 4$ in order for S_1, S_2, and the vertices of P to be pairwise different.

A careful look at the construction of S_1 and S_2 shows that each of the sets S_1 and S_2 can be regarded as a set of vertices of a union of tangency paths for P. Moreover, every edge of P is contained in (exactly) two edges of these tangency paths, and every edge of a tangency path contains an edge of P. The sum of the indices of these tangency paths with respect to any point in P is also easy to calculate (keeping in mind that all of the paths are counterclockwise oriented). Every point of S_1 sees P at an angle of $4\pi/(2k-1)$, and every point of S_2 sees P at an angle of $2(k-1)\pi/(2k-1)$. Therefore the sum of the indices of all paths with respect to P is

$$\frac{\sum_{i=1}^{2k-1}(\pi - 4\pi/(2k-1)) + \sum_{i=1}^{2k-1}(\pi - 2(k-1)\pi/(2k-1))}{2\pi} = k+1.$$

By Lemma 2.6, the set that consists of the vertices of P together with S_1 and S_2 skips both k and $k+1$. This is still true if we add an arbitrary number of points inside P. (Figure 5 shows the resulting construction in Theorem 4.1 for $k = 4$.) □

Next we show, in Theorems 4.2 and 4.3, that the sets $\{2, 3\}$ and $\{3, 4\}$ are not skippable. In the proofs we use the method of flip arrays also known as allowable sequences that was invented by Goodman and Pollack (see, e.g., [GP84, GP93]). We refer the reader to the corresponding section of [PP1] for a detailed description and useful notation. Briefly, one can encode a set of points G in the plane as a sequence of permutation on n elements. This is done by keeping track of the order of the projections of the points from left to right on a directed line that changes its direction graduately counterclockwise until it reverses its direction. If we number the points according to the order of their projection on the line in its initial position, the recorded

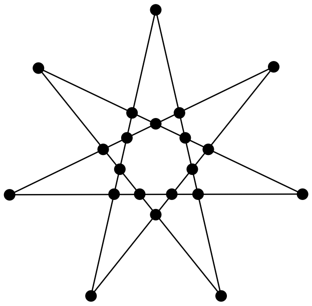


FIG. 5. *A set of points that skips $k = 4, 5$.*

different ordering of the projections of the points form a sequence of permutation on $\{1, \dots, n\}$. The first permutation is the identity, while the last is $(n, n-1, \dots, 1)$. Each permutation is obtained from its predecessor by flipping a contiguous block of monotone increasing elements. Each such flip corresponds to a line determined by the set of points. The important observation is that a flip of the block $[a, b]$ (namely, the elements in the places $a, a+1, \dots, b$ in a permutation) represents a line, determined by G, that passes through $b - a + 1$ points of G. This line has $a - 1$ points of G in one open half-plane bounded by it and $n - b$ points of G in the other. Another important property of the sequence of permutations is that any pair of elements from $\{1, 2, \dots, n\}$ changes order exactly once.

THEOREM 4.2. *If G is a noncollinear set of at least 8 points in the plane, then G cannot skip both $k = 2$ and $k = 3$. In particular, $\{2, 3\}$ is not skippable.*

Proof. Assume to the contrary that $|G| \geq 8$ and G does not have a 2-edge nor a 3-edge. We carefully analyze the flip array of G. Let $n = |G|$. The flip array of G consists of a sequence of permutation in S_n. Each permutation is obtained from its predecessor by flipping a contiguous monotone increasing block of elements. Observe that we are not allowed to flip blocks of the form $[3, b]$ (where $b > 3$) nor $[4, b]$ (where $b > 4$). Indeed, such blocks represent a 2-edge or a 3-edge, respectively, of G.

Define an *interesting flip* as a flip of a block that contains the block $[2, 5]$. Observe that, in view of the forbidden flips, the only way to take an element that is in one of the first 4 places of a permutation to a place within $[5, n]$ is by an interesting flip. The element 1 must eventually move to position n. Hence, there must be at least one interesting flip. The block of the first interesting flip must be of the form $[2, b]$ for some $b \geq 5$, because if it contained the block $[1, 5]$, that would mean that the convex hull of G contains an edge with at least 5 points, and this implies easily that G has an r-edge for every $r < 5$.

Therefore, after the first interesting flip there is at least one element from $\{1, 2, 3, 4\}$ that remains in the region $[1, 4]$. Since $|G| \geq 8$, this element must eventually move to the region $[5, n]$. Therefore, there must be a second interesting flip. The block of the second interesting flip will again contain 3 elements from the region $[1, 4]$. That implies that at least 2 elements that took part in the first interesting flip will take part also in the second interesting flip. We reached a contradiction, as any two elements must change order exactly once. □

THEOREM 4.3. *If G is a noncollinear set of at least 10 points, then G has either a 3-edge or a 4-edge. In particular, the set $\{3, 4\}$ is not skippable.*

Proof. Assume to the contrary that $|G| \geq 10$ and that G does not have a 3-edge nor a 4-edge. Once again we will make use of the flip array of the set G. Let $n = |G|$.

This time the flips of blocks of the form $[4, b]$ and $[5, b]$ are not allowed. We will also make use of the observation that flips of blocks of the form $[a, n-4]$ and $[a, n-5]$ are not allowed.

We will call a flip *interesting of type I* if the block of this flip contains $[3, 6]$. A flip will be called *interesting of type II* if the block of the flip contains $[n-5, n-2]$. Observe that, in view of the forbidden flips, an element can move from the region $[1, 5]$ to the region $[6, n]$ (and vice versa) only by an interesting flip of type I. Similarly an element can move from the region $[n-4, n]$ to the region $[1, n-5]$ (and vice versa) only by an interesting flip of type II.

We claim that it is not possible to have a flip that is an interesting flip of both types I and II. Indeed, this would necessarily mean that the block of such a flip must contain the region $[3, n-2]$ in a permutation. This implies that G determines a line with at least $n-4$ points, having at most two points in each open half-plane bounded by it. It is easily seen by inspection, taking into account that $|G| \geq 10$, that G must then have either a 3-edge or a 4-edge.

Just like in the proof of Theorem 4.2, there must be at least two interesting flips of type I and similarly two interesting flips of type II. Let us consider the first interesting flip of type I. This flip includes three elements from $\{1, 2, 3, 4, 5\}$ in the positions $[3, 5]$. Right after this flip the elements in the region $[3, 5]$ are in decreasing order. Therefore the second interesting flip of type I may include just one of them in its block. The other two elements must therefore be at the positions $[1, 2]$ and remain untouched while the second interesting flip happens. Right after the second interesting flip we have two elements in positions $[1, 2]$ that already changed order with each other and three elements in positions $[3, 5]$, every two of which already changed order. Therefore a third interesting flip of type I is not possible (for it must include three elements from the region $[1, 5]$, no two of which already changed order). Similarly, there are just two interesting flips of type II.

Since $|G| \geq 10$, the elements $\{1, 2, 3, 4, 5\}$ must all end at the region $[n-4, n]$. We know that three elements from $\{1, 2, 3, 4, 5\}$ belong to the block of the first interesting flip. Those three elements must move from the region $[1, n-5]$ to $[n-4, n]$, and this can be done only by interesting flips of type II. There are at most two interesting flips of type II, and we obtain a contradiction, since two of the three elements from $\{1, 2, 3, 4, 5\}$ that were flipped during the first interesting flip of type I must be flipped during the same interesting flip of type II. □

The next theorem will complete the picture as for the skippable sets of two elements.

Theorem 4.4. *For every $k \geq 2$ and every $l \geq k+3$ the set $\{k, l\}$ is skippable.*

Proof. For any k and l that satisfy the conditions in the theorem, we must show that there are arbitrary large sets of points that do not have a k-edge nor an l-edge.

We will make use of Lemma 2.5 to show the validity of our construction. Fix $k \geq 2$, and assume first that $l = k+3$. Let Q_1 be a regular $(2k+2)$-gon. Let Q_2 be the regular $(2k+2)$-gon whose vertices are the intersection points of consecutive diagonals of order k of Q_1. Clearly, the vertices of Q_1 can be regarded as the union of vertices of tangency paths for Q_2. Each vertex of Q_1 sees Q_2 at the angle of $\pi/(k+1)$; thus, the sum of the indices of all tangency paths is $\frac{(2k+2)(\pi - \pi/(k+1))}{2\pi} = k$. Observe that by construction every edge of Q_2 is contained in an edge of some tangency path (just like in the proof of Theorem 3.1). Hence by Lemma 2.5, the set that consists of the union of the vertices of Q_1 and Q_2 skips k. This remains true if we add points inside Q_2. Let Q_3 be the $(2k+2)$-gon whose vertices are the intersection points of consecutive diagonals of order 2 of Q_2. Clearly, $Q_3 \subset Q_2$. The vertices of Q_2 can

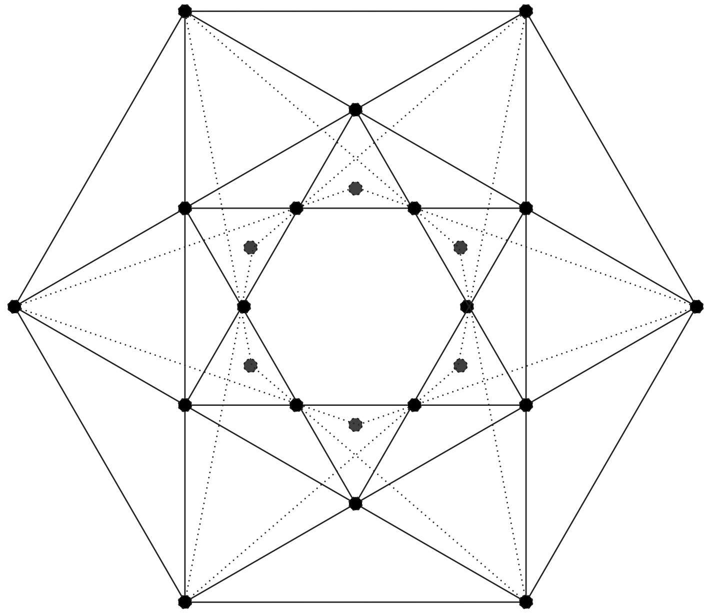


FIG. 6. *Theorem* 4.4 *(the case $k = 2$ and $l = k + 3 = 5$).*

be regarded as the set of vertices of a union of tangency paths to Q_3, and each edge of Q_3 is contained in one edge of some tangency path. Since every point of Q_2 sees Q_3 at angle $\pi(k-1)/(k+1)$, the sum of the indices of these paths with respect to Q_3 is $\frac{(2k+2)(\pi-\pi(k-1)/(k+1))}{2\pi} = 2$. It is thus enough to show that we can add points inside Q_2 but outside Q_3 so that together with the points of Q_1 they will constitute the vertex set of a union of tangency paths of a total index $k+1$. This can indeed be done. For every two opposite points of Q_1 we add two points in $Q_2 \setminus Q_3$ so that they form a tangency path to Q_3 with index 1 for Q_3. This is illustrated in Figure 6.

We thus obtained a set G that skips k and $k+3$. This remains true if we add an arbitrary number of points inside Q_3. This shows that $\{k, k+3\}$ is skippable. If $l > k+3$, we add to G l tangency paths of index 1 that are contained in $Q_2 \setminus Q_3$. This completes the proof of the theorem. □

We can now summarize our results in the following theorem that characterizes all skippable sets of two elements.

THEOREM 4.5. *Let $k < l$ be nonnegative integers. The set $\{k, l\}$ is skippable if and only if one of the following two conditions is satisfied:*

- *$k \geq 2$ and $l \geq k+3$, or*
- *$k \geq 4$ and $l = k+1$.*

REFERENCES

[GP84] J. F. GOODMAN AND R. POLLACK, *On the number of k-sets of a set of n points in the plane*, J. Combin. Theory Ser. A, 36 (1984), pp. 101–104.

[GP93] J. E. GOODMAN AND R. POLLACK, *Allowable sequences and order types in discrete and computational geometry*, in New Trends in Discrete and Computational Geometry, J. Pach, ed., Springer-Verlag, Berlin, 1993, pp. 103–134.

[K93] Y. S. KUPITZ, *Separation of a finite set in* R^d *by spanned hyper-planes*, Combinatorica, 13 (1993), pp. 249–258.

[K94] Y. S. Kupitz, *Spanned k-supporting hyper-planes of finite sets in* R^d, J. Combin. Theory Ser. A, 65 (1994), pp. 117–136.

[KP] Y. S. Kupitz and M. A. Perles, *private communication*.

[PP1] M. A. Perles and R. Pinchasi, *Large sets must have either a* k*-edge or a* $(k+2)$*-edge*, in Towards a Theory of Geometric Graphs, Contemp. Math. 342, Providence, RI, 2004, pp. 225–232.

SIAM J. Discrete Math.
Vol. 21, No. 2, pp. 396–407

THE MIXING SET WITH FLOWS*

MICHELE CONFORTI†, MARCO DI SUMMA†, AND LAURENCE A. WOLSEY‡

Abstract. We consider the mixing set with flows:

$$s + x_t \geq b_t,\ x_t \leq y_t \text{ for } 1 \leq t \leq n;\ \ s \in \mathbb{R}^1_+,\ x \in \mathbb{R}^n_+,\ y \in \mathbb{Z}^n_+.$$

It models a "flow version" of the basic mixing set introduced and studied by Günlük and Pochet [*Math. Program.*, 90 (2001), pp. 429–457], as well as the most simple stochastic lot-sizing problem with recourse. More generally it is a relaxation of certain mixed integer sets that arise in the study of production planning problems. We study the polyhedron defined as the convex hull of the above set. Specifically we provide an inequality description, and we also characterize its vertices and rays.

Key words. mixed integer programming, mixing, lot-sizing

AMS subject classifications. 90C11, 90C57

DOI. 10.1137/05064148X

1. Introduction.

We give an inequality (external) and extreme point and extreme ray (internal) description for the convex hull of the *mixing set with flows* X^{FM}:

$$s + x_t \geq b_t \text{ for } 1 \leq t \leq n, \tag{1.1}$$

$$x_t \leq y_t \text{ for } 1 \leq t \leq n, \tag{1.2}$$

$$s \in \mathbb{R}^1_+,\ x \in \mathbb{R}^n_+,\ y \in \mathbb{Z}^n_+, \tag{1.3}$$

where $0 \leq b_1 \leq \cdots \leq b_n,\ b \in \mathbb{R}^n$.

This set is a relative of the *mixing set* X^{MIX}:

$$s + y_t \geq b_t \text{ for } 1 \leq t \leq n, \tag{1.4}$$

$$s \in \mathbb{R}^1_+,\ y \in \mathbb{Z}^n_+, \tag{1.5}$$

with $b \in \mathbb{R}^n$ introduced formally by Günlük and Pochet [5] and studied by Pochet and Wolsey [7] and Miller and Wolsey [6]. Internal and external descriptions of the convex hull of X^{MIX} are given in [5].

The original motivation for studying X^{FM} was to generalize X^{MIX} by introducing the continuous (flow) variables x, noting that $\text{conv}(X^{MIX})$ is a face of $\text{conv}(X^{FM})$. However, X^{FM} is also closely related to two lot-sizing models that we now present.

[*] Received by the editors September 29, 2005; accepted for publication (in revised form) December 15, 2006; published electronically May 10, 2007. This work was carried out partly within the framework of ADONET, a European network in Algorithmic Discrete Optimization, under contract MRTN-CT-2003-504438. This text presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The scientific responsibility is assumed by the authors.

http://www.siam.org/journals/sidma/21-2/64148.html

[†] Dipartimento di Matematica Pura ed Applicata, Università degli Studi di Padova, Via Trieste 63, 35121 Padova, Italy (conforti@math.unipd.it, mdsumma@math.unipd.it).

[‡] Center of Operations Research and Econometrics (CORE), Université catholique de Louvain. 34, Voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium (wolsey@core.ucl.ac.be).

The constant capacity lot-sizing model can be formulated as

$$s_0 + \sum_{u=1}^{t} w_u \geq \sum_{u=1}^{t} d_u \text{ for } 1 \leq t \leq n, \tag{1.6}$$

$$w_u \leq z_u \text{ for } 1 \leq u \leq n, \tag{1.7}$$

$$s_0 \in \mathbb{R}^1_+,\ w \in \mathbb{R}^n_+,\ z \in \{0,1\}^n, \tag{1.8}$$

where d_t is the demand in period t, s_0 is the initial stock variable, w_t is the amount produced in t bounded above by the capacity C (we take $C = 1$ throughout without loss of generality), and z_t is a 0-1 set-up variable with $z_t = 1$ if $x_t > 0$. Summing the constraints (1.7) over $1 \leq u \leq t$ (for each $t = 1, \ldots, n$) leads to the relaxation

$$s_0 + \sum_{u=1}^{t} w_u \geq \sum_{u=1}^{t} d_u \text{ for } 1 \leq t \leq n,$$

$$\sum_{u=1}^{t} w_u \leq \sum_{u=1}^{t} z_u \text{ for } 1 \leq t \leq n,$$

$$s_0 \in \mathbb{R}^1_+,\ \sum_{u=1}^{t} w_u \in \mathbb{R}^1_+,\ \sum_{u=1}^{t} z_u \in \mathbb{Z}^1_+ \text{ for } 1 \leq t \leq n.$$

With $s := s_0$, $x_t := \sum_{u=1}^{t} w_u$, and $y_t := \sum_{u=1}^{t} z_u$, this is precisely the set X^{FM}.

The second link is to the two-period stochastic lot-sizing model with constant capacities. Specifically, at time 0 one must choose to produce a quantity s at a per unit cost of h. Then in period 1, n different outcomes are possible. For $1 \leq t \leq n$, the probability of event t is ϕ_t, the demand is b_t, and the unit production cost is p_t, with production in batches of size up to $C = 1$. There are also a fixed cost of q_t per batch and a possible bound k_t on the number of batches. If we want to minimize the total expected cost, the resulting problem is

$$\min \left\{ hs + \sum_{t=1}^{n} \phi_t (p_t x_t + q_t y_t) : (s, x, y) \in X^{FM};\ y_t \leq k_t,\ 1 \leq t \leq n \right\}. \tag{1.9}$$

Note that when $k_t = 1,\ 1 \leq t \leq n$, this is the standard lot-sizing variant. Also the uncapacitated case when $b_t \leq 1,\ 1 \leq t \leq n$, has been treated in Guan et al. [4].

It is also interesting to view X^{MIX} and X^{FM} as simple mixed integer sets with special structure. One observation is that the associated constraint matrices are totally unimodular, but the right-hand sides are typically noninteger as $b \in \mathbb{Q}^n$. Miller and Wolsey [6] and Van Vyve [9] have introduced and studied a different extension, called a *continuous mixing set*, again having a totally unimodular system of constraints.

We now describe the contents of this paper, and then end the introduction with some notation. In section 2 we develop a polyhedral result used later to establish that a given polyhedron is "integral" (i.e., its vertices are points of the mixed integer set under consideration). In section 3 we find an external description of conv(X^{FM}) and two closely related sets, and in section 4 we give an internal description that leads to a simple polynomial time algorithm for optimization over the set X^{FM}. We conclude in section 5 with a brief indication of related work on other generalizations of mixing sets.

Notation. Throughout we will use the following notation: $N := \{1, \ldots, n\}$, e_S for the characteristic vector of a subset $S \subseteq N$, $e_i := e_{\{i\}}$ for the ith unit vector, and $\underline{0} := e_\emptyset$ and $\underline{1} := e_N$ for the n-vectors of 0's and 1's, respectively.

2. Some equivalences of polyhedra. In the next section we will relate the polyhedra $\text{conv}(X^{FM})$ and $\text{conv}(X^{MIX})$. To do this, we will need some polyhedral equivalences that we introduce here.

For a nonempty polyhedron P in $\mathbb{R}^n$ and a vector $\alpha \in \mathbb{R}^n$, define $\mu_P(\alpha) := \min\{\alpha x : x \in P\}$ and let $M_P(\alpha)$ be the face $\{x \in P : \alpha x = \mu_P(\alpha)\}$, where $M_P(\alpha) = \emptyset$ whenever $\mu_P(\alpha) = -\infty$.

LEMMA 2.1. *Let $P \subseteq Q$ be two nonempty polyhedra in $\mathbb{R}^n$ and let α be a nonzero vector in $\mathbb{R}^n$. Then the following conditions are equivalent:*

1. *$\mu_P(\alpha) = \mu_Q(\alpha)$;*
2. *$M_P(\alpha) \subseteq M_Q(\alpha)$.*

Proof. Suppose $\mu_P(\alpha) = \mu_Q(\alpha)$. Since $P \subseteq Q$, every point in $M_P(\alpha)$ belongs to $M_Q(\alpha)$. So if 1 holds, then 2 holds as well. The converse is obvious. □

LEMMA 2.2. *Let $P \subseteq Q$ be two nonempty polyhedra in $\mathbb{R}^n$, where P is not an affine variety. Suppose that for every inequality $\alpha x \geq \beta$ that is facet-inducing for P, at least one of the following holds:*

1. *$\mu_P(\alpha) = \mu_Q(\alpha)$;*
2. *$M_P(\alpha) \subseteq M_Q(\alpha)$.*

Then $P = Q$.

Proof. We prove that if $M_P(\alpha) \subseteq M_Q(\alpha)$ for every inequality $\alpha x \geq \beta$ that is facet-inducing for P, then every facet-inducing inequality for P is a valid inequality for Q and every hyperplane containing P also contains Q. This shows $Q \subseteq P$ and therefore $P = Q$. By Lemma 2.1, the conditions $\mu_P(\alpha) = \mu_Q(\alpha)$ and $M_P(\alpha) \subseteq M_Q(\alpha)$ are equivalent and we are done.

Let $\alpha x \geq \beta$ be a facet-inducing inequality for P. Since $M_P(\alpha) \subseteq M_Q(\alpha)$, then $\beta = \mu_P(\alpha) = \mu_Q(\alpha)$ and $\alpha x \geq \beta$ is an inequality which is valid for Q. Now let $\gamma x = \delta$ be a hyperplane containing P. If $Q \not\subseteq \{x : \gamma x = \delta\}$, then there exists $\bar{x} \in Q$ such that $\gamma \bar{x} \neq \delta$. We assume without loss of generality that $\sigma = \gamma\bar{x} - \delta > 0$. Since P is not an affine variety, there exists an inequality $\alpha x \geq \beta$ which is facet-inducing for P (and so it is valid for Q). Then, for $\lambda > 0$ the inequality $(\lambda\alpha - \gamma)x \geq \lambda\beta - \delta$ is also facet-inducing for P, so it is valid for Q. Choosing $\lambda > 0$ such that $\lambda(\alpha\bar{x} - \beta) < \sigma$ gives a contradiction, as $(\lambda\alpha - \gamma)\bar{x} = \lambda\alpha\bar{x} - \gamma\bar{x} < \lambda\beta + \sigma - \gamma\bar{x} = \lambda\beta - \delta$. □

If P is not full-dimensional, for each facet F of P there are infinitely many distinct inequalities that define F (two inequalities are distinct if their associated half-spaces are distinct—that is, if one is not the positive multiple of the other). Observe that the hypotheses of the lemma must be verified for *all* distinct facet-defining inequalities (not just one facet-defining inequality for each facet), otherwise the result is false. For instance, consider the polyhedra $P = \{(x, y) : 0 \leq x \leq 1,\ y = 0\} \subset Q = \{(x, y) : 0 \leq x \leq 1,\ 0 \leq y \leq 1\}$. The hypotheses of Lemma 2.2 are satisfied for the inequalities $x \geq 0$ and $x \leq 1$, which define all the facets of P.

Also note that the assumption that P is not an affine variety cannot be removed: indeed, in such a case P does not have proper faces, so the hypotheses of the lemma are trivially satisfied, even if $P \neq Q$.

COROLLARY 2.3. *Let $P \subseteq Q$ be two pointed polyhedra in $\mathbb{R}^n$, with the property that every vertex of Q belongs to P. Let $Cx \geq d$ be a system of inequalities that are valid for P such that for every inequality $\gamma x \geq \delta$ of the system, $P \not\subseteq \{x \in \mathbb{R}^n : \gamma x = \delta\}$.*

If for every $\alpha \in \mathbb{R}^n$ such that $\mu_P(\alpha)$ is finite but $\mu_Q(\alpha) = -\infty$, $Cx \geq d$ contains an inequality $\gamma x \geq \delta$ such that $M_P(\alpha) \subseteq \{x \in \mathbb{R}^n : \gamma x = \delta\}$, then $P = Q \cap \{x \in \mathbb{R}^n : Cx \geq d\}$.

Proof. We first show that $\dim(P) = \dim(Q)$. If not, there exists a hyperplane $\alpha x = \beta$ containing P but not Q. Without loss of generality we can assume that

$\mu_Q(\alpha) < \beta = \mu_P(\alpha)$. So $\mu_Q(\alpha) = -\infty$, otherwise there would exist an α-optimal vertex $\bar{x}$ of Q such that $\alpha\bar{x} < \beta$, contradicting the fact that $\bar{x} \in P$. Now the system $Cx \geq d$ must contain an inequality $\gamma x \geq \delta$ such that $P = M_P(\alpha) \subseteq \{x \in \mathbb{R}^n : \gamma x = \delta\}$, a contradiction.

Let $Q' = Q \cap \{x \in \mathbb{R}^n : Cx \geq d\}$. Note that $P \subseteq Q' \subseteq Q$; thus $\dim(P) = \dim(Q') = \dim(Q)$. Let $\alpha x \geq \beta$ be a facet-inducing inequality for P. If $\mu_Q(\alpha)$ is finite, then Q contains an α-optimal vertex which is in P and therefore $\beta = \mu_P(\alpha) = \mu_{Q'}(\alpha) = \mu_Q(\alpha)$. If $\mu_Q(\alpha) = -\infty$, the system $Cx \geq d$ contains an inequality $\gamma x \geq \delta$ such that $M_P(\alpha) \subseteq \{x \in \mathbb{R}^n : \gamma x = \delta\}$ and $P \not\subseteq \{x \in \mathbb{R}^n : \gamma x = \delta\}$. It follows that $\gamma x \geq \delta$ is a facet-inducing inequality for P and that it defines the same facet of P as $\alpha x \geq \beta$ (that is, $M_P(\alpha) = M_P(\gamma)$). This means that there exist $\nu > 0$, a vector λ, and a system $Ax = b$ which is valid for P such that $\gamma = \nu\alpha + \lambda A$ and $\delta = \nu\beta + \lambda b$. Since $\dim(P) = \dim(Q')$ and $P \subseteq Q'$, the system $Ax = b$ is valid for Q' as well. As $\gamma x \geq \delta$ is also valid for Q', it follows that $\alpha x \geq \beta$ is valid for Q' (because $\alpha = \frac{1}{\nu}\gamma - \frac{\lambda}{\nu}A$ and $\beta = \frac{1}{\nu}\delta - \frac{\lambda}{\nu}b$). Therefore $\beta = \mu_P(\alpha) = \mu_{Q'}(\alpha)$.

Now assume that P consists of a single point and $P \neq Q$. Then Q is a cone having P as apex. Given a ray α of Q, $\mu_P(\alpha)$ is finite while $\mu_Q(\alpha) = -\infty$, so the system $Cx \geq d$ contains an inequality $\gamma x \geq \delta$ such that $P \subseteq \{x \in \mathbb{R}^n : \gamma x = \delta\}$, a contradiction. So we can assume that P is not a single point and thus P is not an affine variety, as it is pointed. Now we can conclude by applying Lemma 2.2 to the polyhedra P and Q'. □

We remark that in the statement of Corollary 2.3 the condition that the two polyhedra are pointed is not necessary: if we replace the property "every vertex of Q belongs to P" with "every minimal face of Q belongs to P," the proof needs a very slight modification to remain valid. (However, in this case we should assume that P is not an affine variety, so that we can apply Lemma 2.2 in the proof.)

We also observe that the condition "for every inequality $\gamma x \geq \delta$ of the system, $P \not\subseteq \{x \in \mathbb{R}^n : \gamma x = \delta\}$" is necessary. For instance, consider the polyhedra $P = \{(x, y) : 0 \leq x \leq 1,\ y = 0\} \subset Q = \{(x, y) : x \geq 0,\ y = 0\}$ and the system consisting of the single inequality $y \geq 0$.

3. An external description of X^{FM}.

The approach taken to derive an inequality description of $\text{conv}(X^{FM})$ is first outlined briefly. We work with two intermediate mixed integer sets Z and X^{INT} for which we establish several properties. The first two link $\text{conv}(X^{FM})$ and $\text{conv}(Z)$, and the next two provide an external description of $\text{conv}(Z)$:

(i) First we observe that $X^{FM} = Z \cap \{(s, x, y) : \underline{0} \leq x \leq y\}$.

(ii) Using Corollary 2.3, we prove that $\text{conv}(X^{FM}) = \text{conv}(Z) \cap \{(s, x, y) : \underline{0} \leq x \leq y\}$.

(iii) We then show that the polyhedra $\text{conv}(Z)$ and $\text{conv}(X^{INT})$ are in 1-1 correspondence via an affine transformation.

(iv) Finally we note that X^{INT} is the intersection of mixing sets, and therefore external descriptions of $\text{conv}(X^{INT})$ and $\text{conv}(Z)$ are known.

3.1. A relaxation of X^{FM}.

Consider the set Z:

$$s + y_t \geq b_t \text{ for } 1 \leq t \leq n, \tag{3.1}$$

$$s + x_k + y_t \geq b_t \text{ for } 1 \leq k < t \leq n, \tag{3.2}$$

$$s + x_t \geq b_t \text{ for } 1 \leq t \leq n, \tag{3.3}$$

$$s \in \mathbb{R}^1_+,\ x \in \mathbb{R}^n,\ y \in \mathbb{Z}^n_+. \tag{3.4}$$

PROPOSITION 3.1. *Let X^{FM} and Z be defined on the the same vector b. Then $X^{FM} \subseteq Z$ and $X^{FM} = Z \cap \{(s,x,y) : \underline{0} \leq x \leq y\}$.*

Proof. To see that $X^{FM} \subseteq Z$, observe that for $(s,x,y) \in X^{FM}$, $s + y_t \geq s + x_t \geq b_t$, so $s + y_t \geq b_t$ is a valid inequality. Also $s + y_t \geq b_t$ and $x_k \geq 0$ imply that $s + x_k + y_t \geq b_t$ is a valid inequality. The only inequalities that define X^{FM} but do not appear in the definition of Z are the inequalities $\underline{0} \leq x \leq y$. □

Since the left-hand sides of inequalities (1.1)–(1.3) and (3.1)–(3.4) have integer coefficients, the recession cones of X^{FM} and Z coincide with the recession cones of their linear relaxations. Thus we have the following.

Observation 1. The extreme rays of $\mathrm{conv}(X^{FM})$ are the following $2n+1$ vectors: $(1,\underline{0},\underline{0})$, $(0,\underline{0},e_k)$, $(0,e_k,e_k)$. The $2n+1$ extreme rays of $\mathrm{conv}(Z)$ are $(0,\underline{0},e_k)$, $(0,e_k,\underline{0})$, $(1,-\underline{1},\underline{0})$. Therefore both recession cones of $\mathrm{conv}(X^{FM})$ and $\mathrm{conv}(Z)$ are full-dimensional simplicial cones, thus showing that $\mathrm{conv}(X^{FM})$ and $\mathrm{conv}(Z)$ are both full-dimensional polyhedra.

Observation 2. Let (s^*,x^*,y^*) be a vertex of $\mathrm{conv}(Z)$. Then

$$s^* = \max \begin{cases} 0, \\ b_t - y^*_t, \ 1 \leq t \leq n, \\ b_t - x^*_t, \ 1 \leq t \leq n, \\ b_t - y^*_t - x^*_k, \ 1 \leq k < t \leq n, \end{cases}$$

$$x^*_k = \max \begin{cases} b_k - s^*, \\ b_t - s^* - y^*_t, \ k < t \leq n. \end{cases}$$

LEMMA 3.2. *Let (s^*,x^*,y^*) be a vertex of* $\mathrm{conv}(Z)$. *Then $\underline{0} \leq x^* \leq y^*$.*

Proof. Assume $x^*_k < 0$ for some index k. Then $s^* > 0$; otherwise, if $s^* = 0$, the constraints $s + x_k \geq b_k$, $b_k \geq 0$ imply $x^*_k \geq 0$.

We now claim that there is an index $t \in N$ such that $s^* = b_t - y^*_t$. If not, $s^* > b_t - y^*_t$, $1 \leq t \leq n$, and there is an $\varepsilon \neq 0$ such that $(s^*,x^*,y^*) \pm \varepsilon(1,-\underline{1},\underline{0})$ belong to $\mathrm{conv}(Z)$, a contradiction.

So there is an index $t \in N$ such that $s^* = b_t - y^*_t > 0$. Since $b_t - y^*_t \geq b_t - y^*_t - x^*_k$, $1 \leq k < t$, this implies $x^*_k \geq 0$, $1 \leq k < t$. Observation 2 also implies $b_t - y^*_t \geq b_k - x^*_k$, $1 \leq k \leq n$. Together with $y^*_t \geq 0$ and $b_t \leq b_k$, $k \geq t$, this implies $x^*_k \geq y^*_t \geq 0$, $k \geq t$. This completes the proof that $x^* \geq \underline{0}$.

Assume $x^*_k > y^*_k$ for some index k. Then $y^*_k \geq 0$ implies $x^*_k > 0$. Assume $x^*_k = b_k - s^*$. Then $y^*_k \geq b_k - s^*$ implies that $x^*_k \leq y^*_k$, a contradiction. Therefore by Observation 2, $x^*_k = b_t - s^* - y^*_t$ for some $t > k$. Since $x^*_k > 0$, then $b_t - s^* - y^*_t > 0$, a contradiction to $s^* + y^*_t \geq b_t$. This shows $x^* \leq y^*$. □

We now can state the main theorem of this section.

THEOREM 3.3. *Let X^{FM} and Z be defined on the the same vector b. Then* $\mathrm{conv}(X^{FM}) = \mathrm{conv}(Z) \cap \{(s,x,y) : \underline{0} \leq x \leq y\}$.

Proof. By Proposition 3.1, $\mathrm{conv}(X^{FM}) \subseteq \mathrm{conv}(Z)$. By Lemma 3.2 and Proposition 3.1, every vertex of $\mathrm{conv}(Z)$ belongs to $\mathrm{conv}(X^{FM})$.

Let $\alpha = (h,p,q)$, $h \in \mathbb{R}^1$, $p \in \mathbb{R}^n$, $q \in \mathbb{R}^n$, be such that $\mu_{\mathrm{conv}(X^{FM})}(\alpha)$ is finite and $\mu_{\mathrm{conv}(Z)}(\alpha) = -\infty$. Since by Observation 1 the extreme rays of $\mathrm{conv}(Z)$ that are not rays of $\mathrm{conv}(X^{FM})$ are $(0,e_k,\underline{0})$ and $(1,-\underline{1},\underline{0})$, then either $p_k < 0$ for some index k or $h < \sum_{t=1}^n p_t$.

If $p_k < 0$, then $M_{\mathrm{conv}(X^{FM})}(\alpha) \subseteq \{(s,x,y) : x_k = y_k\}$.

If $h < \sum_{t=1}^n p_t$, let $N^+ = \{j \in N : p_j > 0\}$ and $k = \min\{j : j \in N^+\}$. We show that $M_{\mathrm{conv}(X^{FM})}(\alpha) \subseteq \{(s,x,y) : x_k = 0\}$. Suppose that $x_k > 0$ in some optimal solution. As the solution is optimal and $p_k > 0$, we cannot decrease only the variable x_k and remain feasible. Thus $s + x_k = b_k$, which implies that $s < b_k$. However this implies that for all $j \in N^+$, we have $x_j \geq b_j - s > b_j - b_k \geq 0$ as $j \geq k$. Now as $x_j > 0$ for all $j \in N^+$, we can increase s by $\varepsilon > 0$ and decrease x_j by ε for all $j \in N^+$. The new point is feasible in X^{FM} and has lower objective value, a contradiction.

To complete the proof, since $\operatorname{conv}(X^{FM})$ is full-dimensional, the system $\underline{0} \leq x \leq y$ does not contain an improper face of $\operatorname{conv}(X^{FM})$. So we can now apply Corollary 2.3 to $\operatorname{conv}(X^{FM})$, $\operatorname{conv}(Z)$, and the system $\underline{0} \leq x \leq y$. □

3.2. The intersection set. The following set is the *intersection set* X^{INT}:

$$\begin{gathered}\sigma_k + y_t \geq b_t - b_k \text{ for } 0 \leq k < t \leq n,\\ \sigma \in \mathbb{R}^{n+1}_+,\ y \in \mathbb{Z}^n_+,\end{gathered}$$

where $0 = b_0 \leq b_1 \leq \cdots \leq b_n$.

Note that X^{INT} is the intersection of the following $n+1$ mixing sets X^{MIX}_k, each one associated with a single variable σ_k:

$$\begin{gathered}\sigma_k + y_t \geq b_t - b_k \text{ for } k < t \leq n,\\ \sigma_k \in \mathbb{R}^1_+,\ y \in \mathbb{Z}^{n-k}_+.\end{gathered}$$

THEOREM 3.4. *Let X^{INT} be an intersection set and let X^{FM} be defined on the same vector b. The affine transformation $\sigma_0 = s$ and $\sigma_t = s + x_t - b_t$, $1 \leq t \leq n$, maps $\operatorname{conv}(X^{FM})$ into $\operatorname{conv}(X^{INT}) \cap \{(\sigma, y) : 0 \leq \sigma_k - \sigma_0 + b_k \leq y_k,\ 1 \leq k \leq n\}$.*

Proof. Let Z be defined on the same vector b. It is straightforward to check that the affine transformation $\sigma_0 = s$ and $\sigma_t = s + x_t - b_t$, $1 \leq t \leq n$, maps $\operatorname{conv}(Z)$ into $\operatorname{conv}(X^{INT})$. By Theorem 3.3, $\operatorname{conv}(X^{FM}) = \operatorname{conv}(Z) \cap \{(s, x, y) : \underline{0} \leq x \leq y\}$ and the result follows. □

The above theorem shows that an external description of $\operatorname{conv}(X^{FM})$ can be obtained from an external description of $\operatorname{conv}(X^{INT})$. Such a description is already known.

PROPOSITION 3.5 (Günlük and Pochet [5]). *Consider the mixing set X^{MIX} defined in* (1.4)–(1.5). *For $t = 1, \ldots, n$ we define $f_t := b_t - \lfloor b_t \rfloor$. Let $T \subseteq N$ and suppose that $i_1, \ldots, i_{|T|}$ is an ordering of T such that $f_{i_{|T|}} \geq \cdots \geq f_{i_1} \geq f_{i_0} := 0$. Then the* mixing inequalities

$$s \geq \sum_{t=1}^{|T|} (f_{i_t} - f_{i_{t-1}})(\lfloor b_{i_t} \rfloor + 1 - y_{i_t}),$$

$$s \geq \sum_{t=1}^{|T|} (f_{i_t} - f_{i_{t-1}})(\lfloor b_{i_t} \rfloor + 1 - y_{i_t}) + (1 - f_{i_{|T|}})(\lfloor b_{i_1} \rfloor - y_{i_1})$$

are valid for X^{MIX}. Moreover, adding all mixing inequalities to the linear constraints defining X^{MIX} gives the convex hull of X^{MIX}.

PROPOSITION 3.6 (Miller and Wolsey [6]). *Let $X^{MIX}_k(n^k, s^k, y^k, b^k)$ for $1 \leq k \leq m$ be m mixing sets with some or all y variables in common. Let $X^* = \cap_{k=1}^m X^{MIX}_k$. Then*

$$\operatorname{conv}(X^*) = \bigcap_{k=1}^{m} \operatorname{conv}(X^{MIX}_k). \tag{3.5}$$

Observation 3. Günlük and Pochet [5] have shown that there is a compact formulation of the polyhedron $\operatorname{conv}(X^{MIX})$; see also [2]. Therefore it follows from Theorem 3.4 and Proposition 3.6 that a compact formulation of $\operatorname{conv}(X^{FM})$ can be obtained by writing the compact formulations of all the mixing polyhedra $\operatorname{conv}(X_k^{MIX})$, together with the inequalities $0 \leq \sigma_t - \sigma_0 + b_t \leq y_t$, $1 \leq t \leq n$, and then applying the transformation $s = \sigma_0$ and $x_t = -s + \sigma_t + b_t$, $1 \leq t \leq n$.

3.3. Variants of $\boldsymbol{X^{FM}}$. Here for the purpose of comparison we examine the convex hulls of two sets closely related to X^{FM}.

The first is the relaxation obtained by dropping the nonnegativity constraint on the flow variables x. The *unrestricted mixing set with flows* X^{UFM} is the set

$$\begin{gathered} s + x_t \geq b_t \text{ for } 1 \leq t \leq n, \\ x_t \leq y_t \text{ for } 1 \leq t \leq n, \\ s \in \mathbb{R}^1_+,\ x \in \mathbb{R}^n,\ y \in \mathbb{Z}^n_+, \end{gathered}$$

where $0 < b_1 \leq \cdots \leq b_n$, $b \in \mathbb{Q}^n$. Its convex hull turns out to be much simpler, and in fact the unrestricted mixing set with flows and the mixing set are closely related.

Proposition 3.7. *For an unrestricted mixing set with flows X^{UFM} and the mixing set X^{MIX} defined on the same vector b,*

$$\operatorname{conv}(X^{UFM}) = \{(s,x,y) :\ (s,y) \in \operatorname{conv}(X^{MIX});\ b_t - s \leq x_t \leq y_t,\ 1 \leq t \leq n\}.$$

Proof. Let $P = \{(s,x,y) :\ (s,y) \in \operatorname{conv}(X^{MIX});\ b_t - s \leq x_t \leq y_t,\ 1 \leq t \leq n\}$. The inclusion $\operatorname{conv}(X^{UFM}) \subseteq P$ is obvious. In order to show that $P \subseteq \operatorname{conv}(X^{UFM})$, we prove that the extreme rays (resp., vertices) of P are rays (resp., feasible points) of $\operatorname{conv}(X^{UFM})$.

The cone $\{(s,x,y) \in \mathbb{R}^1_+ \times \mathbb{R}^n \times \mathbb{R}^n_+ : -s \leq x_t \leq y_t,\ 1 \leq t \leq n\}$ is the recession cone of both P and $\operatorname{conv}(X^{UFM})$; thus P and $\operatorname{conv}(X^{UFM})$ have the same rays.

We now prove that if (s^*, x^*, y^*) is a vertex of P, then (s^*, x^*, y^*) belongs to $\operatorname{conv}(X^{UFM})$. It is sufficient to show that y^* is integer. We do so by proving that (s^*, y^*) is a vertex of $\operatorname{conv}(X^{MIX})$. If not, there exists a nonzero vector $(u,w) \in \mathbb{R} \times \mathbb{R}^n$ such that $(s^*, y^*) \pm (u,w) \in \operatorname{conv}(X^{MIX})$ and $w_t = -u$ whenever $y^*_t = b_t - s^*$. Define a vector $v \in \mathbb{R}^n$ as follows: If $x^*_t = b_t - s^*$, set $v_t = -u$ and if $x^*_t = y^*_t$, set $v_t = w_t$. (Since x^*_t satisfies at least one of these two equations, this assignment is indeed possible.) It is now easy to check that, for $\varepsilon > 0$ sufficiently small, $(s^*, x^*, y^*) \pm \varepsilon(u,v,w) \in P$, a contradiction. Therefore (s^*, y^*) is a vertex of $\operatorname{conv}(X^{MIX})$ and thus $(s^*, y^*) \in X^{MIX}$. Then $(s^*, x^*, y^*) \in X^{UFM}$ and the result is proved. □

The second set we consider is a restriction of the set X^{FM} in which we add simple bounds and network dual constraints on the integer variables y. Specifically, consider the following inequalities:

$$l_i \leq y_i \leq u_i, \qquad 1 \leq i \leq n, \tag{3.6}$$

$$\alpha_{ij} \leq y_i - y_j \leq \beta_{ij},\ 1 \leq i,j \leq n, \tag{3.7}$$

where $l_i, u_i, \alpha_{ij}, \beta_{ij} \in \mathbb{Z} \cup \{+\infty, -\infty\}$ and define the following set:

$$W = \{(s,x,y) \in \mathbb{R}^1 \times \mathbb{R}^n \times \mathbb{Z}^n : y \text{ satisfies (3.6)–(3.7)}\}.$$

We assume that for every index i, W contains a vector with $y_i > 0$.

Theorem 3.8.

$$\operatorname{conv}(X^{FM} \cap W) = \operatorname{conv}(X^{FM}) \cap W.$$

Proof. The proof uses the same technique as in sections 3.1–3.2, where Z (resp., X^{FM}) has to be replaced with $Z \cap W$ (resp., $X^{FM} \cap W$). We only point out the main differences.

To see that the proof of Theorem 3.3 is still valid, note that the extreme rays of $\operatorname{conv}(Z \cap W)$ are of the following types:

(i) $(1, \underline{0}, \underline{0})$ and $(0, e_k, \underline{0})$;

(ii) $(0, \underline{0}, y)$ for suitable vectors $y \in \mathbb{Z}^n$.

However, the rays of type (ii) are also rays of $\operatorname{conv}(X^{FM} \cap W)$. Also, the condition that for every index i, W contains a vector with $y_i > 0$ shows that none of the inequalities $0 \leq x_i \leq y_i$ defines an improper face of $\operatorname{conv}(X^{FM} \cap W)$ and Corollary 2.3 can still be applied. Thus the proof of Theorem 3.3 is still valid.

Finally, the following extension of (3.5) (due to Miller and Wolsey [6]) is needed: $\operatorname{conv}(X^* \cap W) = \cap_{k=1}^m \operatorname{conv}(X_k^{MIX}) \cap W$. □

Note that since the feasible region of problem (1.9) is of the type $X^{FM} \cap W$, Theorem 3.8 yields a linear inequality description of the feasible region of the two-period stochastic lot-sizing model with constant capacities.

4. An internal description of X^{FM}.

Since the extreme rays of $\operatorname{conv}(X^{FM})$ are described in Observation 1, in order to give a complete internal description of $\operatorname{conv}(X^{FM})$ we only have to characterize its vertices. These will then be used to describe a simple polynomial algorithm for optimizing over X^{FM}.

First we state a result concerning the vertices of any mixed integer set.

LEMMA 4.1. *Let $P = \{(x, y) \in \mathbb{R}^n \times \mathbb{Z}^p : Ax + By \leq c\}$. If (x^*, y^*) is a vertex of* $\operatorname{conv}(P)$*, then x^* is a vertex of the polyhedron $P(y^*) = \{x \in \mathbb{R}^n : Ax \leq c - By^*\}$.*

Proof. If x^* is not a vertex of $P(y^*)$, there exists a nonzero vector $\varepsilon \in \mathbb{R}^n$, $\varepsilon \neq \underline{0}$, such that $A(x^* \pm \varepsilon) \leq c - By^*$. But then $(x^*, y^*) \pm (\varepsilon, \underline{0})$ is in P and thus (x^*, y^*) is not a vertex of $\operatorname{conv}(P)$. □

In the following, given a point $p = (\bar{s}, \bar{x}, \bar{y})$ in $\operatorname{conv}(X^{FM})$, we denote by $f_{\bar{s}}$ the fractional part of $\bar{s}$.

CLAIM 4.2. *Let $v = (s^*, x^*, y^*)$ be a vertex of* $\operatorname{conv}(X^{FM})$*. If $s^* > 0$, there exists $j \in N$ such that $s^* + x_j^* = b_j$, $f_{s^*} = f_j$, and $s^* \leq b_j$.*

Proof. By Lemma 4.1, (s^*, x^*) is a vertex of the polyhedron $P(y^*)$ defined by

$$s + x_t \geq b_t \text{ for } 1 \leq t \leq n, \tag{4.1}$$

$$x_t \leq y_t^* \text{ for } 1 \leq t \leq n, \tag{4.2}$$

$$s \in \mathbb{R}_+^1, x \in \mathbb{R}_+^n. \tag{4.3}$$

Then among the constraints defining $P(y^*)$ there exist $n + 1$ inequalities which are tight for (s^*, x^*) and whose left-hand sides form a nonsingular $(n+1) \times (n+1)$ matrix. Therefore, if $s^* > 0$, there exists an index j such that $s^* + x_j^* = b_j$ and either $x_j^* = y_j^*$ or $x_j^* = 0$. Thus $x_j^* \in \mathbb{Z}$ and thus $f_{s^*} = f_j$. Also $x_j^* \geq 0$ implies $s^* \leq b_j$. □

CLAIM 4.3. *Let $v = (s^*, x^*, y^*)$ be a vertex of* $\operatorname{conv}(X^{FM})$*. Then for $1 \leq t \leq n$*

$$y_t^* = \max\{0, \lceil b_t - s^* \rceil\}. \tag{4.4}$$

Proof. Suppose $b_t - s^* < 0$. Then either $x_t^* = 0$ or $x_t^* = y_t^*$. Now if $y_t^* \geq 1$, in the first case both points $v \pm (0, \underline{0}, e_t)$ are in X^{FM}, and in the second case both points $v \pm (0, e_t, e_t)$ are in X^{FM}, a contradiction.

Suppose $b_t - s^* \geq 0$. If $y_t^* \geq \lceil b_t - s^* \rceil + 1$, then, setting $\varepsilon = \min\{x_t^* - (b_t - s^*), 1\}$, both points $v \pm (0, \varepsilon e_t, e_t)$ are in X^{FM}, a contradiction. □

CLAIM 4.4. *Let $v = (s^*, x^*, y^*)$ be a vertex of* $\text{conv}(X^{FM})$. *Then for $1 \leq t \leq n$*

$$x_t^* = \begin{cases} 0 & \text{if} \quad b_t - s^* < 0, \\ b_t - s^* \quad \text{or} \quad \lceil b_t - s^* \rceil & \text{if} \quad b_t - s^* \geq 0. \end{cases} \tag{4.5}$$

Proof. As (s^*, x^*) is a vertex of the polyhedron $P(y^*)$ defined by (4.1)–(4.3), it is easy to verify as in the proof of Claim 4.2 that for each t one of the following holds: either $s^* + x_t^* = b_t$ or $x_t^* = 0$ or $x_t^* = y_t^* = \max\{0, \lceil b_t - s^* \rceil\}$ (where the last equality follows from Claim 4.3). It follows that if $b_t - s^* < 0$, then $x_t^* = 0$ (otherwise inequality $x_t^* \geq 0$ would be violated), and that if $b_t - s^* \geq 0$, then $x_t^* \in \{b_t - s^*, \lceil b_t - s^* \rceil\}$ (otherwise inequality $s^* + x^* \geq b_t$ would be violated). □

Given a point $p = (\bar{s}, \bar{x}, \bar{y})$ in $\text{conv}(X^{FM})$, we define the following subsets of N:

$$N_p = \{t \in N : -1 < b_t - \bar{s} \leq 0\},$$
$$P_p = \{t \in N : 0 < b_t - \bar{s} < 1\}.$$

CLAIM 4.5. *Let $v = (s^*, x^*, y^*)$ be a vertex of* $\text{conv}(X^{FM})$. *If $s^* \geq 1$, then $N_v \cup P_v \neq \emptyset$. Moreover, if $s^* \geq 1$ and $N_v = \emptyset$, then there exists $t \in P_v$ such that $0 < x_t^* < 1$.*

Proof. Suppose $s^* \geq 1$ and $N_v \cup P_v = \emptyset$. Then $|b_t - s^*| \geq 1$, $1 \leq t \leq n$. Let $I \subseteq N$ be the set of indices t such that $b_t - s^* \geq 1$. Note that if $t \in I$, then $x_t^* \geq 1$ by Claim 4.4, and that if $t \notin I$, then $s^* + x_t^* \geq b_t + 1$. It follows that both points $v \pm (1, -e_I, -e_I)$ are in X^{FM}, a contradiction as v is a vertex of $\text{conv}(X^{FM})$.

Now suppose $s^* \geq 1$ and $N_v = \emptyset$ and assume that for every $t \in P_v$ either $x_t^* = 0$ or $x_t^* \geq 1$. Then Claim 4.4 implies that $x_t^* = 1$ for every $t \in P_v$. If $t \notin P_v$, then either $b_t - s^* \leq -1$ or $b_t - s^* \geq 1$, as $N_v = \emptyset$. Let I be the set of indices t such that $b_t - s^* \geq 1$. Note that if $t \in I$, then $x_t^* \geq 1$, and that if $t \notin P_v \cup I$, then $s^* + x_t^* \geq b_t + 1$. Thus it follows that both points $v \pm (1, -e_{P_v \cup I}, -e_{P_v \cup I})$ are in X^{FM}, again a contradiction. □

We need the following lemma.

LEMMA 4.6. *Let $p = (\bar{s}, \bar{x}, \bar{y}) \in$* $\text{conv}(X^{FM})$. *Suppose that the components of p satisfy both conditions* (4.4) *and* (4.5). *If for every convex combination of points in X^{FM} giving p, all the points appearing with nonzero coefficient have s-component equal to $\bar{s}$, then p is a vertex of* $\text{conv}(X^{FM})$.

Proof. Consider any convex combination of points in X^{FM} giving p and let C be the set of points in X^{FM} appearing with nonzero coefficient in such combination. Given $t \in N$, either $\bar{y}_t = 0$ or $\bar{y}_t = \lceil b_t - \bar{s} \rceil$. If $\bar{y}_t = 0$, then, since all points in C satisfy $y_t \geq 0$, they all satisfy $y_t = 0$. If $\bar{y}_t = \lceil b_t - \bar{s} \rceil$, then, since all points in C satisfy $y_t \geq \lceil b_t - \bar{s} \rceil$, they all satisfy $y_t = \lceil b_t - \bar{s} \rceil$. Thus all points in C have the same y-components. As to the x-components, either $\bar{x}_t = 0$ or $\bar{x}_t = b_t - \bar{s}$ or $\bar{x}_t = \lceil b_t - \bar{s} \rceil$. If $\bar{x}_t = 0$, then, since all points in C satisfy $x_t \geq 0$, they all satisfy $x_t = 0$. If $\bar{x}_t = b_t - \bar{s}$, then, since all points in C satisfy $x_t \geq b_t - \bar{s}$, they all satisfy $x_t = b_t - \bar{s}$. If $\bar{x}_t = \lceil b_t - \bar{s} \rceil$, then $\bar{x}_t = \bar{y}_t$ and so, since all points in C satisfy $x_t \leq y_t$, they all satisfy $x_t = y_t$. Thus all points in C have the same x-components. Therefore all points in C are identical. This shows that p cannot be expressed as a convex combination of points in X^{FM} distinct from p, and thus p is a vertex of $\text{conv}(X^{FM})$. □

CLAIM 4.7. *Let $p = (\bar{s}, \bar{x}, \bar{y}) \in$* $\text{conv}(X^{FM})$. *Suppose that the components of p satisfy both conditions* (4.4) *and* (4.5). *If $\bar{s} = 0$, or $\bar{s} = f_j$ for some $j \in N$, or $\bar{s} = b_j$ for some $j \in N$, then p is a vertex of* $\text{conv}(X^{FM})$.

Proof. Consider an arbitrary convex combination of points in X^{FM} giving p and let C be the set of points appearing with nonzero coefficient in such combination. Suppose $\bar{s} = 0$. Then all points in C satisfy $s = 0$. Thus, by Lemma 4.6, p is a vertex of $\text{conv}(X^{FM})$. Suppose $\bar{s} = f_j$ for some j. Condition (4.4) implies that $\bar{s} + \bar{y}_j = b_j$. Then all points in C satisfy $s + y_j = b_j$ and thus they all have $f_s = f_j$, in particular $s \geq f_j$. It follows that they all satisfy $s = f_j$. The conclusion now follows from Lemma 4.6. Suppose $\bar{s} = b_j$ for some j. Then $\bar{x}_j = 0$, and thus all points in C satisfy $x_j = 0$ and so they satisfy $s \geq b_j$. It follows that they all satisfy $s = b_j$. Again the conclusion follows from Lemma 4.6. □

CLAIM 4.8. *Let* $p = (\bar{s}, \bar{x}, \bar{y}) \in \text{conv}(X^{FM})$. *Let* $\bar{s} = m + f_j$ *for some* $j \in N$*, where* $0 < m < \lfloor b_j \rfloor$, $m \in \mathbb{Z}$. *Suppose that there exists an index* h *such that* $-1 < b_h - \bar{s} < 0$. *Suppose that the components of* p *satisfy both conditions* (4.4) *and* (4.5). *Then* p *is a vertex of* $\text{conv}(X^{FM})$.

Proof. Consider an arbitrary convex combination of points in X^{FM} giving p and let C be the set of points appearing with nonzero coefficient in such a combination. Since $b_j - \bar{s} \geq 0$ by assumption, condition (4.4) implies that $\bar{s} + \bar{y}_j = b_j$; then all points in C satisfy $s + y_j = b_j$ and thus they all have $f_s = f_j = f_{\bar{s}}$. Since $b_h - \bar{s} < 0$, Claim 4.4 implies that $\bar{x}_h = 0$; then all points in C satisfy $x_h = 0$. Suppose that there exists a point in C satisfying $s \neq \bar{s}$. Then there exists a point in C satisfying $s < \bar{s}$, i.e., $s \leq \bar{s} - 1$. Therefore, for such a point, $s + x_h = s \leq \bar{s} - 1 < b_h$, a contradiction. Thus all points in C satisfy $s = \bar{s}$. Lemma 4.6 concludes the proof. □

CLAIM 4.9. *Let* $p = (\bar{s}, \bar{x}, \bar{y}) \in \text{conv}(X^{FM})$. *Let* $\bar{s} = m + f_j$ *for some* $j \in N$*, where* $0 < m < \lfloor b_j \rfloor$, $m \in \mathbb{Z}$. *Suppose that there exists an index* h *such that* $0 < b_h - \bar{s} < 1$. *Suppose that the components of* p *satisfy both conditions* (4.4) *and* (4.5) *and that* $\bar{x}_h = b_h - \bar{s}$. *Then* p *is a vertex of* $\text{conv}(X^{FM})$.

Proof. Consider an arbitrary convex combination of points in X^{FM} giving p and let C be the set of points appearing with nonzero coefficient in such a combination. Since by assumption $b_j - \bar{s} \geq 0$, condition (4.4) implies that $\bar{s} + \bar{y}_j = b_j$; then all points in C satisfy $s + y_j = b_j$ and thus they all have $f_s = f_j = f_{\bar{s}}$. Since $\bar{s} + \bar{x}_h = b_h$, all points in C satisfy $s + x_h = b_h$. Suppose that there exists a point in C satisfying $s \neq \bar{s}$. Then there exists a point in C satisfying $s > \bar{s}$, i.e., $s \geq \bar{s} + 1$ since $f_s = f_{\bar{s}}$. Therefore, for such point, $x_h = b_h - s \leq b_h - \bar{s} - 1 < 0$, a contradiction. Thus all points in C satisfy $s = \bar{s}$. Lemma 4.6 concludes the proof. □

THEOREM 4.10. *The point* $p = (s^*, x^*, y^*)$ *is a vertex of* $\text{conv}(X^{FM})$ *if and only if its components satisfy one of the following conditions:*

(i) $s^* = 0$,

$x^*_t = b_t$ *or* $x^*_t = \lceil b_t \rceil$ *for* $1 \leq t \leq n$,

$y^*_t = \lceil b_t \rceil$ *for* $1 \leq t \leq n$;

(ii) $s^* = f_j$ *for some* $1 \leq j \leq n$,

$$x^*_t = \begin{cases} 0 & \textit{if } b_t - f_j < 0, \\ b_t - f_j \textit{ or } \lceil b_t - f_j \rceil & \textit{if } b_t - f_j \geq 0, \end{cases}$$

$y^*_t = \max\{0, \lceil b_t - f_j \rceil\}$ *for* $1 \leq t \leq n$;

(iii) $s^* = b_j$ *for some* $1 \leq j \leq n$,

$$x^*_t = \begin{cases} 0 & \textit{if } b_t - b_j < 0, \\ b_t - b_j \textit{ or } \lceil b_t - b_j \rceil & \textit{if } b_t - b_j \geq 0, \end{cases}$$

$y^*_t = \max\{0, \lceil b_t - b_j \rceil\}$ *for* $1 \leq t \leq n$;

(iv) *$s^* = m + f_j$ for some $1 \le j \le n$, where $0 < m < \lfloor b_j \rfloor$, $m \in \mathbb{Z}$, and $-1 < b_h - s^* < 0$ for some $1 \le h \le n$,*

$$x_t^* = \begin{cases} 0 & \textit{if } b_t - s^* < 0, \\ b_t - s^* \textit{ or } \lceil b_t - s^* \rceil & \textit{if } b_t - s^* \ge 0, \end{cases}$$

$y_t^* = \max\{0, \lceil b_t - s^* \rceil\}$ for $1 \le t \le n$;

(v) *$s^* = m + f_j$ for some $1 \le j \le n$, where $0 < m < \lfloor b_j \rfloor$, $m \in \mathbb{Z}$, and $0 < b_h - s^* < 1$ for some $1 \le h \le n$,*

$$x_t^* = \begin{cases} 0 & \textit{if } b_t - s^* < 0, \\ b_t - s^* \textit{ or } \lceil b_t - s^* \rceil & \textit{if } b_t - s^* \ge 0 \textit{ and } t \ne h, \\ b_t - s^* & \textit{if } t = h, \end{cases}$$

$y_t^* = \max\{0, \lceil b_t - s^* \rceil\}$ for $1 \le t \le n$.

Proof. Claim 4.7 shows that points of types (i), (ii), and (iii) are vertices of $\operatorname{conv}(X^{FM})$. Claims 4.8 and 4.9 show that points of types (iv) and (v) are vertices of $\operatorname{conv}(X^{FM})$. It remains to prove that there are no other vertices. If $p = (s^*, x^*, y^*)$ is a vertex of $\operatorname{conv}(X^{FM})$, then its components satisfy conditions (4.4) and (4.5). By Claim 4.2, either $s^* = 0$ or $f_{s^*} \in \{f_1, \dots, f_n\}$. If $s^* = 0$, p satisfies the conditions of case (i). If $s^* = f_j$ for some j, then p satisfies the conditions of case (ii). If $s^* = b_j$ for some j, then p satisfies the conditions of case (iii). Otherwise, by Claim 4.2 there exists $j \in N$ such that $f_{s^*} = f_j$ and $1 \le s^* < b_j$. Then $s^* = m + f_j$, where $0 < m < \lfloor b_j \rfloor$, $m \in \mathbb{Z}$. Claim 4.5 implies that $N_p \cup P_p \ne \emptyset$. If $N_p \ne \emptyset$, then p satisfies the conditions of case (iv). Otherwise $P_p \ne \emptyset$ and Claim 4.5 implies the existence of an index $h \in P_p$ such that $0 < x_h^* < 1$. But then necessarily $x_h^* = b_h - s^*$ and thus p satisfies the conditions of case (v). □

COROLLARY 4.11. *The problem of optimizing a rational linear function over the set X^{FM} (defined on a rational vector b) can be solved in polynomial time.*

Proof. Let $\alpha = (h, p, q) \in \mathbb{Q}^1 \times \mathbb{Q}^n \times \mathbb{Q}^n$ and consider the optimization problem

$$\min\{hs + px + qy : (s, x, y) \in X^{FM}\}. \tag{4.6}$$

Observation 1 shows that problem (4.6) is unbounded if and only if $h < 0$ or $p_t + q_t < 0$ or $q_t < 0$ for some $t \in N$. Otherwise there exists an optimal extreme point solution. Let S be the set of all possible values taken by variable s at a vertex of $\operatorname{conv}(X^{FM})$. By Theorem 4.10, $|S| = \mathcal{O}(n^2)$. For each $\bar{s} \in S$, let $V_{\bar{s}}$ be the set of vertices of $\operatorname{conv}(X^{FM})$ such that $s = \bar{s}$ and let $v_{\bar{s}}(\alpha)$ be an optimal solution of the problem

$$\min\{hs + px + qy : (s, x, y) \in V_{\bar{s}}\}.$$

The components of $v_{\bar{s}}(\alpha)$ satisfy $s = \bar{s}$, $y_t = \max\{0, \lceil b_t - \bar{s} \rceil\}$ for $1 \le t \le n$ and

$$x_t^* = \begin{cases} 0 & \text{if } b_t - \bar{s} < 0, \\ b_t - \bar{s} & \text{if } b_t - \bar{s} \ge 0 \text{ and } p_t \ge 0, \\ \lceil b_t - \bar{s} \rceil & \text{if } b_t - \bar{s} \ge 0 \text{ and } p_t < 0 \end{cases}$$

if the value $s = \bar{s}$ corresponds to one of cases (i)–(iv), and similarly for case (v).

Since solving problem (4.6) is equivalent to solving the problem $\min\{\alpha v_{\bar{s}}(\alpha) : \bar{s} \in S\}$, we only need to compute the objective function in $\mathcal{O}(n^2)$ points. This requires $\mathcal{O}(n^3)$ time. □

5. Concluding remarks. Several other generalizations of the mixing set appear to be interesting, some of which are already being investigated.

A common generalization of the set studied in this paper and the continuous mixing set [6, 9] is the *continuous mixing set with flows*:

$$X^{CFM} = \{(s, r, x, y) \in \mathbb{R}^1_+ \times \mathbb{R}^n_+ \times \mathbb{R}^n_+ \times \mathbb{Z}^n_+ : s + r_t + x_t \geq b_t,\ x_t \leq y_t,\ 1 \leq t \leq n\}.$$

Though a compact extended formulation of this set has been found recently [1], the question of finding an inequality description in the original space of variables is still open.

The *mixing-MIR set with divisible capacities*

$$X^{MMIX} = \{(s, y) \in \mathbb{R}^1_+ \times \mathbb{Z}^n : s + C_t y_t \geq b_t\},$$

where $C_1|C_2|\cdots|C_n$, has been studied by de Farias and Zhao [3]. An interesting question is to give a polyhedral description of conv(X^{MMIX}). The special case when the C_i only take two distinct values has been treated in Van Vyve [8].

Another intriguing question is the complexity status of the problem of optimizing a linear function over the *divisible mixing set*

$$X^{DMIX} = \left\{ (s, y) \in \mathbb{R}^1_+ \times \mathbb{Z}^{mn}_+ : s + \sum_{j=1}^{m} C_j y_{jt} \geq b_t \right\},$$

with again $C_1|C_2|\cdots|C_n$. For the case $m = 2$, a compact extended formulation of conv(X^{DMIX}) is given in Conforti and Wolsey [2].

REFERENCES

[1] M. Conforti, M. Di Summa, and L. A. Wolsey, *The intersection of continuous mixing polyhedra and the continuous mixing polyhedron with flows*, Conference on Integer Programming and Combinatorial Optimization (IPCO 2007), Cornell University, Ithaca, NY, 2007.

[2] M. Conforti and L. A. Wolsey, *Compact formulations as unions of polyhedra*, Math. Program., to appear.

[3] I. de Farias and M. Zhao, *The Mixing-MIR Set with Divisible Capacities*, Report, University of Buffalo, Buffalo, NY, 2005 (revised 2006).

[4] Y. Guan, S. Ahmed, A. J. Miller, and G. L. Nemhauser, *On formulations of the stochastic uncapacitated lot-sizing problem*, Oper. Res. Lett., 34 (2006), pp. 241–250.

[5] O. Günlük and Y. Pochet, *Mixing mixed integer inequalities*, Math. Program., 90 (2001), pp. 429–457.

[6] A. Miller and L. A. Wolsey, *Tight formulations for some simple MIPs and convex objective IPs*, Math. Program., 98 (2003), pp. 73–88.

[7] Y. Pochet and L. A. Wolsey, *Lot-sizing with constant batches: Formulation and valid inequalities*, Math. Oper. Res., 18 (1993), pp. 767–785.

[8] M. Van Vyve, *A Solution Approach of Production Planning Problems Based on Compact Formulations for Single-Item Lot-Sizing Models*, Ph.D. thesis, Faculté des Sciences appliquées, Univiersité Catholique de Louvain, Belgium, 2003.

[9] M. Van Vyve, *The continuous mixing polyhedron*, Math. Oper. Res., 30 (2005), pp. 441–452.

SIAM J. DISCRETE MATH.
Vol. 21, No. 2, pp. 408–422

RANDOM 2-SAT DOES NOT DEPEND ON A GIANT*

DAVID KRAVITZ†

Abstract. Here we introduce a new model for random 2-SAT. It is well known that on the standard model there is a sharp phase transition; the probability of satisfiability quickly drops as the number of clauses exceeds the number of variables. The location of this phase transition suggests that there is a direct connection between the appearance of a giant in the corresponding $2n$-vertex graph and satisfiability. Here we show that the giant has nothing to do with satisfiability and that in fact the expected degree of a randomly chosen vertex is the important thing.

Key words. satisfiability, SAT, random processes, Boolean, clauses, literals, online, offline, algorithms

AMS subject classifications. 60C05, 60J85, 82B26, 05C80, 05C90, 68R10

DOI. 10.1137/060662216

1. Introduction.

Let $\{x_1, x_2, \dots, x_n\}$ be a set of n Boolean variables. The corresponding set of literals is

$$\mathbf{X} := \{x_1, \overline{x}_1, \dots, x_n, \overline{x}_n\}.$$

A *k-clause* is a set of k literals from $\mathbf{X}$. We say a clause is *satisfied* by an assignment of the variables if and only if at least one of its literals is true. The model of RANDOM k-SAT takes a family of k-clauses, chosen at random, and asks if there is an assignment to the Boolean variables for which every clause in the family is satisfied. We are interested in what happens as $n \to \infty$.

NOTATION 1. *For any $n, m,$ and k, let $F_k(n, m)$ denote a set of m random k-clauses, where each k-clause is chosen uniformly at random from the set of all $\binom{kn}{k}$ possible k-clauses.*

We consider random 2-SAT. While it appears that the structure of the corresponding graph, in particular the appearance of a giant component in this graph, has a lot to do with satisfiability, we present results that indicate this is not the case.

Random 2-SAT is well understood. (See [6, 10] for a survey of known results.) The following were proven by Chvátal and Reed in [8] and Goerdt in [11] for any fixed constant $\epsilon > 0$:

1. $F_2(n, (1 - \epsilon)n)$ is unsatisfiable whp.[1]
2. $F_2(n, (1 + \epsilon)n)$ is satisfiable whp.

There have also been several other results which strengthened this to the case where $\epsilon = o(1)$ (see [4, 15] and others), but from now on we will assume $\epsilon > 0$ is a constant.

In [8], Chvátal and Reed define a *bicycle* as a formula with at least two distinct variables $x_1, \dots, x_s$ and clauses $C_0, C_1, \dots, C_s$ that have the following structure: There are literals $w_1, \dots, w_s$ such that each w_r is either x_r or $\overline{x}_r$, each C_r with $0 < r < s$ is $\{\overline{w}_r, w_{r+1}\}$, and $C_0 = \{u, w_1\}, C_s = \{\overline{w}_s, v\}$ with literals u, v chosen

*Received by the editors June 6, 2006; accepted for publication (in revised form) January 18, 2007; published electronically May 10, 2007.
http://www.siam.org/journals/sidma/21-2/66221.html

†National Security Agency, 9800 Savage Road, Fort Meade, MD (kravitzd@comcast.net). The work of this author was supported in part by a VIGRE grant from the NSF, DMS-9819950.

[1] An event E happens with high probability, or whp, if $\Pr(E) = 1 - 0_n(1)$.

from $\{x_1, \ldots, x_s, \overline{x}_1, \ldots, \overline{x}_s\}$. They prove that every unsatisfiable family of 2-clauses contains a bicycle.

Each family of clauses F is easily seen to correspond to a graph G_F on $2n$ vertices, where each vertex of G_F corresponds to a literal in F and each edge corresponds to a clause. It is well known (see [9, 5, 12] and many others) that G_F undergoes a major change right when the number of clauses exceeds n. When F has $(1-\epsilon)n$ clauses, the largest connected component of G_F has $O(\log n)$ vertices, and all components are either trees or unicyclic, making a bicycle extremely unlikely. However, when there are $(1+\epsilon)n$ clauses, a *giant component* of size $\Omega(n)$ appears; this component contains a lot of cycles and has a substantial 2-core.

It is very reasonable to think that the appearance of this complex component has something to do with the first appearance of at least one bicycle, and therefore the change in satisfiability, but here we introduce a natural random model in which there is no connection between the appearance of a giant in G_F and satisfiability.

Model. Given any simple graph G on $2n$ vertices, we will make a family of clauses $S(G)$ by randomly assigning labels from $\mathbf{X}$ to the vertices; then each edge corresponds to one clause.

This model is equivalent to a random instance of 2-SAT where one condition on G is the underlying subgraph. This is the most natural way of examining the question of how the structure of the underlying subgraph can affect the probability of satisfiability.

We would like to know the probability that $S(G)$ is satisfiable over the space of all possible assignments to the vertices. This question is equivalent to the one with $F_2(n, m)$ if G is a random graph with m edges; however, we allow G to be *anything* (provided $\Delta(G)$ isn't extremely large). This model does allow clauses $x_i \vee \overline{x}_i$, which are usually excluded in 2-SAT; however, whp we will have $O(1)$ such clauses, which makes no difference in our results.

Note that $S(G)$ is satisfiable if and only if there are exactly n vertices in G which cover $E(G) \cup M$, where M is a random perfect matching added to G. We must take exactly one vertex from each edge in M for an edge cover of size n, and these n vertices must cover every edge in G. Vertices in the edge cover are "true," while vertices out of the edge cover are "false." We will primarily use this model; in most cases we will expose one matching edge at a time by matching a given vertex with a randomly chosen unmatched vertex.

Roughly speaking, we will show that when the number of edges in G is cn, $c = 1$ is a threshold for satisfiability, subject to a few technical conditions. This indicates that there is no connection between a giant component in G and satisfiability; for example, G could be a graph with $(1-\epsilon)$ edges all in one giant component or $(1+\epsilon)$ edges in many small components.

Theorem 1. *If G is a graph with $2n$ vertices, less than $(1-\epsilon)n$ edges for some $\epsilon > 0$, and $\Delta(G) = o(\frac{n^{1/10}}{\log n})$, then $S(G)$ is satisfiable whp.*

This can be thought of as an extension of the result from Chvátal and Reed stated above; in that case G would be a random graph with $2n$ vertices and up to $(1-\epsilon)n$ edges. The necessity of a condition on $\Delta(G)$ is discussed in section 5.

Our result in the case when there are $(1+\epsilon)n$ edges requires an additional condition, namely, that enough of the edges come from vertices of degree less than $O(\log n)$.

Notation 2. *For all $i \geq 0$, define $d_i = d_i(G)$ as the number of vertices of degree i in graph G.*

Theorem 2. *If G is a graph with $2n$ vertices and $\Delta(G) = o(n^{1/8})$, and there is some $\epsilon > 0$ and function $\tau \leq c \log n$ for some constant $c < \frac{3\epsilon}{16}$ such that*

$$\sum_{i=0}^{\tau} i d_i = (1+\epsilon)2n, \tag{1}$$

then $S(G)$ is not satisfiable whp.

This is also an extension of the Chvátal and Reed result because a random graph with $(1+\epsilon)n$ edges will whp satisfy (1) with τ equal to some sufficiently large constant. Theorems 1 and 2 are proven in section 3.

If there is a collection of high-degree vertices incident with more than ϵn edges, the structure of the graph is much more important. However, we do believe the following to be true.

Conjecture 1. *Let $\epsilon > 0$. There exists $\phi > 0$ such that if G is a graph with $2n$ vertices and more than $(1+\epsilon)n$ edges, and $\Delta(G) \leq n^{\phi}$, then $S(G)$ is not satisfiable whp.*

Theorem 2 implies that Conjecture 1 holds when n^{ϕ} is replaced by $c \log n$. In section 6 we discuss some results that lead us to believe Conjecture 1 is true, and in section 5 we show that Conjecture 1 does not hold for $\phi \geq 0.5$.

1.1. Inequalities.

We will make use of an Azuma–Hoeffding type of inequality for supermartingales as discussed in [13, 16, 3]: If $Y_0, Y_1, Y_2, \ldots, Y_t$ is a sequence of random variables such that $E[Y_i | Y_1, Y_2, \ldots, Y_{i-1}] \leq Y_{i-1}$ and $|Y_i - Y_{i-1}| \leq \lambda$ for some constant λ and all $i \leq t$, then for all $\alpha > 0$

$$\Pr(Y_t - Y_0 \geq \alpha) \ \leq \ \exp\left(-\frac{\alpha^2}{2t\lambda^2}\right). \tag{2}$$

The following are easily obtained from the Azuma–Hoeffding inequality: If $\{X_i\}_{i \geq 0}$ is a sequence of random variables such that all differences $X_{k+1} - X_k$ are independent and $|X_{k+1} - X_k| \leq z$ for all $k \geq 0$, then

$$\Pr(X_k - E[X_k] \geq \lambda) \quad \leq \quad \exp\left(-\frac{\lambda^2}{8kz^2}\right) \tag{3}$$

and

$$\Pr(E[X_k] - X_k \geq \lambda) \quad \leq \quad \exp\left(-\frac{\lambda^2}{8kz^2}\right) \tag{4}$$

for all $\lambda > 0$.

2. When d_0 is small.

The proof of Theorem 2 will use the following. If G has few isolated vertices, then it is not satisfiable provided at least some ratio of the vertices has degree 2 or more.

Theorem 3. *If G is a $2n$-vertex graph such that*

$$\sum_{i \geq 2} d_i \quad \gg \quad n^{7/8}\Delta^{1/2} + n^{1/2} d_0^{1/2}, \tag{5}$$

then $S(G)$ is not satisfiable whp.[2]

Note that in this case $\Delta = o(n^{1/4})$ and $d_0 = o(n)$ are implied since $\sum_{i \geq 0} d_i = 2n$.

[2] We say that $f(n) \gg g(n)$ if $\frac{g(n)}{f(n)} = o_n(1)$.

Proof of Theorem 3. Suppose that G is any graph with $2n$ vertices. Begin by iteratively removing any edges which join two vertices of degree at least 3. Note that this doesn't change n or (5), and when finished it will allow us to say that G satisfies:

(a) Every edge in G is incident with at least one vertex of degree 1 or 2.

Now define functions $\alpha(n)$ and $\mu(n)$ such that $\alpha(n) \to \infty$ as $n \to \infty$, and the following are true:

(b) $\Delta(G) \leq \frac{n^{1/4}}{\alpha^2(n)}$;

(c) $d_0 \leq \frac{n}{\alpha^2(n)}$;

(d) Either (b) or (c) is satisfied with equality;

(e) $d_1 = 2n(1-\mu(n))$;

(f) $\alpha(n)\mu(n) \to \infty$ as $n \to \infty$.

Existence of $\alpha(n)$ is clear from the conditions of Theorem 3, and (e) defines $\mu(n)$. To show that (5) also implies (f), note that

$$\alpha\mu \quad \geq \quad \frac{\alpha}{2n}\sum_{i\geq 2} d_i \quad \gg \quad \alpha n^{-1/8}\Delta^{1/2} + \alpha n^{-1/2} d_0^{1/2} \quad \geq \quad 1,$$

with the last inequality coming from (d).

To help with technical details, we will define a vertex as *free* if at least one of its neighbors was set or if it is isolated.

First, we will pick any nonisolated vertex v_0 from G. Start by setting v_0 false; we are going to prove that whp this will lead to a contradiction. To do this, we are going to expose the matching of G one edge at a time and simultaneously keep track of the following three sets:

- T is the set of "active" true vertices, vertices which must be true but are not yet matched. Our contradiction will be a matching edge within T. Initially $T = N(v_0)$ since v_0 is false, and $T \neq \emptyset$ when we start because v_0 is a nonisolated vertex.
- U is the set of all unmatched vertices which are "free." Initially U will be the set of all isolated vertices along with $N(N(v_0))$.
- V is the set of all other unmatched vertices not in $T \cup U$. Initially $V = \mathbf{X}\backslash T\backslash U$.

NOTATION 3. *For any vertex v, we will write $N_2(v) = N(N(v)) - v$.*

So $T \cup U \cup V$ is the set of currently unmatched vertices. As long as $T \neq \emptyset$ we are going to select $v \in T$ and match it with a randomly chosen unmatched vertex $\overline{v}$. Then $N(\overline{v})$ must be true so it goes to T, and $N_2(\overline{v})$ will be declared free. It seems unnatural to not set vertices of $N(\overline{v}_i)$ to be true if they are in U or if they are already matched, but we will show that this is an unimportant detail because the number of such vertices is negligible.

This is the precise algorithm we will follow.

```
1. Start with $i = 0$ and initial sets $T_0, U_0, V_0$ described above.
2. While $T_i \neq \emptyset$ and $i \leq \alpha(n)\sqrt{n}$:
   Pick any vertex $v_i \in T_i$ and match it with a random vertex $\overline{v}_i \in T_i \cup U_i \cup V_i - v_i$.
   Then update $T, U, V$ as follows:
     • If $\overline{v}_i \in T_i$, then STOP; we have our contradiction;
     • If $\overline{v}_i \in U_i$, then $T_{i+1} = T_i - v_i$, $U_{i+1} = U_i - \overline{v}_i \cup N(\overline{v}_i)$;
     • If $\overline{v}_i \in V_i$, then $T_{i+1} = T_i \cup N(\overline{v}_i) - v_i$, $U_{i+1} = U_i \cup N_2(\overline{v}_i)$, $V_{i+1} =$
       $V_i \setminus N_2(\overline{v}_i) \setminus N(\overline{v}_i) - \overline{v}_i$.
     • $i = i + 1$.
3. STOP. (Note that either $T_i = \emptyset$ or $i \geq \alpha(n)\sqrt{n}$.)
```

Note that in this algorithm the graph we work with at step i is the graph induced by $U_i \cup V_i$.

We note some bounds on $|U_i|$ and $|V_i|$ in the course of the algorithm. For any vertex u, we have $N_2(u) \leq 2\Delta$ from (a). Therefore, $|U_{i+1}| - |U_i| \leq 2\Delta$ for all i, and since $i \leq \alpha(n)\sqrt{n}$ through our process, we have

$$|U_i| \;\leq\; 2i\Delta + |U_0| \;\leq\; \frac{2n^{3/4}}{\alpha(n)} + \frac{n}{\alpha(n)^2} \;=\; O\left(\frac{n}{\alpha^2(n)}\right),$$

Similarly, $|V_i| - |V_{i+1}| \leq 3\Delta$ for all i and $|V_0| \geq 2n - 2\Delta$; therefore, for all i

$$|V_i| \;\geq\; 2n - 3(i+1)\Delta \;\geq\; 2n - o(n^{3/4}).$$

Now we look at $|T_i|$. We have $|T_{i+1}| < |T_i|$ only if $\overline{v}_i \in U$, and in this case $|T_{i+1}| = |T_i| - 1$. We have

$$\Pr\left(|T_{i+1}| < |T_i|\right) \;\leq\; \frac{|U_i|}{|T_i \cup V_i \cup U_i| - 1} \;\leq\; \frac{O\left(\frac{n}{\alpha^2(n)}\right)}{2n - o(n^{3/4})} \;=\; O\left(\frac{1}{\alpha^2(n)}\right).$$

Now if $\overline{v}_i \in V_i$, then $|T_{i+1}| - |T_i| = |N(\overline{v}_i)| - 1$; therefore, we increase $|T_i|$ if $\deg(\overline{v}_i) > 1$. Define

$$p_L \;=\; \max_i \Pr(\deg(\overline{v}_i) = 1 \mid \overline{v}_i \in V_i).$$

The number of degree 1 vertices in V never increases through the process, because any vertex which loses an edge is immediately free; therefore, if a vertex of degree 1 is created, it would move from V to U. Thus, we have

$$p_L \;\leq\; \frac{d_1}{\min_i |V_i|} \;\leq\; \frac{2n(1-\mu(n))}{2n - o(n^{3/4})} \;=\; 1 - \mu(n) + o(n^{-1/4}).$$

So

$$\begin{aligned}\Pr(|T_{i+1}| - |T_i| \geq 1) \;&\geq\; (1-p_L)\frac{|V_i|}{|T_i \cup V_i \cup U_i|} \;\geq\; \left[\mu(n) - o(n^{-1/4})\right]\frac{2n - o(n^{3/4})}{2n} \\ &\geq\; \mu(n) - o(n^{-1/4}).\end{aligned}$$

Lemma 1. *With high probability*
(i) $|T_i| \neq 0$ *for all* $i \leq \alpha(n)\sqrt{n}$*;*
(ii) *if* $j = \lfloor \sqrt{n}\alpha(n) \rfloor$*, then* $|T_j| \geq \frac{\mu(n)}{2} j$*.*

We prove this below, but for now assume it is true. So whp our algorithm will end either with $\overline{v}_i \in T_i$ or with $i > \alpha(n)\sqrt{n}$, not with $T_i = \emptyset$. If it ends with $\overline{v}_i \in T_i$, we are done; if not, then Lemma 1 implies that whp we will finish with $|T| \geq \frac{\mu(n)\alpha(n)}{2}\sqrt{n}$. In this case it is extremely likely that a matching edge will occur within T; the probability of no such edge can be bounded above by

$$\begin{aligned}\prod_{i=1}^{|T|}\left(1 - \frac{|T| - i}{2n}\right) \;&\leq\; \exp\left(-\frac{1}{2n}\sum_{i=1}^{|T|-1} |T| - i\right) \;=\; \exp\left(-\Omega\left(\frac{|T|^2}{n}\right)\right) \\ &\leq \exp\left(-\Omega(\mu(n)^2\alpha(n)^2)\right) = o(1).\end{aligned}$$

Thus, from Lemma 1 we can say that whp we will have a matching edge within T; therefore, we have our contradiction.

We have

$$\begin{aligned}\Pr(S(G)\text{ satisfiable}) \;\le\;& \;\Pr(\exists\text{ satisfying assignment with }v_0\text{ false})\\ &+\;\Pr(\exists\text{ satisfying assignment with }\overline{v}_0\text{ false});\end{aligned}$$

therefore,

$$\begin{aligned}\Pr(S(G)\text{ satisfiable}) \;\le\;& \;\Pr(\exists\text{ satisfying assignment with }v_0\text{ false})\\ &+\;\Pr(\overline{v}_0\text{ is isolated})\\ &+\;\Pr(\exists\text{ satis. assignment with }\overline{v}_0\text{ false and not isol.}).\end{aligned}$$

The first and third summands on the right-hand side are $o(1)$ because of our contradiction, and the second is $\frac{1}{n}o(n) = o(1)$ because there are only $o(n)$ isolated vertices in G. Thus, $\Pr(S(G)\text{ is satisfiable}) = o(1)$. □

Proof of Lemma 1. We first note that $\{|T_i|\}_{i\ge 0}$ can be thought of as a series of random variables whose differences aren't quite independent, but clearly there is a series $\{X_i\}_{i\ge 0}$ of random variables such that $X_{i+1} - X_i$ are independent for all $i \ge 0$, and the following are all true:

1. $\{|T_i|\}_{i\ge 0}$ majorizes $\{X_i\}_{i\ge 0}$; i.e., $X_i \le |T_i|$ for all $i \ge 0$.
2. $X_0 = |T_0| \ge 1$ because we chose a nonisolated vertex to start.
3. $\Delta \ge X_{i+1} - X_i \ge -1$ for all $i \ge 0$.
4. $\Pr(X_{i+1} < X_i) = O\left(\frac{1}{\alpha^2(n)}\right)$.
5. $\Pr(X_{i+1} \ge X_i + 1) = (1 - o(1))\mu(n)$.

Let P_1 be the probability that $X_i = 0$ for some $i \le \alpha(n)\sqrt{n}$. Furthermore, define $p_< = \Pr(X_1 < X_0)$ and $p_> = \Pr(X_1 > X_0)$. A simple recursion gives us

$$P_1 \;\le\; p_< + (1 - p_< - p_>)P_1 + p_> P_1^2,$$

which leads to

$$0 \;\le\; (p_< - p_> P_1)(1 - P_1).$$

Certainly $P_1 < 1$; therefore,

$$P_1 \;\le\; \frac{p_<}{p_>} \;\le\; \frac{O\left(\frac{1}{\alpha^2(n)}\right)}{(1 - o(1))\mu(n)} \;=\; O\left(\frac{1}{\alpha(n)^2\mu(n)}\right) \;=\; o(1).$$

Now define P_2 as the probability that (i) is true and (ii) is false. Since

$$E\left[X_{i+1} - X_i\right] \;\ge\; (1 - o(1))\mu(n) - O\left(\frac{1}{\alpha(n)^2}\right) \;=\; (1 - o(1))\mu(n)$$

for all $i \le j$, we have $E[X_j] \ge (1 - o(1))\mu(n)j$. So

$$P_2 \;\le\; \Pr\left(E[X_j] - |X_j| \ge \frac{\mu(n)}{3}j\right) \;=\; \Pr\left(|X_j| - E[X_j] \le -\frac{\mu(n)}{3}j\right).$$

Condition 3 above allows us to use (4):

$$P_2 \;\le\; \exp\left(-\frac{\mu(n)^2 j}{72\Delta^2}\right) \;=\; \exp\left(-\Omega(\alpha(n)^5\mu(n)^2)\right) \;=\; o(1),$$

with the last equality following from (f) and the fact that $\alpha(n) \to \infty$. Since $P_1 + P_2 = o(1)$, we know that (i) and (ii) are true whp. □

3. Proof of Theorem 2. Suppose that G is a graph with $2n$ vertices and $\Delta(G) = o(n^{1/8})$. Also, assume there is some $\epsilon > 0$ and some function $\tau \le c\log n$ for some constant $c < \frac{3\epsilon}{16}$ such that

$$\sum_{i=0}^{\tau} id_i = (1+\epsilon)2n.$$

Notation 4. *For any number x, we will write $x^+ = x + o(1)$.*

Let δ be some small positive function satisfying

$$\exp\left(-\tau[\tfrac{2^+}{\epsilon} + \phi]\right) > \ \delta \ > \ n^{-3/8+\phi}$$

for some $\phi > 0$, a fixed constant; we know such a δ exists because of our assumption on τ.

If v is an isolated vertex in G, then any optimal assignment algorithm can set v to be false and $\overline{v}$ to be true. This defines a procedure which is commonly called *pure literal elimination.* We are going to do pure literal elimination on G and show that whp it leads to a graph which is not satisfiable whp by Theorem 3.

Notation 5. *We will write d_i as a function of s, since it will change throughout the process.*

(a) Set $s = 0$.

(b) While $d_0(s) > 0$ and $s < (1-\delta)n$:
Step s: Choose any isolated vertex v, and then randomly choose its match $\overline{v}$ from all other vertices. Make v false and $\overline{v}$ true, and then delete both vertices from the graph, along with any edges incident with $\overline{v}$.
Increment s by 1.

First, we will show that the ratio between the number of edges and the number of vertices is likely not to decrease too much. Define

$$D_s^T := \sum_{i=0}^{T} id_i(s)$$

for any integer $T \leq \tau$, and

$$V_s := \sum_{i \geq 0} d_i(s) - 1.$$

Note that at any time $V_s = 2n - 2s - 1$. This will be the size of the "pool" of vertices that we have to choose from for $\overline{v}$.

Furthermore, define s_1 to be the step when the above process stops.

Notation 6. *Let $\overline{d}(s) = \{d_0(s), d_1(s), d_2(s), \dots\}$ be the entire degree sequence at Step s.*

Lemma 2. *For any $i \geq 0$ and $s < s_1$, we have*

$$V_s E[d_i(s+1) - d_i(s) \mid \overline{d}(s)] = (i+1)\left(d_{i+1}(s) - d_i(s)\right) - V_s \mathbf{1}_{i=0}.$$

Lemma 3. *With high probability, for all $s < (1-\delta)n$ and $s < s_1$, we have*

$$\sum_{i=2}^{\tau} d_i(s) \quad \geq \quad \frac{\epsilon V_s}{1^+\tau}. \tag{6}$$

Let s_2 denote the first Step s in which (6) does not hold, if such a step exists, and let $\overline{s} = \min\{s_1, s_2\}$ (if s_2 does not exist, then $\overline{s} = s_1$). We will continue our process beyond $s = \overline{s}$ for the sake of defining a martingale, but the graph (and hence the degree sequence) will not change after this point.

LEMMA 4. *With high probability, $\overline{s} < (1-\delta)n$.*

Lemmas 3 and 4 show that whp either we will stop because (6) does not hold or there is some number $\overline{s} < (1-\delta)n$ such that $\overline{s}$ steps of pure literal elimination will lead to a graph with $V_{\overline{s}} \geq 2\delta n$ vertices, $d_0 = 0$, $\Delta = o(n^{1/8})$, and $\sum_{i\geq 2} d_i \geq \Omega(\frac{V_{\overline{s}}}{\tau})$. Theorem 3 shows this is not satisfiable whp whenever n is sufficiently large with respect to $\delta = \delta(T, \epsilon)$ and $\delta \geq n^{-1/2}$. It remains only to prove the lemmas.

Proof of Lemma 2. First, fix any $i \geq 1$. To make notation easier, let S_i, S_{i+1} be the set of all vertices of degree $i, i+1$, respectively, and let w_i, w_{i+1} be arbitrary vertices in their respective sets. We have

$$E[d_i(s+1) - d_i(s) \mid \overline{d}(s)] = E[|S_i(s+1) \setminus S_i(s)| \mid \overline{d}(s)] - E[|S_i(s) \setminus S_i(s+1)| \mid \overline{d}(s)].$$

Choose an arbitrary $w_i \in S_i(s)$. We have

$$\Pr(w_i \in S_i(s) \setminus S_i(s+1)) \quad = \quad \Pr(\overline{v} = w_i \text{ or } \overline{v} \in N(w_i)) \quad = \quad \frac{i+1}{V_s}.$$

Thus,

$$E[S_i(s) \setminus S_i(s+1) \mid \overline{d}(s)] \quad = \quad |S_i|\left(\frac{i+1}{V_s}\right) \quad = \quad d_i\left(\frac{i+1}{V_s}\right).$$

Now the only way a vertex is in $S_i(s+1) \setminus S_i(s)$ is if it had degree $i+1$ and it lost a neighbor. Therefore,

$$\Pr(w_{i+1} \in S_i(s+1) \setminus S_i(s)) \quad = \quad \Pr(\overline{v} \in N(w_{i+1})) \quad = \quad \frac{|N(w_{i+1})|}{V_s} \quad = \quad \frac{i+1}{V_s}.$$

Thus,

$$E[S_i(s+1) \setminus S_i(s) \mid \overline{d}(s)] \quad = \quad |S_{i+1}|\left(\frac{i+1}{V_s}\right) \quad = \quad d_{i+1}\left(\frac{i+1}{V_s}\right).$$

When $i = 0$, the only difference is that $E[S_0(s+1) \setminus S_0(s)]$ is one less because pure literal elimination randomly matches a degree 0 vertex. □

Proof of Lemma 3. We will examine the series of variables $\{\frac{D_i^\tau}{V_i}\}_{i\geq 0}$. First, we bound the expected change. Lemma 2 gives

$$V_s E[D_{s+1}^\tau - D_s^\tau \mid \overline{d}(s)] \quad = \quad \sum_{i=1}^{\tau} i(i+1)\left(d_{i+1}(s) - d_i(s)\right)$$

for all s; therefore,

$$V_s E[D_{s+1}^\tau - D_s^\tau \mid \overline{d}(s)] \quad = \quad -2\sum_{i=1}^{\tau} i d_i(s) + \tau(\tau+1) d_{\tau+1}(s) \quad \geq \quad -2D_s^\tau,$$

because $d_{\tau+1} \geq 0$. Since V_s is known and $V_{s+1} = V_s - 2$, we use this to get

$$\begin{aligned} E\left[\frac{D_{s+1}^\tau}{V_{s+1}} - \frac{D_s^\tau}{V_s} \mid \overline{d}(s)\right] \quad &= \quad \frac{E[D_{s+1}^\tau V_s - D_s^\tau (V_s - 2) \mid \overline{d}(s)]}{V_s(V_s-2)} \\ &= \quad \frac{V_s E[D_{s+1}^\tau - D_s^\tau \mid \overline{d}(s)] + 2D_s^\tau}{V_s(V_s-2)} \quad \geq \quad 0 \end{aligned}$$

for all s during our process. So for all s we have

$$E\left[\frac{D_s^\tau}{V_s}\right] \quad\geq\quad \frac{D_0^\tau}{V_0} \quad=\quad 1+\epsilon. \tag{7}$$

Now we bound the actual difference. Each step deletes at most one nonisolated vertex; therefore, $|D_{s+1}^\tau - D_s^\tau| \leq 2\Delta$. Furthermore, $D_s^\tau \leq V_s\Delta$.

$$\left|\frac{D_{s+1}^\tau}{V_{s+1}} - \frac{D_s^\tau}{V_s}\right| \quad\leq\quad \frac{\left|D_{s+1}^\tau - D_s^\tau\right|}{V_s-2} + \frac{2D_s^\tau}{V_s(V_s-2)} \quad\leq\quad \frac{2\Delta + \frac{D_s^\tau}{V_s}}{V_s-2} \quad\leq\quad \frac{3\Delta}{V_s-2}$$

for all s.

Define $\beta(n) = n^{-\phi/2}$ to be a small positive function, and fix any $s < (1-\delta)n$. We use the above with (7), (2), $s < n$, and $V_s \geq 2\delta n$ to get

$$\begin{aligned}\Pr\left(\frac{D_s^\tau}{V_s} \leq (1+\epsilon) - \beta(n)\right) \quad&\leq\quad \exp\left(-\frac{1}{2s}\left[\frac{\beta(n)V_s}{3\Delta}\right]^2\right) \quad\leq\quad \exp\left(-\frac{1}{2n}\left[\frac{n^{-\phi/2}2\delta n}{3n^{1/8}}\right]^2\right)\\ &=\quad \exp\left(-\tfrac{2}{9}n^{3/4-\phi}\delta^2\right) \quad<\quad \exp\left(-\tfrac{2}{9}n^{\phi}\right).\end{aligned}$$

So the probability that this is true for any $s < (1-\delta)n$ can be bounded from above by $n\exp(-\frac{2}{9}n^\phi) = o(1)$. Therefore, whp we have

$$0 \quad\leq\quad D_s^\tau - (1+\epsilon)V_s + \beta(n)V_s \quad=\quad D_s^\tau - V_s - (1-o(1))\epsilon V_s$$

whp for all $s < (1-\delta)n$; this yields

$$\tau\sum_{i=2}^{\tau} d_i(s) \quad\geq\quad \sum_{i=2}^{\tau}(i-1)d_i(s) \quad\geq\quad D_s^\tau - V_s \quad\geq\quad (1-o(1))\epsilon V_s \quad=\quad \frac{\epsilon V_s}{1^+}. \qquad \square$$

Proof of Lemma 4. We will examine the random variables $\{\frac{d_0(i)}{V_i}\}_{i\geq 0}$. First, we bound the difference for all s. Here we use $V_{s+1} = V_s - 2$ and $|d_0(s+1) - d_0(s)| \leq \Delta$.

$$\left|\frac{d_0(s+1)}{V_{s+1}} - \frac{d_0(s)}{V_s}\right| \quad=\quad \left|\frac{d_0(s+1)-d_0(s)}{V_s-2} + \frac{2d_0(s)}{V_s(V_s-2)}\right| \quad=\quad O\left(\frac{\Delta}{V_s}\right).$$

Now we look at the expected change. First using Lemma 2:

$$E\left[\frac{d_0(s+1)}{V_{s+1}} - \frac{d_0(s)}{V_s} \mid \overline{d}(s)\right] \quad=\quad \frac{d_1(s)+d_0(s)-V_s}{V_s(V_s-2)} \quad\leq\quad \frac{-\sum_{i=2}^{\tau} d_i(s)}{V_s(V_s-2)}.$$

Now we can use Lemma 3 and $V_s = 2(n-s)-1$ (for $s < \overline{s}$) to say that whp either $\overline{s} < s$ or

$$E\left[\frac{d_0(s+1)}{V_{s+1}} - \frac{d_0(s)}{V_s} \mid \overline{d}(s)\right] \quad\leq\quad -\frac{\epsilon}{2^+\tau(n-s)}$$

holds for all $s < (1-\delta)n$. Although the differences in $\{\frac{d_0(i)}{V_i}\}_{i\geq 0}$ are not independent, and the process stops if $d_0(s) = 0$, the "2^+" function clearly leaves room for a series of random variables $\{X_i\}_{i\geq 0}$ such that the differences $X_{i+1} - X_i$ are independent for all $i \geq 0$, and the following are true whp for all $s \in [0,(1-\delta)n]$:

1. Either $d_0(s) = 0$ or $s_2 < \overline{s} < s_1$ or $X_s \geq \frac{d_0(s)}{V_s}$.
2. $X_0 = \frac{d_0(0)}{V_0} \leq 1$.
3. $|X_{s+1} - X_s| = O(\frac{\Delta}{V_s})$.
4. $E[X_{i+1} - X_i] \leq -\frac{\epsilon}{2^+\tau(n-s)}$.

So

$$E\left[X_s\right] \quad \leq \quad X_0 - \frac{\epsilon}{2^+\tau}\sum_{r=1}^{s}\frac{1}{n-r} \quad \leq \quad 1 + \frac{\epsilon}{2^+\tau}\log\left(1 - \frac{s}{n}\right)$$

for all s provided n is sufficiently large. So, if $\delta < \exp(-\frac{2^+\tau}{\epsilon} - \phi\tau)$, we will have $\overline{s} < (1-\delta)n$ satisfying $E[X_s] < -\frac{\epsilon\phi}{2^+}$, a constant. This allows us to use (3), so for any function $\alpha(n) \to \infty$,

$$\begin{aligned}\Pr(X_{\overline{s}} > 0) \quad &\leq \quad \exp\left[-\Omega\left(\frac{V_{\overline{s}}}{\Delta\sqrt{\overline{s}}}\right)^2\right] \quad \leq \quad \exp\left[-\Omega\left(\frac{\delta n}{n^{1/8}\sqrt{n}}\right)^2\right] \\ &\leq \quad \exp\left(-\Omega(n^{2\phi})\right) \quad = \quad o(1).\end{aligned}$$

Therefore, whp we have $s_1 < (1-\delta)n$ or $s_2 < (1-\delta)n$. In either case we have $\overline{s} < (1-\delta)n$. □

4. Proof of Theorem 1. Suppose that G is any graph with $2n$ vertices and $\epsilon > 0$ satisfies the following:

1. G has less than $(1-\epsilon)n$ edges;
2. $\Delta(G) \leq \frac{n^{1/10}}{\alpha(n)}$, where $\alpha(n) = \frac{5}{\epsilon^2}\log n$.

To make notation easier, we will define

$$i_* := \left\lfloor \Delta^2\alpha(n)\right\rfloor.$$

We begin by selecting any vertex $v_0 \in G$ and setting it false; a set T will give rise to a process similar to section 2. However, now that the expected degree is less than 1, we will show that whp there will be no contradiction, and we will most likely finish with $T = \emptyset$ instead of an edge within T or $i > i_*$.

Here is the exact procedure we will follow. In section 2, we kept track only of U so we could find a lower bound of $|T|$. Since we are interested only in an upper bound, we will not keep track of it here. All neighborhoods are in the "current" graph, i.e., not including vertices which have been removed from consideration.

```
1. Choose any vertex $v_0$, and set $i = 0$, $T_0 = N(v_0)$, $V = \mathbf{X} \setminus N(v_0) \setminus v_0$.
2. While $T_i \neq \emptyset$ and $i \leq i_*$:
   Pick any vertex $v_i \in T_i$, and match it with a random vertex $\overline{v}_i \in T_i \cup V_i - v_i$.
     • If $\overline{v}_i \in T_i$, then STOP; we have a contradiction.
     • If $\overline{v}_i \in V_i$, then $T_{i+1} = T_i \cup N(\overline{v}_i) - v_i$, $V_{i+1} = V_i \setminus N(\overline{v}_i) - \overline{v}_i$.
     • $i = i + 1$.
3. STOP; either $T_i = \emptyset$ or $i \geq i_*$.
```

The only thing that can raise the expected degree of $\overline{v}_i$ above $1-\epsilon$ is deleting isolated vertices, as deletion of any other vertices will also delete edges. However, we have

$$|V_i| \ > \ 2n - (i_* + 1)\Delta - (\Delta + 1) \quad = \quad 2n - O(i_*\Delta) \quad \leq \quad 2n - o(n).$$

Since we start with at least $2\epsilon n$ isolated vertices and won't lose more than $o(n)$ of them, we know that the increase in expected degree must be small; namely,

$$E[|N(\overline{v}_i)|] \leq 1 - \epsilon + o(1)$$

for all $i_* \geq i \geq 0$ is easily obtained when we divide by $2n$. So we bound $E[|T_i|]$ with the following:

$$E[|T_i| \mid T_{i-1}] \quad = \quad |T_{i-1}| - 1 + [1 - \epsilon + o(1)] \quad = \quad |T_{i-1}| - \epsilon + o(1). \tag{8}$$

Much like the proof of Lemma 1, we take the random variables $\{|T_i|\}_{i\geq 0}$ and note that the $o(1)$ term in (8) clearly leads to a series of random variables $\{X_i\}_{i\geq 0}$ such that for all i in our process we have $X_i \geq |T_i|$, $|X_{i+1} - X_i| \leq \Delta$, and all differences $X_{i+1} - X_i$ are independent. Furthermore, the X_i variables can "continue" even after $T_i = \emptyset$ and our process stops, so we have

$$E[X_i] \leq -\epsilon i + \Delta + o(i) \text{ for all } i \leq i_*. \tag{9}$$

For any vertex $v \in V(G)$, we have defined a process which begins by setting v false and continues keeping track of set T (as defined in the proof of Theorem 3) until either $T = \emptyset$, $\overline{v}_i \in T$, or $i = i_*$. Let E_v be the event that this process does not end with $T = \emptyset$, and define Z_v to be the set of all vertices which appear in the corresponding T at any time.

Lemma 5. *For any $v \in V(G)$, $\Pr(E_v) = O(n^{-3/5})$.*

Lemma 6. *If u is fixed and $\overline{u}$ is chosen randomly from $V(G)$, then*

$$\Pr(E_u \wedge E_{\overline{u}}) \quad = \quad o\left(\frac{1}{n}\right).$$

Lemmas 5 and 6 are proven below.

Consider an instance not in the union

$$\bigcup_u E_u \wedge E_{\overline{u}}.$$

By Lemma 6, the probability of such an instance is $1 - o(\frac{1}{n})O(n) = 1 - o(1)$ by the union bound. Therefore this deterministic entity has a satisfying assignment. (We can iteratively choose a pair of vertices $u, \overline{u}$ and set one of them false because this instance is not in $E_u \wedge E_{\overline{u}}$.) So we are done once we prove Lemmas 5 and 6.

Proof of Lemma 6. Assume Z_u is fixed. When we choose $\overline{v}$ (the partner of v) we need $\overline{v} \notin Z_u$ and $N(\overline{v}) \cap Z_u = \emptyset$. The probability of a problem is bounded above by

$$\frac{(\Delta+1)|Z_u|}{n} \quad = \quad O\left(\frac{\Delta^2 i_*}{n}\right)$$

for any randomly chosen $\overline{u} \in V(G)$, whether E_u is true or not. We make i_* choices in the formation of $Z_{\overline{u}}$, so the probability of a problem is bounded from above by

$$O\left(\frac{\Delta^2 i_*^2}{n}\right) \quad = \quad O\left(\frac{\Delta^6 \alpha(n)^2}{n}\right) \quad = \quad O\left(\frac{n^{-2/5}}{\alpha(n)^4}\right) \quad = \quad o(n^{-2/5}).$$

Therefore,

$$\Pr(Z_u \cap Z_{\overline{u}} \neq \emptyset | E_u) = o(n^{-2/5}). \tag{10}$$

Define $A = A_{u,\overline{u}}$ to be the event that $Z_u \cap Z_{\overline{u}} = \emptyset$. We have

$$\begin{aligned} \Pr(E_u \wedge E_{\overline{u}}) \quad &= \quad \Pr(E_u)\left[\Pr(E_{\overline{u}}|A, E_u)\Pr(A|E_u) + \Pr(E_{\overline{u}}|\overline{A}, E_u)\Pr(\overline{A}|E_u)\right] \\ &\leq \quad \Pr(E_u)\left[\Pr(E_{\overline{u}}|A, E_u) + \Pr(\overline{A}|E_u)\right]. \end{aligned}$$

For the second term, note that being given A, E_u ensures that the process starting at $\overline{u}$ avoids Z_u at all times. Therefore, the exact same proof of Lemma 5 with $G - Z_u$ in place of G tells us that $\Pr(E_{\overline{u}}|A, E_u) = O(n^{-3/5})$. So, using Lemma 5 and (10), we see that

$$\Pr(E_u \wedge E_{\overline{u}}) \quad \leq \quad O(n^{-3/5}) \left[O(n^{-3/5}) + o(n^{-2/5}) \right] \quad = \quad o\left(\frac{1}{n}\right). \qquad \square$$

Proof of Lemma 5. We will prove that all of the following are true with probability $1 - O(n^{-3/5})$:

(a) $|T_i| = 0$ for some $i \leq i_*$.

(b) $|T_i| \leq 2i_*\alpha(n)$ for all $i \leq i_*$.

(c) No edges will occur within T.

We have $i_* \gg \Delta$, so (9) tells us that $E[X_{i_*}] \leq -\frac{\epsilon}{1^+} i_*$. We use (3) with this and the fact that $|X_{i+1} - X_i| \leq \Delta$ for all $i \geq 0$:

$$\begin{aligned} \Pr((\text{a}) \text{ false}) \quad &\leq \quad \Pr(X_{i_*} > 0) \quad \leq \quad \exp\left(-\frac{\epsilon^2 i_*}{8^+\Delta^2}\right) \quad \leq \quad \exp\left(-\frac{\epsilon^2\alpha(n)}{8^+}\right) \\ &= \quad \exp\left(-\frac{\epsilon^2}{8^+}\frac{5}{\epsilon^2}\log n\right) \quad = \quad n^{-5/8^+} \quad \leq \quad n^{-3/5}. \end{aligned}$$

For (b), it is easy to see that

$$X_i > 2i_*\alpha(n) \quad \Rightarrow \quad X_i - E[X_i] \geq i_*\alpha(n),$$

because $E[X_i] \leq \Delta + o(1) \ll i_*\alpha(n)$. So by (3):

$$\Pr\left(X_i > 2i_*\alpha(n)\right) \quad \leq \quad \exp\left(-\frac{(i_*\alpha(n))^2}{8\Delta^2 i}\right) \quad \leq \quad \exp\left(-\frac{1}{8}\alpha(n)^3\right) \quad = \quad o\left(\frac{1}{n}\right)$$

for all $i \leq i_*$; therefore, the probability of this happening for any $i \leq i_*$ is actually $o(n^{-4/5})$. Finally, if (b) is true, then we have for all $i \leq i_*$

$$\Pr(\overline{v}_i \in T_i) = \frac{|T_i|}{|T_i| + |V_i| - 1} = \frac{|T_i|}{2n - o(n)} < \frac{X_i}{n} < \frac{2i_*\alpha(n)}{n}.$$

Therefore, the probability that (b) is true and (c) is false is bounded by

$$\sum_{i=0}^{i_*} \Pr(\overline{v}_i \in T_i) \quad \leq \quad \left(\frac{2i_*\alpha(n)}{n}\right) i_* \quad = \quad O\left(\frac{\Delta^4\alpha(n)^3}{n}\right) \quad = \quad O\left(\frac{n^{-3/5}}{\alpha(n)}\right). \qquad \square$$

5. Why the maximum degree condition is needed.

If the maximum degree is large, then the satisfiability depends much more on where the large degree vertices are matched and less on the actual graph. One example of this is a graph G which is the union of $K_{\alpha\sqrt{n}}$ and $2n - \alpha\sqrt{n}$ isolated vertices. Note that $S(G)$ is *not* satisfiable if and only if two or more of the matching edges end up within the complete graph $K_{\alpha\sqrt{n}}$. So

$\Pr(S(G)$ **is** satisfiable)

$$\begin{aligned} &= \quad \prod_{i=0}^{\alpha\sqrt{n}-1} \frac{2n - \alpha\sqrt{n} - i}{2n - 1 - 2i} \quad + \quad \binom{\alpha\sqrt{n}}{2} \frac{1}{2n-1} \prod_{i=2}^{\alpha\sqrt{n}-1} \frac{2n - \alpha\sqrt{n} - (i-2)}{2n - 1 - 2i} \\ &\approx \quad \left(1 + \frac{\alpha^2}{4}\right) \prod_{i=2}^{\alpha\sqrt{n}-1} \frac{2n - \alpha\sqrt{n} - (i-2)}{2n - 1 - 2i}. \end{aligned}$$

By taking the logarithm of the product and using $\log(1-x) \approx -x$ for $x \approx 0$, we can approximate the value of the product, and we arrive at the following:

$$\Pr(S(G) \text{ is satisfiable}) \quad \approx \quad \left(1+\frac{\alpha^2}{4}\right)\exp\left(-\frac{\alpha^2}{4}\right).$$

Thus, G has about $\alpha^2 n$ edges, but the probability of satisfiability of $S(G)$ does not have a threshold; it is a smooth function of α.

6. Concerning Conjecture 1.

Here we present two graphs G_1, G_2 with $(1+\epsilon)n$ edges each but which violate (1), and both $S(G_1)$ and $S(G_2)$ are not satisfiable whp. Since these graphs are vastly different and they are both such extreme cases, we believe that this is strong evidence that Conjecture 1 should be true.

6.1. Two examples.

Graph G_1. Fix $\log n \ll \alpha(n) \leq n^{\phi}$. Let G_α be any $\alpha(n)$-regular graph with $2(1+\epsilon)\frac{n}{\alpha}$ vertices. Let G_1 be G_α plus $2n - |V_{G_\alpha}|$ isolated vertices. We give the following "informal" argument to show that $S(G_1)$ is satisfiable whp.

Match all vertices which start out isolated; those which are matched may be "deleted" because they are no longer relevant. We will be left with an induced subgraph of G_α, say, G'_α, where $v \in V(G_\alpha)$ exists in G'_α with probability $\frac{|V(G_\alpha)|}{2n-1} \approx \frac{1+\epsilon}{\alpha}$. Also, $e \in E(G_\alpha)$ makes it to G'_α only if both of its vertices survive, which happens with probability close to $\left(\frac{1+\epsilon}{\alpha}\right)^2$. So whp G'_α has about $2(1+\epsilon)^2\frac{n}{\alpha^2}$ vertices and whp

$$\frac{|E(G'_\alpha)|}{|V(G'_\alpha)|} \quad \approx \quad \frac{\left(\frac{1+\epsilon}{\alpha}\right)^2 |E_{G_\alpha}|}{\left(\frac{1+\epsilon}{\alpha}\right)|V_{G_\alpha}|} \quad = \quad \left(\frac{1+\epsilon}{\alpha}\right)\frac{\alpha}{2} \quad = \quad \frac{1+\epsilon}{2}.$$

Also, we can most likely say a lot more about the degrees of the vertices. It is extremely unlikely that G_α has many high-degree vertices; in fact, whp G'_α satisfies (1) with τ equal to some sufficiently large constant; therefore, G_1 is not satisfiable whp by Theorem 2.

Graph G_2. Again fix $\log n \ll \alpha(n) \leq n^{\phi}$, and assume that $\phi < \frac{1}{4}$. Take $(1+\epsilon)\frac{n}{\alpha}$ disjoint stars, each with α leaves and then add $(1-\epsilon)n - (1+\epsilon)\frac{n}{\alpha}$ isolated vertices to make G_2. We can use a procedure similar to that of section 2, starting at any nonisolated vertex and stopping if $i \geq \sqrt{n}$. With stars we know exactly what we are working with, for any $\overline{v}_i$ we have a clearly defined $N(\overline{v}_i)$, $N_2(\overline{v}_i)$, and we know that declaring $N_2(v_i)$ free doesn't assume anything; leaves whose parent is deleted are indeed isolated. It is easy to see that for all $i \leq \alpha^3$ (since each step involves moving at most $\alpha+1$ vertices) we have

$$|U_i| \leq (1-\epsilon)\,n + o(n) \text{ and } |V_i| \geq (1+\epsilon)\,n - o(n).$$

If $\overline{v}_i \in T_i$ for any i, we are done. Otherwise, $|T_i|$ behaves as follows:

$$|T_{i+1}| - |T_i| \quad = \quad \begin{cases} -1 & \text{prob. } \frac{1-\epsilon}{2} - o(1), \\ +\alpha - 1 & \text{prob. } \frac{1}{\alpha}(\frac{1+\epsilon}{2} - o(1)), \\ 0 & \text{otherwise.} \end{cases}$$

The first of the three cases above corresponds to when $\overline{v}_i \in U_i$, so the only change to T is v_i is removed. The second case corresponds to when $\overline{v}_i \in V_i$, and $\overline{v}_i$ is a star center; therefore, v_i gets removed from T and α leaves get added. The third case is when $\overline{v}_i \in V_i$ is a leaf; therefore, v_i is removed from T but the center of the respective star is added.

Regardless of our what our nonisolated starting vertex is, for some constant c we have $\Pr(|T_{\lfloor\sqrt{n}\log n\rfloor}| \gg \sqrt{n}) \geq c$, because on every step the expected change in $|T|$ is a positive constant. Since whp $|T| \gg \sqrt{n}$ forces an edge within T, whp we have unsatisfiability.

6.2. Starting with a bounded degree. Suppose we run the pure literal algorithm on a graph with a bounded degree sequence whose degree sum is at least $(1+\epsilon)$ times its number of vertices. Here we show that as Step s approaches n in the pure literal algorithm, the degrees fall exponentially. It seems likely that this should continue even if the largest degree is at least n^{ϕ} for some ϕ, so long as $\phi < \frac{1}{2}$. If this is the case, then Conjecture 1 is true, because we can begin by running the pure literal algorithm and then create a graph which will meet the conditions Theorem 2.

During the pure literal algorithm, we started with $s = 0$, and we increased s until it was something close to n. If we let $t = \frac{s}{n}$ and $v_i(t) = \frac{1}{n}d_i(s)$ for all i, then we can look at this as a function of t, as t goes from 0 to 1. If the maximum degree starting out is a constant T, then we can use Lemma 2 along with methods discussed in [16] to create a system of differential equations, which whp is accurate within $O(n^{-1/2})$. Here is what the system looks like for $T = 4$; the pattern should be clear:

$$(11) \qquad 2(1-t)\begin{bmatrix} v_1'(t) \\ v_2'(t) \\ v_3'(t) \\ v_4'(t) \end{bmatrix} = \begin{bmatrix} -2 & 2 & 0 & 0 \\ 0 & -3 & 3 & 0 \\ 0 & 0 & -4 & 4 \\ 0 & 0 & 0 & -5 \end{bmatrix} \begin{bmatrix} v_1(t) \\ v_2(t) \\ v_3(t) \\ v_4(t) \end{bmatrix}.$$

This can be solved using the diagonalization $M\Lambda M^{-1}$ of the square matrix. In this case $\Lambda_{ii} = -(i+1)$ for all i, and M is an upper triangular matrix defined by

$$M_{i,j} = \begin{cases} (-1)^{i+j}\binom{j}{i} & i \leq j, \\ 0 & \text{otherwise.} \end{cases}$$

As it turns out, $(M^{-1})_{ij} = |M_{ij}|$ for all i, j. The solution to this system is

$$(12) \qquad \begin{bmatrix} v_1(t) \\ v_2(t) \\ v_3(t) \\ v_4(t) \end{bmatrix} = M\mathrm{diag}\left[M^{-1}\overline{d}(0)\right] \begin{bmatrix} (1-t) \\ (1-t)^{3/2} \\ (1-t)^2 \\ (1-t)^{5/2} \end{bmatrix},$$

where diag(w) for any vector w is the diagonal matrix W, where $W_{ii} = w_i$ for all i. (Again the pattern should be clear for any T, not just $T = 4$.)

To see that this is indeed the solution, let $\mu(t)$ be the last vector on the right-hand side of (12). Note that $2(1-t)\mu'(t) = \Lambda\mu(t)$, where Λ is defined above. Using this, it is easy to see that (11) is satisfied.

So if the largest degree is bounded to start, then so are the binomial coefficients; thus we have

$$v_i(t) = \Theta\left((1-t)^{(i+1)/2}\right)$$

for all i. This implies that for any N there exists $\tau > 0$ such that by time $1-\tau$, whp we have

$$\frac{d_i(\tau)}{d_{i+1}(\tau)} > N.$$

Although it seems much more difficult to prove, we believe that this nice distribution will continue even if the starting degree is larger than a constant T but not quite as large as $\sqrt{n}$. If this is the case, then Conjecture 1 is true.

REFERENCES

[1] D. Achlioptas, A. Naor, and Y. Peres, *On the maximum satisfiability of random formulas*, in Proceedings of the 44th Symposium on Foundations of Computer Science (FOCS 2003), IEEE, New York, 2003, pp. 362–370.

[2] D. Achlioptas and Y. Peres, *The threshold for random k-SAT is* $2^k(\ln 2 + o(1))$, J. Amer. Math. Soc., 17 (2004), pp. 947–973.

[3] N. Alon, J. H. Spencer, and P. Erdõs, *The Probabilistic Method*, John Wiley and Sons, New York, 1992.

[4] B. Bollobás, C. Borgs, J. Chayes, J. H. Kim, and D. B. Wilson, *The scaling window of the 2-SAT transition*, Random Structures Algorithms, 18 (2001), pp. 201–256.

[5] B. Bollobás, *Modern Graph Theory*, Springer, New York, 1998.

[6] D. Coppersmith, D. Gamarnik, M. Hajiaghayi, and G. Sorkin, *Random MAX SAT, random MAX CUT, and their phase transitions*, Random Structures Algorithms, 24 (2004), pp. 502–545.

[7] C. Cooper, A. Frieze, and G. Sorkin, *Random 2-SAT with Prescribed Literal Degrees*, in Proceedings of the 13th Annual Symposium on ACM-SIAM Discrete Algorithms, ACM, New York, 2002, pp. 316–320.

[8] V. Chvátal and B. Reed, *Mick gets some (the odds are on his side)*, in Proceedings of the 33rd Annual Symposium on Foundations of Computer Science, Pittsburgh, PA, 1992, IEEE, Los Alamitos, CA, 1992, pp. 620–627.

[9] P. Erdős and A. Rényi, *On the Evolution of Random Graphs*, Pub. Math. Inst. Hungarian Acad. Sci., 5, pp. 17–61.

[10] W. F. de la Vega, *Random 2-SAT: Results and problems*, Theoret. Comput. Sci., 265 (2001), pp. 131–146.

[11] A. Goerdt, *A threshold for unsatisfiability*, J. Comput. System Sci., 53 (1996), pp. 469–486.

[12] S. Janson, T. Luczak, and A. Ruciński, *Random Graphs*, Wiley-Intersc. Ser. Discrete Math. Optim., Wiley-Interscience, New York, 2000.

[13] C. McDiarmid, *Concentration*, in Probabilistic Methods for Algorithmic Discrete Mathematics, Springer, New York, 1998, pp. 195–248.

[14] M. Molloy, *When does the giant component bring unsatisfiability?*, Combinatorica, to appear.

[15] Y. Verhoeven, *Random 2-SAT and unsatisfiability*, Inform. Proc. Lett., 72 (1999), pp. 119–123.

[16] N. C. Wormald, *Differential Equations for random processes and random graphs*, Ann. Appl. Probab., 5 (1995), pp. 1217–1235.

SIAM J. DISCRETE MATH.
Vol. 21, No. 2, pp. 423–441

ON THE METRIC DIMENSION OF CARTESIAN PRODUCTS OF GRAPHS[*]

JOSÉ CÁCERES[†], CARMEN HERNANDO[‡], MERCÈ MORA[§], IGNACIO M. PELAYO[¶], MARÍA L. PUERTAS[†], CARLOS SEARA[§], AND DAVID R. WOOD[§]

Abstract. A set of vertices S *resolves* a graph G if every vertex is uniquely determined by its vector of distances to the vertices in S. The *metric dimension* of G is the minimum cardinality of a resolving set of G. This paper studies the metric dimension of cartesian products $G \,\Box\, H$. We prove that the metric dimension of $G \,\Box\, G$ is tied in a strong sense to the minimum order of a so-called doubly resolving set in G. Using bounds on the order of doubly resolving sets, we establish bounds on $G \,\Box\, H$ for many examples of G and H. One of our main results is a family of graphs G with bounded metric dimension for which the metric dimension of $G \,\Box\, G$ is unbounded.

Key words. graph, distance, resolving set, metric dimension, metric basis, cartesian product, Hamming graph, Mastermind, coin weighing, Djoković–Winkler relation

AMS subject classification. 05C12

DOI. 10.1137/050641867

1. Introduction.

A set of vertices S *resolves* a graph if every vertex is uniquely determined by its vector of distances to the vertices in S. This paper undertakes a general study of resolving sets in cartesian products of graphs.

All the graphs considered are finite, undirected, simple, and connected. The vertex set and edge set of a graph G are denoted by $V(G)$ and $E(G)$. The distance between vertices $v, w \in V(G)$ is denoted by $d_G(v, w)$, or $d(v, w)$ if the graph G is clear from the context. A vertex $x \in V(G)$ *resolves* a pair of vertices $v, w \in V(G)$ if $d(v, x) \neq d(w, x)$. A set of vertices $S \subseteq V(G)$ *resolves* G, and S is a *resolving set* of G, if every pair of distinct vertices of G is resolved by some vertex in S. A resolving set S of G with the minimum cardinality is a *metric basis* of G, and $|S|$ is the *metric dimension* of G, denoted by $\beta(G)$.

The *cartesian product* of graphs G and H, denoted by $G \,\Box\, H$, is the graph with vertex set $V(G) \times V(H) := \{(a, v) : a \in V(G), v \in V(H)\}$, where (a, v) is adjacent to (b, w) whenever $a = b$ and $\{v, w\} \in E(H)$, or $v = w$ and $\{a, b\} \in E(G)$. Where there is no confusion the vertex (a, v) of $G \,\Box\, H$ will be written av. Observe that if G and H are connected, then $G \,\Box\, H$ is connected. In particular, $d(av, bw) = d_G(a, b) + d_H(v, w)$ for all vertices av, bw of $G \,\Box\, H$. Assuming isomorphic graphs are equal, the cartesian product is associative, and $G_1 \,\Box\, G_2 \,\Box\, \cdots \,\Box\, G_d$ is well-defined.

[*] Received by the editors October 4, 2005; accepted for publication (in revised form) November 20, 2006; published electronically May 14, 2007. This research was supported by projects JA-PAI-FQM 305, MTM-2004-07891-C02-01, MTM-2005-08441-C02-01, MCYT-FEDER-BFM2003-00368, MCYT-HU2002-0010, MEC-SB2003-0270, and Gen. Cat 2001SGR00224.
http://www.siam.org/journals/sidma/21-2/64186.html

[†] Departamento de Estadística y Matemática Aplicada, Universidad de Almería, Almería, Spain (jcaceres@ual.es, mpuertas@ual.es).

[‡] Departament de Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Barcelona, Spain (carmen.hernando@upc.edu).

[§] Departament de Matemàtica Aplicada II, Universitat Politècnica de Catalunya, Barcelona, Spain (merce.mora@upc.edu, carlos.seara@upc.edu, david.wood@upc.edu). DRW's research was supported by a Marie Curie Fellowship of the European Community under contract 023865.

[¶] Departament de Matemàtica Aplicada III, Universitat Politècnica de Catalunya, Barcelona, Spain (ignacio.m.pelayo@upc.edu).

Resolving sets in general graphs were first defined by Harary and Melter [24] and Slater [42], although, as we shall see, resolving sets in hypercubes were studied earlier under the guise of a coin weighing problem [1, 5, 6, 7, 16, 19, 23, 26, 29, 30, 31, 32, 34, 44]. Resolving sets have since been widely investigated [4, 8, 9, 10, 11, 12, 14, 17, 18, 27, 35, 36, 37, 38, 39, 40, 41, 42, 43, 45, 46, 48] and arise in many diverse areas, including network discovery and verification [2], robot navigation [27, 41], connected joins in graphs [40], and strategies for the Mastermind game [3, 13, 20, 21, 22, 26].

Part of our motivation for studying the metric dimension of cartesian products is that in two of the above-mentioned applications, namely, Mastermind strategies and coin weighing, the graphs that arise are in fact cartesian products. These connections are explained in sections 2 and 6, respectively.

The main contributions of this paper are based on the notion of doubly resolving sets, which are introduced in section 4. We prove that the minimum order of a doubly resolving set in a graph G is tied in a strong sense to $\beta(G \,\square\, G)$. Thus doubly resolving sets are essential in the study of metric dimension of cartesian products. We then give a number of examples of bounds on the metric dimension of cartesian products through doubly resolving sets. In particular, sections 5, 6, 7, 8, and 9, respectively, study complete graphs, Hamming graphs, paths and grids, cycles, and trees. One of our main results here is a family of (highly connected) graphs with bounded metric dimension for which the metric dimension of the cartesian product is unbounded.

2. Coin weighing and hypercubes. The *hypercube* Q_n is the graph whose vertices are the n-dimensional binary vectors, where two vertices are adjacent if they differ in exactly one coordinate. It is well known that

$$Q_n = \underbrace{K_2 \,\square\, K_2 \,\square\, \cdots \,\square\, K_2}_{n}.$$

It is easily seen that $\beta(Q_n) \leq n$; see equation (7.4). The first case when this bound is not tight is $n = 5$. A laborious calculation verifies that Q_5 is resolved by the 4-vertex set $\{00000, 00011, 00101, 01001\}$. We have determined $\beta(Q_n)$ for small values of n by computer search.

n	2	3	4	5	6	7	8	10	15
$\beta(Q_n)$	2	3	4	4	5	6	6	≤ 7	≤ 10

The asymptotic value of $\beta(Q_n)$ turns out to be related to the following coin weighing problem first posed by Söderberg and Shapiro [44]. (See [23] for a survey on various coin weighing problems.) Given n coins, each with one of two distinct weights, determine the weight of each coin with the minimum number of weighings. We are interested in the static variant of this problem, where the choice of sets of coins to be weighed is determined in advance. Weighing a set S of coins determines how many light (and heavy) coins are in S, and no further information. It follows that the minimum number of weighings differs from $\beta(Q_n)$ by at most 1 [26, 40]. A lower bound on the number of weighings by Erdős and Rényi [16] and an upper bound by Lindström [29] imply that

$$\lim_{n\to\infty} \beta(Q_n) \cdot \frac{\log n}{n} = 2,$$

where, as always in this paper, logarithms are binary. Note that Lindström's proof is constructive. He gives an explicit scheme of $2^k - 1$ weighings that suffice for $k \cdot 2^{k-1}$ coins.

3. Projections. Let S be a set of vertices in the cartesian product $G \square H$ of graphs G and H. The *projection* of S onto G is the set of vertices $a \in V(G)$ for which there exists a vertex $av \in S$. Similarly, the *projection* of S onto H is the set of vertices $v \in V(H)$ for which there exists a vertex $av \in S$. A *column* of $G \square H$ is the set of vertices $\{av : v \in V(H)\}$ for some vertex $a \in V(G)$, and a *row* of $G \square H$ is the set of vertices $\{av : a \in V(G)\}$ for some vertex $v \in V(H)$. Observe that each row induces a copy of G, and each column induces a copy of H. This terminology is consistent with a representation of $G \square H$ by the points of the $|V(G)| \times |V(H)|$ grid.

LEMMA 3.1. *Let $S \subseteq V(G \square H)$ for graphs G and H. Then every pair of vertices in a fixed row of $G \square H$ is resolved by S if and only if the projection of S onto G resolves G. Similarly, every pair of vertices in a fixed column of $G \square H$ is resolved by S if and only if the projection of S onto H resolves H.*

Proof. Consider two vertices av and aw in a common column. For every other vertex bx of $G \square H$, we have $d(av, bx) - d(aw, bx) = d_H(v, x) - d_H(w, x)$. Thus $d(av, bx) \neq d(aw, bx)$ if and only if $d_H(v, x) \neq d_H(w, x)$. That is, av and aw are resolved by bx if and only if v and w are resolved by x in H. Hence av and aw are resolved by S if and only if v and w are resolved by the projection of S onto H. We have the analogous result for the projection onto G by symmetry. □

COROLLARY 3.2. *For all graphs G and H, and for every resolving set S of $G \square H$, the projection of S onto G resolves G, and the projection of S onto H resolves H. In particular, $\beta(G \square H) \geq \max\{\beta(G), \beta(H)\}$.* □

4. Doubly resolving sets. Many of the results that follow are based on the following definitions. Let $G \neq K_1$ be a graph. Two vertices $v, w \in V(G)$ are *doubly resolved* by $x, y \in V(G)$ if

$$d(v, x) - d(w, x) \neq d(v, y) - d(w, y).$$

Note that this definition generalizes the Djoković–Winkler relation Θ, which can be defined as follows: two edges $xy, vw \in E(G)$ are in Θ if and only if v, w are doubly resolved by x, y; see [25, 15, 47].

A set of vertices $S \subseteq V(G)$ *doubly resolves* G, and S is a *doubly resolving set*, if every pair of distinct vertices $v, w \in V(G)$ is doubly resolved by two vertices in S. Every graph with at least two vertices has a doubly resolving set. Let $\psi(G)$ denote the minimum cardinality of a doubly resolving set of a graph $G \neq K_1$. Note that if x, y doubly resolves v, w, then $d(v, x) - d(w, x) \neq 0$ or $d(v, y) - d(w, y) \neq 0$, and at least one of x and y (singly) resolves v, w. Thus a doubly resolving set is also a resolving set, and

$$\beta(G) \leq \psi(G).$$

Our interest in doubly resolving sets is based on the following upper bound.

THEOREM 4.1. *For all graphs G and $H \neq K_1$,*

$$\beta(G \square H) \leq \beta(G) + \psi(H) - 1.$$

Proof. Let S be a metric basis of G. Let T be a doubly resolving set of H with $|T| = \psi(H)$. Fix vertices $s \in S$ and $t \in T$. Let

$$X := \{sv : v \in T\} \cup \{at : a \in S\}.$$

Observe that $|X| = |S| + |T| - 1$. To prove that X resolves $G \square H$, consider two vertices av and bw of $G \square H$. By Lemma 3.1, if $a = b$ then av and bw are resolved since the projection of X onto H is T. Similarly, if $v = w$ then av and bw are resolved since the projection of X onto G is S. Now assume that $a \neq b$ and $v \neq w$. Since T is doubly resolving for H, there are two vertices $x, y \in T$ such that

$$d_H(v, x) - d_H(w, x) \neq d_H(v, y) - d_H(w, y).$$

Thus for at least one of x and y, say x,

$$d_H(v, x) - d_H(w, x) \neq d_G(b, s) - d_G(a, s).$$

Hence

$$d(av, sx) = d_G(a, s) + d_H(v, x) \neq d_G(b, s) + d_H(w, x) = d(bw, sx).$$

That is, $sx \in X$ resolves av and bw. □

The relationship between resolving sets of cartesian products and doubly resolving sets is strengthened by the following lower bound.

Lemma 4.2. *Suppose that S resolves $G \square G$ for some graph G. Let A and B be the two projections of S onto G. Then $A \cup B$ doubly resolves G. In particular,*

$$\beta(G \square G) \geq \tfrac{1}{2}\psi(G).$$

Proof. For any two vertices $v, w \in V(G)$, there is a vertex $pq \in S$ that resolves vw, wv. That is, $d(vw, pq) \neq d(wv, pq)$. Thus $d(v, p) + d(w, q) \neq d(w, p) + d(v, q)$, which implies $d(v, p) - d(w, p) \neq d(v, q) - d(w, q)$. Thus p, q doubly resolves v, w in G. Now $p \in A$ and $q \in B$. Hence $A \cup B$ doubly resolves G. If, in addition, S is a metric basis of $G \square G$, then $\psi(G) \leq |A \cup B| \leq |A| + |B| \leq 2|S| = 2 \cdot \beta(G \square G)$. □

Observe that Theorem 4.1 and Lemma 4.2 prove that $\beta(G \square G)$ is always within a constant factor of $\psi(G)$. In particular,

$$\tfrac{1}{2}\psi(G) \leq \beta(G \square G) \leq \psi(G) + \beta(G) - 1 \leq 2\psi(G) - 1. \tag{4.1}$$

Thus doubly resolving sets are essential in the study of the metric dimension of cartesian products.

A natural candidate for a resolving set of $G \square G$ is $S \times S$ for a well-chosen set $S \subseteq V(G)$. It follows from Lemma 4.2 and the proof technique employed in Theorem 4.1 that $S \times S$ resolves $G \square G$ if and only if S doubly resolves G.

Now consider the following elementary bound on $\psi(G)$.

Lemma 4.3. *Every graph G with $n \geq 3$ vertices satisfies $\psi(G) \leq n - 1$.*

Proof. Clearly G has a vertex x of degree at least two. Let $S := V(G) \setminus \{x\}$. To prove that S doubly resolves G, consider two vertices $u, v \in V(G)$. If both $u, v \in S$, then the pair u, v doubly resolves itself. Otherwise, without loss of generality, $u \in S$ and $v = x$. Since $\deg(x) \geq 2$, there is a neighbor $y \neq u$ of x. Now $d(u, u) - d(v, u) \leq 0 - 1 = -1$ and $d(u, y) - d(v, y) \geq 1 - 1 = 0$. Thus $u, y \in S$ doubly resolve u, v. Hence S doubly resolves G. □

Note that if G is a graph with $n \geq 3$ vertices, then Theorem 4.1 and Lemma 4.3 imply that $\beta(G \square H) \leq \beta(H) + n - 2$ for every graph H.

5. Complete graphs.

Let K_n denote the complete graph on $n \geq 1$ vertices. It is well known [9, 27] that for every n-vertex graph G,

$$\beta(G) = n - 1 \iff G = K_n. \tag{5.1}$$

LEMMA 5.1. *$\psi(K_n) = \max\{n-1, 2\}$ for all $n \geq 2$.*

Proof. Since $\psi(G) \geq 2$ for every graph $G \neq K_1$, we have $\psi(K_2) = 2$. Now suppose that $n \geq 3$. By Lemma 4.3, $\psi(K_n) \leq n-1$. Conversely, $\psi(K_n) \geq \beta(K_n) = n-1$ by equation (5.1). □

Theorem 4.1 and Lemma 5.1 imply that every graph G satisfies

$$\beta(K_n \,\square\, G) \leq \beta(G) + \max\{n-2, 1\}. \tag{5.2}$$

In certain cases, this result can be improved as follows.

LEMMA 5.2. *For every graph G and for all $n \geq 1$,*

$$\beta(K_n \,\square\, G) \leq \max\{n-1, 2 \cdot \beta(G)\}.$$

Proof. Let S be a metric basis of G. Fix a vertex r of K_n. As illustrated in Figure 5.1, there is a set T of $\max\{n-1, 2|S|\}$ vertices of $K_n \,\square\, G$ such that

(a) for all vertices $a \in V(K_n) \setminus \{r\}$, there is at least one vertex $x \in S$ for which $ax \in T$; and

(b) for all $x \in S$, there are at least two vertices $a, b \in V(K_n)$ for which $ax \in T$ and $bx \in T$.

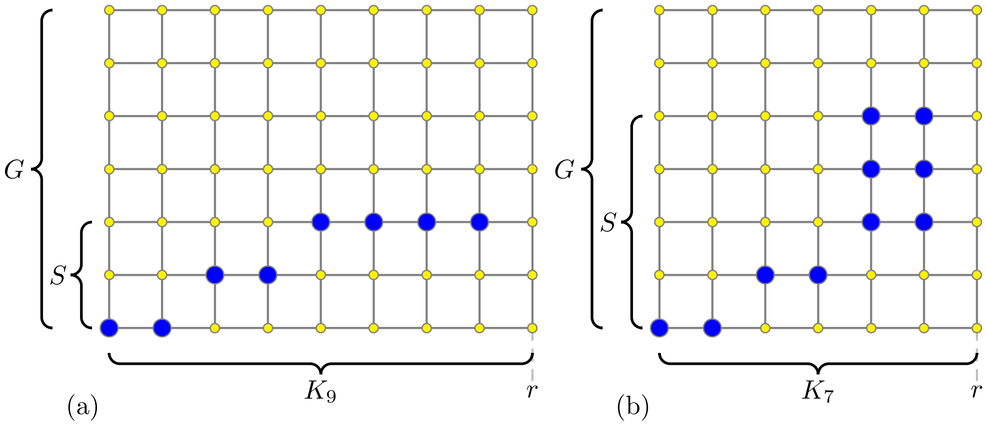


FIG. 5.1. *The resolving set T of $K_n \,\square\, G$ in Lemma 5.2:* (a) $n-1 \geq 2\beta(G)$ *and* (b) $n-1 \leq 2\beta(G)$.

To prove that T resolves $K_n \,\square\, G$, consider two vertices av and bw of $K_n \,\square\, G$. If $v = w$ then since the projection of T onto G is the resolving set S, by Lemma 3.1, av and bw are resolved by T. Now suppose that $v \neq w$. Then there is a vertex $x \in S$ that resolves v and w in G. Hence $d_G(v,x) < d_G(w,x)$ without loss of generality. By (b) there are distinct vertices $c, d \in V(K_n)$ for which $cx \in T$ and $dx \in T$. If $c \neq a$ and $c \neq b$, then

$$d(av, cx) = d_G(v,x) + 1 < d_G(w,x) + 1 = d(bw, cx));$$

that is, cx resolves av and bw in $K_n \,\square\, G$. Similarly, if $d \neq a$ and $d \neq b$, then dx resolves av and bw. Otherwise $c = a$ or $c = b$, and $d = a$ or $d = b$. Since $c \neq d$, without loss of generality $c = a$ and $d = b$. Then

$$d(av, cx) = d_G(v,x) < d_G(w,x) < d_G(w,x) + 1 = d(bw, cx),$$

and again cx resolves av and bw in $K_n \,\square\, G$. □

When n is large in comparison with $\beta(G)$ we know $\beta(K_n \,\square\, G)$ exactly.

THEOREM 5.3. *For every graph G and for all $n \geq 2 \cdot \beta(G) + 1$,*

$$\beta(K_n \,\square\, G) = n - 1.$$

Proof. The lower bound $\beta(K_n \,\square\, G) \geq n - 1$ follows from Corollary 3.2 and (5.1). The upper bound $\beta(K_n \,\square\, G) \leq n - 1$ is a special case of Lemma 5.2. □

6. Mastermind and Hamming graphs. *Mastermind* is a game for two players, the *code setter* and the *code breaker.*[1] The code setter chooses a secret vector $s = [s_1, s_2, \ldots, s_n] \in \{1, 2, \ldots, k\}^n$. The task of the code breaker is to infer the secret vector by a series of questions, each a vector $t = [t_1, t_2, \ldots, t_n] \in \{1, 2, \ldots, k\}^n$. The code setter answers with two integers, the first being the number of positions in which the secret vector and the question agree, denoted by $a(s,t) = |\{i : s_i = t_i, 1 \leq i \leq n\}|$. The second integer $b(s,t)$ is the maximum of $a(\tilde{s}, t)$, where $\tilde{s}$ ranges over all permutations of s.

In the commercial version of the game, $n = 4$ and $k = 6$. The secret vector and each question is represented by four pegs each colored with one of six colors. Each answer is represented by $a(s,t)$ black pegs and $b(s,t) - a(s,t)$ white pegs. Knuth [28] showed that four questions suffice to determine s in this case. Here the code breaker may determine each question in response to the previous answers. *Static mastermind* is the variation in which all the questions must be supplied at once. Let $g(n,k)$ denote the maximum, taken over all vectors s, of the minimum number of questions required to determine s in this static setting.

The *Hamming graph* $H_{n,k}$ is the cartesian product of cliques

$$H_{n,k} = \underbrace{K_k \,\square\, K_k \,\square\, \cdots \,\square\, K_k}_{n}.$$

Note that the hypercube $Q_n = H_{n,2}$. The vertices of $H_{n,k}$ can be thought of as vectors in $\{1, 2, \ldots, k\}^n$, with two vertices being adjacent if they differ in precisely one coordinate. Thus the distance $d_H(v,w)$ between two vertices v and w is the number of coordinates in which their vectors differ. That is,

$$d_H(v,w) = n - a(v,w).$$

Suppose for the time being that we remove the second integer $b(s,t)$ from the answers given by the code setter in the static mastermind game. Let $f(n,k)$ denote the maximum, taken over all vectors s, of the minimum number of questions required to determine s without $b(s,t)$ in the answers. For the code breaker to correctly infer the secret vector s from a set of questions T, s must be uniquely determined by the values $\{a(s,t) : t \in T\}$. Equivalently, for any two vertices v and w of $H_{n,k}$, there is a $t \in T$ for which $a(v,t) \neq a(w,t)$; that is, the distances $d_H(v,t) \neq d_H(w,t)$. Hence the secret vector can be inferred if and only if T resolves $H_{n,k}$. Thus

$$g(n,k) \leq f(n,k) = \beta(H_{n,k}).$$

Chvátal [13] proved the upper bound

$$\beta(H_{n,k}) = f(n,k) \leq (2 + \epsilon)n \,\frac{1 + 2\log k}{\log n - \log k}$$

[1] Chvátal [13] referred to the code setter and code breaker as S.F. and P.G.O.M. (in honor of P.E.).

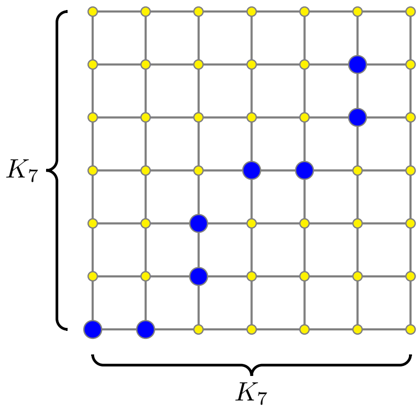


FIG. 6.1. *Resolving set of $K_7 \square K_7$ with one empty row, one empty column, and no lonely vertex.*

for large $n > n(\epsilon)$ and small $k < n^{1-\epsilon}$. For $k \in \{3, 4\}$, improvements to the constant in the above upper bound are stated without proof by Kabatianski, Lebedev, and Thorpe [26]. They also state that a "straightforward generalization" of the lower bound on $\beta(Q_n)$ by Erdős and Rényi [16] gives, for large n,

$$\beta(H_{n,k}) \geq g(n,k) \geq (2 + o(1)) \frac{n \log k}{\log n}.$$

Here we study $\beta(H_{n,k})$ for large values of k rather than for large values of n. A similar approach is taken for static mastermind by Goddard [20, 21], who proved that $g(2,k) = \lceil \frac{2}{3}k \rceil$ and $g(3,k) = k - 1$. Our contribution is to determine the exact value of $\beta(H_{2,k})$. We show that for all $k \geq 1$,

$$\beta(H_{2,k}) = \left\lfloor \tfrac{2}{3}(2k-1) \right\rfloor. \tag{6.1}$$

Equation (6.1) is a special case (with $m = n = k$) of the following more general result.

THEOREM 6.1. *For all $n \geq m \geq 1$,*

$$\beta(K_n \square K_m) = \begin{cases} \left\lfloor \frac{2}{3}(n+m-1) \right\rfloor & \text{if } m \leq n \leq 2m-1, \\ n-1 & \text{if } n \geq 2m-1. \end{cases}$$

We prove Theorem 6.1 by a series of lemmas. First note that two vertices of $K_n \square K_m$ are adjacent if and only if they are in a common row or column. Otherwise they are at distance two. Fix a set S of vertices of $K_n \square K_m$. With respect to S, a row or column is *empty* if it contains no vertex in S, and a vertex $v \in S$ is *lonely* if v is the only vertex of S in its row and in its column. As illustrated in Figure 6.1, we have the following characterization of resolving sets in $K_n \square K_m$.

LEMMA 6.2. *For $m, n \geq 2$, a set S of vertices resolves $K_n \square K_m$ if and only if*

(a) *there is at most one empty row and at most one empty column;*

(b) *there is at most one lonely vertex; and*

(c) *if there is an empty row and an empty column, then there is no lonely vertex.*

Proof. ($\Longrightarrow$) First suppose that S resolves $K_n \square K_m$. By Corollary 3.2, the projections of S, respectively, resolve K_m and K_n. By (5.1), there is at most one empty row and at most one empty column. Thus (a) holds.

Suppose on the contrary that v and w are two lonely vertices in S. Thus v and w are in distinct rows and columns, and no other vertex of S is in a row or column that contains v or w. Let x be the vertex in the row of v and the column of w. Let y be the vertex in the column of v and the row of w. Then $d(x,v) = d(y,v) = 1$, $d(x,w) = d(y,w) = 1$, and $d(x,u) = d(y,u) = 2$ for every vertex $u \in S \setminus \{v,w\}$. Thus S does not resolve x and y. This contradiction proves that S satisfies (b).

Finally, suppose that there is an empty row, an empty column, and a lonely vertex $v \in S$. Let x be the vertex in the row of v and in the empty column. Let y be the vertex in the column of v and in the empty row. We have $d(x,v) = d(y,v) = 1$, and $d(x,u) = d(y,u) = 2$ for every vertex $u \in S \setminus \{v\}$. Thus S does not resolve x and y. This contradiction proves that S satisfies (c).

($\Longleftarrow$) Now suppose that S is a set of vertices satisfying (a), (b), and (c). We will prove that S resolves any two vertices x and y. If $x \in S$ then x resolves x, y. If $y \in S$ then y resolves x, y. Now suppose that $x \notin S$ and $y \notin S$.

If x and y are in the same row, then at least one of the columns of x and y contains a vertex $v \in S$. Suppose v is in the column of x. Thus $d(x,v) = 1$ and $d(y,v) = 2$, and v resolves x, y. Similarly, if x and y are in the same column, then some $v \in S$ resolves x, y.

Suppose now that x and y are in distinct rows and columns. Then there is a vertex of S in the column of x or in the column of y. Suppose $v \in S$ is in the column of x. If v is not in the row of y, $d(x,v) = 1 \neq 2 = d(y,v)$, and v resolves x, y. If v is in the row of y, by (b) and (c), at least one of the vertices in the rows and columns of x and y, but not in the intersection of two of them, is in S. This vertex resolves x and y. □

Lemma 6.3. *For all $n, m \geq 3$, if S resolves $K_n \,\Box\, K_m$, then there exists a resolving set S^* of $K_n \,\Box\, K_m$ such that $|S^*| \leq |S|$, and S contains two vertices v and w in the same row or column, such that v and w are the only vertices in S^* in the row(s) and column(s) that contain v and w.*

Proof. By Lemma 6.2, there are two vertices $v, w \in S$ in the same row or column. By symmetry, we can suppose that v and w are in the same row. If v and w are the only vertices in S^* in the row and columns that contain v and w, then we are done. Otherwise there is a vertex $x \in S$ in the row or columns that contain v and w. It suffices to prove that x can be deleted from S, or replaced in S by some other vertex not in the row or columns that contain v and w, such that S still satisfies the conditions of Lemma 6.2, and thus resolves $K_n \,\Box\, K_m$. We can then repeat this step to obtain the desired set S^*.

First suppose that x is in the same row as v and w. If all the vertices of the column of x are in S, then delete x from S; clearly S still satisfies the conditions of Lemma 6.2. Otherwise, let y be a vertex not in S such that y is in the column containing x, and if x is the only vertex in its column that is in S, then y is in a row that contains at least one vertex of S. This is always possible, since S satisfies condition (a). Then $(S \setminus \{x\}) \cup \{y\}$ satisfies the conditions of Lemma 6.2.

Now suppose that x is in the column of v or w. If every vertex in the row containing x is in S, then delete x from S; clearly S still satisfies the conditions of Lemma 6.2. Otherwise, proceeding as in the preceding case, let y be a vertex in the same row as x, but not in the columns of v and w, such that there is at least one other vertex of S in the row or column that contains y. Then $(S \setminus \{x\}) \cup \{y\}$ satisfies the conditions of Lemma 6.2. This completes the proof. □

LEMMA 6.4. *For all $n, m \geq 3$,*

$$\beta(K_n \,\square\, K_m) = 2 + \min\{\beta(K_{n-2} \,\square\, K_{m-1}), \beta(K_{n-1} \,\square\, K_{m-2})\}.$$

Proof. We first prove that

$$\beta(K_n \,\square\, K_m) \leq 2 + \min\{\beta(K_{n-2} \,\square\, K_{m-1}), \beta(K_{n-1} \,\square\, K_{m-2})\}. \tag{6.2}$$

Without loss of generality $\beta(K_{n-2} \,\square\, K_{m-1}) \leq \beta(K_{n-1} \,\square\, K_{m-2})$. Let S be a metric basis of $K_{n-2} \,\square\, K_{m-1}$. Construct $S' \subseteq V(K_n \,\square\, K_m)$ from S by adding two new vertices that are positioned in one new row and in two new columns. The number of empty rows, empty columns, and lonely vertices is the same in S and S'. Since S resolves $K_{n-2} \,\square\, K_{m-1}$, S' resolves $K_n \,\square\, K_m$ by Lemma 6.2. Thus $\beta(K_n \,\square\, K_m) \leq |S'| = |S| + 2 = 2 + \beta(K_{n-2} \,\square\, K_{m-1})$, which implies (6.2). It remains to prove that

$$\min\{\beta(K_{n-2} \,\square\, K_{m-1}), \beta(K_{n-1} \,\square\, K_{m-2})\} \leq \beta(K_n \,\square\, K_m) - 2. \tag{6.3}$$

Let S be a metric basis of $K_n \,\square\, K_m$. By Lemma 6.3, we can assume that S contains two vertices v and w in the same row or column, such that v and w are the only vertices in S in the row(s) and column(s) that contain v and w. Without loss of generality, v and w are in the same row. Construct $S' \subseteq V(K_{n-2} \,\square\, K_{m-1})$ from S by deleting the row containing v and w, and by deleting the two columns containing v and w. The number of empty rows, empty columns, and lonely vertices is the same in S and S'. Since S resolves $K_n \,\square\, K_m$, S' resolves $K_{n-2} \,\square\, K_{m-1}$ by Lemma 6.2. Thus $\beta(K_{n-2} \,\square\, K_{m-1}) \leq |S'| \leq |S| - 2 = \beta(K_n \,\square\, K_m) - 2$, which implies (6.3). □

Proof of Theorem 6.1. We proceed by induction on $n + m$ in increments of 3. (Formally speaking, we are doing induction on $\lfloor \frac{1}{3}(n+m) \rfloor$.)

First observe that for $m = 1$, we know that $\beta(K_n \,\square\, K_m) = n - 1$. For $m = 2$, we have $\beta(K_2 \,\square\, K_2) = 2 = \lfloor \frac{2}{3}(2 + 2 - 1) \rfloor$, $\beta(K_3 \,\square\, K_2) = 2 = \lfloor \frac{2}{3}(3 + 2 - 1) \rfloor$, and $\beta(K_n \,\square\, K_2) = n - 1$ for all $n \geq 3$. Thus the assertion is true for $m \leq 2$. Now suppose that $m \geq 3$. By Lemma 6.4,

$$\beta(K_n \,\square\, K_m) = 2 + \min\{\beta(K_{n-2} \,\square\, K_{m-1}), \beta(K_{n-1} \,\square\, K_{m-2})\}. \tag{6.4}$$

Case 1. $n \geq 2m - 1$: Then $n \geq 2 \cdot \beta(K_m) + 1$ by (5.1), and $\beta(K_n \,\square\, K_m) = n - 1$ by Theorem 5.3 with $G = K_m$.

Case 2. $n = 2m - 2$: First consider $K_{n'} \,\square\, K_{m'}$, where $n' = n - 1 = 2m - 3$ and $m' = m - 2$. Then $m' \leq n'$ and $n' \geq 2m' - 1$. By induction,

$$\beta(K_{n'} \,\square\, K_{m'}) = n' - 1 = n - 2 = \lfloor \tfrac{2}{3}(n + m - 1) \rfloor - 2.$$

Now consider $K_{n'} \,\square\, K_{m'}$, where $m' = m - 1$ and $n' = n - 2 = 2m - 4$. Then $m' \leq n' \leq 2m' - 1$. By induction

$$\beta(K_{n'} \,\square\, K_{m'}) = \lfloor \tfrac{2}{3}(n' + m' - 1) \rfloor = \lfloor \tfrac{2}{3}(n + m - 1) \rfloor - 2.$$

By (6.4), $\beta(K_n \,\square\, K_m) = \lfloor \frac{2}{3}(n + m - 1) \rfloor$.

Case 3. $n = 2m - 3$: First consider $K_{n'} \,\square\, K_{m'}$, where $m' = m - 2$ and $n' = n - 1 = 2m - 4$. Then $m' \leq n'$ and $n' \geq 2m' - 1$. By induction,

$$\beta(K_{n'} \,\square\, K_{m'}) = n' - 1 = n - 2 = \lfloor \tfrac{2}{3}(n + m - 1) \rfloor - 2.$$

Now consider $K_{n'} \,\square\, K_{m'}$, where $m' = m-1$, $n' = n-2 = 2m-5$. For $m \geq 4$, we have $m' \leq n' \leq 2m'-1$. By induction

$$\beta(K_{n'} \,\square\, K_{m'}) = \lfloor \tfrac{2}{3}(n'+m'-1) \rfloor = \lfloor \tfrac{2}{3}(n+m-1) \rfloor - 2.$$

For $m = 3$, we have $n = 2m-3 = 3$. It is easily verified that $\beta(K_3 \,\square\, K_3) = 3 = \lfloor \frac{2}{3}(3+3-1) \rfloor$. In all cases we obtain $\beta(K_n \,\square\, K_m) = \lfloor \frac{2}{3}(n+m-1) \rfloor$ by (6.4).

Case 4. $n \leq 2m-4$: First consider $K_{n'} \,\square\, K_{m'}$, where $m' = m-2$ and $n' = n-1 \leq 2m-5$. Then $m' \leq n' \leq 2m'-1$. By induction,

$$\beta(K_{n'} \,\square\, K_{m'}) = \lfloor \tfrac{2}{3}(n'+m'-1) \rfloor = \lfloor \tfrac{2}{3}(n+m-1) \rfloor - 2.$$

Now consider $K_{n'} \,\square\, K_{m'}$, where $m' = m-1$ and $n' = n-2 \leq 2m-6$. If $m \leq n-1$, then $m' \leq n' < 2m'-1$, and by induction

$$\beta(K_{n'} \,\square\, K_{m'}) = \lfloor \tfrac{2}{3}(n'+m'-1) \rfloor = \lfloor \tfrac{2}{3}(n+m-1) \rfloor - 2.$$

If $m = n \geq 4$, then $n' \leq m' \leq 2n'-1$, and by induction

$$\beta(K_{m'} \,\square\, K_{n'}) = \lfloor \tfrac{2}{3}(m'+n'-1) \rfloor = \lfloor \tfrac{2}{3}(n+m-1) \rfloor - 2.$$

Finally, if $m = n = 3$, then $\beta(K_{n'} \,\square\, K_{m'}) = \beta(K_2 \,\square\, K_1) = 1 = \lfloor \frac{2}{3}(3+3-1) \rfloor - 2$. In all cases, we obtain $\beta(K_n \,\square\, K_m) = \lfloor \frac{2}{3}(m+n-1) \rfloor$ by (6.4). □

7. Paths and grids. Let P_n denote the path on $n \geq 1$ vertices. Khuller, Raghavachari, and Rosenfeld [27] and Chartrand et al. [9] proved that an n-vertex graph G has

$$\beta(G) = 1 \iff G = P_n. \tag{7.1}$$

Thus, by Theorem 5.3, for all $n \geq 3$,

$$\beta(K_n \,\square\, P_m) = n-1. \tag{7.2}$$

Minimum doubly resolving sets in paths are easily characterized.

Lemma 7.1. *For all $n \geq 2$ we have $\psi(P_n) = 2$. Moreover, the two endpoints of P_n are in every doubly resolving set of P_n.*

Proof. By definition $\psi(G) \geq 2$ for every graph $G \neq K_1$. Let $P_n = (v_1, v_2, \ldots, v_n)$. For all $1 \leq i < j \leq n$, we have $d(v_i, v_1) - d(v_j, v_1) = (i-1) - (j-1) = i - j$, and $d(v_i, v_n) - d(v_j, v_n) = (n-i) - (n-j) = j - i$. Thus $\{v_1, v_n\}$ doubly resolve P_n, and $\psi(P_n) = 2$. Finally, observe that v_1 is in every doubly resolving set, as otherwise v_1 and v_2 would not be doubly resolved. Similarly v_n is in every doubly resolving set. □

Lemma 7.2. *If $\beta(G \,\square\, H) = 2$, then G or H is a path.*

Proof. Say $S = \{av, bw\}$ resolves $G \,\square\, H$. Suppose that $a = b$. Then the projection of S onto G is a single vertex. By Lemma 3.1, the projection of S onto G resolves G, and by (7.1), only paths have singleton resolving sets. Thus G is a path, and we are done. Similarly, if $v = w$ then H is a path, and we are done. Now suppose that $a \neq b$ and $v \neq w$. Let c be the neighbor of b on a shortest path from a to b. Note that c may equal a. Then $d_G(a,c) + 1 = d_G(a,b)$ and $d_G(b,c) = 1$. Similarly, let x be the neighbor of w on a shortest path from v to w. Then $d_H(v,x) + 1 = d_H(v,w)$ and $d_H(x,w) = 1$. This implies that S does not resolve bx and cw, since

$$d(bx, av) = d_G(a,b) + d_H(x,v) = d_G(a,c) + d_H(v,w) = d(cw, av)$$

and

$$d(bx, bw) = d_H(x, w) = 1 = d_G(b, c) = d(cw, bw).$$

This contradiction proves the result. $\square$

Theorem 4.1 and Lemma 7.1 imply that every graph G satisfies

$$\beta(G) \leq \beta(G \,\square\, P_n) \leq \beta(G) + 1, \tag{7.3}$$

as proved by Chartrand et al. [9] in the case that $n = 2$.

An n-dimensional *grid* is a cartesian product of paths $P_{m_1} \,\square\, P_{m_2} \,\square\, \cdots \,\square\, P_{m_n}$. Equations (7.1) and (7.3) imply that

$$\beta(P_{m_1} \,\square\, P_{m_2} \,\square\, \cdots \,\square\, P_{m_n}) \leq n, \tag{7.4}$$

as proved by Khuller, Raghavachari, and Rosenfeld [27], who in addition claimed that

$$\beta(P_{m_1} \,\square\, P_{m_2} \,\square\, \cdots \,\square\, P_{m_n}) = n.$$

They wrote "we leave it for the reader to see why n is a lower bound." This claim is false if every $m_i = 2$ and n is large, since $\beta(P_2 \,\square\, P_2 \,\square\, \cdots \,\square\, P_2) \to 2n/\log n$, as discussed in section 2. Sebő and Tannier [40] claimed without proof that "using a result of Lindström [30]" one can prove that

$$\limsup_{n\to\infty} \beta(\underbrace{P_k \,\square\, P_k \,\square\, \cdots \,\square\, P_k}_{n}) \cdot \frac{\log n}{n \log k} \leq 2. \tag{7.5}$$

8. Cycles.

Let C_n denote the cycle on $n \geq 3$ vertices. Two vertices v and w of C_n are *antipodal* if $d(v, w) = \frac{n}{2}$. Note that no two vertices are antipodal in an odd cycle.

LEMMA 8.1 (see [27, 39]). *$\beta(C_n) = 2$ for all $n \geq 3$. Moreover, two vertices resolve C_n if and only if they are not antipodal.*

LEMMA 8.2. *For all $n \geq 3$,*

$$\psi(C_n) = \begin{cases} 2 & \textit{if } n \textit{ is odd}, \\ 3 & \textit{if } n \textit{ is even}. \end{cases}$$

Proof. We have $\psi(C_n) \geq 2$ by definition. Now we prove the upper bound. Denote $C_n = (v_1, v_2, \ldots, v_n)$. Let $k := \lfloor \frac{n}{2} \rfloor$. Consider two vertices v_i and v_j of C_n. Without loss of generality $i < j$.

Case 1. $1 \leq i < j \leq k + 1$: Then $d(v_i, v_1) - d(v_j, v_1) = (i - 1) - (j - 1) = i - j$ and $d(v_i, v_{k+1}) - d(v_j, v_{k+1}) = (k + 1 - i) - (k + 1 - j) = j - i \neq i - j$. Thus v_1, v_{k+1} doubly resolve v_i, v_j.

Case 2. $k+1 \leq i < j \leq n$: Then $d(v_i, v_1) - d(v_j, v_1) = (n+1-i) - (n+1-j) = j-i$ and $d(v_i, v_{k+1}) - d(v_j, v_{k+1}) = (i - k - 1) - (j - k - 1) = i - j \neq j - i$. Thus v_1, v_{k+1} doubly resolve v_i, v_j.

Case 3. $1 \leq i \leq k + 1 < j \leq n$: Suppose that v_1, v_{k+1} does not doubly resolve v_i, v_j. That is, $d(v_i, v_1) - d(v_j, v_1) = d(v_i, v_{k+1}) - d(v_j, v_{k+1})$. Thus $(i-1) - (n+1-j) = (k + 1 - i) - (j - k - 1)$. Hence $n = 2i + 2j - 2k - 4$ is even.

Therefore for odd n, $\{v_1, v_{k+1}\}$ doubly resolves C_n, and $\psi(C_n) = 2$.

For even n, in Case 3, suppose that v_1, v_2 does not doubly resolve v_i, v_j. That is, $d(v_i, v_1) - d(v_j, v_1) = d(v_i, v_2) - d(v_j, v_2)$. Thus $(i-1) - (n+1-j) = (i-2) - (n+2-j)$

and $-2 = -4$, a contradiction. Hence for even n, $\{v_1, v_2, v_{k+1}\}$ doubly resolve C_n, and $\psi(C_n) \le 3$.

It remains to prove that $\psi(C_n) \ge 3$ for even n. Suppose that $\psi(C_n) \le 2$ for some even $n = 2k$. By symmetry we can assume that $\{v_1, v_i\}$ doubly resolves C_n for some $2 \le i \le k+1$.

Case 1. $2 \le i \le k-1$: Then $d(v_{i+1}, v_1) - d(v_{i+2}, v_1) = i - (i+1) = -1$ and $d(v_{i+1}, v_i) - d(v_{i+2}, v_i) = 1 - 2 = -1$. Thus v_1, v_i does not resolve v_{i+1}, v_{i+2}.

Case 2. $i = k$: Then $d(v_2, v_1) - d(v_{n-1}, v_1) = 1 - 2 = -1$ and $d(v_2, v_i) - d(v_{n-1}, v_i) = (k-2) - (k-1) = -1$. Thus v_1, v_i does not resolve v_2, v_{n-1}.

Case 3. $i = k+1$: Then $d(v_2, v_1) - d(v_n, v_1) = 1 - 1 = 0$ and $d(v_2, v_i) - d(v_n, v_i) = (k-1) - (k-1) = 0$. Thus v_1, v_i does not resolve v_2, v_n.

In each case we have derived a contradiction. Thus $\psi(C_n) \ge 3$ for even n. □

Theorem 4.1 and Lemma 8.2 imply that every graph G satisfies

$$\beta(G) \le \beta(G \,\square\, C_n) \le \begin{cases} \beta(G) + 1 & \text{if } n \text{ is odd,} \\ \beta(G) + 2 & \text{if } n \text{ is even.} \end{cases} \tag{8.1}$$

THEOREM 8.3. *For every graph G and for all $n \ge 3$, we have $\beta(G \,\square\, C_n) = 2$ if and only if G is a path and n is odd.*

Proof. ($\Longleftarrow$) Since G is a path, $\beta(G) = 1$ by (7.1). Since n is odd, $\psi(C_n) = 2$ by Lemma 8.2. Thus $\beta(G \,\square\, C_n) \le \psi(C_n) + \beta(G) - 1 = 2$ by Theorem 4.1.

($\Longrightarrow$) Suppose that $\beta(G \,\square\, C_n) = 2$. Say $S = \{av, bw\}$ resolves $G \,\square\, C_n$. Then G is a path by Lemma 7.2. It remains to show that n is odd. Suppose on the contrary that $n = 2r$ is even. Let $C = C_n$. By Corollary 3.2, the projection $\{v, w\}$ of S onto C resolves C. By Lemma 8.1, we have $\beta(C) = 2$, and thus $v \ne w$. Moreover, v and w are not antipodal. That is, $d_C(v, w) \le r - 1$. Hence there is a neighbor x of w in C with $d_C(v, x) = d_C(v, w) + 1$. Now consider G. If $a \ne b$, then using the argument from the proof of Lemma 7.2, we can construct a pair of vertices that are not resolved by S. So now assume $a = b$. That is, our resolving set is contained in a single column of $G \,\square\, C_n$. Let p be a neighbor of a in G. Then S does not resolve pw and ax, since $d(pw, bw) = 1 = d(ax, bw)$ and $d(pw, av) = 1 + d_C(v, w) = d_C(x, v) = d(ax, av)$. This contradiction proves the result. □

By Lemma 8.2 and (7.1), we have $\beta(P_m \,\square\, C_n) \le \psi(C_n) + \beta(P_m) - 1 \le 3 + 1 - 1 = 3$. Thus Theorem 8.3 implies that for all $m \ge 2$ and $n \ge 3$ we have

$$\beta(P_m \,\square\, C_n) = \begin{cases} 2 & \text{if } n \text{ is odd,} \\ 3 & \text{if } n \text{ is even.} \end{cases} \tag{8.2}$$

THEOREM 8.4. *For all $m, n \ge 3$,*

$$\beta(C_m \,\square\, C_n) = \begin{cases} 3 & \textit{if } m \textit{ or } n \textit{ is odd,} \\ 4 & \textit{otherwise.} \end{cases}$$

Proof. We have $\beta(C_m \,\square\, C_n) \ge 3$ by Theorem 8.3. If m or n is odd, then $\beta(C_m \,\square\, C_n) \le 3$ by (8.1) and since $\beta(C_m) = 2$. It remains to prove that $\beta(C_m \,\square\, C_n) \ge 4$ when m and n are even. Let $G := C_{2r} \,\square\, C_{2s}$. We denote each vertex U of G by $u_1 u_2$, where $u_1 \in C_{2r}$ and $u_2 \in C_{2s}$.

Observe that in C_{2r}, every vertex u is antipodal with a unique vertex v; thus $d(x, u) + d(x, v) = r$ for every vertex x of C_{2r}.

Two vertices U and V of G are *antipodal* if u_1 and v_1 are antipodal in C_{2r} and u_2 and v_2 are antipodal in C_{2s}. Suppose that U and V are antipodal. Then for every vertex W of G,

$$(8.3)\quad d_G(W,U)+d_G(W,V)=d(w_1,u_1)+d(w_2,u_2)+d(w_1,v_1)+d(w_2,v_2)=r+s.$$

CLAIM 8.5. *Let U be a vertex in a resolving set S of G. Say U and V are antipodal. Then the set S' obtained by replacing U by V in S also resolves G.*

Proof. Suppose on the contrary that S' does not resolve G. Thus there exist vertices X, Y of G such that $d_G(X,Z)=d_G(Y,Z)$ for every vertex $Z \in S'$. In particular, $d_G(X,V)=d_G(Y,V)$. By (8.3), $d_G(X,U)-r-s=d_G(Y,U)-r-s$, implying $d_G(X,U)=d_G(Y,U)$. Thus $d_G(X,Z)=d_G(Y,Z)$ for every vertex $Z \in S$; that is, X and Y are not resolved by S. This contradiction proves the claim. □

Suppose on the contrary that $S=\{U,V,W\}$ is a resolving set of G. Represent G by the points of a $2r \times 2s$ grid. Consecutive points in the same row or column are adjacent, and the first and last points of the same row or column are adjacent. Observe that antipodal vertices of G are in opposite quadrants of the grid. Thus, by the above claim, we can assume that U, V, W are in one of the four halves of the grid. Without loss of generality, U, V, W are in the left half of the grid. This implies that $d(u_1,v_1)<r$, $d(u_1,w_1)<r$ and $d(v_1,w_1)<r$. Furthermore, U, V, W are in at least two different rows and two different columns, since the projections of S resolve C_{2r} and C_{2s}.

By symmetry, it suffices to consider the following cases:

1. U, V, W are in different rows and different columns.
2. U, V, W are in different rows, but U, V are in the same column.
3. U, V are in the same column and V, W in the same row.

In each case we will find vertices X, Y such that $d(X,U)=d(Y,U)$, $d(X,V)=d(Y,V)$, and $d(X,W)=d(Y,W)$; that is, S does not resolve the pair X, Y.

Case 1. Assume that if one of the vertices u_2, v_2, w_2 is in the shortest path determined by the other two vertices, then that vertex is v_2. It is then possible to draw the grid in such a way that the projections u_2, v_2, w_2 appear from bottom to top in C_{2s}, $d(u_2,v_2)<s$, and $d(v_2,w_2)<s$. Now, if v_1 is in the shortest path between u_1 and w_1 in C_{2r}, then let X, Y be the two neighbors of V lying in shortest paths between V and W; see Figure 8.1(a). Otherwise, assume that u_1 is in the shortest path between v_1 and w_1. Let Z be the vertex u_1v_2. Let X, Y be the neighbors of Z in shortest paths between Z and W; see Figure 8.1(b). It is easy to verify that in both cases $d(X,U)=d(Y,U)$, $d(X,V)=d(Y,V)$, and $d(X,W)=d(Y,W)$.

Case 2. Observe that at least two of the distances $d(u_2,v_2)$, $d(v_2,w_2)$, and $d(u_2,w_2)$ in C_{2s} must be less than s. If u_2, v_2 are not antipodal in C_{2s} and w_2 is not in the shortest path between u_2 and v_2 in C_{2s}, then $d(u_2,w_2)<s$ or $d(v_2,w_2)<s$. Let us assume that $d(v_2,w_2)<s$. Let X, Y be the vertices adjacent to V lying in a shortest path between V and W; see Figure 8.2(a). If u_2, v_2 are not antipodal in C_{2s} and w_2 is in the shortest path between u_2 and v_2 in C_{2s}, then let X, Y be the neighbors of V not lying in a shortest path between V and W; see Figure 8.2(b). Finally, if u_2, v_2 are antipodal in C_{2s}, consider the vertices X, Y at distance two from V; see Figure 8.2(c). It is easy to verify that in all cases $d(X,U)=d(Y,U)$, $d(X,V)=d(Y,V)$, and $d(X,W)=d(Y,W)$.

Case 3. In this case, $d(u_2,v_2)<s$ since the projection $\{u_2,v_2,w_2\}=\{u_2,v_2\}$ resolves C_{2s}. Let $Z:=(w_1,u_2)$. Let X, Y be the neighbors of Z not lying in a shortest

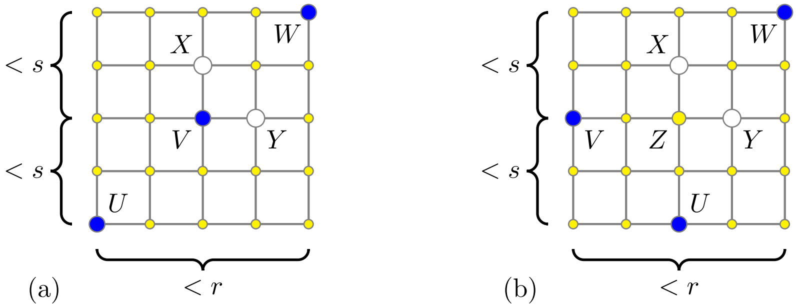


FIG. 8.1. *Illustration for Case* 1 *in the proof of Theorem* 8.4.

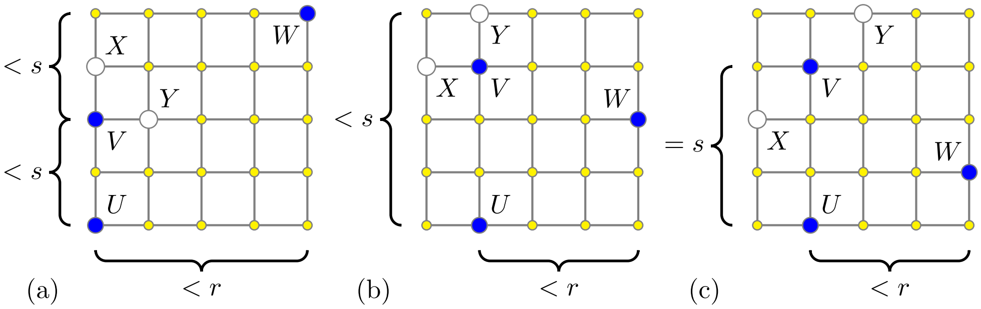


FIG. 8.2. *Illustration for Case* 2 *in the proof of Theorem* 8.4.

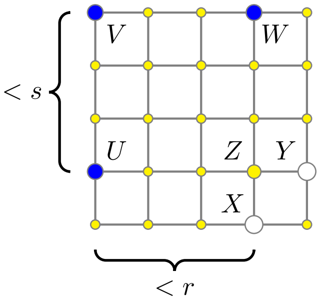


FIG. 8.3. *Illustration for Case* 3 *in the proof of Theorem* 8.4.

path between Z and V; see Figure 8.3. It is easy to verify that $d(X,U) = d(Y,U)$, $d(X,V) = d(Y,V)$, and $d(X,W) = d(Y,W)$. □

THEOREM 8.6. *For all $n \geq 1$ and $m \geq 3$,*

$$\beta(K_n \,\Box\, C_m) = \begin{cases} 2 & \text{if } n = 1, \\ 2 & \text{if } n = 2 \text{ and } m \text{ is odd}, \\ 3 & \text{if } n = 2 \text{ and } m \text{ is even}, \\ 3 & \text{if } n = 3, \\ 3 & \text{if } n = 4 \text{ and } m \text{ is even}, \\ 4 & \text{if } n = 4 \text{ and } m \text{ is odd}, \\ n-1 & \text{if } n \geq 5. \end{cases}$$

Proof. The case $n \geq 2\beta(C_n) + 1 = 5$ is an immediate corollary of Theorem 5.3 and Lemma 8.1. The case $n = 3$ is a special case of Theorem 8.4 since $K_3 = C_3$. The case $n = 2$ is a special case of (8.2) since $K_2 = P_2$. The case $n = 1$ is a repetition of Lemma 8.1. It remains to prove the case $n = 4$. Say $V(K_4) = \{a, b, c, d\}$. First note that $\beta(K_4 \,\square\, C_m) \geq \beta(K_4) = 3$ by Corollary 3.2 and (5.1). By Lemma 5.1 we have $\psi(K_4) = 3$. Thus $\beta(K_4 \,\square\, C_m) \leq 4$ by Lemma 8.1 and Theorem 4.1 with $H = K_4$. For even m, it is easily verified that $\{av, bv, cw\}$ resolves $K_4 \,\square\, C_m$ for any edge vw of C_m.

It remains to prove that $\beta(K_4 \,\square\, C_m) \geq 4$ for odd $m = 2h + 1$. Consider the vertices of $K_4 \,\square\, C_m$ to be in a $4 \times m$ grid, where two vertices in the same row are adjacent, and two vertices in the same column are adjacent if and only if they are consecutive rows or they are in the first and last rows. Suppose on the contrary that $S = \{u, v, w\}$ resolves $K_4 \,\square\, C_m$. Then u, v, w are in three different columns and in at least two different rows (by considering the projections of S onto K_4 and C_m).

Case 1. Suppose that two vertices in S, say u and v, are in the same row. Consider the grid centered at the row of u, v. Without loss of generality, u and v are in the first and second columns, and w is in a row above u and v. Let x and y be the vertices shown in Figure 8.4(a). Then $d(x, u) = d(y, u) = h + 1$, $d(x, v) = d(y, v) = h + 1$, and $d(x, w) = d(y, w) = p$. Thus S does not resolve x and y, which is the desired contradiction.

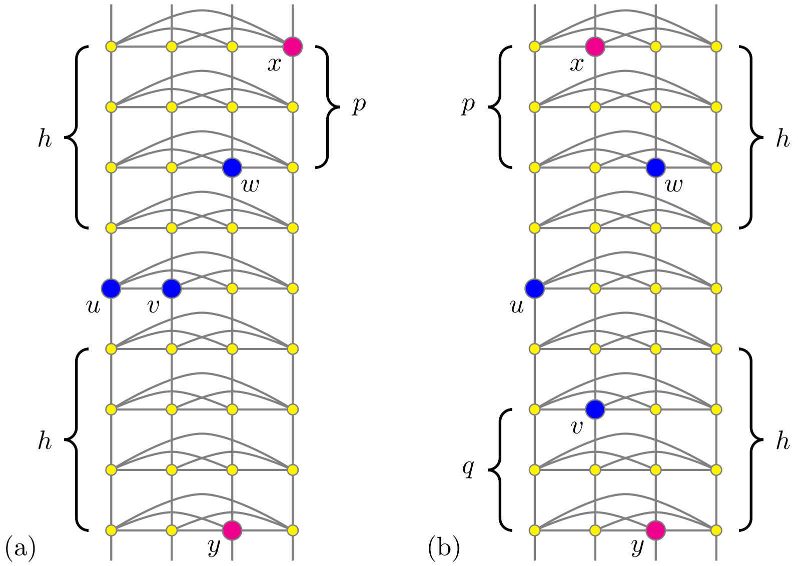


FIG. 8.4. *Illustration for Theorem* 8.6.

Case 2. Now suppose that u, v, w are in different rows. Without loss of generality, u is in the middle row and the first column, and v is in the second column and in a row below u, and w is in the third column and in a row above u. Let x and y be the vertices shown in Figure 8.4(b). Then $d(x, u) = d(y, u) = h + 1$, $d(x, v) = d(y, v) = q$, and $d(x, w) = d(y, w) = p$. Thus S does not resolve x and y, which is the desired contradiction. □

9. Trees. Let v be a vertex of a tree T. Let ℓ_v be the number of components of $T \setminus v$ that are (possibly edgeless) paths. Slater [42], and subsequently a number of other authors [9, 24, 27], proved that for every tree T that is not a path,

$$\beta(T) = \sum_{v \in V(T)} \max\{\ell_v - 1, 0\}. \tag{9.1}$$

A *leaf* of a graph is a vertex of degree 1. The following result for doubly resolving sets in trees is a generalization of Lemma 7.1 for paths.

Lemma 9.1. *The set of leaves L is the unique minimum doubly resolving set for a tree T, and $\psi(T) = |L|$.*

Proof. Every pair of vertices v, w of T lies on a path whose endpoints are leaves x, y. Clearly x, y doubly resolve v, w. Thus L is a doubly resolving set. Say v is a leaf of T whose neighbor is w. Every shortest path from v passes through w. Thus v, w can only be doubly resolved by a pair including v. Thus v is in every doubly resolving set of T. The result follows. □

Theorem 4.1 and Lemma 9.1 imply that for every tree T with k leaves and for every graph G,

$$\beta(T \,\square\, G) \leq \beta(G) + k - 1. \tag{9.2}$$

Moreover, many leaves force up the metric dimension of a cartesian product.

Lemma 9.2. *Every graph G with $k \geq 2$ leaves satisfies $\beta(G \,\square\, G) \geq k$.*

Proof. Let S be a metric basis of $G \,\square\, G$. Let b and w be distinct leaves of G adjacent to a and v, respectively. There is a vertex $xy \in S$ that resolves aw and bv. Suppose on the contrary that $x \neq b$ and $y \neq w$. Thus $d_G(b, x) = d_G(a, x) + 1$ and $d_G(w, y) = d_G(v, y) + 1$. Hence $d_G(a, x) - d_G(b, x) = d_G(v, y) - d_G(w, y) = -1$, which implies that $d_G(a, x) + d_G(w, y) = d_G(b, x) + d_G(v, y)$. That is, $d(aw, xy) = d(bv, xy)$. Thus xy does not resolve aw and bv. This contradiction proves that $x = b$ or $y = w$. Thus for every pair of leaves b, w there is a vertex by or xw in S. Suppose that for some leaf b, there is no vertex $by \in S$. Then for every leaf w, there is a vertex $xw \in S$, and $|S| \geq k$. Otherwise for every leaf b, there is a vertex $by \in S$, and again $|S| \geq k$. □

The following result implies that ψ is not bounded by any function of metric dimension.

Theorem 9.3. *For every integer $n \geq 4$ there is a tree B_n with $\beta(B_n) = 2$ and*

$$n = \psi(B_n) \leq \beta(B_n \,\square\, B_n) \leq n + 1.$$

Proof. Let B_n be the *comb* graph obtained by attaching one leaf at every vertex of P_n. Now $\ell_v = 0$ for every leaf v of B_n, and $\ell_w = 1$ for every other vertex w of B_n, except for the two vertices x and y indicated in Figure 9.1, for which $\ell_x = \ell_y = 2$. Thus $\beta(B_n) = 2$ by (9.1). Since B_n has n leaves, we have $\psi(B_n) = n$ by Lemma 9.1. Moreover, $\beta(B_n \,\square\, B_n) \geq n$ by Lemma 9.2. The upper bound $\beta(B_n \,\square\, B_n) \leq n + 1$ follows from Theorem 4.1. □

Given that the proof of Theorem 9.3 is heavily dependent on the presence of leaves in B_n, it is tempting to suspect that such behavior does not occur among more highly connected graphs. This is not the case.

Theorem 9.4. *For all $k \geq 1$ and $n \geq 2$ there is a k-connected graph $G_{n,k}$ for which $\beta(G_{n,k}) \leq 2k$ and $\beta(G_{n,k} \,\square\, G_{n,k}) \geq n$.*

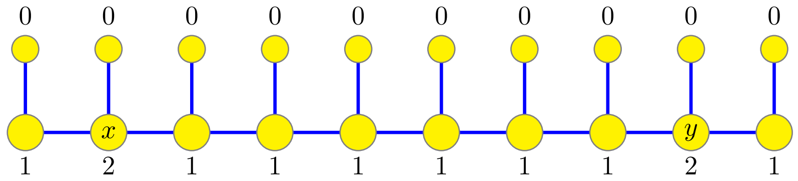


FIG. 9.1. *An illustration of the comb graph B_{10} showing the ℓ-values at each vertex.*

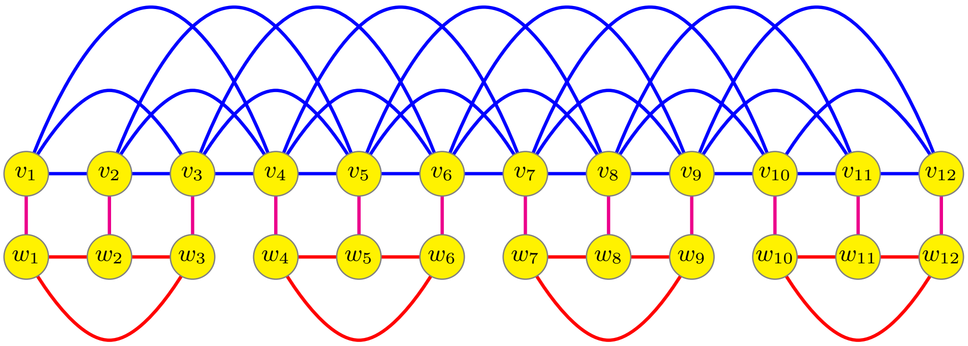


FIG. 9.2. *The construction in Theorem 9.4 with $k = 3$ and $n = 2$.*

Proof. As illustrated in Figure 9.2, let $G_{n,k}$ be the graph with vertex set $\{v_i, w_i : 1 \leq i \leq 2kn\}$, where every v_iw_i is an edge, v_iv_j is an edge whenever $|i - j| \leq k$, and w_iw_j is an edge whenever $\lceil i/k \rceil = \lceil j/k \rceil$. Note that $G_{n,1} = B_{2n}$. Clearly $G_{n,k}$ is k-connected. It is easily seen that $\{v_i, v_{2kn+1-i} : 1 \leq i \leq k\}$ resolves $G_{n,k}$. Thus $\beta(G_{n,k}) \leq 2k$.

Say S doubly resolves $G_{n,k}$. On the contrary, suppose that

$$S \cap \{w_{\ell k+1}, w_{\ell k+2}, \ldots, w_{\ell k+k}\} = \emptyset$$

for some ℓ with $0 \leq \ell \leq 2n-1$. This implies that $d(w_{\ell k+1}, x) = d(v_{\ell k+1}, x)+1$ for every vertex $x \in S$. Hence S does not doubly resolve $w_{\ell k+1}$ and $v_{\ell k+1}$. This contradiction proves that $S \cap \{w_{\ell k+1}, w_{\ell k+2}, \ldots, w_{\ell k+k}\} \neq \emptyset$ for every ℓ with $0 \leq \ell \leq 2n - 1$. Thus $|S| \geq 2n$ and $\psi(G_{n,k}) \geq 2n$. That $\beta(G_{n,k} \,\square\, G_{n,k}) \geq n$ follows from Lemma 4.2. □

We conclude that for all $k \geq 1$, there is no function f such that $\beta(G \,\square\, H) \leq f(\beta(G), \beta(H))$ for all k-connected graphs G and H.

Note added in proof. The metric dimension of the cartesian product of a cycle and a graph was independently studied by Peters-Fransen and Oellermann [33]. They independently proved (8.1), Theorem 8.4, and Theorem 8.6 with $n = 2$.

REFERENCES

[1] N. ALON, D. N. KOZLOV, AND V. H. VU, *The geometry of coin-weighing problems*, in Proceedings of the 37th Annual Symposium on Foundations of Computer Science (FOCS '96), IEEE, 1996, pp. 524–532.

[2] Z. BEERLIOVA, F. EBERHARD, T. ERLEBACH, A. HALL, M. HOFFMANN, M. MIHALAK, AND L. S. RAM, *Network discovery and verification*, in Proceedings of the 31st Workshop on Graph Theoretic Concepts in Computer Science (WG '05), Lecture Notes in Comput. Sci. 3787, D. Kratsch, ed., Springer, Berlin, 2005, pp. 127–138.

[3] A. Bogomolny and D. Greenwell, *Cut the Knot: Invitation to Mastermind*, http://www.maa.org/editorial/knot/Mastermind.html (1999).

[4] R. C. Brigham, G. Chartrand, R. D. Dutton, and P. Zhang, *Resolving domination in graphs*, Math. Bohem., 128 (2003), pp. 25–36.

[5] P. S. Buczkowski, G. Chartrand, C. Poisson, and P. Zhang, *On k-dimensional graphs and their bases*, Period. Math. Hungar., 46 (2003), pp. 9–15.

[6] D. G. Cantor, *Determining a set from the cardinalities of its intersections with other sets*, Canad. J. Math., 16 (1964), pp. 94–97.

[7] D. G. Cantor and W. H. Mills, *Determination of a subset from certain combinatorial properties*, Canad. J. Math., 18 (1966), pp. 42–48.

[8] G. G. Chappell, J. Gimbel, and C. Hartman, *Bounds on the Metric and Partition Dimensions of a Graph*, http://www.cs.uaf.edu/~chappell/papers/metric/ (2003).

[9] G. Chartrand, L. Eroh, M. A. Johnson, and O. R. Oellermann, *Resolvability in graphs and the metric dimension of a graph*, Discrete Appl. Math., 105 (2000), pp. 99–113.

[10] G. Chartrand, C. Poisson, and P. Zhang, *Resolvability and the upper dimension of graphs*, Comput. Math. Appl., 39 (12) (2000), pp. 19–28.

[11] G. Chartrand and P. Zhang, *The forcing dimension of a graph*, Math. Bohem., 126 (2001), pp. 711–720.

[12] G. Chartrand and P. Zhang, *The theory and applications of resolvability in graphs. A survey*, in Proceedings of the 34th Southeastern International Conference on Combinatorics, Graph Theory and Computing, Congr. Numer., 160 (2003), pp. 47–68.

[13] V. Chvátal, *Mastermind*, Combinatorica, 3 (1983), pp. 325–329.

[14] J. Currie and O. R. Oellermann, *The metric dimension and metric independence of a graph*, J. Combin. Math. Combin. Comput., 39 (2001), pp. 157–167.

[15] D. Ž. Djoković, *Distance-preserving subgraphs of hypercubes*, J. Combin. Theory Ser. B, 14 (1973), pp. 263–267.

[16] P. Erdős and A. Rényi, *On two problems of information theory*, Magyar Tud. Akad. Mat. Kutató Int. Közl., 8 (1963), pp. 229–243.

[17] M. Fehr, S. Gosselin, and O. R. Oellermann, *The metric dimension of Cayley digraphs*, Discrete Math., 306 (2006), pp. 31–41.

[18] M. Fehr, S. Gosselin, and O. R. Oellermann, *The partition dimension of Cayley digraphs*, Aequationes Math., 71 (2006), pp. 1–18.

[19] P. Frank and R. Silverman, *Remarks on detection problems*, Amer. Math. Monthly, 74 (1967), pp. 171–173.

[20] W. Goddard, *Static mastermind*, J. Combin. Math. Combin. Comput., 47 (2003), pp. 225–236.

[21] W. Goddard, *Mastermind revisited*, J. Combin. Math. Combin. Comput., 51 (2004), pp. 215–220.

[22] D. L. Greenwell, *Mastermind*, J. Recr. Math., 30 (1999–2000), pp. 191–192.

[23] R. K. Guy and R. J. Nowakowski, *Coin-weighing problems*, Amer. Math. Monthly, 102 (1995), p. 164.

[24] F. Harary and R. A. Melter, *On the metric dimension of a graph*, Ars Combin., 2 (1976), pp. 191–195.

[25] W. Imrich and S. Klavžar, *Product Graphs*, Wiley, New York, 2000.

[26] G. Kabatianski, V. S. Lebedev, and J. Thorpe, *The Mastermind game and the rigidity of Hamming spaces*, in Proceedings of the IEEE International Symposium on Information Theory (ISIT '00), IEEE, 2000, p. 375.

[27] S. Khuller, B. Raghavachari, and A. Rosenfeld, *Landmarks in graphs*, Discrete Appl. Math., 70 (1996), pp. 217–229.

[28] D. E. Knuth, *The computer as master mind*, J. Recr. Math., 9 (1976/1977), pp. 1–6.

[29] B. Lindström, *On a combinatory detection problem.* I, Magyar Tud. Akad. Mat. Kutató Int. Közl., 9 (1964), pp. 195–207.

[30] B. Lindström, *On a combinatorial problem in number theory*, Canad. Math. Bull., 8 (1965), pp. 477–490.

[31] B. Lindström, *On a combinatory detection problem.* II, Studia Sci. Math. Hungar., 1 (1966), pp. 353–361.

[32] W. H. Mow, *Multiuser coding based on detecting matrices for synchronous-CDMA systems*, in Proceedings of the Cryptography and Coding, Lecture Notes in Comput. Sci. 1355, Springer, Berlin, 1997, pp. 251–257.

[33] J. Peters-Fransen and O. R. Oellermann. *The metric dimension of the cartesian product of graphs*, Utilitas Math., 69 (2006), pp. 33–41.

[34] N. Pippenger, *An information-theoretic method in combinatorial theory*, J. Combin. Theory Ser. A, 23 (1977), pp. 99–104.

[35] C. Poisson and P. Zhang, *The metric dimension of unicyclic graphs*, J. Combin. Math. Combin. Comput., 40 (2002), pp. 17–32.

[36] V. Saenpholphat and P. Zhang, *Connected resolvability of graphs*, Czechoslovak Math. J., 53 (2003), pp. 827–840.

[37] V. Saenpholphat and P. Zhang, *Conditional resolvability in graphs: A survey*, Int. J. Math. Math. Sci., 37–40 (2004), pp. 1997–2017.

[38] V. Saenpholphat and P. Zhang, *Detour resolvability of graphs*, in Proceedings of the 35th Southeastern International Conference on Combinatorics, Graph Theory and Computing, Congr. Numer., 169 (2004), pp. 3–21.

[39] V. Saenpholphat and P. Zhang, *On connected resolving decompositions in graphs*, Czechoslovak Math. J., 54 (2004), pp. 681–696.

[40] A. Sebő and E. Tannier, *On metric generators of graphs*, Math. Oper. Res., 29 (2004), pp. 383–393.

[41] B. Shanmukha, B. Sooryanarayana, and K. S. Harinath, *Metric dimension of wheels*, Far East J. Appl. Math., 8 (2002), pp. 217–229.

[42] P. J. Slater, *Leaves of trees*, in Proceedings of the 6th Southeastern Conference on Combinatorics, Graph Theory, and Computing, Congr. Numer., 14 (1975), pp. 549–559.

[43] P. J. Slater, *Dominating and reference sets in a graph*, J. Math. Phys. Sci., 22 (1988), pp. 445–455.

[44] S. Söderberg and H. S. Shapiro, *A combinatory detection problem*, Amer. Math. Monthly, 70 (1963), p. 1066.

[45] B. Sooryanarayana, *On the metric dimension of a graph*, Indian J. Pure Appl. Math., 29 (1998), pp. 413–415.

[46] B. Sooryanarayana and B. Shanmukha, *A note on metric dimension*, Far East J. Appl. Math., 5 (2001), pp. 331–339.

[47] P. M. Winkler, *Isometric embedding in products of complete graphs*, Discrete Appl. Math., 7 (1984), pp. 221–225.

[48] S. V. Yushmanov, *Estimates for the metric dimension of a graph in terms of the diameters and the number of vertices*, Vestnik Moskov. Univ. Ser. I Mat. Mekh., 103 (1987), pp. 68–70.

SIAM J. DISCRETE MATH.
Vol. 21, No. 2, pp. 442–460

RAMSEY PROPERTIES OF RANDOM k-PARTITE, k-UNIFORM HYPERGRAPHS*

VOJTĚCH RÖDL[†], ANDRZEJ RUCIŃSKI[‡], AND MATHIAS SCHACHT[§]

Abstract. We investigate the threshold probability for the property that every r-coloring of the edges of a random binomial k-uniform hypergraph $\mathbb{G}^{(k)}(n,p)$ yields a monochromatic copy of some fixed hypergraph G. In this paper we solve the problem for arbitrary $k \geq 3$ and k-partite, k-uniform hypergraphs G.

Key words. random hypergraphs, Ramsey properties

AMS subject classifications. Primary, 05C55; Secondary, 05C65, 05C80

DOI. 10.1137/060657492

1. Introduction.

Given two hypergraphs, G and F, we write $F \longrightarrow G$ if every two-coloring of the edges of F results in a monochromatic copy of G. We then say that F has *the Ramsey property with respect to* G. Note that for fixed G, this property, viewed as the family of hypergraphs $\{F : F \longrightarrow G\}$, is increasing; that is, it is closed under taking superhypergraphs.

In this paper we study Ramsey properties of random k-uniform hypergraphs. Given $k \geq 2$ and $0 \leq p = p(n) \leq 1$, let $\mathbb{G}^{(k)}(n,p)$ be a random hypergraph obtained by declaring each k-element subset of $\{1, 2, \ldots, n\} = [n]$ an edge, independently, with probability p.

We say that $\mathbb{G}^{(k)}(n,p)$ possesses a property $\mathcal{P}$ *asymptotically almost surely* (*a.a.s.*) if $\mathbb{P}\left(\mathbb{G}^{(k)}(n,p) \in \mathcal{P}\right) \to 1$ as $n \to \infty$. For an increasing hypergraph property $\mathcal{P}$, the most relevant question in the theory of random hypergraphs is to find a threshold sequence $\widehat{p}(n)$ above which the random hypergraph possesses $\mathcal{P}$ a.a.s., while below which it does not possess $\mathcal{P}$ a.a.s. More precisely, we say that property $\mathcal{P}$ has a threshold $\widehat{p}(n)$ if

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbb{G}^{(k)}(n,p) \in \mathcal{P}\right) = \begin{cases} 0 & \text{if} \quad p = o(\widehat{p}), \\ 1 & \text{if} \quad \widehat{p} = o(p)\,. \end{cases}$$

The two parts of the above definition will be referred to as the 0-*statement* and the 1-*statement*, respectively. It is known that for increasing set properties a threshold always exists (see [3] and [12]).

In [18] and [20] thresholds for Ramsey properties of graphs have been found. To state this and other results we need further notation.

*Received by the editors April 18, 2006; accepted for publication (in revised form) November 10, 2006; published electronically May 22, 2007.

http://www.siam.org/journals/sidma/21-2/65749.html

[†] Department of Mathematics and Computer Science, Emory University, Atlanta, GA 30322 (rodl@mathcs.emory.edu). This author was partially supported by NSF grant DMS 0300529.

[‡] Department of Discrete Mathematics, Adam Mickiewicz University, Poznań, Poland (rucinski @amu.edu.pl). This author worked on this project during his frequent, long-term visits to the Department of Mathematics and Computer Science, Emory University, Atlanta, and his research was supported in part by KBN grant 2 P03A 015 23.

[§] Institut für Informatik, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany (schacht@informatik.hu-berlin.de). This author was supported by DFG grant SCHA 1263/1-1.

The numbers of vertices and edges of a hypergraph G will be denoted by v_G and e_G (or $e(G)$), respectively. For a k-uniform hypergraph G with at least one edge, we define the parameters

$$d_G^{(k)} = \begin{cases} \dfrac{e_G - 1}{v_G - k} & \text{if} \quad e_G > 1, \\ \dfrac{1}{k} & \text{if} \quad e_G = 1 \end{cases}$$

and

$$m_G^{(k)} = \max\left\{ d_G^{(k)} : \ H \subseteq G \text{ and } e_H \geq 1 \right\} .$$

Note that for all hypergraphs G with at least one edge

- $m_G^{(k)} > 0$,
- $m_G^{(k)} = 1/k$ if and only if $\Delta(G) = 1$, that is, G consists of isolated edges and vertices,
- $m_G^{(k)} \geq 1/(k-1)$ otherwise.

The parameter $m_G^{(k)}$ is defined in such a way that for $p = \Omega(n^{-1/m_G^{(k)}})$, we have $n^{v_H} p^{e_H} = \Omega(n^k p)$ for each $H \subseteq G$ with $e_H > 0$; that is, the expected number of copies of each subgraph H of G in $\mathbb{G}^{(k)}(n,p)$ is at least of the order of magnitude of the expected number of edges. This seems to be a necessary condition for the property $\mathbb{G}^{(k)}(n,p) \longrightarrow G$ to hold a.a.s. (see [18] for a proof in the graph case).

Below is an abridged version of the threshold theorem for Ramsey properties of random graphs ($k = 2$). For the full version see [12, Theorem 8.1].

THEOREM 1 (see [18, 20]). *Given a graph G, other than a forest, the threshold for the Ramsey property with respect to G is $p(n) = n^{-1/m_G^{(2)}}$. Moreover, there exist constants c and $C > 0$ such that*

$$(1) \qquad \lim_{n\to\infty} \mathbb{P}\left(\mathbb{G}^{(2)}(n,p) \longrightarrow G \right) = \begin{cases} 0 & \textit{if} \quad p \leq cn^{-1/m_G^{(2)}}, \\ 1 & \textit{if} \quad p \geq Cn^{-1/m_G^{(2)}}. \end{cases}$$

As a by-product of our approach in this paper, we obtain a simple proof of the 1-statement of Theorem 1. This proof will be outlined in section 4. As opposed to [20], where a stronger version of Theorem 1 with arbitrarily many colors was shown, our current proof does not require any use of the regularity lemma from [23].

Note that (1) is stronger than what the definition of the threshold says. It has been recently shown in [7] that in the case $G = K_3$ the threshold is even sharper.

The only result about Ramsey properties of random hypergraphs was obtained in [21]. There it is shown that $\lim_{n\to\infty} \mathbb{P}(\mathbb{G}^{(3)}(n,p) \longrightarrow K_4^{(3)}) = 1$ if $p \gg n^{-1/3}$, where $K_4^{(3)}$ is the complete 3-uniform hypergraph on four vertices. (Note that $m_{K_4^{(3)}}^{(3)} = 3$.) That proof used a recent regularity lemma for hypergraphs from [6] and the ideas from [19] and [20].

In this paper we extend the 1-statement of Theorem 1 to the class of k-partite, k-uniform hypergraphs for all $k \geq 2$. Recall that a k-uniform hypergraph is *k-partite* if its vertex set can be partitioned into k nonempty sets in such a way that every edge intersects every set of the partition in exactly one vertex.

Theorem 2. *For all $k \geq 2$ and every k-uniform, k-partite hypergraph G with $\Delta(G) \geq 2$, there exists $C > 0$ such that for every sequence $p = p(n) \geq Cn^{-1/m_G^{(k)}}$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbb{G}^{(k)}(n,p) \longrightarrow G\right) = 1\,.$$

For hypergraphs G with $\Delta(G) = 1$, Theorem 2 is not true (see discussion after the statement of Theorem 9 below). For some other special hypergraphs G, such as stars, the actual threshold is lower than $n^{-1/m_G^{(k)}}$. More precisely, for integers $k, t \geq 2$ and $s \geq 1$ let $S_{s,t}^{(k)}$ denote the star (or Δ-system or sunflower) with t edges in which every two edges intersect in precisely the same set of s vertices (e.g., $S_{1,t}^{(2)} = K_{1,t}$). Clearly, a hypergraph has the Ramsey property with respect to $S_{1,t}^{(2)} = K_{1,t}$ if it contains a copy of $S_{s,2t-1}^{(k)}$ as a subhypergraph. Hence, if $p \gg n^{-k+s-s/(2t-1)}$, then

$$\lim_{n\to\infty} \mathbb{P}\big(\mathbb{G}^{(k)} \longrightarrow S_{s,t}^{(k)}\big) = 1.$$

On the other hand, for $t \geq 2$ we have for $S = S_{s,t}^{(k)}$

$$n^{-1/m_S^{(k)}} = n^{-(t(k-s)+s-k)/(t-1)} \gg n^{-k+s-s/(2t-1)}.$$

However, we believe that the corresponding 0-statement is true with C replaced by a smaller constant c for "most" hypergraphs G. For $k = 2$, a tedious proof was given in [18]. We are convinced that it will be possible to extend it for $k > 2$ and we hope to get back to this in the near future. In fact, in [21] a similar proof of the 0-statement of an analogous threshold result in the vertex-coloring case for hypergraphs was given.

We provide a complete proof of Theorem 2 in section 3. In this proof, the special structure of k-partite hypergraphs allows for replacing the regularity lemma by an old result of Erdős; see Lemma 8 below. As a technical tool, we will actually be proving a stronger theorem, Theorem 9. Being stronger, it will easily imply yet another related result.

We write $F \xrightarrow{\text{ind}} G$ if every two-coloring of the edges of F results in an *induced* monochromatic copy of G. Note that this property is not monotone.

Theorem 3. *For all $\varepsilon > 0$, $k \geq 2$, and every k-uniform, k-partite hypergraph G with $\Delta(G) \geq 2$ there exists $C > 0$ such that for every sequence $p = p(n)$ with $Cn^{-1/m_G^{(k)}} \leq p(n) \leq 1 - \varepsilon$*

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbb{G}^{(k)}(n,p) \xrightarrow{\text{ind}} G\right) = 1\,.$$

All three results, Theorems 1, 2, and 3, have generalizations to an arbitrary number $r \geq 2$ of colors. Interestingly enough, the parameter r does not influence the order of magnitude of the threshold (only the constant C depends on r). In section 4 we outline the proof of such a generalization of Theorem 2. This is possible, because in Theorem 2 we restrict ourselves to k-partite hypergraphs only. In general, even for graphs the proofs for $r \geq 3$ colors are technically more involved. In particular, we are unable to simplify the proof of the r-color version of Theorem 1 from [20], as we do here for $r = 2$.

2. Preliminaries. Unless noted otherwise, throughout the paper all hypergraphs are k-uniform for a fixed integer $k \geq 2$. We will use notation $G - f$ for a hypergraph obtained from G by removing the edge f, and $G + f$ for the hypergraph obtained from G by adding the edge f, where f is a fixed set of k vertices of G.

2.1. Exponentially small probabilities. Let Γ be a finite set, $|\Gamma| = N$, and let $0 \leq p \leq 1$ and $0 \leq M \leq N$, where M is an integer. Then the random subset Γ_p is obtained by including in Γ_p each element of Γ, independently, with probability p. The random subset Γ_M is obtained by selecting uniformly at random one M-element subset of Γ.

By Chernoff's bound (see, e.g., [12, inequality (2.6), page 26]), we have

$$\mathbb{P}(|\Gamma_p| \leq Np - t) \leq \exp\{-t^2/(2Np)\}. \tag{2}$$

Further, let $\mathcal{S}$ be a family of subsets of Γ, and for $A \in \mathcal{S}$, let I_A be the indicator random variable equal to 1 if $A \subset \Gamma_p$ and equal to 0 otherwise. Finally, let $X = \sum_{A\in\mathcal{S}} I_A$ be the random variable counting all subsets belonging to $\mathcal{S}$ which are present in Γ_p. The following inequality may be thought of as an extension of (2) to sums of dependent indicators.

LEMMA 4 (Janson's inequality [11]). *With the above notation, let*

$$\bar{\Delta} = \sum_{A\in\mathcal{S}} \sum_{B\in\mathcal{S}:A\cap B\neq\emptyset} \mathbb{E}(I_A I_B).$$

Then, for all $t \geq 0$

$$\mathbb{P}(X \leq \mathbb{E}X - t) \leq \exp\left\{-\frac{t^2}{2\bar{\Delta}}\right\}. \tag{3}$$

As a useful illustration, let $\Gamma \subset \binom{[n]}{k}$ be a k-uniform hypergraph, and let $\mathcal{S}$ be the family of the edge sets of all copies of a given hypergraph G present in Γ. Then Γ_p is a random subhypergraph of Γ, and X counts the copies of G in Γ_p. Set

$$\Psi_G = n^{v_G} p^{e_G}$$

and

$$\Phi_G = \min\{\Psi_{G'}:\ G' \subseteq G \text{ and } e_{G'} \geq 1\}.$$

Assume that $|\mathcal{S}| \geq cn^{v_G}$ for some $c > 0$. Then $\mathbb{E}X \geq cn^{v_G}p^{e_G} = c\Psi_G$. Note that for every $H \subseteq G$, there are in Γ no more than $n^{2v_G - v_H}$ pairs of copies of G which intersect in a subgraph isomorphic to H. Thus,

$$\bar{\Delta} \leq \sum_{H\subseteq G, e_H\geq 1} n^{2v_G - v_H} p^{2e_G - e_H} \leq 2^{e_G}\frac{\Psi_G^2}{\Phi_G},$$

and, for every $\varepsilon > 0$, by (3) with $t = \varepsilon\mathbb{E}X$,

$$\mathbb{P}(X \leq (1-\varepsilon)\mathbb{E}X) \leq \exp\left\{-\frac{1}{2}\varepsilon^2 c^2 2^{-e_G}\Phi_G\right\}. \tag{4}$$

In our main proof we will also need a stronger property to be held by Γ_p, namely, that with probability very close to one, the number of copies of a given hypergraph remains large even after deleting from Γ_p a fraction of its edges. To this end, for an increasing family $\mathcal{P}$ of subsets of Γ and a nonnegative integer s, define

$$\mathcal{P}_s = \{A \in \mathcal{P} \colon \ \forall B \subseteq A \text{ with } |B| \leq s, \ A \setminus B \in \mathcal{P}\}.$$

Note that $\mathcal{P}_s$ is also increasing.

The following lemma has appeared already in different forms in [20] and [12]. We provide the short proof for completeness.

Lemma 5. *Let Γ be a set of size N. For every $0 < p < 1$, every $0 < \delta < 1$ and $b > 0$ such that $\delta(2 + \log(1/\delta)) \leq b$, every $0 < s \leq \delta Np/2$, and every increasing family $\mathcal{P}$ of subsets of Γ, if $N/\log n \gg p$ and*

$$\mathbb{P}(\Gamma_{(1-\delta)p} \in \neg\mathcal{P}) < \mathrm{e}^{-bNp},$$

then

$$\mathbb{P}(\Gamma_p \in \neg\mathcal{P}_s) < \mathrm{e}^{-0.1\delta^2 Np}.$$

Proof. We will switch to the uniform model Γ_M and utilize the relations between the two probability spaces. Without loss of generality, we may assume that $s = \delta Np/2$. By Chernoff's bound (2),

$$\mathbb{P}(|\Gamma_p| \leq Np - s) \leq \mathrm{e}^{-\delta^2 Np/8}.$$

Hence, for every increasing property $\mathcal{P}$,

$$\mathbb{P}(\Gamma_p \in \neg\mathcal{P}) \leq \mathbb{P}(\Gamma_{Np-s} \in \neg\mathcal{P}) + \mathrm{e}^{-\delta^2 Np/8}.$$

After applying the above inequality to $\mathcal{P}_s$, it remains to estimate $\mathbb{P}(\Gamma_{Np-s} \in \neg\mathcal{P}_s)$. To do this, it is convenient to view Γ_M as a random sequence of M elements, chosen one by one from Γ, uniformly and without replacements. Observe that any subsequence of length $M' \leq M$ generates a random copy of $\Gamma_{M'}$ of its own.

If $\Gamma_{Np-s} \in \neg\mathcal{P}_s$, then, by the definition of $\mathcal{P}_s$, there exists a subsequence of length $Np - 2s$ such that the set of the elements of this subsequence does not have property $\mathcal{P}$. Thus, by Boole's inequality,

$$\mathbb{P}(\Gamma_{Np-s} \in \neg\mathcal{P}_s) \leq \binom{Np-s}{Np-2s}\mathbb{P}(\Gamma_{Np-2s} \in \neg\mathcal{P}).$$

Since $\binom{n}{k} \leq (en/k)^k$ for all n and k, the binomial term can be bounded by

$$\binom{Np-s}{Np-2s} = \binom{Np-s}{s} \leq (2e/\delta)^s.$$

By Pittel's inequality (see, e.g., [12, page 17]),

$$\mathbb{P}(\Gamma_{Np-2s} \in \neg\mathcal{P}) \leq 3\sqrt{N}\mathbb{P}(\Gamma_{(1-\delta)p} \in \neg\mathcal{P}).$$

Hence, by our assumption on δ and b,

$$\begin{aligned}\mathbb{P}(\Gamma_p \in \neg\mathcal{P}_s) &\leq (2e/\delta)^s 3\sqrt{N}\mathrm{e}^{-bNp} + \mathrm{e}^{-\delta^2 Np/8} \\ &\leq 3\sqrt{N}\mathrm{e}^{-bNp/2} + \mathrm{e}^{-\delta^2 Np/8} \leq \mathrm{e}^{-0.1\delta^2 Np},\end{aligned}$$

where the last inequality holds for sufficiently large N. □

2.2. Intersecting copies. Next we prove an elementary result about the number of small subhypergraphs of $\mathbb{G}^{(k)}(n,p)$, with a special structure relevant to our proof of Theorem 2. We will need a simple fact first.

Given a hypergraph H, let X_H be the number of copies of H in $\mathbb{G}^{(k)}(n,p)$. We recall from section 2.1 that the expectation of X_H can be well upper-bounded by

$$\Psi_H = n^{v_H} p^{\mathrm{e}_H}$$

and that

$$\Phi_H = \min\{\Psi_{H'} :\ H' \subseteq H \text{ and } e_{H'} \geq 1\}.$$

Claim 6. *If $\Phi_H \to \infty$, then a.a.s. $X_H \leq 2\mathbb{E}X_H$.*

Proof. By estimates similar to those in the case of random graphs (see, e.g., [12, Lemma 3.5]),

$$\operatorname{Var} X_H = O\left(\sum_{H' \subseteq H, e_{H'} > 0} \frac{(\mathbb{E}X_H)^2}{\Psi_{H'}} \right),$$

and so, by Chebyshev's inequality,

$$\mathbb{P}(X_H > 2\mathbb{E}X_H) \leq \mathbb{P}(|X_H - \mathbb{E}X_H| > \mathbb{E}X_H) \leq \frac{\operatorname{Var} X_H}{(\mathbb{E}X_H)^2} = o(1). \qquad \square$$

Now we are ready to prove the main result of this subsection.

Lemma 7. *Let G be a k-uniform hypergraph with $\Delta(G) \geq 2$, and let T be a union of two copies G_1 and G_2 of G, intersecting in at least one edge. Furthermore, let $\widetilde{T}$ be obtained from T by removing an edge $f \in G_1 \cap G_2$. If $p = p(n) \geq n^{-1/m_G^{(k)}}$, then a.a.s.*

$$X_{\widetilde{T}} \leq 2n^{2v_G - k} p^{2e_G - 2}.$$

Proof. Set $I = G_1 \cap G_2$ and, for every $H \subseteq T$, set $\widetilde{H} = H - f$, regardless of whether $f \in H$ or not. Then

$$\Psi_{\widetilde{T}} = \frac{\Psi_{\widetilde{G}_1} \Psi_{\widetilde{G}_2}}{\Psi_{\widetilde{I}}}. \tag{5}$$

The probability $p = p(n)$ was chosen in such a way that for every $H \subseteq G$, $e_H \geq 1$, we have

$$\Psi_H = n^{v_H} p^{e_H - 1} p \geq n^{v_H} n^{-(e_H - 1)/m_H^{(k)}} p \geq n^k p, \tag{6}$$

and, in particular, for $H = I$, we get

$$\Psi_{\widetilde{I}} = \frac{1}{p} \Psi_I \geq n^k$$

and, consequently,

$$\mathbb{E}X_{\widetilde{T}} \leq \Psi_{\widetilde{T}} \leq n^{2v_G - k} p^{2e_G - 2}.$$

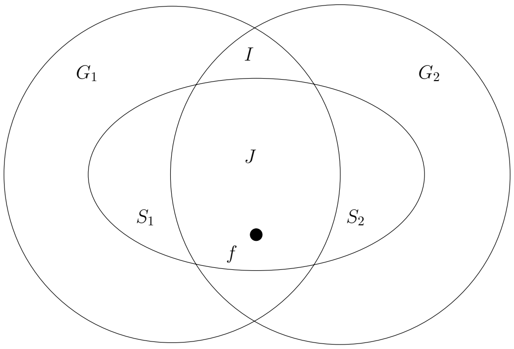


FIG. 1. *Illustration for the proof of Lemma* 7.

Hence, if $\Phi_{\widetilde{T}} \to \infty$, then we are done by Claim 6. On the other hand, by (5), if $n^{-k}\Psi_{\widetilde{I}} \to \infty$, then $\mathbb{E}X_{\widetilde{T}} = o\left(n^{2v_G-k}p^{2e_G-2}\right)$, and, by Markov's inequality,

$$\mathbb{P}(X_{\widetilde{T}} > 2n^{2v_G-k}p^{2e_G-2}) = o(1).$$

It remains to show that either $\Phi_{\widetilde{T}} \to \infty$ or $n^{-k}\Psi_{\widetilde{I}} \to \infty$. Quite arbitrarily, suppose that $\Phi_{\widetilde{T}} \leq \sqrt{n}$. Note that for every $H \subseteq T$ we have

$$\Psi_H = \begin{cases} \Psi_{\widetilde{H}} & \text{if} \quad f \notin H, \\ p\Psi_{\widetilde{H}} & \text{if} \quad f \in H \end{cases}$$

and thus, $\Psi_H \leq \Psi_{\widetilde{H}}$ and $\Phi_T \leq \Phi_{\widetilde{T}} \leq \sqrt{n}$.

Let $\Phi_T = \Psi_S$; that is, S is a subhypergraph of T which achieves the minimum in Φ_T. Set $S_i = S \cap G_i$, $i = 1, 2$, and $J = S_1 \cap S_2$ (see Figure 1). Note that $S \cap I = J$. Note also that $e(S_i) \geq 1$, $i = 1, 2$, since otherwise S would consist of a subgraph S' of G and, possibly, some isolated vertices. However, then we would have, by (6), $\Psi_S \geq \Psi_{S'} \geq n^k p$, and since $\Delta(G) \geq 2$ we have

$$n^k p \geq n^{k-1/m_G^{(k)}} \geq n^{k-(k-1)} = n, \tag{7}$$

a contradiction with the choice of S.

But then, again by (6), we have

$$\Psi_S = \frac{\Psi_{S_1}\Psi_{S_2}}{\Psi_J} \geq \frac{(n^k p)^2}{\Psi_J},$$

which yields that

$$\Psi_J \geq \frac{(n^k p)^2}{\sqrt{n}}.$$

Finally, observe that

$$\Psi_S \leq \Psi_{S\cup I} = \frac{\Psi_S \Psi_I}{\Psi_J},$$

so $\Psi_I \geq \Psi_J$ and, consequently,

$$\Psi_{\widetilde{I}} = \frac{1}{p}\Psi_I \geq n^k \frac{n^k p}{\sqrt{n}} \overset{(7)}{\geq} n^k \sqrt{n}. \qquad \square$$

2.3. Erdős' k-partite counting lemma. For two hypergraphs, F and H, let $N(F,H)$ stand for the number of *labeled copies* of H in F, that is, the number of injective mappings $f : V(H) \to V(F)$ such that if $e \in E(H)$, then $f(e) \in E(F)$. For a fixed labeling on $V(H)$, say, $v_1, \dots, v_{v_H}$, we will identify a labeled copy f of H in F with the sequence $(f(v_1), \dots, f(v_{v_H}))$. We use labeled copies just for convenience, noting that the number of ordinary copies of H in F, that is, the number of subhypergraphs of F which are isomorphic to H, equals $N(F,H)/\mathrm{aut}(H)$.

LEMMA 8. *For every integer $k \geq 2$, every $d > 0$, and every k-uniform, k-partite hypergraph H, there exist $c > 0$ and n_0 such that for every k-uniform hypergraph F on $n \geq n_0$ vertices with $e_F \geq dn^k$, we have $N(F,H) \geq cn^{v_H}$.*

A similar statement was first proved by Erdős in [4] (see also [5]). For completeness we give a short proof.

Proof. It suffices to prove Lemma 8 for complete k-uniform, k-partite hypergraphs H. Let $k \geq 2$ and $d > 0$ be given and let $H = K(\ell_1, \dots, \ell_k)$ be the complete k-uniform, k-partite hypergraph with vertex classes $W_1 \dot\cup \cdots \dot\cup W_k = V(H)$ of sizes $W_i = \ell_i$. (For the sake of defining labeled copies of H in F, we impose on $V(H)$ an arbitrary labeling in which each vertex of W_i precedes each vertex of W_{i+1}, $i = 1, \dots, k-1$.)

Let L_H be the set of indices i for which $\ell_i = |W_i| > 1$, i.e.,

$$L_H = \{i \in [k] \colon \ \ell_i \geq 2\}\,.$$

The proof is by induction on $|L_H|$. The induction base is trivial, as for $|L_H| = 0$ the hypergraph $H = K(1, \dots, 1)$ contains only one edge and we can choose $c = dk!$.

Suppose $|L_H| = \ell > 0$ and Lemma 8 holds for hypergraphs H' with $|L_{H'}| < \ell$. Without loss of generality assume that $k \in L_H$ and consider the subhypergraph $H' = K(\ell_1, \dots, \ell_{k-1}, 1)$. Clearly, $|L_{H'}| = \ell - 1$, and from the induction assumption we infer that

$$N(F,H') \geq c' n^{v(H')} = c' n^{v(H) - \ell_k + 1} \tag{8}$$

for some constant $c' = c'(k,d,H')$. Set $\tilde{\ell} = v(H) - \ell_k$ and consider the set $\mathcal{X}$ of all $\tilde{\ell}$-element sequences of distinct vertices of F. Note that

$$|\mathcal{X}| = n(n-1)\cdots(n-\tilde{\ell}+1) = (n)_{\tilde{\ell}} < n^{\tilde{\ell}}. \tag{9}$$

For a sequence $X = (v_1, \dots, v_{\tilde{\ell}}) \in \mathcal{X}$ we define

$$\deg(X) = \big|\{v \in V(F) \colon \ (v_1, \dots, v_{\tilde{\ell}}, v) \text{ is an labeled copy of } H' \text{ in } F\}\big|\,.$$

Therefore,

$$N(F,H') = \sum_{X \in \mathcal{X}} \deg(X). \tag{10}$$

By Jensen's inequality and by (8), (9), and (10) we conclude that

$$N(F,H) = \sum_{X \in \mathcal{X}} (\deg(X))_{\ell_k} \geq |\mathcal{X}| \left(\frac{N(F,H')}{|\mathcal{X}|}\right)_{\ell_k} \geq (n)_{\tilde{\ell}} \left(\frac{c' n^{\tilde{\ell}+1}}{n^{\tilde{\ell}}}\right)_{\ell_k} \geq cn^{v(H)}$$

for some suitably chosen $c = c(c', H', H) = c(k,d,H)$ and n sufficiently large. □

3. Proof of Theorem 2.

3.1. The idea of the proof. The underlying idea of our proof comes from classical Ramsey theory, where often to force a monochromatic object, a coloring process is put into a dead end. A simplest and best known illustration of this strategy is the proof of the "six-person-party theorem," which says that every 2-coloring of the edges of K_6 results in a monochromatic triangle. In that proof, at some point a vertex is known to be connected to three others by edges of the same color, while the edges between the three neighbors are yet uncolored. But then no matter how they are colored, a monochromatic triangle is guaranteed.

To facilitate this idea in the context of random hypergraphs, we employ the two-round exposure technique (see [12, section 1.1]), where the random hypergraph $\mathbb{G}^{(k)}(n,p)$ is generated in two installments; that is, it is expressed as the union of two independent random hypergraphs $\mathbb{G}_1 = \mathbb{G}^{(k)}(n,p_1)$ and $\mathbb{G}_2 = \mathbb{G}^{(k)}(n,p_2)$ with p_1 and p_2 suitably chosen and such that

$$p_1 + p_2 - p_1 p_2 = p\,. \tag{11}$$

From now on, by a coloring we will always mean a 2-coloring where the colors are *blue* and *red*. For every instance of $\mathbb{G}_1$ and every coloring χ of its edges, we will consider a hypergraph $\Gamma_\chi = \Gamma_\chi(\mathbb{G}_1)$ consisting of all edges $f \notin \mathbb{G}_1$ such that $\mathbb{G}_1 + f$ contains a copy G_f of G, where one edge is f and all other edges are of the same color (in fact, $\mathbb{G}_1$ will contain many such copies; see the precise definition later). Depending on the color of $G_f - f$, we may refer to each $f \in \Gamma_\chi$ as "blue-closing" or "red-closing," and thus express Γ as a union

$$\Gamma_\chi = \Gamma_\chi^{\text{blue}} \cup \Gamma_\chi^{\text{red}}\,,$$

with the obvious meaning of the superscripts. Note that $\Gamma_\chi^{\text{blue}}$ and Γ_χ^{red} are not necessarily disjoint, as an edge f may close blue and red copies of $G - f$ at the same time. We think of Γ_χ as the hypergraph of "closing edges" after round one or, alternatively, as the hypergraph of "useful" edges for round two.

The ultimate goal of the first round is to show that a.a.s. for every χ, either $\Gamma_\chi^{\text{blue}}$ or Γ_χ^{red} contains many copies of G. Say it is the case of Γ_χ^{red}. Then, in the second round, we focus exclusively on the random subhypergraph of Γ_χ^{red}, that is, on

$$\left(\Gamma_\chi^{\text{red}}\right)_{p_2} = \Gamma_\chi^{\text{red}} \cap \mathbb{G}_2,$$

and argue that, with probability very close to 1, at least one copy G_0 of G in Γ_χ^{red} (in fact, many) will be present in $\mathbb{G}_2$. But if this is the case, then there is no way to extend χ without creating a monochromatic copy of G. Indeed, either every edge of G_0 is blue, or an edge $f \in G_0$ is red, turning G_f into a red copy of G.

There is one important refinement to the above simplified argument. Whatever we claim to hold in round two must hold for all colorings χ of the outcome of the first round, $\mathbb{G}_1$. Thus, it must hold with probability so close to 1 that the probability of failure, multiplied by the number of colorings, still converges to 0. Since a.a.s. $e(\mathbb{G}_1) = \Theta(n^k p_1)$, the number of colorings of $\mathbb{G}_1$ is $2^{\Theta(n^k p_1)}$, and we need the probability of having a copy of G in $(\Gamma_\chi^{\text{red}})_p$ to be $1 - \exp\{-\Theta(n^k p_2)\}$. To achieve this goal we will prove first that a.a.s. the number of copies of G in Γ_χ^{red} is $\Theta(n^{v_G})$ and then apply Janson's inequality.

It remains to explain how we prove that a.a.s. Γ_χ^{red} contains $\Theta(n^{v_G})$ copies of G. Since G is k-partite, by Erdős' k-partite counting lemma, Lemma 8, it is enough to show that a.a.s. $|\Gamma_\chi| = \Theta(n^k)$, and then apply Lemma 8 to the majority color class, Γ_χ^{red} or $\Gamma_\chi^{\text{blue}}$.

To show that $|\Gamma_\chi| = \Theta(n^k)$, we will argue that a.a.s. for every coloring χ of $\mathbb{G}_1$ there are $\Theta(n^{v_G})$ monochromatic copies of $\widetilde{G} := G - f_0$, a hypergraph obtained by removing from G one fixed edge f_0. This is how we come across the idea of using induction on e_G. But the induction hypothesis must be stronger than the theorem itself, claiming not one but $\Theta(n^{v_G}p^{e_G})$ monochromatic copies of G in every coloring (see Theorem 9 below).

As a consequence of strengthening Theorem 2, our argument has to be modified slightly. First, we should request that $f \in \Gamma_\chi$ if $\mathbb{G}_1 + f$ contains not one but $\Theta(n^{v_G-k}p_1^{e_G-1})$ copies of G which contain f, and, except for f, are monochromatic. Assume, again, that red is the majority color. Then, after $\mathbb{G}_2$ is exposed, either an extension of coloring χ colors at least $\Theta(n^kp_2)$ edges of $\Gamma_\chi^{\text{red}} \cap \mathbb{G}_2$ red, creating

$$\Theta(n^kp_2 \times n^{v_G-k}p_1^{e_G-1}) = \Theta(n^{v_G}p^{e_G})$$

red copies of G, or not. In the latter case, though, Janson's inequality combined with Lemma 5 guarantees, with probability $1 - \exp\{-\Theta(n^kp_2)\}$, that the remaining blue part of $\Gamma_\chi^{\text{red}} \cap \mathbb{G}_2$ contains $\Theta(n^{v_G}p_2^{e_G})$ copies of G.

Returning to round one, it is a bit tedious to show that having $\Theta(n^{v_G})$ monochromatic copies of $\widetilde{G}$ in $\mathbb{G}_1$ implies $|\Gamma_\chi| = \Theta(n^k)$. The proof involves Jensen's inequality and an upper tail estimate for the number of pairs of copies of $\widetilde{G}$ in $\mathbb{G}_1$ sharing the same nonedge.

As an example, suppose $k = 2$ and $G = C_4$, the four-cycle. Then $\widetilde{G} = P_4$, the path on four vertices, and an edge f belongs to Γ_χ if together with some $\Theta(n^2p_1^3)$ monochromatic copies of P_4 in $\mathbb{G}_1$ it forms a copy of C_4. Hence, many monochromatic copies of P_4 in $\mathbb{G}_1$ will give rise to many edges in Γ_χ, provided that not too many P_4's will "sit" on the same edge f. One way to forbid this is to bound the number of six-cycles C_6 in $\mathbb{G}_1$, which can be viewed as pairs of copies of P_4 that share the same "closing nonedge" but are otherwise disjoint.

3.2. The strengthening. We will, in fact, be proving by induction on e_G the following strengthening of Theorem 2. For a real number $a > 0$, we write $F \xrightarrow{a} G$ if every coloring of the edges of F results in at least $aN(F, G)$ monochromatic copies of G. For example, it is well known that $K_6 \xrightarrow{0.1} K_3$, since every two-coloring of K_6 yields two monochromatic triangles. Note that for given G and a, property $F \xrightarrow{a} G$ is not a monotone property of F.

Theorem 9. *For all $k \geq 2$ and every k-uniform, k-partite hypergraph G with at least one edge there exist $C \geq 1$ and $a > 0$ such that if $p = p(n) > Cn^{-1/m_G^{(k)}}$, $n^kp \to \infty$ but $p \to 0$, then*

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbb{G}^{(k)}(n,p) \xrightarrow{a} G\right) = 1\,.$$

By a standard application of the second moment method, it can be easily proved that in the above range of p, $\mathbb{G}^{(k)}(n,p)$ contains at least one copy (in fact, $\Theta(n^{v_G}p^{e_G})$ copies) of G. Hence, Theorem 9 does, indeed, imply Theorem 2. (We may assume that $p \to 0$, since the Ramsey property in Theorem 2 is increasing.) Although Theorem 9 is about a nonmonotone property, it is also true for p constant, a fact which we will not need here.

Another consequence of Theorem 9 is Theorem 3—the induced version of Theorem 2. Indeed, if $p \to 0$, then a.a.s. only $o(n^{v_G} p^{e_G})$ copies of G in $\mathbb{G}^{(k)}(n,p)$ are not induced. Thus, in view of Theorem 9, a.a.s. for every coloring of $\mathbb{G}^{(k)}(n,p)$ there is at least one (in fact, many) induced copy of G which is monochromatic.

To prove Theorem 3 also for $p < 1$ constant, we argue as follows. By the result from [1, 17], there exists a hypergraph F such that $F \xrightarrow{\text{ind}} G$. For p constant it is easy to show that a.s.s. there is at least one induced copy of F in $\mathbb{G}^{(k)}(n,p)$ (see [2] for the graph case), and thus every coloring of $\mathbb{G}^{(k)}(n,p)$ produces an induced, monochromatic copy of G.

Our proof of Theorem 9 is by induction on e_G, the number of edges in G, and it is convenient to begin with the case $e_G = 1$. (This is why, unlike in Theorem 2, we included here the case $\Delta(G) = 1$.) But then $m_G^{(k)} = 1/k$ and thus, for $p = \Theta(n^{-1/m_G^{(k)}})$, the expected number of edges in $\mathbb{G}^{(k)}(n,p)$ equals $\Theta(n^k p) = \Theta(1)$. This is why we added the assumption that $n^k p \to \infty$. Note that in this case C is irrelevant. As another convenience, in Theorem 9 we require that $C \geq 1$, which is not a restriction at all.

3.3. The case $\boldsymbol{\Delta(G) = 1}$. To begin the inductive proof of Theorem 9, let $\Delta(G) = 1$, which includes the case $e_G = 1$. The following two properties are true for all $p = p(n)$ satisfying $n^k p \to \infty$. The random variable $e(\mathbb{G}^k(n,p))$ has the binomial distribution with $\mathbb{E}e(\mathbb{G}^k(n,p)) = \binom{n}{k}p < n^k p$ and $\operatorname{Var} e(\mathbb{G}^k(n,p)) = \binom{n}{k}p(1-p)$. Hence, by Chebyshev's inequality

$$(12)\qquad \mathbb{P}\left(|e(\mathbb{G}^k(n,p)) - p\tbinom{n}{k}| > \tfrac{1}{2}p\tbinom{n}{k}\right) \leq \frac{\operatorname{Var} e(\mathbb{G}^k(n,p))}{(\frac{1}{2}p\binom{n}{k})^2} < \frac{4}{p\binom{n}{k}} = o(1).$$

For each $\ell \geq 2$, let X_ℓ be the number of (unordered) ℓ-tuples of distinct edges of $\mathbb{G}^{(k)}(n,p)$, not all of which are pairwise disjoint. We have $\mathbb{E}X_\ell \leq n^{\ell k - 1} p^\ell$, and by Markov's inequality,

$$\mathbb{P}(X_\ell > (n^k p)^\ell / \sqrt{n}) \leq 1/\sqrt{n}.$$

Hence, a.a.s., we have $e(\mathbb{G}^k(n,p)) > \frac{1}{2}\binom{n}{k}p$, and, taking $\ell = e_G$, $X_{e_G} \leq (n^k p)^{e_G}/\sqrt{n}$. Consequently, a.a.s., after coloring the edges of $\mathbb{G}^{(k)}(n,p)$, the number of e_G-tuples of edges of $\mathbb{G}^{(k)}(n,p)$ which are pairwise disjoint and monochromatic (in the majority color alone) is, a.a.s., at least

$$\binom{\frac{1}{4}\binom{n}{k}p}{e_G} - \frac{1}{\sqrt{n}}(n^k p)^{e_G} = (1 - o(1))\binom{\frac{1}{4}\binom{n}{k}p}{e_G}.$$

Each set of e_G pairwise disjoint edges can be extended to $\binom{n - ke_G}{v_G - ke_G}$ copies of G, by adding $v_G - ke_G$ arbitrary vertices. Therefore, a.a.s., for every coloring there are at least

$$(1 - o(1))\binom{\frac{1}{4}\binom{n}{k}p}{e_G} \times \binom{n - ke_G}{v_G - ke_G} > an^{v_G} p^{e_G}$$

monochromatic copies of G for some constant $a > 0$ independent of n. This proves Theorem 9 for all graphs with $\Delta(G) = 1$.

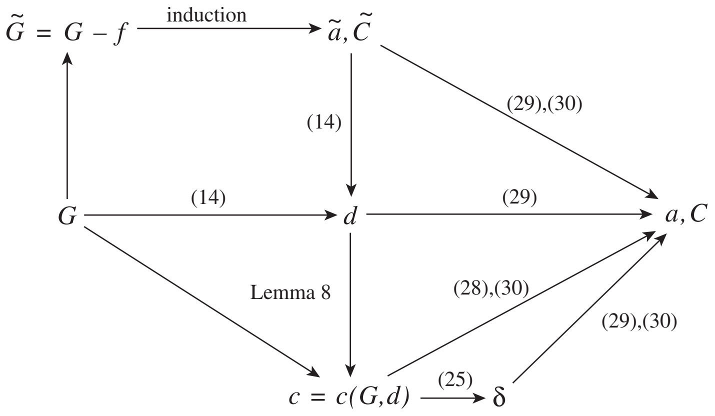


FIG. 2. *Flowchart of constants for the proof of Theorem* 9.

3.4. Mainstream proof. The proof of the induction step requires several constants. We will specify those constants later in the proof instead of defining them all up front. We believe this eases the reading. The dependencies of the main constants are given in Figure 2.

Assume that $e_G \geq 2$ and $p \geq Cn^{-1/m_G^{(k)}}$, where C will be specified later (see (30)). Let p_1 and p_2 be suitably chosen (see (27) and (31)), so that p_2 is sufficiently larger than p_1 and (11) holds. Throughout the proof we will assume that p, p_1, and p_2 all tend to 0 as $n \to \infty$. As before, we will be using abbreviated notation $\mathbb{G} = \mathbb{G}^{(k)}(n,p)$, $\mathbb{G}_1 = \mathbb{G}^{(k)}(n,p_1)$, and $\mathbb{G}_2 = \mathbb{G}^{(k)}(n,p_2)$.

For a suitably selected constant $a > 0$ (see (28) and (29) below), let BAD be the event that there is a coloring of the edges of $\mathbb{G}$ with less than $an^{v_G}p^{e_G}$ monochromatic copies of G. Since by (12), a.a.s. $e(\mathbb{G}) = \Theta(n^k p)$, Theorem 9 is equivalent to the fact that $\mathbb{P}(\text{BAD}) = o(1)$ for some $a > 0$.

Fix an arbitrary edge f_0 of G and let $\widetilde{G}$ be the hypergraph obtained from G by removing the edge f_0. By the induction assumption applied to $\widetilde{G}$, there exist $\widetilde{a}$ and $\widetilde{C}$ such that if $p \geq \widetilde{C}n^{-1/m_{\widetilde{G}}^{(k)}}$, then $\lim_{n\to\infty} \mathbb{P}(\mathbb{G}^{(k)}(n,p) \xrightarrow{\widetilde{a}} \widetilde{G}) = 1$.

For a copy $\widetilde{G}'$ of $\widetilde{G}$ in $\mathbb{G}_1$, let $\text{cl}(\widetilde{G}')$ be the set of edges $f \in K_n^{(k)}$ such that $\widetilde{G}' + f$ is isomorphic to G. For a coloring χ of $\mathbb{G}_1$, we define the auxiliary hypergraph

$$\Gamma_\chi^{\text{blue}} = \{f \in K_n^{(k)} \setminus \mathbb{G}_1\colon\ |\{\widetilde{G}' \subseteq \mathbb{G}_1\colon\ f \in \text{cl}(\widetilde{G}') \text{ and } \widetilde{G}' \text{ is blue copy of } \widetilde{G}\}| \geq z\},$$

where

$$z = \frac{\widetilde{a}}{2} n^{v_G - k} p_1^{e_G - 1}\,. \tag{13}$$

We set

$$\Gamma_\chi = \Gamma_\chi^{\text{blue}} \cup \Gamma_\chi^{\text{red}}\,,$$

where Γ_χ^{red} is defined as $\Gamma_\chi^{\text{blue}}$ with the word "blue" replaced by "red." Further, let

$$d = \frac{\widetilde{a}^2}{2^{\binom{2v_G - k}{k} + 8} v_G^k}\,, \tag{14}$$

and let GOOD be the event that

$$e(\mathbb{G}_1) < n^k p_1 \tag{15}$$

and, for every coloring χ of $\mathbb{G}_1$,

$$\max\{|\Gamma_\chi^{\text{red}}|, |\Gamma_\chi^{\text{blue}}|\} \geq dn^k\,. \tag{16}$$

Note that GOOD is not the complement of BAD.

Conditioning on $\mathbb{G}_1$ and fixing some coloring χ of $\mathbb{G}_1$, let BAD_χ be the event that there is an extension of χ, $\bar{\chi}\colon\ \mathbb{G} \to \{\text{blue, red}\}$, with fewer than $an^{v_G}p^{e_G}$ monochromatic copies of G in both colors. We will later verify the following two facts.

Fact 10. *The event* GOOD *holds a.a.s.*

Fact 11. *For every* $\mathbb{G}_1 \in$ GOOD *and every coloring* χ *of* $\mathbb{G}_1$,

$$\mathbb{P}(\text{BAD}_\chi \mid \mathbb{G}_1) \leq \mathrm{e}^{-n^k p_1}\,.$$

Assuming these two facts, we may easily complete the proof of Theorem 9. Indeed, we have

$$\mathbb{P}(\text{BAD}) \leq \mathbb{P}(\neg\text{GOOD}) + \sum_{\mathbb{G}_1 \in \text{GOOD}} \mathbb{P}(\text{BAD} \mid \mathbb{G}_1)\mathbb{P}(\mathbb{G}_1)\,,$$

and, for every $\mathbb{G}_1 \in$ GOOD,

$$\mathbb{P}(\text{BAD} \mid \mathbb{G}_1) = \sum_{\chi} \mathbb{P}(\text{BAD}_\chi \mid \mathbb{G}_1) \leq 2^{n^k p_1}\mathbb{P}(\text{BAD}_{\chi_0} \mid \mathbb{G}_1)\,,$$

where the summation is taken over all, at most $2^{n^k p_1}$, colorings χ of the edges of $\mathbb{G}_1$, and χ_0 maximizes the conditional probability. Therefore, by Facts 10 and 11,

$$\mathbb{P}(\text{BAD}) \leq o(1) + (2/e)^{n^k p_1} = o(1).$$

3.5. Round one: Proof of Fact 10. Due to the choice of p_1 (cf. (31)) and the concentration of the number of edges in $\mathbb{G}_1$ as given in (12), $\mathbb{G}_1$ a.a.s. contains at most $n^k p_1$ edges as claimed in (15). For the rest of this subsection our goal will be to prove that a.a.s. (16) also holds.

Since $m^{(k)}_{\widetilde{G}} \leq m^{(k)}_G$, we have by (31) $p_1 \geq \widetilde{C} n^{-1/m^{(k)}_{\widetilde{G}}}$, and we are in position to apply the induction assumption, that is, Theorem 9, to $\widetilde{G}$. Consequently, a.a.s., for every coloring χ of the edges of $\mathbb{G}_1$ there is a color (say, red) such that at least

$$\ell := \frac{\widetilde{a}}{2} n^{v_G} p_1^{e_G - 1} \tag{17}$$

copies of $\widetilde{G}$ are colored red in $\mathbb{G}_1$. Note that, by (13), (31), and (14),

$$\ell = zn^k \geq \frac{\widetilde{a}}{2} n^k \geq 8dn^k. \tag{18}$$

For every $f \in K_n^{(k)}$, let x_f be the number of red copies $\widetilde{G}'$ of $\widetilde{G}$ in $\mathbb{G}_1$ for which $f \in \text{cl}(\widetilde{G}')$. Then, a.a.s.

$$\sum_{f \in K_n^{(k)}} x_f \geq \ell. \tag{19}$$

Let $\mathcal{T} = \{T_1, T_2, \ldots, T_t\}$ be the family of all pairwise nonisomorphic hypergraphs which are the unions of two copies of $\widetilde{G}$, say $\widetilde{G}' \cup \widetilde{G}''$, with the property that there is a k-tuple f which makes both $\widetilde{G}' + f$ and $\widetilde{G}'' + f$ isomorphic to G. We will say that f is a *common G-closing nonedge* of $\widetilde{G}'$ and $\widetilde{G}''$. Clearly, $|\mathcal{T}|$ does not exceed the number of all graphs on $2v_G - k$ vertices, that is,

$$t := |\mathcal{T}| \leq 2^{\binom{2v_G-k}{k}}.$$

Then, by Lemma 7, $\mathbb{G}_1$ contains a.a.s. at most

$$2tn^{2v_G-k}p_1^{2e_G-2} \tag{20}$$

copies of members of $\mathcal{T}$. As $|\operatorname{cl}(\widetilde{G})| \leq \binom{v_G}{k} < v_G^k$, a particular copy of a graph from $\mathcal{T}$ may be obtained as a union of two copies of $\widetilde{G}$ with a common G-closing nonedge in at most v_G^k ways. Thus, a.a.s.

$$\sum_{f \in K_n^{(k)}} \binom{x_f}{2} < 2tv_G^k n^{2v_G-k}p_1^{2e_G-2}. \tag{21}$$

Let $Z \subseteq K_n^{(k)}$ be the set of edges f such that $x_f \geq z$. We are going to show that

$$|Z| \geq 2dn^k. \tag{22}$$

First, observe that

$$\sum_{f \in K_n^{(k)} \setminus Z} x_f < z\binom{n}{k} \leq \frac{\ell}{2}$$

and consequently, in view of (19),

$$\sum_{f \in Z} x_f \geq \frac{\ell}{2}. \tag{23}$$

If $|Z| \geq \ell/4$, then, by (18), inequality (22) holds. Assuming that $|Z| \leq \ell/4$, we derive by Jensen's inequality and by (23) that

$$\sum_{f \in Z} \binom{x_f}{2} \geq |Z| \binom{\frac{\sum_{f \in Z} x_f}{|Z|}}{2} \geq |Z| \binom{\frac{\ell}{2|Z|}}{2} \geq \frac{\ell^2}{16|Z|} = \frac{\widetilde{a}^2 n^{2v_G} p_1^{2e_G-2}}{64|Z|},$$

which, by (21) and (14), yields (22) again. Thus, by (15) and the fact that $p_1 \to 0$, a.a.s.

$$|\Gamma_\chi^{\text{red}}| \geq 2dn^k - |\mathbb{G}_1| > (2d - p_1)n^k \geq dn^k,$$

and the property GOOD holds.

3.6. Round two: Proof of Fact 11. We condition on the event that $\mathbb{G}_1$ satisfies property GOOD. Let a coloring χ of the edges of $\mathbb{G}_1$ be given. According to property (16), let, say, $|\Gamma_\chi^{\text{red}}(\mathbb{G}_1)| \geq dn^k$. Set $\Gamma^{\text{red}} = \Gamma_\chi^{\text{red}}(\mathbb{G}_1)$.

Let $c = c(G, d)$ be given by Lemma 8. Hence $N(\Gamma^{\mathrm{red}}, G) \geq cn^{v_G}$. Later we use Lemma 5. Therefore, we first consider a random subhypergraph Γ^{red}_q, with

$$q := (1-\delta)p_2\,, \tag{24}$$

where $\delta > 0$ is so small that

$$\delta(2 - \log\delta) < b := \frac{c^2}{400 \cdot 2^{e_G}}. \tag{25}$$

We want to apply Janson's inequality (in the form of inequality (4)) with $\varepsilon = 0.1$, $\Gamma = \Gamma^{\mathrm{red}}$, and $X = N(\Gamma^{\mathrm{red}}_q, G)$. Note that $\mathbb{E}[X] \geq cn^{v_G}q^{e_G}$. By (24), (31), and (27), we have $q \geq n^{-1/m_G^{(k)}}$ and, consequently, $n^{v_K}q^{e_K} \geq n^k q$ for every $K \subseteq G$ with $e_K \geq 1$. Hence,

$$\begin{aligned}\mathbb{P}(X \leq 0.9cn^{v_G}q^{e_G}) &\leq \mathbb{P}(X \leq 0.9\mathbb{E}X)\\ &\leq \exp\left(-\frac{c^2n^kq}{200 \cdot 2^{e_G}}\right) \leq \mathrm{e}^{-bn^kp_2} \leq \mathrm{e}^{-b|\Gamma^{\mathrm{red}}|p_2}\,,\end{aligned}$$

where we also use the fact that $\delta < 1/2$.

Next we apply Lemma 5 with

$$s := \delta(dn^k)p_2/2 \leq \delta|\Gamma^{\mathrm{red}}|p_2/2\,. \tag{26}$$

We conclude that, for sufficiently large n, with probability at least

$$1 - 3\sqrt{|\Gamma^{\mathrm{red}}|}\mathrm{e}^{-b|\Gamma^{\mathrm{red}}|p_2/2} - \mathrm{e}^{-\delta^2|\Gamma^{\mathrm{red}}|p_2/8} \geq 1 - \mathrm{e}^{-\delta^2cn^kp_2/10} \geq 1 - \mathrm{e}^{-n^kp_1}$$

we have

$$N(\Gamma^{\mathrm{red}}_{p_2} \setminus D, G) \geq 0.9cn^{v_G}q^{e_G} \geq 0.8cn^{v_G}p_2^{e_G}$$

for all $D \subseteq \Gamma^{\mathrm{red}}_{p_2}$ of size $|D| \leq s$. For the last inequality in the above bound on probability, we need the relation

$$p_2 \geq 10p_1/(c\delta^2). \tag{27}$$

We will now verify that, with probability at least $1 - \mathrm{e}^{-n^kp_1}$, for every extension $\bar{\chi}$ of the coloring χ, $\Gamma^{\mathrm{red}}_{p_2}$ either contains at least $an^{v_G}p^{e_G}$ blue copies of G or it completes at least $an^{v_G}p^{e_G}$ red copies of G in $\mathbb{G}$.

Let D be the set of edges of $\Gamma^{\mathrm{red}}_{p_2}$ colored red by $\bar{\chi}$. If $|D| < s$, then, by the above property and with a suitably chosen a, there are at least

$$0.8cn^{v_G}p_2^{e_G} \geq an^{v_G}p^{e_G} \tag{28}$$

copies of G in $\Gamma^{\mathrm{red}}_{p_2} \setminus D$, all of them blue.

If, on the other hand, $|D| \geq s$, then, as each edge of Γ^{red} closes at least z red copies of $\widetilde{G}$ in $\mathbb{G}_1$, there are, with a suitably chosen a, at least

$$\frac{s \times z}{v_G^k} \geq \frac{\delta dn^kp_2 \times \widetilde{a}n^{v_G-k}p^{e_G-1}}{4v_G^k} \geq an^{v_G}p^{e_G} \tag{29}$$

red copies of G in $\mathbb{G}$.

To complete the proof, we choose

$$C = \widetilde{C}\left(\frac{10}{c\delta^2} + 1\right), \tag{30}$$

where δ is defined by (25) and $c = c(G, d)$ comes from Lemma 8. Then, with

$$p_1 = \widetilde{C} n^{-1/m_G^{(k)}}, \tag{31}$$

(27) is satisfied. We leave the determination of the constant a for an interested reader.

4. Outlines of other proofs.

4.1. Theorem 1: Two colors. The proof we present here follows the main strategy of the proof from [20] but avoids the use of the regularity lemma. Therefore, rather than outlining the whole proof, we just point out how it differs from the original argument. To this end, we first give a sketch of the original proof in [20], in a simplified version for $r = 2$ colors.

One of the ingredients of the proof of Theorem 1 in [20] was the following simple result which could be viewed as an extension of Lemma 8 to the nonpartite case, but limited to graphs only.

For $0 < d \leq 1$ and $0 < \rho \leq 1$ we call an n-vertex graph F (ρ, d)*-dense* if every induced subgraph of F on $v = \lceil \rho n \rceil$ vertices contains at least $d\binom{v}{2}$ edges.

LEMMA 12 (see [20]). *For every $d > 0$ and every graph H there exist $\rho > 0$ and $c > 0$ such that for every n-vertex (ρ, d)-dense graph F we have $N(F, H) \geq cn^{v_H}$.*

Thus, Lemma 8 specifies that for bipartite graphs H, Lemma 12 holds with $\rho = 1$.

The original proof of Theorem 1, similarly to the above presented proof of Theorem 9, was based on induction on e_G and the two-round exposure technique. Applying the induction assumption to all induced subgraphs of $\mathbb{G}(n, p_1)$ on ρn vertices, viewed as random graphs on their own, resulted in showing that a.a.s., for every coloring χ, the graph Γ_χ was (ρ, d)-dense.

Then, by an application of Szemerédi's regularity lemma for graphs, it was shown that either $\Gamma_\chi^{\text{blue}}$ or Γ_χ^{red} contained a (ρ', d')-dense subgraph F with some new parameters. By Lemma 12 with $H := G$, the graph F contained lots of copies of G and from that point on, the proof went along the same lines as the proof of Theorem 9.

Now, we describe how one can avoid the use of the regularity lemma. The crucial change is to apply Lemma 12 directly to the graph $F = \Gamma_\chi$ with $H = K_R$, where $R = R(G)$ is the Ramsey number for the graph G. As a result, Γ_χ contains $\Theta(n^R)$ copies of K_R. Consequently, by the definition of $R(G)$, the partition $\Gamma_\chi = \Gamma_\chi^{\text{blue}} \cup \Gamma_\chi^{\text{red}}$ contains $\Theta(n^{v_G})$ copies of G in one class, say Γ_χ^{red}, and the proof can be completed as before.

4.2. Theorem 2: More colors. As we have mentioned in section 1, Theorems 1, 2, and 3 remain true for $r \geq 3$ colors, but the proofs become more technical. While for $r > 2$, the r-colored version of Theorem 1 seems to be much harder to prove than the 2-colored version, for Theorem 2 the proofs of these two cases do not differ essentially.

Below we outline the proof of the r-colored version of Theorem 2, $r \geq 2$. We write $F \longrightarrow (G, r)$ if every r-coloring of the edges of F results in a monochromatic copy of G.

Theorem 13. *For all $k \geq 2$ and $r \geq 2$ and for every k-uniform, k-partite hypergraph G with $\Delta(G) \geq 2$ there exists $C > 0$ such that for every sequence $p = p(n) \geq Cn^{-1/m_G^{(k)}}$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbb{G}^{(k)}(n,p) \longrightarrow (G,r)\right) = 1\,.$$

For two colors we argued that in round two, either $\Gamma_{p_2}^{\text{red}}$ had many edges colored red, or it contained many copies of G colored blue. With more colors we may only claim that either $\Gamma_{p_2}^{\text{red}}$ has many edges colored red, or not so many. Since, in view of Lemma 5, these few red edges can be deleted, this calls for induction on the number of colors r.

To make this idea work, we have to generalize and strengthen the statement of Theorem 13 in three ways. First, note that $\Gamma_{p_2}^{\text{red}}$ is a random subhypergraph of an incomplete hypergraph Γ^{red}. Hence, for the sake of induction, we must generalize our statement to random subhypergraphs F_p of dense hypergraphs F. But then, not every closing nonedge is in F, and we more appropriately restrict our attention to those monochromatic copies of $\widetilde{G}$ in F whose complements are also in F. We call such copies of G *nested.* We write

$$F \xrightarrow[\text{nested}]{a} (G,r)$$

if every r-coloring of the edges of F results in at least $aN(F,G)$ nested, monochromatic copies of G.

Finally, since the second round will now be successful if our statement holds for $r-1$ colors, the probability of the failure must be, as all failures in round two, exponentially small (to beat the number of colorings χ from the first round). All in all, we are to prove the following statement.

Theorem 14. *For all integers $k \geq 2$ and $r \geq 1$, every k-uniform, k-partite hypergraph G with at least one edge, and for every real $0 < d \leq 1$, there exist positive numbers a, b, C, and n_0 such that if*

(i) *$n > n_0$,*

(ii) *F is a k-uniform hypergraph with $e_F \geq dn^k$, and*

(iii) *$p = p(n) > Cn^{-1/m_G^{(k)}}$,*

then

$$\mathbb{P}\left(F_p \xrightarrow[\text{nested}]{a} (G,r)\right) > 1 - \mathrm{e}^{-be_F p}\,.$$

The proof of Theorem 14 is by double induction on r and e_G. The case $e_G = 1$ or, more generally, $\Delta(G) = 1$, is practically the same as in the proof of Theorem 9, while the case $r = 1$ relies on Lemma 8 and Janson's inequality (4).

The proof of the induction step boils down to showing analogues of Facts 10 and 11, except that now also Fact 10 must hold with probability exponentially close to 1. The most difficult part is then to prove that (20) holds with probability exponentially close to 1, for which we apply a technique for bounding upper tails of subgraph counts called the deletion method (see Lemma 2.51 in [12] and also [13]), combined with Lemma 5.

We also employ Lemma 5, as before, inside the proof of the analogue of Fact 11. This is no longer preceded by Janson's inequality, but, instead, the induction's hypothesis with $r-1$ colors. In a sense, Janson's inequality is equivalent to Theorem 14 for $r = 1$.

5. Open problems.

The main problem which remains open is to prove Theorem 2 for arbitrary (not necessarily k-partite) k-uniform hypergraphs G. To do so, we need to find the right notion of a *dense* hypergraph F, for which, on the one hand, an extension of Lemma 12 holds, while on the other hand, it could be proved that Γ_χ (cf. section 4.1) is dense in the sense of that new concept.

Another related problem is to find threshold probabilities for the Turán properties of $\mathbb{G}^{(k)}(n,p)$. For a k-uniform hypergraph G, let

$$\mathrm{ex}(n,G) = \max\{\mathrm{e}(F)\colon\ F \text{ is a } k\text{-uniform hypergraph, } G \nsubseteq F, \text{ and } v(F)=n\},$$

and let $\pi(G) = \lim_{n\to\infty} \mathrm{ex}(n,F)/\binom{n}{k}$. It is well known that the limit $\pi(G)$ exists for every G (see, e.g., [14]). For example, Lemma 8 implies that $\pi(G)=0$ for every k-partite, k-uniform hypergraph G.

Given a hypergraph G, we say that a family of hypergraphs $\mathcal{F}$ has the *Turán property* if for every $\delta > 0$ every sufficiently large hypergraph $F \in \mathcal{F}$ has the property that every subhypergraph F' of F with $e(F') \geq (\pi(G)+\delta)e(F)$ contains a copy of G. In the case of random graphs, i.e., $\mathcal{F} = \{\mathbb{G}(n,p)\colon\ n \in \mathbb{N}\}$, thresholds for Turán properties were established so far only for very few cases, including odd and even cycles [10, 9], and small cliques K_4 and K_5 [8, 15] (see also [22, 16] for weaker bounds for general graphs G). This experience with random graphs suggests that Turán thresholds should coincide with those for Ramsey properties.

As opposed to Ramsey properties, the 0-statements for Turán properties are rather easy. Indeed, we know that for $p = o(n^{-1/m_G^{(k)}})$, there are in $\mathbb{G}^{(k)}(n,p)$ a.a.s. $o(n^k p)$ copies of the least likely (the densest) subgraph H of G. These copies, and thus all copies of G in $\mathbb{G}^{(k)}(n,p)$, can be destroyed by removing $o(n^k p)$ edges. This shows that for $p = o(n^{-1/m_G^{(k)}})$, the Turán property with respect to G does not hold a.a.s. (see [12, section 8.1] for the case $k=2$).

The real challenge is the 1-statement, but, in view of Lemma 8, we believe that similarly to Ramsey properties, the case of k-partite G is somewhat easier. In particular, the following conjecture seems to be true.

CONJECTURE 15. *For all integers $k \geq 2$, for every k-partite, k-uniform hypergraph G, and for all $\delta > 0$ there exists $C > 0$ such that if $p \geq Cn^{-1/m_G^{(k)}}$, then a.a.s. every subhypergraph F of $\mathbb{G}^{(k)}(n,p)$ with $e(F) \geq \delta e(\mathbb{G}^{(k)}(n,p))$ contains a copy of G. In particular, if $pn^{1/m_G^{(k)}} \to \infty$, then a.a.s. $\mathbb{G}^{(k)}(n,p)$ has the Turán property with respect to G.*

For $k=2$ (graphs), the conjecture was proved only for even cycles in [9]. It would be most interesting to settle it for $G = K_{3,3}$. For $k \geq 3$ nothing is known, except that for p constant, Lemma 8 implies the conclusion of the above conjecture for all k-partite, k-uniform hypergraphs G, $k \geq 2$.

REFERENCES

[1] F. G. Abramson and L. A. Harrington, *Models without indiscernibles*, J. Symbolic Logic, 43 (1978), pp. 572–600.

[2] B. Bollobás, *Random Graphs*, Cambridge Stud. Adv. Math. 73, 2nd ed., Cambridge University Press, Cambridge, UK, 2001.

[3] B. Bollobás and A. Thomason, *Threshold functions*, Combinatorica, 7 (1987), pp. 35–38.

[4] P. Erdős, *On extremal problems of graphs and generalized graphs*, Israel J. Math., 2 (1964), pp. 183–190.

[5] P. Erdős and J. Spencer, *Probabilistic Methods in Combinatorics*, Probability and Mathematical Statistics 17, Academic Press, New York, London, 1974.

[6] P. Frankl and V. Rödl, *Extremal problems on set systems*, Random Structures Algorithms, 20 (2002), pp. 131–164.

[7] E. Friedgut, V. Rödl, A. Ruciński, and P. Tetali, *A sharp threshold for random graphs with a monochromatic triangle in every edge coloring*, Mem. Amer. Math. Soc., 179 (2006), pp. vi–66.

[8] S. Gerke, T. Schickinger, and A. Steger, *K_5-free subgraphs of random graphs*, Random Structures Algorithms, 24 (2004), pp. 194–232.

[9] P. E. Haxell, Y. Kohayakawa, and T. Łuczak, *Turán's extremal problem in random graphs: Forbidding even cycles*, J. Combin. Theory Ser. B, 64 (1995), pp. 273–287.

[10] P. E. Haxell, Y. Kohayakawa, and T. Łuczak, *Turán's extremal problem in random graphs: Forbidding odd cycles*, Combinatorica, 16 (1996), pp. 107–122.

[11] S. Janson, *Poisson approximation for large deviations*, Random Structures Algorithms, 1 (1990), pp. 221–229.

[12] S. Janson, T. Łuczak, and A. Ruciński, *Random Graphs*, Wiley-Interscience, New York, 2000.

[13] S. Janson and A. Ruciński, *The deletion method for upper tail estimates*, Combinatorica, 24 (2004), pp. 615–640.

[14] G. Katona, T. Nemetz, and M. Simonovits, *On a problem of Turán in the theory of graphs*, Mat. Lapok, 15 (1964), pp. 228–238.

[15] Y. Kohayakawa, T. Łuczak, and V. Rödl, *On K^4-free subgraphs of random graphs*, Combinatorica, 17 (1997), pp. 173–213.

[16] Y. Kohayakawa, V. Rödl, and M. Schacht, *The Turán theorem for random graphs*, Combin. Probab. Comput., 13 (2004), pp. 61–91.

[17] J. Nešetřil and V. Rödl, *Partitions of finite relational and set systems*, J. Combin. Theory Ser. A, 22 (1977), pp. 289–312.

[18] V. Rödl and A. Ruciński, *Lower bounds on probability thresholds for Ramsey properties*, in Combinatorics, Paul Erdős Is Eighty, Vol. 1, Bolyai Soc. Math. Stud., János Bolyai Math. Soc., Budapest, 1993, pp. 317–346.

[19] V. Rödl and A. Ruciński, *Random graphs with monochromatic triangles in every edge coloring*, Random Structures Algorithms, 5 (1994), pp. 253–270.

[20] V. Rödl and A. Ruciński, *Threshold functions for Ramsey properties*, J. Amer. Math. Soc., 8 (1995), pp. 917–942.

[21] V. Rödl and A. Ruciński, *Ramsey properties of random hypergraphs*, J. Combin. Theory Ser. A, 81 (1998), pp. 1–33.

[22] T. Szabó and V. H. Vu, *Turán's theorem in sparse random graphs*, Random Structures Algorithms, 23 (2003), pp. 225–234.

[23] E. Szemerédi, *Regular partitions of graphs*, in Problèmes Combinatoires et Théorie des Graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976), Colloq. Internat. CNRS 260, CNRS, Paris, 1978, pp. 399–401.

SIAM J. DISCRETE MATH.
Vol. 21, No. 2, pp. 461–465

CIRCULAR COLORING THE PLANE*

MATT DEVOS[†], JAVAD EBRAHIMI[†], MOHAMMAD GHEBLEH[†], LUIS GODDYN[†], BOJAN MOHAR[†], AND REZA NASERASR[‡]

Abstract. The *unit distance graph* $\mathcal{R}$ is the graph with vertex set $\mathbb{R}^2$ in which two vertices (points in the plane) are adjacent if and only if they are at Euclidean distance 1. We prove that the circular chromatic number of $\mathcal{R}$ is at least 4, thus improving the known lower bound of 32/9 obtained from the fractional chromatic number of $\mathcal{R}$.

Key words. graph coloring, circular coloring, unit distance graph

AMS subject classifications. 05C15, 05C10, 05C62

DOI. 10.1137/060664276

1. Introduction.

The *unit distance graph* $\mathcal{R}$ is defined to be the graph with vertex set $\mathbb{R}^2$ in which two vertices (points in the plane) are adjacent if and only if they are at Euclidean distance 1. Every subgraph of $\mathcal{R}$ is also said to be a *unit distance graph*. It is known that (cf. [1, 2])

$$4 \leqslant \chi(\mathcal{R}) \leqslant 7$$

and that (cf. [3, pp. 59–65])

$$\frac{32}{9} \leqslant \chi_f(\mathcal{R}) \leqslant 4.36.$$

Here $\chi(\mathcal{R})$ denotes the chromatic number of $\mathcal{R}$, and $\chi_f(\mathcal{R})$ is the fractional chromatic number of $\mathcal{R}$ defined as follows: a *b–fold coloring* of a graph G is an assignment of sets of b colors to the vertices of G. The *fractional chromatic number* of G, denoted $\chi_f(G)$, is defined by

$$\chi_f(G) = \inf\left\{\frac{a}{b} \mid G \text{ has a } b\text{–fold coloring using } a \text{ colors}\right\}.$$

In this paper we study the circular chromatic number of the unit distance graph $\mathcal{R}$.

Let $r \geqslant 2$, $a, b \in [0, r)$, and $a \leqslant b$. We define the *circular distance* of a and b, denoted by $\delta(a, b) = \delta_r(a, b)$, to be $\min\{b - a, r + a - b\}$. One may identify the interval $[0, r)$ with a circle C^r having circumference r, and then $\delta(a, b)$ will be the distance between a and b in C^r. It is easy to see that δ satisfies the triangle inequality.

If $a, b \in [0, r)$ (or equivalently $a, b \in C^r$), we define the *circular interval from a to b*, denoted $[a, b]$, as follows (see Figure 1.1):

$$[a, b] = \begin{cases} \{x \mid a \leqslant x \leqslant b\} & \text{if } a \leqslant b, \\ \{x \mid 0 \leqslant x \leqslant b \text{ or } a \leqslant x < r\} & \text{if } a > b. \end{cases}$$

*Received by the editors July 4, 2006; accepted for publication (in revised form) January 31, 2007; published electronically May 22, 2007.

http://www.siam.org/journals/sidma/21-2/66427.html

[†] Department of Mathematics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada (mdevos@sfu.can, javad.ebrahim@gmail.com, ghebleh@gmail.com, goddyn@math.sfu.ca, mohar@sfu.ca).

[‡] Department of Combinatorics and Optimization, Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada (naserasr@math.uwaterloo.ca).

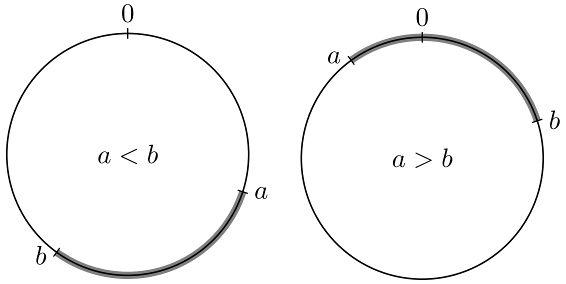


FIG. 1.1. *Circular intervals (clockwise direction is the positive direction).*

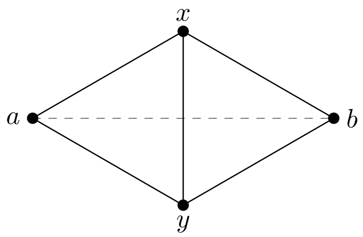


FIG. 2.1. *The unit distance graph $H_{a,b}$.*

An *r-circular coloring* of a graph G is a function $c : V(G) \to C^r$ such that for every edge xy in G, $\delta(c(x), c(y)) \geqslant 1$. The *circular chromatic number* of G, denoted by $\chi_c(G)$, is

$$\chi_c(G) = \inf\{r \mid G \text{ admits an } r\text{-circular coloring}\}.$$

It is well known [4] that for every graph G, $\chi_f(G) \leqslant \chi_c(G) \leqslant \chi(G)$. For the unit distance graph $\mathcal{R}$, these inequalities give

$$\frac{32}{9} \leqslant \chi_f(\mathcal{R}) \leqslant \chi_c(\mathcal{R}) \leqslant \chi(\mathcal{R}) \leqslant 7.$$

We improve the lower bound for $\chi_c(\mathcal{R})$ to 4. We give two proofs of this result. The second one is constructive and gives a construction of finite unit distance graphs whose circular chromatic numbers are arbitrarily close to 4.

2. Proof. Let a and b be two points in the plane, and let $d(a, b)$ denote the Euclidean distance between a and b. If $d(a, b) = \sqrt{3}$, then we may find points x and y in the plane such that the subgraph of $\mathcal{R}$ induced on the set $\{a, b, x, y\}$ is isomorphic to the graph H obtained by deleting one edge from K_4 (see Figure 2.1). We denote this unit distance graph by $H_{a,b}$. On the other hand, it is easy to see that, in any embedding of H as a unit distance graph in the plane, the Euclidean distance between the two vertices of degree 2 in H is $\sqrt{3}$.

LEMMA 2.1. *Let $0 < \varepsilon < 1$ and let $a, b \in \mathbb{R}^2$ with $d(a, b) = \sqrt{3}$. Let c be a $(3 + \varepsilon)$-circular coloring of $H_{a,b}$. Then $\delta(c(a), c(b)) \leqslant \varepsilon$.*

Proof. Without loss of generality, we may assume $c(a) = 0$. Since a, x, y form a triangle in $H_{a,b}$, we have $c(x) \in [1, 1 + \varepsilon]$ and $c(y) \in [2, 2 + \varepsilon]$ up to symmetry. On the other hand, b is adjacent to both x and y. Thus

$$\begin{aligned} c(b) &\in [c(x)+1, c(x)-1] \cap [c(y)+1, c(y)-1] \\ &\subseteq [2, \varepsilon] \cap [-\varepsilon, 1+\varepsilon] \\ &= [-\varepsilon, \varepsilon]. \end{aligned}$$

The last equality is true since $1+\varepsilon < 2$. $\square$

THEOREM 2.2. $\chi_c(\mathcal{R}) \geqslant 4$.

Proof. Suppose that c is a $(3+\varepsilon)$-circular coloring of $\mathcal{R}$ where $0 \leqslant \varepsilon < 1$. Let

$$\mu = \sup\{\delta(c(a), c(b)) \mid a, b \in \mathbb{R}^2 \text{ and } d(a,b) = \sqrt{3}\}.$$

By Lemma 2.1, $\mu \leqslant \varepsilon$. By the definition of μ, for every $0 < \mu' < \mu$, there exist points a and b at distance $\sqrt{3}$ in the plane such that $\delta(c(a), c(b)) > \mu'$. Consider the graph $H_{a,b}$ as in Figure 2.1. Without loss of generality we may assume

$$0 = c(a) \leqslant c(b) < c(x) < c(y) \leqslant 2 + \varepsilon.$$

Since $3+\varepsilon < 4$, we have

$$\delta(c(a), c(x)) = c(x) = \delta(c(a), c(b)) + \delta(c(b), c(x)) > \mu' + 1.$$

On the other hand, since a and x are at distance 1, there exists a point z which is at distance $\sqrt{3}$ from both a and x. Therefore

$$1 + \mu' < \delta(c(a), c(x)) \leqslant \delta(c(a), c(z)) + \delta(c(z), c(x)) \leqslant 2\mu.$$

Since this is true for every $\mu' < \mu$, we have $\mu \geqslant 1$. This is a contradiction since $\mu \leqslant \varepsilon < 1$. $\square$

3. A constructive proof. The graph $G_0 = K_2$ is obviously a unit distance graph. In our construction of graphs G_n ($n \geqslant 0$) we distinguish two vertices in each of them. To emphasize the distinguished vertices x and y of G_n, we write $G_n^{x,y}$. We identify subgraphs of $\mathcal{R}$ with their geometric representation given by their vertex set.

For $n \geqslant 0$, the graph G_{n+1} is constructed recursively from four copies of G_n. Let $S = V(G_n^{x,y}) \subseteq \mathbb{R}^2$. Let us rotate the set S in the plane about the point x, so that the image y' of y under this rotation is at distance 1 from y. Let S' be the image of S under this rotation. Let T be the set of all points in $S \cup S'$ and their reflections across the line yy'. In particular let $z \in T$ be the reflection of x across the line yy'. We define $G_{n+1}^{x,z}$ to be the subgraph of $\mathcal{R}$ induced on T. This construction is depicted in Figure 3.1.

Note that G_1 is the graph $H_{a,b}$ of Figure 2.1 and G_2 contains the Moser graph shown in Figure 3.2 as a subgraph. The Moser graph, also known as the spindle graph, was the first 4–chromatic unit distance graph discovered [2].

LEMMA 3.1. *For every $n \geqslant 1$, $\chi_c(G_n) \geqslant 4 - 2^{1-n}$. Moreover, for every $r = 4 - 2^{1-n} + \varepsilon$ with $0 \leqslant \varepsilon < 2^{1-n}$ and every circular r-coloring c of $G_n^{x,z}$, we have $\delta(c(x), c(z)) \leqslant 2^{n-1}\varepsilon$.*

Proof. We use induction on n. The nontrivial part of the case $n = 1$ is proved in Lemma 2.1. Let $n \geqslant 1$ and $G_{n+1}^{x,z}$ be as shown in Figure 3.1. Let $r = 4 - 2^{1-n} + \varepsilon$ for some $\varepsilon \geqslant 0$, and let c be a circular r-coloring of $G_{n+1}^{x,z}$. Without loss of generality we may assume that $c(x) = 0$. By the induction hypothesis, $\delta(0, c(y))$ and $\delta(0, c(y'))$ are both at most $2^{n-1}\varepsilon$. Hence $\delta(c(y), c(y')) \leqslant 2^n\varepsilon$. On the other hand, since y and

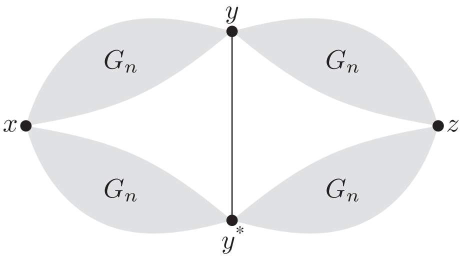


FIG. 3.1. *Construction of G_{n+1} from G_n.*

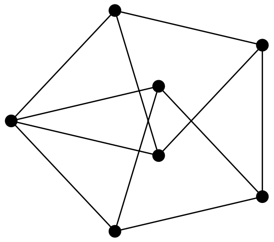


FIG. 3.2. *The Moser (spindle) graph.*

y' are adjacent in $G_{n+1}^{x,z}$, we have $\delta(c(y), c(y')) \geqslant 1$. Therefore $\varepsilon \geqslant 2^{-n}$, and we have $\chi_c(G_{n+1}) \geqslant 4 - 2^{1-n} + 2^{-n} = 4 - 2^{-n}$.

Now let $r = 4 - 2^{-n} + \varepsilon$ for some $0 \leqslant \varepsilon < 2^{-n}$, and let c be a circular r-coloring of G_{n+1} with $c(x) = 0$. Note that $r = 4 - 2^{1-n} + \varepsilon'$, with $\varepsilon' = 2^{-n} + \varepsilon < 2^{1-n}$. By the induction hypothesis, $\delta(0, c(y))$, $\delta(0, c(y'))$, $\delta(c(z), c(y))$, and $\delta(c(z), c(y'))$ are all at most $2^{n-1}\varepsilon' < 1$. Therefore we have

$$c(y), c(y') \in [-2^{n-1}\varepsilon', 2^{n-1}\varepsilon']$$

and

$$c(z) \in [c(y) - 2^{n-1}\varepsilon', c(y) + 2^{n-1}\varepsilon'] \cap [c(y') - 2^{n-1}\varepsilon', c(y') + 2^{n-1}\varepsilon'].$$

Since $\delta(c(y), c(y')) \geqslant 1$, one of $c(y)$ and $c(y')$, say $c(y)$, is in the circular interval $[-2^{n-1}\varepsilon', 2^{n-1}\varepsilon' - 1]$, and $c(y') \in [-2^{n-1}\varepsilon' + 1, 2^{n-1}\varepsilon']$. Therefore

$$[c(y) - 2^{n-1}\varepsilon', c(y) + 2^{n-1}\varepsilon'] \subseteq [-2^{n}\varepsilon', 2^{n}\varepsilon' - 1] = [-2^{n}\varepsilon', 2^{n}\varepsilon]$$

and

$$[c(y') - 2^{n-1}\varepsilon', c(y') + 2^{n-1}\varepsilon'] \subseteq [-2^{n}\varepsilon' + 1, 2^{n}\varepsilon'] = [-2^{n}\varepsilon, 2^{n}\varepsilon'].$$

Finally, since $\varepsilon' < 2^{1-n}$, we have $2^n\varepsilon' < r - 2^n\varepsilon'$. Hence

$$c(z) \in [-2^{n}\varepsilon', 2^{n}\varepsilon] \cap [-2^{n}\varepsilon, 2^{n}\varepsilon'] = [-2^{n}\varepsilon, 2^{n}\varepsilon].$$

This completes the induction step. ▯

Let us observe that, when constructing G_{n+1} from four copies of G_n, it may happen that vertices in distinct copies of G_n correspond to the same points in the plane. Additionally, it may happen that some edges between vertices in distinct copies of G_n are introduced. We may define in the same way a sequence of abstract graphs H_n, where neither of these two issues occur. Clearly $\chi_c(G_n) \geqslant \chi_c(H_n)$, but we cannot argue equality in general. The proof of Lemma 3.1 applied to the graphs H_n gives slightly more, as follows.

THEOREM 3.2. *For every $n \geqslant 0$, $\chi_c(H_n) = 4 - 2^{1-n}$.*

Proof. The cases $n = 0, 1$ are trivial. Let $n \geqslant 1$, and let H_{n+1} be as in Figure 3.1. Let $r = 4 - 2^{-n} = 4 - 2^{1-n} + 2^n$. By the proof of Lemma 3.1, $H_n^{x,y}$ admits a circular r-coloring c_1, with $c_1(x) = 0$ and $c_1(y) = \frac{1}{2}$. Similarly the graphs $H_n^{x,y'}$, $H_n^{y,z}$, and $H_n^{y',z}$ admit circular r-colorings c_2, c_3, and c_4, respectively, with $c_2(x) = 0$, $c_2(y') = c_4(y') = -\frac{1}{2}$, $c_3(y) = \frac{1}{2}$, and $c_3(z) = c_4(z) = 0$. Now a circular r-coloring c of H_{n+1} can be obtained by combining the partial colorings c_1, c_2, c_3, c_4. $\square$

The construction of this section gives an infinite subgraph of $\mathcal{R}$ with a circular chromatic number of at least 4. It remains open whether or not $\mathcal{R}$ has a finite subgraph with the same property.

REFERENCES

[1] H. HADWIGER AND H. DEBRUNNER, *Combinatorial Geometry in the Plane*, Holt, Rinehart and Winston, New York, 1964.

[2] L. MOSER AND W. MOSER, *Solution to problem* 10, Canad. Math. Bull., 4 (1961), pp. 187–189.

[3] E. R. SCHEINERMAN AND D. H. ULLMAN, *Fractional Graph Theory*, John Wiley & Sons, New York, 1997.

[4] X. ZHU, *Circular chromatic number: A survey*, Discrete Math., 229 (2001), pp. 371–410.

SIAM J. DISCRETE MATH.
Vol. 21, No. 2, pp. 466–481

THE LEMPEL–ZIV COMPLEXITY OF FIXED POINTS OF MORPHISMS*

SORIN CONSTANTINESCU[†] AND LUCIAN ILIE[‡]

Abstract. The Lempel–Ziv complexity is a fundamental measure of complexity for words, closely connected with the famous LZ77 compression algorithm. We investigate this complexity measure for one of the most important families of infinite words in combinatorics, namely, the fixed points of morphisms. We give a complete characterization of the complexity classes which are $\Theta(1)$, $\Theta(\log n)$, and $\Theta(n^{1/k})$, $k \in \mathbb{N}$, $k \geq 2$, depending on the periodicity of the word and the growth function of the morphism. The relation with the well-known classification of Ehrenfeucht, Lee, Rozenberg, and Pansiot for factor complexity classes is also investigated. The two measures complete each other, giving an improved picture for the complexity of these infinite words.

Key words. combinatorics on words, infinite words, Lempel–Ziv complexity, fixed points, morphisms, factors

AMS subject classifications. 68R15, 68P30

DOI. 10.1137/050646846

1. Introduction.

Before publishing their famous papers introducing the well-known compression schemes LZ77 and LZ78 in [36] and [37], resp., Lempel and Ziv introduced a complexity measure for words in [21] which attempted to detect "sufficiently random looking" sequences. In contrast with the fundamental measures of Kolmogorov [19] and Chaitin [4], Lempel and Ziv's measure is computable. The definition is purely combinatorial; its basic idea, splitting the word into minimal never-seen-before factors, proved to be at the core of the well-known compression algorithm LZ77, as well as subsequent variations. Another, closely related variant is to decompose the word into maximal already-seen factors, as introduced by Crochemore [7] as a tool for algorithm design.

Lempel–Ziv-type complexity and factorizations have important applications in many areas, such as data compression [36, 37], string algorithms [7, 20, 25, 32], cryptography [26], molecular biology [5, 15, 16], and neural computing [1, 34, 35].

Lempel and Ziv [21] investigate various properties which are expected from a complexity measure which intends to detect randomness. They prove it to be subadditive and also that most (but not too many) sequences are complex; see [21] for details. Also, they test it against de Bruijn words [3] as a well-established case of complex words—de Bruijn words contain as factors all words of a given length within the minimum possible space. Therefore, they establish the first connection with the factor complexity, which is also one of our topics.

In this paper, we investigate the Lempel–Ziv complexity from the combinatorial point of view and not from an information theoretical perspective. Nevertheless, some implications of our results to data compression are obtained. We shall consider the Lempel–Ziv complexity for one of the most important classes of infinite words in

*Received by the editors December 5, 2005; accepted for publication (in revised form) January 26, 2007; published electronically June 19, 2007.
http://www.siam.org/journals/sidma/21-2/64684.html

[†] Department of Computer Science, University of Western Ontario, London, Ontario, N6A 5B7, Canada (sorinco@csd.uwo.ca).

[‡] Corresponding author. Department of Computer Science, University of Western Ontario, London, Ontario, N6A 5B7, Canada (ilie@csd.uwo.ca). The research of this author was partially supported by NSERC.

combinatorics, namely, the fixed points of morphisms. Many famous infinite words, such as Fibonacci or Thue–Morse, belong to this family; see, e.g., [23].

The fundamental nature of this measure allows for a complete characterization of the complexity of infinite fixed points of morphisms. The lowest complexity, constant, or $\Theta(1)$, is encountered for the simplest words, that is, ultimately periodic. For nonperiodic words, the complexity depends on the growth function of the underlying morphism for the letter on which the morphism is iterated. Thus, for polynomial growth we obtain $\Theta(n^{1/k})$, $k \in \mathbb{N}$, $k \geq 2$, whereas for exponential growth the complexity is $\Theta(\log n)$. We give examples for which each of the above complexities is reached.

An interesting by-product of this research is the observation that LZ77 will succeed in compressing these infinite fixed points down to 0 bits/symbol asymptotically which is desirable of any good compression algorithm since the underlying mechanism generating these infinite words has only a finite amount of information.

Our results are similar with the well-known ones of Ehrenfeucht, Lee, and Rozenberg [9], Ehrenfeuct and Rozenberg [10, 11, 12, 13, 14, 30], and Pansiot [27, 28], who provided the same characterization for the factor complexity. Comparing the two characterizations, we find out that they complete each other in an interesting way. While theirs distinguishes four complexity classes for the exponential case, ours gives an infinite hierarchy (given by the parameter k above) in the polynomial case, corresponding to their quadratic complexity.

The paper is structured as follows. After some basic definitions in the next section, we introduce the Lempel–Ziv complexity and related concepts in section 3. Section 4 contains an important intermediate result which characterizes the complexity of powers of a morphism. Using it, our complete characterization is proved in section 5, where examples which reach each complexity involved are shown. The comparison with the characterization of factor complexity is included in section 6. Many problems need to be investigated about the Lempel–Ziv complexity. We propose several in the last section.

2. Basic notations. We introduce here the basic definitions and concepts we need. For further details we refer the reader to [6, 22, 23, 24].

Let Σ be an alphabet (finite nonempty set) and denote by Σ^* the free monoid generated by Σ, that is, the set of all finite words over Σ. The elements of Σ are called letters, and the empty word is denoted ε. The length of a word w is denoted $|w|$ and represents the number of letters in w; e.g., $|\mathsf{abaab}| = 5$ and $|\varepsilon| = 0$.

Given the words $w, x, y, z \in \Sigma^*$ such that $w = xyz$, x is called a *prefix*, y is a *factor* and z a *suffix* of w; we use the notation $x \leq w$. If moreover $x \neq w$, then x is a *proper* prefix of w, denoted $x < w$. The prefix of length n of w is denoted $\mathsf{pref}_n(w)$.

An infinite word is a function $w : \mathbb{N} \setminus \{0\} \to \Sigma$. A finite word can be viewed as a function $w : \{1, 2, \ldots, |w|\} \to \Sigma$. In either case, the factor of w starting at position i and ending at position j will be denoted by $w(i, j) = w_i w_{i+1} \ldots w_j$. The set of all factors of w is $F(w)$. The set of letters of Σ that actually occur in w is denoted $\Sigma(w)$. The set of infinite words over Σ is denoted Σ^ω. An infinite word w is *ultimately periodic* if $w = uvvv\ldots$, for some $u, v \in \Sigma^*$, $v \neq \varepsilon$. When we say w is nonperiodic, we mean it is not ultimately periodic.

A morphism is a function $h : \Sigma^* \to \Delta^*$ such that $h(\varepsilon) = \varepsilon$ and $h(uv) = h(u)h(v)$ for all $u, v \in \Sigma^*$. Clearly, a morphism is completely defined by the images of the letters in the domain. For all of our morphisms, $\Sigma = \Delta$.

A morphism $h : \Sigma^* \to \Sigma^*$ is called *nonerasing* if $h(a) \neq \varepsilon$ for all $a \in \Sigma$, *uniform* if $|h(a)| = |h(b)|$ for all $a, b \in \Sigma$, and *prolongable* on $a \in \Sigma$ if $a < h(a)$.

If h is prolongable on a, then $h^n(a)$ is a proper prefix of $h^{n+1}(a)$ for all $n \in \mathbb{N}$. Therefore, the sequence $(h^n(a))_{n\geq 0}$ of words defines an infinite word $h^\infty(a) \in \Sigma^\omega$ that is a fixed point of h. Formally, the ith letter of $h^\infty(a)$ is defined as being the ith letter of a power $h^n(a)$ whose length is greater than i. The fact that $h^\infty(a)$ is a well-defined fixed point of h is easily verified. Also, for h and a fixed, the fixed point is unique.

It is possible to have finite strings as fixed points of morphisms, and one can also consider erasing morphisms, but the interesting case is that of nonerasing prolongable morphisms. Therefore, when we say *fixed point* $h^\infty(a)$, we mean an infinite word obtained by iterating a nonerasing morphism h that is prolongable on a.

3. Word histories and Lempel–Ziv complexity. Let w be a (possibly infinite) word. We now introduce a fundamental notion for the Lempel–Ziv complexity. We define the operator π that removes the final letter of a finite word w:

$$\pi(w) = w(1, |w| - 1) \ .$$

A *history* $H = (u_1, u_2, \ldots, u_n)$ of $w \neq \varepsilon$ is a factorization of w, $w = u_1u_2 \ldots u_n$, having the property that $u_1 \in \Sigma$ and

$$\pi(u_i) \in F(\pi^2(u_1u_2 \ldots u_i))$$

for all $2 \leq i \leq n$. We assume also that all u_i's are nonempty. This definition requires that any new factor u_i, excepting its last letter, appears before in the word. However, it is still possible that the whole u_i does occur before in w, or $u_i \in F(\pi(u_1u_2 \ldots u_i))$. In this case u_i is called *reproductive*. Otherwise, u_i is *innovative*.

Example 1. Consider the word $w = \mathsf{aaabaabbaba}$. A possible history of w is $(\mathsf{a}, \mathsf{aab}, \mathsf{aab}, \mathsf{bab}, \mathsf{a})$. The second and fourth components are innovative, whereas the third and fifth are reproductive.

By definition, n is called the length of the history H and is denoted by $|H|$.

Two kinds of history are important to us. The first, directly connected to the definition of Lempel–Ziv complexity, is the *exhaustive* history. A history H is exhaustive if all u_i, $2 \leq i \leq |H| - 1$, are innovative. In other words, the whole new factor u_i does not occur before in the word even if all of its proper prefixes do. Clearly, the exhaustive history of a word is unique. Sometimes (e.g., in [2]) the exhaustive history is called *Lempel–Ziv factorization.*

By contrast with the exhaustive history, a *reproductive history* requires that all of its factors have occurred before (they are reproductive), with the necessary exceptions of never-seen-before letters: A history $H = (u_1, u_2, \ldots, u_n)$ is reproductive if either

$$u_i \in F(\pi(u_1u_2 \ldots u_i)) \quad \text{or} \quad u_i \notin F(\pi(u_1u_2 \ldots u_i)) \text{ but then } u_i \in \Sigma.$$

The innovative factors in a reproductive history are single letters. A reproductive history need not be unique.

Example 2. For the word in Example 1, $(\mathsf{a}, \mathsf{aab}, \mathsf{aabb}, \mathsf{aba})$ is the exhaustive history, whereas $(\mathsf{a}, \mathsf{aa}, \mathsf{b}, \mathsf{aa}, \mathsf{b}, \mathsf{ba}, \mathsf{ba})$ and $(\mathsf{a}, \mathsf{aa}, \mathsf{b}, \mathsf{aab}, \mathsf{ba}, \mathsf{ba})$ are two reproductive histories.

The following result, due to [21], relates the exhaustive history with all other histories of a word.

LEMMA 1. *The exhaustive history of a word is the shortest history of that word.*

By definition, the *Lempel–Ziv complexity* of a finite word w, denoted by $\text{LZ}(w)$, is the length of the exhaustive history of w, that is, the number of factors in the Lempel–Ziv factorization. Therefore, by Lemma 1, for any history H, $\text{LZ}(w) \leq |H|$.

The *Lempel–Ziv complexity of an infinite word* w is the function $\textsc{lz}_w : \mathbb{N} \to \mathbb{N}$ defined by

$$\textsc{lz}_w(n) = \textsc{lz}(\mathsf{pref}_n(w))$$

as the complexity of finite prefixes of w.

Remark 1. The Lempel–Ziv complexity of finite words can be computed in linear time by using suffix trees; see [7, 17].

4. The complexity of powers. The main result of this section is that the complexity of $h^n(a)$, as a function of n, is either linear or bounded for a nonerasing morphism h prolongable on a. That is, either $\textsc{lz}(h^n(a)) = \Theta(n)$ or $\textsc{lz}(h^n(a)) = \Theta(1)$. Throughout this section, a is fixed, and h is nonerasing and prolongable on a.

Given the morphism h, we can assume, without loss of generality, that each letter of Σ occurs in $h^\infty(a)$, the fixed point of h. If that is not the case, h can be restricted to the set of those letters that do occur in w, and the fixed point of the restriction will still be the same.

4.1. Maximal reproductive history. We show first that the complexity of powers is at most linear. To this end, we define the *maximal reproductive history*[1] of a finite word w, denoted $RH(w)$. For $w = w_1 w_2 \dots w_{|w|}$, $w_i \in \Sigma$, we define $RH(w) = (u_1, u_2, \dots, u_n)$ as follows:

- $u_1 = w_1$, the first letter of w;
- $$u_{i+1} = \begin{cases} w_{|u_1 u_2 \dots u_i|+1} & \text{if } w_{|u_1 u_2 \dots u_i|+1} \notin \Sigma(u_1 u_2 \dots u_i), \\ \text{longest } w \text{ with } w \in F(\pi(u_1 u_2 \dots u_i w)) & \\ & \text{if } w_{|u_1 u_2 \dots u_i|+1} \in \Sigma(u_1 u_2 \dots u_i) \end{cases}$$

  for all $i \geq 2$.

With the exception of new single letters, $RH(w)$ is created by taking at each step the maximal factor that has occurred before. For the word in Example 1, the maximal reproductive history is $(\mathsf{a}, \mathsf{aa}, \mathsf{b}, \mathsf{aab}, \mathsf{ba}, \mathsf{ba})$.

From the definition it is clear that $RH(w)$ is a reproductive history. It follows from Lemma 1 that $|RH(w)| \geq \textsc{lz}(w)$.

Remark 2. The maximally reproductive history has been introduced independently by Crochemore [7] as a tool for algorithm design. It is more natural than the Lempel–Ziv factorization. Indeed, most applications we mentioned above use Crochemore's factorization. On the other hand, the two factorizations are very closely related. For historical reasons, we defined the Lempel–Ziv complexity as the number of factors in the Lempel–Ziv factorization, but our asymptotical results hold as well for Crochemore's factorization. This can be seen directly by looking at the proofs or from the following lemma, which connects the lengths of the two histories.

LEMMA 2. *For any $w \in \Sigma^*$, we have*

$$\textsc{lz}(w) \leq |RH(w)| \leq 2\,\textsc{lz}(w) - 1.$$

Proof. The first inequality follows by Lemma 1. For the second, we show first that the maximal reproductive history is the shortest among all reproductive histories. Denote $RH(w) = (u_1, \dots, u_n)$, and consider another reproductive history, $(v_1, \dots, v_m)$. First, for all $1 \leq i \leq \min(n, m)$, we have $|v_1 \dots v_i| \leq |u_1 \dots u_i|$. Indeed, if this is not the case, consider the smallest i_0 for which it does not hold. In this case, $i_0 \geq 2$ and

[1] This is called *s-factorization* in [7, 25], *f-factorization* in [8], *Lempel–Ziv factorization* in [32], and *Crochemore factorization* in [2].

u_{i_0} appears in v_{i_0} as a factor but not at the end of v_{i_0}. Thus $|v_{i_0}| \geq 2$, so v_{i_0} is not a letter, and, by the definition of the reproductive histories, v_{i_0} must have occurred before. Therefore, u_{i_0} is not the longest prefix of $u_{i_0} \dots u_n$ which has occurred before, a contradiction. It follows immediately that $n \leq m$.

Consider then the exhaustive history of w: $(t_1, \dots, t_k)$. Put, for all $2 \leq i \leq k$, $t_k = s_k a_k$, $a_k \in \Sigma$. We construct the history H obtained from the factorization $(t_1, s_2, a_2, s_3, a_3, \dots, s_k, a_k)$ by removing the empty factors, if any. We then have $|H| \leq 2k - 1 = 2\,\textsc{lz}(w) - 1$. By the above, $|RH(w)| \leq |H|$, which concludes the proof. □

Notice that Lemma 1 can be easily proved in a similar way.

4.2. Morphic images of histories. The next step is to iterate reproductive histories through a morphism h. We will show a way to create a reproductive history of $h(w)$, given a reproductive history of w.

Let w be a word and $H = (v_0, v_1, \dots, v_n)$ be a reproductive history of w. Let $1 = i_1 < i_2 < \cdots < i_{|\Sigma(w)|}$ be the indexes corresponding to the single letter factors of H that have not occurred before. We define a factorization of $h(w)$, denoted $h(H)$, by replacing all factors of w that have occurred before by their image through h and the single letters v_{i_j}, by the history $RH(h(v_{i_j}))$. We claim that this is a reproductive history of $h(w)$.

Example 3. Let us consider the Thue–Morse morphism

$$\begin{aligned} t(\mathsf{a}) &= \mathsf{ab}\ , \\ t(\mathsf{b}) &= \mathsf{ba}\ , \end{aligned}$$

and the word from Example 1, $w = \mathsf{aaabaabbaba}$. A reproductive history H (in fact, $RH(w)$) and its image through t, $t(H)$, are

$$\begin{array}{rcllllll} H & = & (\mathsf{a}\ , & \mathsf{aa}\ , & \mathsf{b}\ , & \mathsf{aab}\ , & \mathsf{ba}\ , & \mathsf{ba})\ , \\ h(H) & = & (\mathsf{a}, \mathsf{b}, & \mathsf{abab}, & \mathsf{b}, \mathsf{a}, & \mathsf{ababba}, & \mathsf{baab}, & \mathsf{baab})\ . \end{array}$$

LEMMA 3. *If H is a reproductive history of w, then $h(H)$ is a reproductive history of $h(w)$.*

Proof. There are two kinds of factors in $h(H)$. One originates from a factor of H that has already occurred. If a factor u has already occurred in w, then its image $h(u)$ will have also occurred in $h(w)$.

Also, each factor of the history $RH(h(v_{i_j}))$ is either a new single letter or has already occurred in the factor $h(v_{i_j})$ of w and therefore has occurred in w.

By selecting the first occurrence of all of the single letters in $h(w)$, we conclude that each factor of $h(H)$ is either a factor that has already occurred or a letter that has not been previously seen. Equivalently, $h(H)$ is a reproductive history of $h(w)$. □

4.3. Linear upper bound. With respect to the length of $h(H)$, we note that each letter in $\Sigma(w)$, originally a stand-alone factor of H, is transformed into the factorization $RH(h(v_{i_j}))$, and, consequently, each letter x of w prompts a $|RH(h(x))| - 1$ increase in the length of $h(H)$:

$$|h(H)| \leq |H| + \sum_{x \in \Sigma(w)} (|RH(h(x))| - 1)\ .$$

If we assume that all letters of Σ occur in w, then the increase in length is constant, which leads us to the following result.

Proposition 1. *If $h : \Sigma^* \to \Sigma^*$ is nonerasing and $a < h(a)$, $a \in \Sigma$, then* $\textsc{lz}(h^n(a)) = O(n)$.

Proof. We will use the above method for iteratively creating histories for $h^n(a)$ that will have a linearly increasing length.

Let n_0 be the first integer for which $h^{n_0}(a)$ contains all letters of Σ:

$$n_0 = \min\{n \in \mathbb{N} \mid \Sigma(h^n(a)) = \Sigma\},$$

and let $H_0 = RH(h^{n_0}(a))$.

Applying the above method, $h(H_0)$ is a valid history for $h^{n_0+1}(a)$ and

$$|h(H_0)| = |H_0| + \sum_{x\in\Sigma}(|RH(h(x))| - 1) \ .$$

Iterating for $n \geq n_0$, we get

$$|h^{n-n_0}(H_0)| = |H_0| + (n - n_0)\sum_{x\in\Sigma}(|RH(h(x))| - 1)$$

or

$$|h^{n-n_0}(H_0)| = A \cdot n + B,$$

where $B = |H_0| - n_0 \sum_{x\in\Sigma}(|RH(h(x))| - 1)$ and $A = \sum_{x\in\Sigma}(|RH(h(x))| - 1)$.

Since $h^{n-n_0}(H_0)$ is a valid history for $h^n(a)$, it follows that

$$\textsc{lz}(h^n(a)) \leq An + B$$

or $\textsc{lz}(h^n(a)) = O(n)$. □

The next result gives the inferior asymptotic limit for $\textsc{lz}(h^n(a))$. It is obvious that $\textsc{lz}(h^n(a))$, as a function of n, is increasing since $h^n(a) < h^{n+1}(a)$. The remaining part of this section is dedicated to showing that the growth of the Lempel–Ziv complexity of powers is at least linear unless the fixed point word is ultimately periodic.

Throughout the rest of this section, the word u is defined by $h(a) = au$.

4.4. Some technical results. We prove next two technical lemmas to be used later in the proof of the lower bound.

Lemma 4. *If $h^p(u)h^{p+1}(u)$ occurs at most $|h^p(u)|$ positions before its last occurrence in*

$$h^{p+2}(a) = auh(u)\dots h^p(u)h^{p+1}(u) \ ,$$

then $h^\infty(a)$ is ultimately periodic.

Proof. Let $\alpha = h^p(u)$. Since $\alpha h(\alpha)$ occurs at most $|\alpha|$ positions from the end of $t = h^{p+2}(a)$, there exists v, with $|v| \leq |\alpha|$, such that $v\alpha h(\alpha)$ is a suffix of t and also $\alpha h(\alpha)$ is a prefix of $v\alpha h(\alpha)$. Let v be the minimal word that satisfies this property—in other words, v marks the occurrence of $\alpha h(\alpha)$ that is the closest to the end of $\pi(t)$; see Figure 1.

Both α and $\alpha h(\alpha)$ are fractional powers of v:

$$\alpha = v^n v', \text{ with } v' \leq v, n \geq 1, \tag{1}$$

$$\alpha h(\alpha) = v^m v'', \text{ with } v'' \leq v, m \geq 2. \tag{2}$$

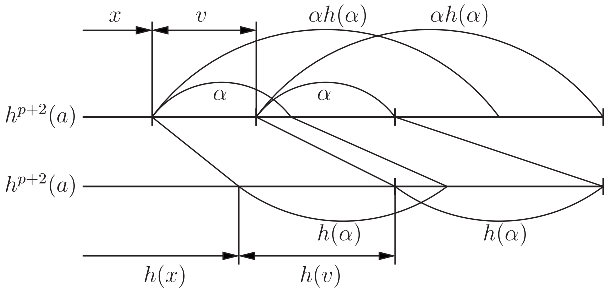


FIG. 1. *The occurrence of $\alpha h(\alpha)$ that has v as a prefix is the closest to the end of $\pi(h^{p+2}(a))$.*

Let x be defined by $t = xv\alpha h(\alpha)$. Therefore $h(xv) = xv\alpha$. Since h is nonerasing, $|x| \le |h(x)|$, which implies that $h(v)$ is a suffix of $v\alpha$.

By applying h to (1), we get that $h(\alpha) = h(v)^n h(v')$. This indicates that $h(v)h(\alpha)$ has period $|h(v)|$. However, $h(v)h(\alpha)$ is a suffix of $v\alpha h(\alpha)$, which has period $|v|$. By Fine and Wilf's theorem (see [6, 23]) $h(v)h(\alpha)$ has the period $d = \gcd(|v|, |h(v)|)$.

If $d < |v|$, then $h(v)$ has period d. Since α has period $|v|$, all factors of α of length $|v|$ are circular shifts of v. Consequently, the circular shift of v occurring at the end of $v\alpha$ is completely covered by $h(v)$, and, therefore, that particular circular shift of v has period d. However, the length of v is a multiple of d, so v is a nontrivial power of one of its proper prefixes of length d. In this case, we could find another occurrence of $\alpha h(\alpha)$ closer to the end t which contradicts the choice of v.

Therefore $d = |v|$, or $|v|$ divides $|h(v)|$. Furthermore $h(v)$ is a factor of some power of v since it is a factor of $v\alpha$, a fractional power of v. Let $r \in \mathbb{N}$ be defined by $r|v| = |h(v)|$. It follows that $h(v)$ is a circular shift of v^r. Inductively, if $h^s(v)$ is a circular shift of v^{r^s}, then $h^{s+1}(v)$ is a circular shift of $h(v)^{r^s}$, which is a circular shift of $(v^r)^{r^s} = v^{r^{s+1}}$. This implies that $|h^s(v)| = r^s|v|$ for all $s \ge 0$.

Because $h(v)$ is a suffix of $v\alpha$, it follows that $h^{s+1}(v)$ is a suffix of $h^s(v)h^s(\alpha)$. Inductively, if $h^s(v)$ has period $|v|$, then it is a power of some word of length $|v|$. Since $h^s(v)h^s(\alpha)$ is a fractional power of $h^s(v)$ by (1), it must also have period $|v|$, which implies that $h^{s+1}(v)$ has period v.

We have established that all $h^s(v)$ have period $|v|$ and their lengths are all multiples of $|v|$. We can now apply h^s to (1) and obtain

$$h^s(\alpha) = h^s(v)^n h^s(v'), \text{ with } v' \le v, n \ge 1.$$

Since $h^s(v)$ has period $|v|$ and its length is a multiple of that period, $h^s(\alpha)$ must also have the period $|v|$.

By a similar argument, using (2), $h^s(\alpha)h^{s+1}(\alpha)$ has period $|v|$. As this holds for all $s \ge 0$ and $|h^s(\alpha)| \ge |v|$, it follows that $\alpha h(\alpha) \dots h^s(\alpha) \dots$ has period $|v|$. □

LEMMA 5. *If $h^q(u)h^{q+1}(u)h^{q+2}(u)$ occurs before its last occurrence in*

$$h^{q+3}(a) = auh(u) \dots h^q(u)h^{q+1}(u)h^{q+2}(u)$$

and $|h^q(u)| < |h^{q+1}(u)|$, then $h^\infty(a)$ is ultimately periodic.

Proof. Let $\alpha = h^q(u)$ and $t = h^{q+3}(a)$. If $\alpha h(\alpha)h^2(\alpha)$ occurs at most $|\alpha|$ positions from its last occurrence in t as a suffix, then, by Lemma 4, $h^\infty(a)$ is ultimately periodic.

Otherwise, there exist words x and y such that

$$t = x\alpha y\alpha h(\alpha)h^2(\alpha) = h(x)h(\alpha)h(y)h(\alpha)h^2(\alpha)$$

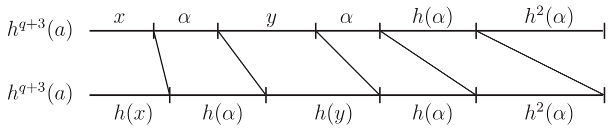


FIG. 2. *Here $h(\alpha)h^2(\alpha)$ is a prefix of $y\alpha h(\alpha)h^2(\alpha)$, and so $h^2(\alpha)$ is a prefix of $h(y)h(\alpha)h^2(\alpha)$.*

and $h(\alpha)h^2(\alpha)$ is a prefix of $y\alpha h(\alpha)h^2(\alpha)$; see Figure 2. By taking the lengths of the two factorizations of t, we have

$$|x| + |\alpha| + |y| + |\alpha| = |h(x)| + |h(\alpha)| + |h(y)|$$

or

$$|h(x)| - |x| = |\alpha| - (|h(\alpha)| - |\alpha|) - (|h(y)| - |y|) \ .$$

Since h is nonerasing, $|h(x)| - |x| < |\alpha|$. However, $h(\alpha)h^2(\alpha)$ is a prefix of $y\alpha h(\alpha)h^2(\alpha)$, so $h^2(\alpha)$ is a prefix of $h(y)h(\alpha)h^2(\alpha)$. This leads to $h(\alpha)h^2(\alpha)$ occurring at position $|h(x)|-|x|$ in the first occurrence of $\alpha h(\alpha)h^2(\alpha)$. Consequently, $h(\alpha)h^2(\alpha)$ occurs in t at a distance less than $|h(\alpha)|$ symbols before its last occurrence in t, which makes Lemma 4 applicable for $p = q + 1$, since $h(\alpha)h^2(\alpha) = h^p(u)h^{p+1}(u)$ occurs at most $|h(\alpha)| = |h^p(u)|$ positions before its last occurrence in $t = h^{q+3}(a) = h^{p+2}(a)$. □

4.5. Growth functions. In order to be able to use Lemma 5, we need to find values of q for which $|h^q(u)| < |h^{q+1}(u)|$. It is clear that $|h^q(u)| \leq |h^{q+1}(u)|$, and, if there exists a letter z in $h^q(u)$ satisfying $|h(z)| \geq 2$, the inequality is strict. We shall prove that such powers must exist or else the fixed point is ultimately periodic. We need more definitions and results.

The *growth function* of the letter $x \in \Sigma$ in h is the function $h_x : \mathbb{N} \to \mathbb{N}$ defined by

$$h_x(n) = |h^n(x)| \ .$$

The following result from [31, 33] is very useful.

LEMMA 6. *There exist an integer $e_a \geq 0$ and an algebraic real number $\rho_a \geq 1$ such that*

$$h_a(n) = \Theta(n^{e_a}\rho_a^n) \ .$$

The pair (e_a, ρ_a) is called the *growth index* of a in h. We say that h_a (and a as well) is called *bounded*, *polynomial*, and *exponential* if a's growth index w.r.t. h is $(0, 1)$, $(> 0, = 1)$, $(\geq 0, > 1)$, resp.

Example 4. All letters of a uniform morphism with images of length k share the same growth index: $(0, k)$. For instance, the growth index of a for the Thue–Morse morphism of Example 3 is $(0, 2)$.

Let us consider the morphism h defined by

$$\begin{aligned} h(\mathsf{a}) &= \mathsf{ab} \ , \\ h(\mathsf{b}) &= \mathsf{bc} \ , \\ h(\mathsf{c}) &= \mathsf{c} \ . \end{aligned}$$

The growth index of a is $(2, 1)$, the growth index of b is $(1, 1)$, and, finally, the growth index of c is $(0, 1)$.

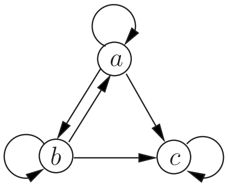


FIG. 3. *The graph G^h for Example* 5.

4.6. The associated graph. We introduce the following graph, which is very useful for some proofs. Given a morphism $h : \Sigma^* \to \Sigma^*$, we denote the sets of bounded, polynomial, and exponential letters by Σ_B, Σ_P, and Σ_E, resp. The *graph associated with* h is the directed graph

$$G^h = (\Sigma, \{(a, b) \mid b \in F(h(a))\}) \ .$$

Thus, the vertices of G^h are the letters of the alphabet, and there is an edge from a to b if b appears in the image of a.

Consider its subgraphs G^h_X, induced by the sets Σ_X, $X \in \{B, P, E\}$, of vertices, resp. A few observations about the graphs we just defined are in order:

1. Any letter a belonging to two distinct cycles of G^h is exponential, as some power $h^r(a)$ would contain at least two a's.
2. Let us fix the order $B < P < E$. Then for any X and any $a \in \Sigma_X$, the image $h(a)$ of a must contain at least one letter from Σ_X and cannot contain any letter from Σ_Y for any $Y > X$.
3. The above observation implies that, as soon as Σ_X is nonempty, there is a cycle (which might be a loop) in G^h_X and from each vertex in G^h_X there is a path leading to a vertex in a cycle (everything in G^h_X).

Example 5. Consider the morphism h:

$$h(\mathsf{a}) = \mathsf{acb} \ ,$$
$$h(\mathsf{b}) = \mathsf{bca} \ ,$$
$$h(\mathsf{c}) = \mathsf{c} \ .$$

The graph G^h is shown in Figure 3. This is also the graph of a different morphism: $\mathsf{a} \mapsto \mathsf{abc}$, $\mathsf{b} \mapsto \mathsf{bac}$, $\mathsf{c} \mapsto \mathsf{c}$, which indicates that different morphisms can produce isomorphic graphs.

4.7. Linear lower bound for nonperiodic words. We need only one more lemma before proving the main result of this section.

LEMMA 7. *If $h(a) = au$, $u \in \Sigma^*$, $u \neq \varepsilon$, then there exist $m, p \in \mathbb{N}$ such that $|h^{m+jp}(a)| < |h^{m+jp+1}(a)|$, for all $j \geq 0$, or else $h^\infty(a)$ is ultimately periodic.*

Proof. Since h is prolongable on a, it means a is not bounded. Assume $a \in \Sigma_P$; the case $a \in \Sigma_E$ is similar. Denote also $u = u_1 u_2 \ldots u_{|u|}$, $u_i \in \Sigma$. We have in G^h the edges (a, a) and (a, u_i) for all $1 \leq i \leq |u|$.

If all u_i's are in Σ_B, then $|h^n(u)|$ is bounded as $|h^n(u)| = \sum_{i=1}^{|u|} h_{u_i}(n)$. Hence, we can find n and r such that $h^n(u) = h^{n+r}(u)$, implying that $h^\infty(a) = auh(u)h^2(u)\ldots$ is ultimately periodic.

Assume $u_i \in \Sigma_P$ for some i. By the above properties of G^h, we can find in G^h_P a path from u_i to a vertex which belongs to a cycle which is also in G^h_P. There must be a vertex, say, z, in that cycle, whose outdegree is at least two; otherwise, all vertices in the cycle would be bounded. If we denote the length of the path from u_i to z

by m and the length of the cycle by p, then $|h^{m+jp}| < |h^{m+jp+1}|$ for all $j \geq 0$ as claimed. $\square$

Using Lemmas 5 and 7, we obtain for all $j \geq 0$ that either

$$h^{m+jp}(u)h^{m+jp+1}(u)h^{m+jp+2}(u) \tag{3}$$

has never occurred before or w is ultimately periodic.

If we assume $w = h^{\infty}(a)$ to be nonperiodic, then all factors of the form (3) can never occur before their last occurrence. This shows that there must exist a factor in the exhaustive history of w that ends within each distinct factor of the above-mentioned form. It follows that $\text{LZ}(h^n(a)) \geq \frac{1}{k}(n - n_0) + \text{LZ}(h^{n_0+1}(a))$ or $\text{LZ}(h^n(a)) = \Omega(n)$.

Combining this result with Proposition 1 we obtain that $\text{LZ}(h^n(a))$ is either constant or linear. On the other hand, the fact that ultimate periodicity is equivalent to a bounded Lempel–Ziv complexity has been mentioned in [18]. Therefore, we proved the main result of this section.

PROPOSITION 2. *For a nonerasing morphism h that admits the fixed point $h^{\infty}(a)$, $\text{LZ}(h^n(a))$ is either $\Theta(1)$, if $h^{\infty}(a)$ is ultimately periodic, or $\Theta(n)$, otherwise.*

5. Growth functions and infinite word complexity. Let w be an infinite word generated by iterating a nonerasing morphism h, $w = h^{\infty}(a)$. The prefix of a given length m of w will fall between two consecutive powers of h:

$$h^{n(m)}(a) \leq \mathsf{pref}_m(w) < h^{n(m)+1}(a) \tag{4}$$

for a $n(m) \in \mathbb{N}$. If $\text{LZ}(h^n(a))$ is bounded, then $\text{LZ}_w(n)$ is bounded. This establishes our first case for the complexity of $\text{LZ}_w(\cdot)$, $\Theta(1)$.

When $\text{LZ}(h^n(a))$ is not bounded, it has to be linear, by Proposition 2. Then a is not bounded, and hence, by Lemma 6, we distinguish two cases:

1. $\rho_a = 1$ (h_a is polynomial). Then $|h^n(a)| = \Theta(n^{e_a})$ or $n(m) = \Theta(m^{1/e_a})$. Since, by (4), $\text{LZ}(h^{n(m)}(a)) \leq \text{LZ}(\mathsf{pref}_m(w)) \leq \text{LZ}(h^{n(m)+1}(a))$ and $\text{LZ}(h^n(a)) = \Theta(n)$, it follows that $\text{LZ}_w(m) = \Theta(m^{1/e_a})$.
2. $\rho_a > 1$ (h_a is exponential). There exist ρ_1 and ρ_2 positive numbers such that $\rho_1^n \leq |h^n(a)| \leq \rho_2^n$, which means that $n(m) = \Theta(\log m)$. By the same argument, $\text{LZ}_w(m) = \Theta(\log m)$.

Notice, however, that h_a growing does not imply $\text{LZ}_w(\cdot)$ unbounded. For instance, if $h(\mathsf{a}) = \mathsf{ab}$, $h(\mathsf{b}) = \mathsf{b}$, then h_{a} is polynomial, but $w = h^{\infty}(\mathsf{a}) = \mathsf{abbb}\ldots$ has bounded $\text{LZ}_w(\cdot)$. For the exponential case we can take $h(\mathsf{a}) = \mathsf{aa}$, whose fixed point also has bounded Lempel–Ziv complexity.

Also, in the first case above, we cannot have $e_a = 1$ as this implies bounded Lempel–Ziv complexity, contradicting the assumption on $\text{LZ}(h^n(a))$. Indeed, $e_a = 1$ implies $|h^n(a)| = \Theta(n)$, and so $|h^{n+1}(a)| - |h^n(a)|$ is bounded. Assuming $h(a) = au$, $u \neq \varepsilon$, we have $h^n(a) = auh(u)h^2(u)\ldots h^{n-1}(u)$. Consequently $|h^n(u)|$ is bounded; hence, we can find $h^n(u) = h^{n+p}(u)$, which implies that $w = h^{\infty}(a)$ is ultimately periodic.

We have just proved the main result of the paper.

THEOREM 1. *For a fixed point infinite word $w = h^{\infty}(a)$ of a nonerasing morphism h, we have the following:*

1. *The Lempel–Ziv complexity of w is $\Theta(1)$ if and only if w is ultimately periodic.*
2. *If w is not ultimately periodic, then the Lempel–Ziv complexity of w is $\Theta(\log n)$ or $\Theta(n^{1/k})$, $k \in \mathbb{N}$, $k \geq 2$, depending on whether h_a is exponential or polynomial, resp.*

Notice that the logarithmic Lempel–Ziv complexity in the exponential case was already proved in a different context by Ilie, Yu, and Zhang [18, Lemma 12].

Remark 3. Notice that the Lempel–Ziv complexity of fixed points is lower than the maximal Lempel–Ziv complexity, in the sense that there is no fixed point whose Lempel–Ziv complexity is of the order $\Theta(\frac{n}{\log n})$, which is the order of the maximum Lempel–Ziv complexity for finite words of length n, as proved by Lempel and Ziv [21].

Furthermore, since the LZ77-compressed size of a word w is $\Theta(\textsc{lz}(w)\log|w|)$, it follows that the LZ77 compression algorithm will succeed in compressing the fixed points down to 0 bits/symbol asymptotically, which is desirable of any good compression algorithm, since the underlying mechanism generating these infinite words has only a finite amount of information. Therefore, this is a positive conclusion regarding the usage of this algorithm to find random sequences, stating that the algorithm won't misclassify the infinite words considered in this paper.

Remark 4. For a morphism h prolongable on a, it is decidable to which of the classes in Theorem 1 its Lempel–Ziv complexity function belongs. First of all, a test for ultimate periodicity can be found in [29]. Assuming that the fixed point is not ultimately periodic, h_a is exponential if and only if there exists some letter b, accessible from a, deriving in a number of steps a word containing two occurrences of b (see [31]). As noted above, this is equivalent to b belonging to two different cycles in the associated graph. This can be easily tested for each letter. An algorithm that decides whether or not h_a is exponential needs only to check if any of the letters belonging to two different cycles are reachable from a.

5.1. Examples. We give next examples showing that all of the above complexities are indeed possible.

Example 6. The highest Lempel–Ziv complexity is realized for $k = 2$, that is, $O(\sqrt{n})$, for the three letter morphism h_3 given by

$$\begin{aligned} h_3(\mathsf{a}) &= \mathsf{ab}\ ,\\ h_3(\mathsf{b}) &= \mathsf{bc}\ ,\\ h_3(\mathsf{c}) &= \mathsf{c}\ , \end{aligned}$$

for which $h_3^n(\mathsf{a}) = \mathsf{abc}^0\mathsf{bc}^1\ldots\mathsf{bc}^{n-1}$. Clearly, the growth function of a, $(h_3)_{\mathsf{a}}$, is quadratic, whereas the complexity of powers is exactly linear, which gives a final Lempel–Ziv complexity of $\sqrt{n}$; this can be checked directly by constructing the exhaustive history of $h_3^\infty(\mathsf{a})$:

$$(\mathsf{a},\mathsf{b},\mathsf{bc},\mathsf{bc}^2,\mathsf{bc}^3,\ldots)\ .$$

This example can be easily extended to k letters. Let

$$h_k : \{\mathsf{a}_1,\mathsf{a}_2,\ldots,\mathsf{a}_k\}^* \to \{\mathsf{a}_1,\mathsf{a}_2,\ldots,\mathsf{a}_k\}^*$$

be defined by

$$\begin{aligned} h_k(\mathsf{a}_1) &= \mathsf{a}_1\mathsf{a}_2\ ,\\ h_k(\mathsf{a}_2) &= \mathsf{a}_2\mathsf{a}_3\ ,\\ &\vdots\\ h_k(\mathsf{a}_{k-1}) &= \mathsf{a}_{k-1}\mathsf{a}_k\ ,\\ h_k(\mathsf{a}_k) &= \mathsf{a}_k\ . \end{aligned}$$

We have that $(h_k)_{\mathsf{a}_1}$ is a polynomial of degree $k-1$ (see [31, Theorem 3.5]). We can also see that directly, as follows. Note that h_k restricted to $\{\mathsf{a}_2, \mathsf{a}_3, \ldots, \mathsf{a}_k\}^*$ is actually h_{k-1} modulo the renaming $\mathsf{a}_2 = \mathsf{a}_1, \mathsf{a}_3 = \mathsf{a}_2, \ldots, \mathsf{a}_k = \mathsf{a}_{k-1}$. Since

$$|(h_k)_{\mathsf{a}_1}(n)| = |\mathsf{a}_1\mathsf{a}_2 h(\mathsf{a}_2)\ldots h_k^{n-1}(\mathsf{a}_2)| = 1 + \sum_{i=0}^{n-1} |(h_{k-1})_{\mathsf{a}_1}(n)| \ ,$$

we conclude inductively that, if $(h_{k-1})_{\mathsf{a}_1}(n) = \Theta(n^{k-2})$, then $(h_k)_{\mathsf{a}_1}(n) = \Theta(n^{k-1})$. The base case follows from the previous example for $k = 3$.

Consequently, the Lempel–Ziv complexity of the fixed point $h_k^\infty(\mathsf{a}_1)$ is $\Theta(\sqrt[k-1]{n})$. These examples illustrate the polynomial case.

Example 7. With respect to the exponential case, any uniform morphism with images of length k has a growth function of exactly k^n. Since the complexity of powers is linear for nonperiodic words, the Lempel–Ziv complexity of the fixed point will be $\Theta(\log n)$.

Such an example is the famous Thue–Morse morphism (see Example 3), which fits the requirements for $k = 2$. Both fixed points $t^\infty(\mathsf{a})$ and $t^\infty(\mathsf{b})$ are nonperiodic, and the growth functions associated with both letters are exactly 2^n. Their Lempel–Ziv complexity is $\Theta(\log n)$.

Example 8. Another famous example is given by the Fibonacci morphism

$$\begin{aligned} f(\mathsf{a}) &= \mathsf{ab} \ , \\ f(\mathsf{b}) &= \mathsf{a} \ , \end{aligned}$$

for which we can precisely compute the value of $\text{LZ}(f^n(\mathsf{a})) = n+1$. The powers of the Fibonacci morphism grow exponentially, at the rate $(\frac{1-\sqrt{5}}{2})^n + (\frac{1+\sqrt{5}}{2})^n$, and therefore the Lempel–Ziv complexity of the infinite word is again $\Theta(\log n)$.

6. Comparison with factor complexity. We dedicate the final section to a comparison between the Lempel–Ziv complexity and the factor complexity for infinite words generated by morphisms. The factor complexity is a natural function defined as the number of factors of a certain length occurring in an infinite word. For a word $w \in \Sigma^\omega$, this is

$$f_w(n) = \text{card}(\{u \in \Sigma^* \mid u \in F(w), |u| = n\}) \ .$$

The investigation of factor complexity for the fixed points of morphisms has been initiated by Ehrenfeucht, Lee, and Rozenberg in [9] (they actually considered the closely related D0L systems) and continued by Ehrenfeucht and Rozenberg in a series of papers; see [10, 11, 12, 13, 14, 30]. The classification was completed by Pansiot [27, 28], who also found the missing complexity class $\Theta(n \log\log n)$.

The following definitions appear, with different names, in [6]. The morphism h is called[2]

- *nongrowing* if there exists a bounded letter in Σ;
- *u-exponential* if $\rho_a = \rho_b > 1$, $e_a = e_b = 1$ for all $a, b \in \Sigma$;
- *p-exponential* if $\rho_a = \rho_b > 1$ for all a, b and $e_a > 1$ for some a; and
- *e-exponential* if $\rho_a > 1$ for all a and $\rho_a > \rho_b$ for some a, b.

The characterization of Ehrenfeucht, Lee, Rozenberg, and Pansiot is as follows.

[2] What we call *u*-, *p*-, and *e-exponential* are *quasi-uniform*, *polynomially diverging*, and *exponentially diverging*, resp., in [6, 27, 28]. We changed the terminology so that it does not conflict with the corresponding notations for h_a.

Theorem 2 (Ehrenfeucht, Lee, Rozenberg, and Pansiot). *Let $w = h^\infty(a)$ be an infinite nonperiodic word of factor complexity $f_w(\cdot)$.*

1. *If h is growing, then $f_w(n)$ is either $\Theta(n)$, $\Theta(n \log\log n)$, or $\Theta(n \log n)$, depending on whether h is u-, p-, or e-exponential, resp.*
2. *If h is not growing, then either*
   (a) *w has arbitrarily large factors over the set of bounded letters, and then $f_w(n) = \Theta(n^2)$, or*
   (b) *w has finitely many factors over the set of bounded letters, and then $f_w(n)$ can be any of $\Theta(n)$, $\Theta(n \log\log n)$, or $\Theta(n \log n)$.*

In order to establish a correspondence with our hierarchy, we note that, in the first case of Theorem 2, the function h_a is exponential, which implies a logarithmic Lempel–Ziv complexity. However, a logarithmic Lempel–Ziv complexity does not necessarily imply one of the n, $n \log\log n$, or $n \log n$ cases for the factor complexity as is illustrated by the following example.

Example 9. Consider the morphism h given by

$$h(\mathsf{a}) = \mathsf{abc}\ ,\quad h(\mathsf{b}) = \mathsf{bac}\ ,\quad h(\mathsf{c}) = \mathsf{c}\ .$$

Since h_{a} grows exponentially, $\textsc{lz}(h^\infty(\mathsf{a}))$ is, by Theorem 1, logarithmic. However, there exist arbitrarily large factors of $h^\infty(\mathsf{a})$ of the form c^n (c is bounded), which implies a $\Theta(n^2)$ factor complexity.

On the other hand, a radical-type Lempel–Ziv complexity does imply a quadratic factor complexity. To prove this, we again need the associated graph.

Lemma 8. *Assume $h : \Sigma^* \to \Sigma^*$ is a nonerasing morphism prolongable on $a \in \Sigma$. If h_a is polynomial, then there exist arbitrarily large factors over Σ_B in $h^\infty(a)$.*

Proof. Consider the associated graph introduced above. First, since h_a is polynomial, G_E^h must be empty.

By the properties of G^h, there exists at least one cycle in G_P^h, say, C. If there is a vertex of C which has other outgoing edges (different from the one in C) in G_P^h, then any path starting with such an edge cannot go back to C (this would make the letters of C exponential). Therefore, further cycles can be constructed. As Σ_P is finite, there must be a cycle C' in G_P^h which has no outgoing edges in G_P^h except for those in the cycle. On the other hand, at least one vertex (letter) of C', say, b, has an outgoing edge to a vertex in G_B^h. We then have $h(b) = ubv$, $uv \in \Sigma_B^*$, $uv \neq \varepsilon$. The letter b will create in $h^\infty(a)$ arbitrarily long factors from Σ_B^*, as claimed. □

Therefore, Theorems 1 and 2, Example 9, and Lemma 8 imply the correspondence between Lempel–Ziv and factor complexities for fixed points of morphisms shown in Table 1, where all intersections are indeed possible.

We see that both measures of complexity recognize ultimately periodic words as having bounded complexity, the lowest class of complexity.

In the nontrivial case of nonperiodic fixed points, the Lempel–Ziv complexity groups together all words $h^\infty(a)$ with the h_a exponential, whereas the factor complexity distinguishes four different complexities. On the other hand, the factor complexity does not make any distinction among words with the h_a polynomial, whereas Lempel–Ziv gives an infinite hierarchy.

7. Further research.

Most combinatorial aspects of the Lempel–Ziv complexity need to be investigated. We mention a few problems below:

Table 1
Lempel–Ziv vs. factor complexity.

	Lempel–Ziv complexity	Factor complexity
$h^\infty(a)$ is ultimately periodic	$\Theta(1)$	$\Theta(1)$
$h^\infty(a)$ is not ultimately periodic and h_a is polynomial	$\Theta(n^{\frac{1}{2}})$	$\Theta(n^2)$
	$\Theta(n^{\frac{1}{3}})$	
	$\vdots$	
	$\Theta(n^{\frac{1}{k}})$	
	$\vdots$	
$h^\infty(a)$ is not ultimately periodic and h_a is exponential	$\Theta(\log n)$	$\Theta(n^2)$
		$\Theta(n \log n)$
		$\Theta(n \log\log n)$
		$\Theta(n)$

1. Characterize the fixed points of morphisms in each Lempel–Ziv complexity class (especially $\Theta(n^{\frac{1}{k}})$).
2. What is the connection between k in $\Theta(n^{1/k})$ and card(Σ)?
3. Investigate the relations, in general, between Lempel–Ziv complexity and other complexity measures, especially the factor complexity.
4. How is Lempel–Ziv complexity affected by operations on words? For concatenation, it is subadditive, that is, $\text{LZ}(uv) \leq \text{LZ}(u) + \text{LZ}(v)$, as proved by Lempel and Ziv [21]. Also, it is easy to see that it is monotonic for prefixes, that is, $\text{LZ}(u) \leq \text{LZ}(uv)$. But the same is not true for suffixes. Here is a counterexample: $\text{LZ}(\mathsf{a.ab.aaba}) = 3$, $\text{LZ}(\mathsf{a.b.aa.ba}) = 4$. Also, the behavior with respect to the reversal operation (already questioned in [18]) should be investigated, that is, the relation between the Lempel–Ziv complexity of w and that of w^R, the reversal of w.
5. Another complexity measure can be defined naturally from the factorization used in the LZ78 compression algorithm, which is $w = u_1.u_2.\cdots.u_n$ such that, for all $i \geq 2$, u_i is the shortest prefix of $u_i u_{i+1} \ldots u_n$ that does not belong to the set $\{u_1, u_2, \ldots, u_{i-1}\}$. That means u_i may have appeared as a factor of $\pi(u_1 u_2 \ldots u_i)$ but not as a member of the factorization so far. In particular, this factorization is a history. Denoting the new complexity by $\text{LZ}_{78}(w)$ we have by Lemma 1 that $\text{LZ}(w) \leq \text{LZ}_{78}(w)$. Investigating this complexity measure is certainly of interest. The precise relation between the two complexity measures is not obvious, and it may be that different techniques are required for investigating LZ_{78}.

Acknowledgments. The authors thank the anonymous referees for very careful reading of the paper and for useful comments which helped improve the clarity of the presentation. Also, the second part of Remark 3 was suggested by one of the referees.

REFERENCES

[1] J. M. Amigó, J. Szczepański, E. Wajnryb, and M. V. Sanchez-Vives, *Estimating the entropy rate of spike trains via Lempel-Ziv complexity*, Neural Comput., 16 (2004), pp. 717–736.

[2] J. Berstel and A. Savelli, *Crochemore factorization of Sturmian and other infinite words*, in Proceedings of MFCS'06, Lecture Notes in Comput. Sci. 4162, Springer, Berlin, 2006, pp. 157–166.

[3] N.G. de Bruijn, *A combinatorial problem*, Nederl. Akad. Wetensch. Proc., 49 (1946), pp. 758–764.

[4] G. Chaitin, *On the length of programs for computing finite binary sequences*, J. Assoc. Comput. Mach., 13 (1966), pp. 547–569.

[5] X. Chen, S. Kwong, and M. Li, *A compression algorithm for DNA sequences*, IEEE Eng. Med. Biol., 20 (2001), pp. 61–66.

[6] C. Choffrut and J. Karhumäki, *Combinatorics on words*, in Handbook of Formal Languages, Vol. I, G. Rozenberg and A. Salomaa, eds., Springer, Berlin, 1997, pp. 329–438.

[7] M. Crochemore, *Recherche linéaire d'un carré dans un mot*, C. R. Math. Acad. Sci. Paris, 296 (1983), pp. 781–784.

[8] M. Crochemore and W. Rytter, *Text Algorithms*, Oxford University Press, New York, 1994.

[9] A. Ehrenfeucht, K.P. Lee, and G. Rozenberg, *Subword complexities of various classes of deterministic developmental languages without interaction*, Theoret. Comput. Sci., 1 (1975), pp. 59–75.

[10] A. Ehrenfeucht and G. Rozenberg, *On the subword complexities of square-free D0L-languages*, Theoret. Comput. Sci., 16 (1981), pp. 25–32.

[11] A. Ehrenfeucht and G. Rozenberg, *On the subword complexities of D0L-languages with a constant distribution*, Theoret. Comput. Sci., 13 (1981), pp. 108–113.

[12] A. Ehrenfeucht and G. Rozenberg, *On the subword complexities of homomorphic images of languages*, RAIRO Inf. Théor., 16 (1982), pp. 303–316.

[13] A. Ehrenfeucht and G. Rozenberg, *On the subword complexities of locally catenative D0L-languages*, Inform. Process. Lett., 16 (1982), pp. 7–9.

[14] A. Ehrenfeucht and G. Rozenberg, *On the subword complexities of m-free D0L-languages*, Inform. Process. Lett., 17 (1983), pp. 121–124.

[15] M. Farach, M.O. Noordewier, S.A. Savari, L.A. Shepp, A.D. Wyner, and J. Ziv, *On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence*, in Proceedings of SODA'95, 1995, pp. 48–57.

[16] V.D. Gusev, V.A. Kulichkov, and O.M. Chupakhina, *The Lempel-Ziv complexity and local structure analysis of genomes*, Biosystems, 30 (1993), pp. 183–200.

[17] D. Gusfield, *Algorithms on Strings, Trees, and Sequences. Computer Science and Computational Biology*, Cambridge University Press, Cambridge, 1997.

[18] L. Ilie, S. Yu, and K. Zhang, *Word complexity and repetitions in words*, Internat. J. Found. Comput. Sci., 15 (2004), pp. 41–55.

[19] A.N. Kolmogorov, *Three approaches to the quantitative definition of information*, Probl. Inf. Transm., 1 (1965), pp. 1–7.

[20] R. Kolpakov and G. Kucherov, *Finding maximal repetitions in a word in linear time*, in Proceedings of the 40th Annual Symposium on Foundations of Computer Science, IEEE, Los Alamitos, CA, 1999, pp. 596–604.

[21] A. Lempel and J. Ziv, *On the complexity of finite sequences*, IEEE Trans. Inform. Theory, 92 (1976), pp. 75–81.

[22] M. Lothaire, *Combinatorics on Words*, Addison-Wesley, Reading, MA, 1983 (reprinted with corrections, Cambridge University Press, Cambridge, 1997).

[23] M. Lothaire, *Algebraic Combinatorics on Words*, Cambridge University Press, Cambridge, 2002.

[24] M. Lothaire, *Applied Combinatorics on Words*, Cambridge University Press, Cambridge, 2005.

[25] M.G. Main, *Detecting leftmost maximal periodicities*, Discrete Appl. Math., 25 (1989), pp. 145–153.

[26] S. Mund, *Ziv-Lempel complexity for periodic sequences and its cryptographic application*, in Advances in Cryptology–EUROCRYPT '91, Lecture Notes in Comput. Sci. 547, Springer, Berlin, 1991, pp. 114–126.

[27] J.-J. Pansiot, *Bornes inférieures sur la complexité des facteurs des mots infinis engendrés par morphismes itérés*, in Proceedings of STACS'84, Lecture Notes in Comput. Sci. 166, Springer, Berlin, 1984, pp. 230–240.

[28] J.-J. Pansiot, *Complexité des facteurs des mots infinis engendrés par morphismes itérés*, in Proceedings of ICALP'84, Lecture Notes in Comput. Sci. 172, Springer, Berlin, 1984, pp. 380–389.

[29] J.-J. Pansiot, *Decidability of periodicity for infinite words*, Theor. Inform. Appl., 20 (1986), pp. 43–46.

[30] G. Rozenberg, *On subwords of formal languages*, in Proceedings of Fundamentals of Computation Theory, Lecture Notes in Comput. Sci. 117, Springer, Berlin, 1981, pp. 328–333.

[31] G. Rozenberg and A. Salomaa, *The Mathematical Theory of L Systems*, Academic, New York, 1980.

[32] W. Rytter, *Application of Lempel-Ziv factorization to the approximation of grammar-based compression*, Theoret. Comput. Sci., 302 (2003), pp. 211–222.

[33] A. Salomaa and M. Soittola, *Automata-Theoretic Aspects of Formal Power Series*, Springer, New York, 1978.

[34] J. Szczepański, M. Amigó, E. Wajnryb, and M.V. Sanchez-Vives, *Application of Lempel-Ziv complexity to the analysis of neural discharges*, Network: Comput. Neural Syst., 14 (2003), pp. 335–350.

[35] J. Szczepański, J. M. Amigó, E. Wajnryb, and M. V. Sanchez-Vives, *Characterizing spike trains with Lempel-Ziv complexity*, Neurocomputing, 58-60 (2004), pp. 79–84.

[36] J. Ziv and A. Lempel, *A universal algorithm for sequential data compression*, IEEE Trans. Inform. Theory, 23 (1977), pp. 337–343.

[37] J. Ziv and A. Lempel, *Compression of individual sequences via variable-rate coding*, IEEE Trans. Inform. Theory, 24 (1978), pp. 530–536.

SIAM J. DISCRETE MATH.
Vol. 21, No. 2, pp. 482–503

VIRTUAL PRIVATE NETWORK DESIGN: A PROOF OF THE TREE ROUTING CONJECTURE ON RING NETWORKS[*]

C. A. J. HURKENS[†], J. C. M. KEIJSPER[†], AND L. STOUGIE[†‡]

Abstract. A basic question in virtual private network (VPN) design is if the symmetric version of the problem always has an optimal solution which is a tree network. An affirmative answer would imply that the symmetric VPN problem is solvable in polynomial time. We give an affirmative answer in case the communication network, within which the VPN must be created, is a circuit. This seems to be an important step towards answering the general question. The proof relies on a dual pair of linear programs and actually implies an even stronger property of VPNs. We show that this property also holds for some other special cases of the problem, in particular when the network is a tree of rings.

Key words. network design, duality, combinatorial optimization, multicommodity flows

AMS subject classifications. 90C27, 90C46, 90C35

DOI. 10.1137/050626259

1. Introduction.

In this paper, we consider a problem emerging in telecommunication known as the *symmetric virtual private network* problem. Think of a large communication network represented by an undirected graph $G = (V, E)$, with a vertex for each user and an edge for each link in the network. Within this network, a subgroup $W \subseteq V$ of the users wishes to reserve capacity on the links of the network for communication among themselves: they wish to establish a virtual private network (VPN). Vertices in W are also called *terminals.*

On each link, capacity (bandwidth) has a certain price per unit, $c : E \to \mathbb{R}_+$. The problem is to select one or more communication paths between every pair $\{i, j\}$ of users in W and to reserve enough capacity on the edges of the selected paths to accommodate any possible communication pattern among the users in W. Possible communication patterns are defined through an upper bound on the amount to be communicated (transmitted and received) for each node in W, specified by $b : W \to \mathbb{R}_+$. More precisely, a *communication scenario* for the symmetric VPN problem can be defined as a symmetric matrix $D = (d_{ij})_{\{i,j\}\subseteq W}$ with zeros on the diagonal, specifying for each *unordered* pair of distinct nodes $\{i, j\} \subseteq W$ the amount of communication $d_{ij} \geq 0$ between i and j. A communication scenario $D = (d_{ij})_{\{i,j\}\subseteq W}$ is said to be *valid* if $\sum_{j\in W\setminus\{i\}} d_{ij} \leq b(i)\ \forall i \in W$. We denote the collection of valid communication scenarios by $\mathcal{D}$.

We call a network consisting of the selected communication paths with enough capacity reserved on the edges to accommodate every valid communication scenario a *feasible VPN.* The (symmetric) VPN problem is to find the cheapest feasible VPN. There are several variants of the problem emerging from additional routing requirements.

[*] Received by the editors March 8, 2005; accepted for publication (in revised form) December 4, 2006; published electronically June 28, 2007. An extended abstract appeared in *Proceedings of the Integer Programming and Combinatorial Optimization Conference (IPCO)* 2005. The work was partially supported by the Dutch BSIK-BRICKS project and by the FET project ARRIVAL of the EC under contract FP6-021235-2.

http://www.siam.org/journals/sidma/21-2/62625.html

[†] Department of Mathematics and Computer Science, Technische Universiteit Eindhoven, Eindhoven S600 MB, The Netherlands (wscor@win.tue.nl, j.c.m.keijsper@tue.nl).

[‡] CWI, Kruislaan 413, NL 1098, Amsterdam, The Netherlands (leen.stougie@cwi.nl).

(i) SPR (*single path routing*): For each pair $\{i,j\} \subseteq W$, exactly one path $P_{ij} \subseteq E$ is to be selected to accommodate all possible demand between i and j. The problem is to choose the paths P_{ij} so as to minimize $\{\sum_{e\in E} c(e)x_e \mid x_e \geq \sum_{\{i,j\}:e\in P_{ij}} d_{ij}\ \forall e \in E\ \forall D = (d_{ij}) \in \mathcal{D}\}$.

(ii) TTR (*terminal tree routing*): This is SPR with the additional restriction that $\cup_{j\in W} P_{ij}$ should form a tree in $G\ \forall i \in W$.

(iii) TR (*tree routing*): This is SPR with the extra restriction that $\cup_{\{i,j\}\subseteq W} P_{ij}$ is a tree in G.

(iv) MPR (*multipath routing*): For each pair $\{i,j\} \subseteq W$, and for each possible path between i and j, the fraction of communication between i and j to be routed along that path has to be specified.

(v) FR (*flexible routing*): No communication paths have to be selected beforehand. Different demand scenarios are allowed to use different sets of paths.

The following lemma summarizes the rather obvious relations between the optimal solution values of these variants. By $OPT(\text{SPR})$ we denote the cost of an optimal solution for the SPR variant of the VPN problem. Similar notation is used for the other optimal values.

Lemma 1.1.

$$OPT(\text{FR}) \leq OPT(\text{MPR}) \leq OPT(\text{SPR}) \leq OPT(\text{TTR}) \leq OPT(\text{TR}).$$

Proof. SPR is the MPR problem with the extra restriction that all fractions must be 0 or 1. The other inequalities are similarly trivial. □

A prominent open question in VPN design is whether SPR is polynomially solvable (SPR $\in P$); cf. Italiano, Leonardi, and Oriolo [12]. This question would be answered affirmatively if one could prove that $OPT(\text{SPR}) = OPT(\text{TR})$, since Kumar et al. [13] have shown that TR $\in P$ (see also [9]). Gupta et al. [9] showed that $OPT(\text{TR}) = OPT(\text{TTR})$ and that $OPT(\text{TR}) \leq 2OPT(\text{FR})$. To the best of our knowledge the complexity of FR is unresolved. There are instances (even on circuits) where $OPT(\text{FR}) < OPT(\text{MPR})$: if we take for G a triangle, $c \equiv 1$, $b \equiv 1$, then for the optimal solution to FR it suffices to buy all three edges with capacity 1/2, whereas for MPR it is optimal to buy two edges with capacity 1. Erlebach and Rüegg [6] proved that MPR $\in P$, which also follows from our LP-formulation in section 2. They also mention that no VPN instance has been found so far for which even $OPT(\text{MPR}) < OPT(\text{TR})$. Indeed, our conjecture is that $OPT(\text{MPR}) = OPT(\text{TR})$, from which SPR $\in P$ would follow. $OPT(\text{SPR}) = OPT(\text{TR})$ was not known to be true for any class of graphs other than trees. It seems to be a crucial step forward to prove it for circuits, which is implied by the main result in this paper.

Theorem 1.2. *Let $G = (V,E)$ be a circuit. Then* $OPT(\text{MPR}) = OPT(\text{TR})$.

This theorem is proved in section 3. The proof boils down to showing that the cost of an optimal solution to TR equals the value of an optimal dual solution in a formulation of MPR as a linear program (LP). The LP for MPR is given in section 2, where the conjecture $OPT(\text{MPR}) = OPT(\text{TR})$ is restated in terms of this LP.

In section 4, we proceed to prove our conjecture, $OPT(\text{MPR}) = OPT(\text{TR})$, for some other special cases. We prove it for any graph G and any cost function c if the communication bound of some terminal is larger than the sum of the bounds of the other terminals. We also prove it for any graph on at most 4 vertices, and for any complete graph if the cost function c is identical to 1. We also prove that the property $OPT(\mathrm{MPR}) = OPT(\mathrm{TR})$ is preserved under taking 1-sums of graphs, implying a common generalization of all the aforementioned results.

The model of the VPN problem presented above was proposed for the first time by Fingerhut, Suri, and Turner [7], and later independently by Duffield et al. [3]. They also formulated the asymmetric version of the problem in which for each node there is a distinction between a bound $b^- : W \to \mathbb{R}_+$ for incoming communication and a bound $b^+ : W \to \mathbb{R}_+$ for outgoing communication. Gupta et al. [9] proved that even the TR problem is NP-hard for the VPN problem with asymmetric communication bounds. However, the TR problem is solvable in polynomial time if $b^-(v) = b^+(v)$ for all $v \in W$. Italiano, Leonardi, and Oriolo [12] showed that this is true already if $\sum_{v \in W} b^-(v) = \sum_{v \in W} b^+(v)$. Gupta et al. [9] claimed that FR is co-NP-hard for the asymmetric problem. The polynomial time algorithm for MPR by Erlebach and Rüegg [6] has been derived for the asymmetric problem. Altin et al. [1] presented an LP-formulation of the general asymmetric MPR VPN problem of polynomial size, immediately implying polynomial solvability of this problem. Independently, we found a similar formulation, which we present in a technical report [11]. The LP-formulation in that report covers MPR-variants with four types of asymmetry: asymmetric bounds ($b^-(v) \neq b^+(v)$), asymmetric costs ($c_{uv} \neq c_{vu}$), asymmetric routing, and asymmetric communication scenarios ($d_{ij} \neq d_{ji}$). If $b^-(v) = b^+(v)\ \forall v$, and either cost or routing is symmetric, then attention can be restricted to symmetric communication scenarios, and Theorem 1.2 still holds. For symmetric routing this is easy to see. In section 6 of [11] it is argued that allowing asymmetric routing under symmetric arc costs does not yield any advantage; there is always an optimal LP-solution with *symmetric* routing patterns.

As soon as both cost and routing are allowed to be asymmetric, Theorem 1.2 is false, even for symmetric bounds: if we consider a (bidirected) circuit where clockwise arcs have zero cost and counterclockwise arcs have cost 1, then buying all clockwise arcs is cheaper than buying any tree.

Gupta et al. [9], Gupta, Kumar, and Roughgarden [10], and Eisenbrand et al. [4] studied approximation algorithms for NP-hard versions of the VPN problem. More hardness results appear in Chekuri et al. [2].

The challenge remains to prove or disprove that SPR is polynomially solvable on any graph.

2. A linear programming formulation.

Let $G = (V, E)$ be a graph, $W \subseteq V$ a set of terminals, $b : W \to \mathbb{R}_+$ communication (upper) bounds, and $c : E \to \mathbb{R}_+$ unit edge costs. To facilitate the exposition, we will regard b as a function on V rather than on W, defining $b(v) = 0$ for $v \notin W$, and simply identify W with the set of vertices $\{v \in V \mid b(v) > 0\}$. Thus, the triple (G, b, c) defines an *instance* of the MPR or SPR or TR problem. We also use the notation b_v for $b(v)$ and c_e for $c(e)$.

The set of all paths between vertices i and j in W is denoted by $\mathcal{P}_{ij}$. Let $\mathcal{P} = \cup_{\{i,j\} \subseteq W} \mathcal{P}_{ij}$. We introduce the variable x_p for each $p \in \mathcal{P}$. In the SPR VPN problem we are to select one path for each pair $\{i, j\}$ of distinct nodes in W; i.e., we are to select values for the x-variables that satisfy $\sum_{p \in \mathcal{P}_{ij}} x_p = 1\ \forall \{i, j\} \subseteq W$ and $x_p \in \{0, 1\}$ $\forall p \in \mathcal{P}$. In the MPR VPN problem values for x_p are allowed to be fractional, $\forall p \in \mathcal{P}$, still satisfying $\sum_{p \in \mathcal{P}_{ij}} x_p = 1\ \forall \{i, j\} \subseteq W$.

Once paths have been selected, i.e., values for x_p, $p \in \mathcal{P}$, have been set, the computation of the capacity that has to be reserved on the edges, z_e, $e \in E$, is straightforwardly formulated in the following LP. Let $\alpha_p^e = 1$ if edge e is on path p, and 0 otherwise.

$$\begin{aligned} z_e = \max \quad & \sum_{\{i,j\}\subseteq W} \sum_{p\in\mathcal{P}_{ij}} \alpha_p^e x_p d_{ij} \\ \text{s.t.} \quad & \sum_{j\in W} d_{ij} \le b_i \qquad && \forall\ i \in W, \\ & d_{ij} \ge 0 && \forall\ \{i,j\} \subseteq W, \end{aligned}$$

which, by strong duality, is equal to

$$\begin{aligned} z_e = \min \quad & \sum_{i\in W} b_i y_i^e \\ \text{s.t.} \quad & y_i^e + y_j^e \ge \sum_{p\in\mathcal{P}_{ij}} \alpha_p^e x_p \quad && \forall\ \{i,j\} \subseteq W, \\ & y_i^e \ge 0 && \forall\ i \in W. \end{aligned}$$

The MPR problem is to make a feasible choice for the variables x_p such as to minimize total reservation costs $\sum_{e\in E} c_e z_e$, which by the above duality can be formulated as

$$\begin{aligned} \min \quad & \sum_{e\in E} c_e \sum_{i\in W} b_i y_i^e \\ \text{s.t.} \quad & y_i^e + y_j^e - \sum_{p\in\mathcal{P}_{ij}} \alpha_p^e x_p \ge 0 \quad && \forall\ \{i,j\} \subseteq W,\ \forall\ e \in E, \\ & \sum_{p\in\mathcal{P}_{ij}} x_p = 1 && \forall\ \{i,j\} \subseteq W, \\ & y_i^e \ge 0 && \forall\ i \in W,\ \forall\ e \in E, \\ & x_p \ge 0 && \forall\ p \in \mathcal{P}. \end{aligned} \tag{1}$$

In the SPR problem, of which MPR is the LP-relaxation, all variables x_p are restricted to be 0 or 1. The dual of MPR is given by

$$\begin{aligned} \max \quad & \sum_{\{i,j\}\subseteq W} \mu_{ij} \\ \text{s.t.} \quad & \sum_{j\in W} \lambda_{ij}^e \le c_e b_i \qquad && \forall\ i \in W,\ \forall\ e \in E, \\ & \mu_{ij} - \sum_{e\in E} \alpha_p^e \lambda_{ij}^e \le 0 && \forall\ \{i,j\} \subseteq W,\ \forall\ p \in \mathcal{P}_{ij} \\ & \lambda_{ij}^e \ge 0 && \forall\ \{i,j\} \subseteq W,\ \forall\ e \in E. \end{aligned} \tag{2}$$

At this point, let us note that the separation problem over the dual polytope can be solved in polynomial time. There is only a polynomial number of constraints of the first type. For the second set of constraints, suppose we are given λ_{ij}^e $\forall\{i,j\} \subseteq W$, $\forall e \in E$, and μ_{ij} $\forall\{i,j\} \subseteq W$. Take for any pair $\{i,j\} \subseteq W$ the constraints $\forall p \in \mathcal{P}_{ij}$ together. To check if they are satisfied is a matter of computing a shortest path between i and j in W, where each edge $e \in E$ has weight λ_{ij}^e. There are only polynomially many i,j pairs in W.

It follows, using the ellipsoid method (see [8]), that MPR can be solved in polynomial time, which was also proved in [6]. The formulation above, as well as the formulation in [6], is not of polynomial size, however. As mentioned before, in [1] and [11] LP-formulations are presented of the general asymmetric version of the MPR VPN problem of polynomial size, immediately implying polynomial solvability of this problem.

A possible economic interpretation of the dual is the following. Consider λ_{ij}^e to be the price at which the competition of the current provider of the VPN offers to accommodate all communication between i and j along (an alternative for) link e. Evidently this price should be nonnegative, but in order to be competitive, it should not be too high. Here, not too high means that $\sum_j \lambda_{ij}^e$ should not exceed $b_i c_e$ because this is the maximum amount i is willing to pay for the use of this edge in all his communication (given the prices c_e of the current provider). Setting a price λ_{ij}^e on edge e of course causes the pair $\{i,j\}$ to choose the cheapest i–j path, with cost μ_{ij}, for their communication. Now, the optimum dual value is the maximum revenue the competing provider can expect from accommodating this particular VPN.

We finish this section by investigating some properties of the problem. A *tree solution* for the instance (G,b,c) is a solution to the TR problem with these parameters: a tree solution is a Steiner tree in G spanning the set of terminals $W = \{v \in V \mid b(v) > 0\}$, together with optimal capacity reservations on the edges of the tree.

By weak duality, and by $OPT(\text{MPR}) \leq OPT(\text{TR})$, we have for any feasible solution (λ,μ) to (2) that $\sum_{\{i,j\}\subseteq W} \mu_{ij}$ is at most the cost of any tree solution. Thus, the following conjecture is equivalent to the conjecture that $OPT(\text{MPR}) = OPT(\text{TR})$.

Conjecture 2.1. *For any instance (G,b,c), the cost of an optimal tree solution equals the value of an optimal solution of the dual problem* (2).

In this paper, we will show that Conjecture 2.1 holds in several special cases, the most important one being the case where G is a circuit, and b and c are arbitrary.

The next paragraph, which is essentially extracted from [9], summarizes how to compute the cost of a given tree solution. Let (G,b,c) be given, and let W be the set of terminals. We write $b(U)$ for $\sum_{v\in U} b(v)$, $U \subseteq V$. Given a tree $T \subseteq E$ spanning a vertex set $V(T) \supseteq W$, a directed tree can be constructed by directing the edges of T towards the *lighter* side: if L_e and R_e are the components of $T - e$, and if $b(L_e) < b(R_e)$, direct e towards L_e; if $b(L_e) = b(R_e)$, direct e away from some fixed leaf l of the tree (the latter is a correction of what is written in [9]). This directed tree has a unique vertex r of in-degree zero which is what we call a *balance-point* of the tree: every edge in the directed tree is directed away from r. The cost of the tree T is clearly equal to

$$\sum_e \min\{b(L_e), b(R_e)\} c(e). \tag{3}$$

Another expression for the cost of the tree is given in the following proposition from [9]. Here, we denote by $d_G^c(u,v)$ the distance from u to v in a graph G with respect to the length function c.

Proposition 2.2 (see [9]). *Let $G = (V,E), b : V \to \mathbb{R}_+, c : E \to \mathbb{R}_+$ be given.*

Then the cost of an optimal tree solution T equals

$$\sum_{v\in W} b(v)d^c_T(r,v)$$

for some balance-point r. This cost is bounded from below by $\sum_{v\in W} b(v)d^c_G(r,v)$, and bounded from above by $\sum_{v\in W} b(v)d^c_T(u,v)$ for any $u \in V(T)$.

As a consequence, we have that an optimal tree solution can be found by computing a shortest path tree T_u from every vertex $u \in V$ and taking the one with minimal cost $\sum_{v\in W} b(v)d^c_{T_u}(u,v) = \sum_{v\in W} b(v)d^c_G(u,v)$. Hence TR is solvable in polynomial time.

3. The circuit. In this section, we prove Conjecture 2.1 for circuits, that is, we prove Theorem 1.2. We will restrict ourselves to circuits $G = (V,E)$ with $|V|$ even and $b(v) = 1\ \forall v \in V$. To show that this is not an essential restriction, we prove a few preliminary lemmas. Some of the results here will be used in section 4 as well.

Notice that, for a fixed graph G, the optimum values of the various (integer) LPs are continuous functions in b and c. Hence, for proving the conjecture, we may restrict ourselves to rational vectors b and c.

LEMMA 3.1. *Let G be a fixed graph. If for any rational-valued b and c the cost of an optimal tree solution to (G,b,c) equals the value of an optimal dual solution, then the same is true for any instance (G,b,c), where b and c are real-valued.*

The next lemma claims that scaling of b or c is allowed when proving Conjecture 2.1.

LEMMA 3.2. *For any $\beta \in \mathbb{R}_+$, the instance $(G,\beta b,c)$ has a feasible dual solution of value βK if and only if the instance (G,b,c) has a feasible dual solution of value K. Moreover, $(G,\beta b,c)$ has a tree solution of cost βK if and only if (G,b,c) has a tree solution of cost K. A similar statement holds if c is scaled instead of b.*

Proof. Multiply all λ and μ values by β to obtain a feasible dual solution for the scaled instance from a feasible dual solution of the original instance. The cost of any tree changes by a factor β in the new situation as well (see (3) for the cost of a tree solution). ◻

The next lemma claims that edges of zero cost may be *contracted* or *decontracted* when proving Conjecture 2.1. Contraction of $e = \{u', v'\}$, by identifying the two vertices u' and v' with one new vertex w', transforms G into $G/e = (V \setminus \{u',v'\} \cup \{w'\}, E')$, with

$$\begin{aligned} E' := \{\{u,v\} \in E \mid \{u,v\} \cap \{u',v'\} = \emptyset\} &\cup \{\{w',v\} \mid \{u',v\} \in E, v \neq v'\} \\ \cup\{\{w',v\} \mid \{v',v\} \in E, v \neq u'\}&. \end{aligned}$$

By the contraction of e in the instance (G,b,c) we mean the instance (G',b',c'), where $G' = G/e$, $b'(v) = b(v)$ for $v \neq u',v'$, and $b'(w') = b(u') + b(v')$, and moreover $c'(\{u,v\}) = c(\{u,v\})$ if $w' \notin \{u,v\} \in E'$, and $c'(\{w',v\}) = c(\{u',v\})$ or $c(\{v',v\})$, or both values occur, in case parallel edges arise (edges in G' can be identified with those in $E \setminus \{e\}$). We will denote this contraction (G',b',c') by $(G,b,c)/e$.

LEMMA 3.3. *Let $(G = (V,E),b,c)$ be an instance, where $e \in E$ has $c(e) = 0$. Then (G,b,c) has a feasible dual solution of value K if and only if the contraction $(G,b,c)/e$ has a feasible dual solution of value K. Moreover, (G,b,c) has an optimal tree solution of cost K if and only if $(G,b,c)/e$ has an optimal tree solution of cost K.*

Proof. For the proof of the first statement, consider a feasible dual solution (λ,μ) for $(G,b,c)/e$. Then the dual solution $(\hat\lambda,\hat\mu)$ for (G,b,c), defined by $\hat\lambda^e_{ij} := 0$,

$\hat{\lambda}^f_{ij} := \lambda^f_{ij}$, $f \neq e$, $\hat{\mu}_{ij} := \mu_{ij}$, is evidently feasible too. Conversely, $\mu := \hat{\mu}$, $\lambda^f := \hat{\lambda}^f$ for $f \neq e$ also maintains feasibility, because each feasible dual solution $(\hat{\lambda}, \hat{\mu})$ for (G, b, c) has the property that $0 \leq \hat{\lambda}^e_{ij} \leq c(e)b_i = 0$. Solutions (λ, μ) and $(\hat{\lambda}, \hat{\mu})$ have the same value.

One implication of the second statement of the lemma is obvious: if we decontract the edge e, we obtain from a tree solution in $(G, b, c)/e$ of cost K a tree solution in (G, b, c) of the same cost (see Proposition 2.2).

The other implication is also obvious if edge e is in an optimal tree solution for (G, b, c), or if at most one of its end points is covered by the tree. The remaining case is one in which the tree solution $T = (V', E')$ of cost K does not contain edge $e = \{u, v\}$, but $\{u, v\} \subset V'$, so contraction would lead to a cycle. Let T have balance-point r, and let T_r be a shortest path tree rooted at r for the graph $(V', E' \cup \{e\})$, which does contain e. T_r has tree cost $K_r \leq K$. As T is optimal, T_r must also have cost K. Contraction of e in T_r yields a tree solution for $(G, b, c)/e$ of cost K. □

The next lemma says that vertices with communication bound 0 and degree 2 in an instance $(G = (V, E), b, c)$ can be neglected when proving Conjecture 2.1. Suppose $v' \in V$ has degree 2 in V, and $e_1 = \{u, v'\}, e_2 = \{v', w\}$ are the two edges incident with v'. Let $e_3 := \{u, w\}$ be a new edge. Then *shortcutting* v' in (G, b, c) results in the instance (G', b', c'), where $G' = (V \setminus v', E \setminus \{e_1, e_2\} \cup \{e_3\})$, $b'(v) = b(v)\ \forall v \in V \setminus v'$, and $c'(e_3) = c(e_1) + c(e_2)$, $c'(e) = c(e)\ \forall e \in E \setminus \{e_1, e_2\}$.

LEMMA 3.4. *Let $(G = (V, E), b, c)$ be an instance, and $v \in V$ a vertex of degree 2 in G with $b(v) = 0$. Denote the instance obtained from (G, b, c) by shortcutting v by (G', b', c'). Then (G', b', c') has a feasible dual solution of value K if and only if (G, b, c) has a feasible dual solution of value K. Moreover, (G', b', c') has a tree solution of cost K if and only if the same holds for (G, b, c).*

Proof. For the first statement, consider the dual solution (λ, μ) for (G, b, c). Define $(\hat{\lambda}, \hat{\mu})$ by $\hat{\lambda}^{e_3}_{ij} := \lambda^{e_1}_{ij} + \lambda^{e_2}_{ij}$, $\hat{\lambda}^e_{ij} := \lambda^e_{ij}\ \forall e \in E \setminus \{e_1, e_2\}$, and $\hat{\mu}_{ij} = \mu_{ij}\ \forall \{i, j\} \subseteq W$. For feasible (λ, μ) this yields a feasible $(\hat{\lambda}, \hat{\mu})$. Similarly, for a feasible dual solution $(\hat{\lambda}, \hat{\mu})$ for (G', b', c'), define a feasible (λ, μ) by taking, in particular, $\lambda^{e_1}_{ij} := \hat{\lambda}^{e_3}_{ij} c(e_1)/c'(e_3)$, and $\lambda^{e_2}_{ij} := \hat{\lambda}^{e_3}_{ij} c(e_2)/c'(e_3)$. A tree solution for (G', b', c') (not) using edge e_3 can be translated into a tree solution of the same cost for (G, b, c) (not) using both e_1 and e_2, and vice versa. □

We are now ready to justify that we restrict ourselves to *even* circuits in which each vertex has communication bound 1 (an even circuit is a circuit with an even number of vertices).

LEMMA 3.5. *If Conjecture* 2.1 *holds for every instance (G, b, c) where G is an even circuit and $b \equiv 1$, then it holds for every instance (G, b, c) where G is a circuit and b is arbitrary.*

Proof. Consider a general circuit instance $(G = C_n, b, c)$. From Lemma 3.4 we know that without loss of generality $b(v) > 0$ for all nodes v. Lemma 3.1 implies that we may assume that b is rational. Dividing b by $\gcd\{\frac{1}{2}b(v) \mid v \in V\}$ is allowed by Lemma 3.2. So we may assume that each $b(v)$ is a positive even integer. Finally, by Lemma 3.3, any path u, v, w, for v with $b(v) \neq 1$, may be substituted by a path $u, v_1, v_2, \ldots, v_N, w$, where $N = b(v)$, setting $b(v_1) = \cdots = b(v_N) = 1$, and $c(\{u, v_1\}) = c(\{u, v\})$, $c(\{v_k, v_{k+1}\}) = 0$, $c(\{v_N, w\}) = c(\{v, w\})$. Thus, we arrive at an even circuit with $b \equiv 1$. □

3.1. The even circuit with unit bounds: Properties.

Given an even circuit $G = C_{2n} = (V, E)$ on which all vertices have communication bound 1, we number the vertices starting at some vertex and following the circuit in a counterclockwise direction $0, 1, 2, \ldots, 2n-1$. All vertex and edge labels are taken modulo $2n$. The edge $\{i-1, i\}$ is denoted by e_i, or by i in case it is used as an index and no confusion with vertices is possible, $i = 0, \ldots, 2n-1$; e.g., we will write c_k for the unit cost $c(e_k)$ of edge e_k. On even circuits each edge e_k has an *opposite edge* e_{k+n}.

The cost of the tree solution on the circuit obtained from deleting the edge e_k is denoted by $C(e_k; c)$; we explicitly indicate dependence on c as the unit cost function of the edges, since we will use other unit cost functions later. Applying (3) yields $C(e_k; c) = \sum_{i=1}^{n-1} i(c_{k-i} + c_{k+i}) + nc_{k+n}$, which, by regrouping of terms, can be written as

$$C(e_k; c) = \sum_{j=k}^{k+n-1} \sum_{i=j+1}^{j+n} c_i = \sum_{j=k}^{k+n-1} H(j; c), \tag{4}$$

where $H(j; c) := \sum_{i=j+1}^{j+n} c_i$ is the so-called *half-sum* for vertex j, i.e., the sum of the unit costs of the n edges on the path starting in j, going in a counterclockwise direction, and ending with the edge opposite to e_j. From this expression the following equations are easily derived:

$$C(e_k; c) - C(e_{k+1}; c) = H(k; c) - H(k+n; c) \quad \forall\ k = 0, 1, \ldots, 2n-1; \tag{5}$$

$$H(j; c) - H(j-1; c) = c_{j+n} - c_j \quad \forall\ j = 0, 1, \ldots, 2n-1. \tag{6}$$

Now suppose that the tree obtained by deleting edge e_k has minimum cost among all spanning trees of the circuit, in other words $C(e_k; c) = \min_{e \in E} C(e; c)$. Then, using (5), we have

$$H(k; c) - H(k+n; c) = C(e_k; c) - C(e_{k+1}; c) \leq 0, \tag{7}$$

$$H(k-1; c) - H(k+n-1; c) = C(e_{k-1}; c) - C(e_k; c) \geq 0. \tag{8}$$

Subtracting (8) from (7) and applying (6) yields

$$\begin{aligned} 2c_{k+n} - 2c_k &= H(k; c) - H(k+n; c) - H(k-1; c) + H(k+n-1; c) \\ &= 2C(e_k; c) - C(e_{k+1}; c) - C(e_{k-1}; c) \leq 0. \end{aligned} \tag{9}$$

Consequently, if $c_k = 0$, then $c_{k+n} = 0$ and $2C(e_k; c) - C(e_{k+1}; c) - C(e_{k-1}; c) = 0$. Hence,

$$\begin{aligned} &c_k = 0 \wedge C(e_k; c) = \min\nolimits_{e \in E} C(e; c) \qquad \Rightarrow \\ &C(e_{k+1}; c) = C(e_{k-1}; c) = C(e_k; c) = \min\nolimits_{e \in E} C(e; c), \end{aligned} \tag{10}$$

i.e., if e_k minimizes $C(e; c)$ and $c_k = 0$, then e_{k-1} and e_{k+1} also minimize $C(e; c)$.

In the case of an even circuit with $b \equiv 1$, the constraints of the dual LP (2) reduce to the following. (In a circuit there are only two possible paths between any pair of vertices.)

$$\begin{aligned} &\sum_{j \in V \setminus \{i\}} \lambda_{ij}^e \leq c_e && \forall\ i \in V,\ \forall\ e \in E, \\ &\mu_{ij} \leq \sum_{l=i+1}^{j} \lambda_{ij}^{e_l} && \forall\ \{i, j\} \subseteq V, \\ &\mu_{ij} \leq \sum_{l=j+1}^{i} \lambda_{ij}^{e_l} && \forall\ \{i, j\} \subseteq V, \\ &\lambda_{ij}^e \geq 0 && \forall\ \{i, j\} \subseteq V,\ \forall\ e \in E. \end{aligned} \tag{11}$$

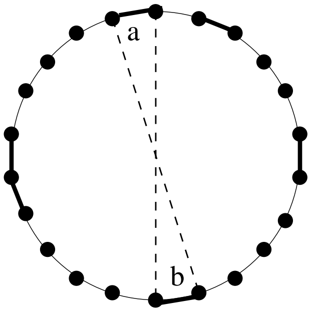


FIG. 1. *Example of a singular subset, denoted by bold lines.*

3.2. The even circuit with unit bounds: Proof. Given a cost function $c : E \to \mathbb{R}_+$, we call the set of edges with nonzero unit cost the *support* of c, i.e., $\operatorname{supp}(c) = \{e \in E | c_e > 0\}$. The following lemma is crucial to the main result.

LEMMA 3.6. *Let $G = (V, E) = C_{2n}$ be an even circuit, and let $b \equiv 1$. Let F be a nonempty subset of E. Then there exist a nonnegative cost function $\hat{c} : E \to \mathbb{R}_+$, not identical to* 0, *with* $\operatorname{supp}(\hat{c}) \subseteq F$, *and a constant K, such that $\forall f \in F$, $K = C(f; \hat{c}) = \min_{e \in E} C(e; \hat{c})$. Moreover, there is a dual solution $(\hat{\lambda}, \hat{\mu})$ for the problem with cost function $\hat{c}$, with value K.*

Proof. The proof is by induction on $|F|$. The theorem is clearly true if $|F| = 1$. For suppose $F = \{e_k\}$. Then we can take $\hat{c}_k = 1$ and $\hat{c}_i = 0$ for other i. Clearly, $\min_{e \in E} C(e; \hat{c}) = C(e_k; \hat{c}) = 0$. A feasible dual solution with value 0 is $\hat{\lambda}^e_{ij} = 0$, $\hat{\mu}_{ij} = 0$ $\forall e \in E$, $\forall \{i, j\} \subseteq V$. For $|F| > 1$ we distinguish three cases.

Case 1. There exists a k such that $e_k \in F$ and its opposite $e_{k+n} \in F$.

In this case we call the edge set F *singular* (see Figure 1).

Consider the cost function $\hat{c} : E \to \mathbb{R}_+$ defined by $\hat{c}_k = \hat{c}_{k+n} = 1$ and $\hat{c}_i = 0$ otherwise. It satisfies $C(e; \hat{c}) = n$ $\forall e \in E$. The following dual solution is feasible with respect to this cost function $\hat{c}$ (see (11)) and has objective value $\sum_{i<j} \hat{\mu}_{ij} = n$:

$$\begin{aligned}
\hat{\lambda}^{e_k}_{i,i+n} &:= 1 \quad && \forall\ i = 0, 1, \dots, n-1,\\
\hat{\lambda}^{e_{k+n}}_{i,i+n} &:= 1 \quad && \forall\ i = 0, 1, \dots, n-1,\\
\hat{\lambda}^{e}_{ij} &:= 0 \quad && \text{otherwise},\\
\hat{\mu}_{i,i+n} &:= 1, \quad && i = 0, \dots, n-1,\\
\hat{\mu}_{ij} &:= 0 \quad && \text{otherwise}.
\end{aligned}$$

Case 2. F is not singular, and there exist k and m, $k < m < k + n$, such that $e_k \in F$, $e_m \in F$, $e_l \notin F$ $\forall k < l < m$, and $e_l \notin F$ $\forall k + n \leq l \leq m + n$.

See Figure 2 for an example, with $e = e_k$ and $f = e_m$.

We *contract* the edges $e_k, \dots, e_m$ to a new edge $e' = (k - 1, m)$, and the edges $e_{k+n}, \dots, e_{m+n}$ to a new edge $\bar{e}' = (k - 1 + n, m + n)$, to arrive at a new even cycle (V', E'), with $|V'| = |E'| = 2(n - m + k)$. We maintain the vertex labels $V' = \{m, m + 1, \dots, k - 1 + n, m + n, \dots, k - 1\}$. The new edge set is $E' = \{e_{m+1}, \dots, e_{k-1+n}, \bar{e}', e_{m+n+1}, \dots, e_{k-1}, e'\}$. Note that edges that were opposite before contraction remain opposite after contraction. Also, the new edges e' and $\bar{e}'$ are opposite.

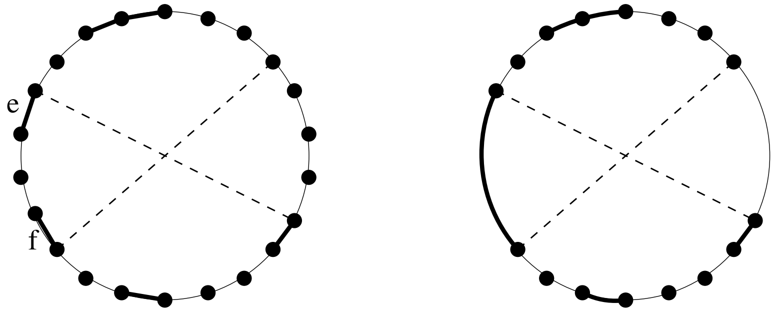


FIG. 2. *Contracting the circuit to a smaller one.*

Consider the subset of edges $F' = F\backslash\{e_k, e_m\} \cup \{e'\}$. As $0 < |F'| < |F|$, we can apply the induction hypothesis. Thus, there exist a cost function $c' : E' \to \mathbb{R}_+$, not identical to 0, with $\text{supp}(c') \subseteq F'$, a constant $K' = \min_{e\in E'} C(e; c') = C(f; c')\ \forall f \in F'$, and a dual solution (λ, μ), with value K'.

Since e' is a minimizer of $C(e; c')$, using (9) and the fact that $c'(\bar{e}') = 0$, we have

$$c'(e') = \frac{1}{2}(H'(m + n; c') - H'(m; c')) + \frac{1}{2}(H'(k - 1; c') - H'(k + n - 1; c')),$$

where $H'(j; c')$ is the half-sum for vertex j on the smaller, contracted, circuit with its corresponding cost function c'. Now we define the cost $\hat{c} : E \to \mathbb{R}_+$ from c' as follows:

$$\begin{aligned}
&\hat{c}_m := \tfrac{1}{2}(H'(m + n; c') - H'(m; c')), \\
&\hat{c}_k := \tfrac{1}{2}(H'(k - 1; c') - H'(k + n - 1; c')), \\
&\hat{c}_i := c'_i && \forall\ \{i - 1, i\} \in F,\ i \neq k, m, \\
&\hat{c}_i := 0 && \forall\ \{i - 1, i\} \in E\backslash F.
\end{aligned}$$

It follows from (7) and (8) that $\hat{c}_m \geq 0$ and $\hat{c}_k \geq 0$. Note that $\hat{c}_k + \hat{c}_m = c'(e')$ and

$$\begin{aligned}
\hat{c}_k + H'(k + n - 1; c') = \hat{c}_m + H'(m; c') &= \frac{1}{2}\sum_{e\in E'} c'_e = \frac{1}{2}\sum_{f\in F'} c'_f \\
&= \frac{1}{2}\sum_{f\in F} \hat{c}(f) = \frac{1}{2}\sum_{e\in E} \hat{c}(e).
\end{aligned}$$

Hence, for the half-sums in the larger circuit, we have

$$\begin{aligned}
&H(j; \hat{c}) = \hat{c}_m + H'(m; c') = \frac{1}{2}\sum_{e\in E} \hat{c}(e), && j = k, \dots, m - 1, \\
&H(j; \hat{c}) = \hat{c}_k + H'(k + n - 1; c') = \frac{1}{2}\sum_{e\in E} \hat{c}(e), && j = k + n, \dots, m - 1 + n, \\
&H(j; \hat{c}) = H'(j; c') && \text{otherwise.}
\end{aligned}$$

Using this in (4) yields $C(e_k; \hat{c}) = C(e_m; \hat{c})$, and $C(e_k; \hat{c}) - C(f; \hat{c}) = C(e'; c') - C(f; c')$ $\forall f \in F' \setminus \{e'\}$. Hence, $C(f; \hat{c}) = \min_{e\in E} C(e; \hat{c}) = K\ \forall f \in F$, with $K = K' + \frac{1}{2}(m - k)\sum_{e\in E} \hat{c}(e)$.

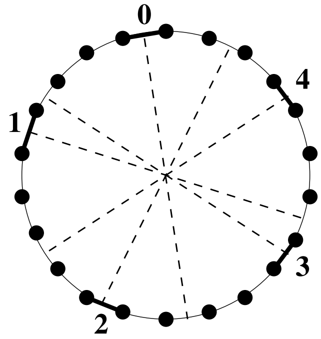


FIG. 3. *Example of an edge set with alternating edges and opposites.*

To define the feasible dual solution with value K we represent V as $V = V' \cup V_{e'} \cup V_{\bar{e}'}$, with $V_{e'} = \{k, k+1, \dots, m-1\}$ and $V_{\bar{e}'} = \{k+n, k+n+1, \dots, m+n-1\}$. Again we construct the dual solution $(\hat{\lambda}, \hat{\mu})$ for $(V, E, \hat{c})$ from the solution (λ, μ) associated with (V', E', c'):

$$\begin{aligned}
\hat{\lambda}^e_{ij} &:= \lambda^e_{ij} && \text{for } e \in F \setminus \{e_k, e_m\}, \{i,j\} \subseteq V', \\
\hat{\lambda}^{e_k}_{ij} &:= \frac{\lambda^{e'}_{ij}\hat{c}_k}{c'(e')} && \text{for } \{i,j\} \subseteq V', \\
\hat{\lambda}^{e_m}_{ij} &:= \frac{\lambda^{e'}_{ij}\hat{c}_m}{c'(e')} && \text{for } \{i,j\} \subseteq V', \\
\hat{\lambda}^e_{ij} &:= \hat{c}(e) && \text{for } e \in E, i \in V_{e'}, j = i+n, \\
\hat{\lambda}^e_{ij} &:= 0 && \text{otherwise,} \\
\hat{\mu}_{ij} &:= \mu_{ij} && \text{for } \{i,j\} \subseteq V', \\
\hat{\mu}_{ij} &:= \frac{1}{2}\sum_{e \in E} \hat{c}(e) && \text{for } i \in V_{e'}, j = i+n, \\
\hat{\mu}_{ij} &:= 0 && \text{otherwise.}
\end{aligned}$$

Case 2 is settled by easy verification that $(\hat{\lambda}, \hat{\mu})$ satisfies the dual constraints (11) and has value

$$\sum_{i,j \in V, i<j} \hat{\mu}_{ij} = \sum_{i,j \in V', i<j} \mu_{ij} + \sum_{i \in V_{e'}} \hat{\mu}_{i,i+n} = K' + \frac{1}{2}(m-k)\sum_{e \in E} \hat{c}(e) = K.$$

Case 3. F is not singular, and Case 2 does not apply, meaning that between each pair of consecutive edges from F, e and f say, there is exactly one opposite edge $\bar{g}$ of some edge $g \in F$.

See Figure 3 for an example.

Hence F consists of an odd number of edges, $2k+1$ say, such that these edges and their opposites are perfectly alternating. As a consequence, an edge $e \in F$ and its opposite $\bar{e}$ split the other edges of F into two groups of size k. Therefore, we call this a *uniform* configuration. Let $F = \{f_0, f_1, \ldots, f_{2k}\}$, with $f_i = (m_i - 1, m_i)$, and denote the number of nodes between edge f_i and f_{i+1} by ν_i, i.e., $\nu_i = m_{i+1} - m_i$ modulo $2n$. We have that $\sum_{i=0}^{2k} \nu_i = 2n$. Please note that from now on subscripts i for ν, f, etc., are taken modulo $2k+1$. Define

$$\tau_i := \sum_{j=0}^{k} \nu_{i+j} - \sum_{j=k+1}^{2k} \nu_{i+j} = 2(m_{i+k+1} - m_i) \bmod 2n,\ i = 0, 1, \ldots, 2k.$$

By uniformity, $\sum_{j=0}^{k} \nu_{i+j} > \sum_{j=k+1}^{2k} \nu_{i+j}$, whence $\tau_i > 0$, $i = 0, \ldots, 2k$. Notice that $\frac{1}{2}\tau_i$ is equal to the number of nodes between $\bar{f}_i$ and f_{i+k+1} (or equivalently between f_i and $\bar{f}_{i+k+1}$). Note that, by definition, $\nu_{i+k} = \frac{1}{2}(\tau_i + \tau_{i+k})$. We propose the cost function

$$\hat{c}(f_i) := \frac{\nu_{i+k}}{\tau_i \tau_{i+k}} = \frac{1}{2}\left(\frac{1}{\tau_i} + \frac{1}{\tau_{i+k}}\right),\ i = 0, 1, \ldots, 2k; \quad \hat{c}(e) := 0,\ e \notin F.$$

Observe the following identity: for $i = 0, 1, \ldots, 2k$,

$$\begin{aligned}
&\sum_{j=i}^{i+k} \hat{c}(f_j) - \sum_{j=i+k+1}^{i+2k} \hat{c}(f_j) \\
&\qquad = \sum_{j=i}^{i+k-1} (\hat{c}(f_j) - \hat{c}(f_{j+k+1})) + \hat{c}(f_{i+k}) \\
&\qquad = \frac{1}{2}\left(\sum_{j=i}^{i+k-1}\left(\left(\frac{1}{\tau_j} + \frac{1}{\tau_{j+k}}\right) - \left(\frac{1}{\tau_{j+k+1}} + \frac{1}{\tau_j}\right)\right) + \left(\frac{1}{\tau_{i+k}} + \frac{1}{\tau_{i+2k}}\right)\right) \qquad (12) \\
&\qquad = \frac{1}{\tau_{i+k}} > 0.
\end{aligned}$$

As a consequence, we have that consecutive edges from F yield the same cost, since

$$\begin{aligned}
&C(f_i; \hat{c}) - C(f_{i+1}; \hat{c}) \\
&\qquad = C(f_i; \hat{c}) - C(\bar{f}_{i+k+1}; \hat{c}) + C(\bar{f}_{i+k+1}; \hat{c}) - C(f_{i+1}; \hat{c}) \\
&\qquad = \frac{1}{2}\tau_i(\hat{c}(f_{i+1}) + \cdots + \hat{c}(f_{i+k}) - \hat{c}(f_{i+k+1}) - \cdots - \hat{c}(f_{i+2k+1})) \qquad (13) \\
&\qquad\quad + \frac{1}{2}\tau_{i+k+1}(\hat{c}(f_{i+1}) + \cdots + \hat{c}(f_{i+k+1}) - \hat{c}(f_{i+k+2}) - \cdots - \hat{c}(f_{i+2k+1})) \\
&\qquad = \frac{1}{2}\tau_i\left(-\frac{1}{\tau_i}\right) + \frac{1}{2}\tau_{i+k+1}\frac{1}{\tau_{i+k+1}} = 0.
\end{aligned}$$

Using (10), we conclude that $C(f_i; \hat{c}) = \min_{e \in E} C(e; \hat{c})\ \forall f_i \in F$.

We introduce the following entities to facilitate the exposition of the dual solution that we propose:

$$\Lambda_i^e := \begin{cases} \dfrac{1}{2}\dfrac{\tau_i}{\nu_i\nu_{i+k}}\hat{c}(e)\left(1+\dfrac{1}{\tau_i}\dfrac{1}{\sum_j \hat{c}(f_j)}\right) & \text{if } f_i < e \leq f_{i+k}, \\ \\ \dfrac{1}{2}\dfrac{\tau_i}{\nu_i\nu_{i+k}}\hat{c}(e)\left(1-\dfrac{1}{\tau_i}\dfrac{1}{\sum_j \hat{c}(f_j)}\right) & \text{if } f_{i+k} < e \leq f_{i+2k+1}. \end{cases}$$

Now we define the dual solution as

$$\hat{\lambda}_{st}^e := \begin{cases} \Lambda_i^e & \forall\ e \in E,\ \ \forall\ i = 0, 1, \ldots, 2k,\ \forall\ s, t : m_i \leq s < m_{i+1}, m_{i+k} \leq t < m_{i+k+1}, \\ \\ 0 & \text{otherwise}, \end{cases}$$

$$\hat{\mu}_{st} := \sum_{e:s<e\leq t} \hat{\lambda}_{st}^e \ \ \forall\ \{s,t\} \subseteq V.$$

To verify feasibility of this solution, first notice that from (12) we have

$$0 < \frac{1}{\tau_{i+k}}\frac{1}{\sum_j \hat{c}(f_j)} \leq \frac{\sum_{j=i}^{i+k}\hat{c}(f_j)}{\sum_j \hat{c}(f_j)} \leq 1 \quad \forall\ i,$$

whence $\lambda_{uv}^e \geq 0$ is satisfied $\forall u, v \in V$ and $\forall e \in E$. Simple algebraic computations show that, $\forall e \in E$, $\forall i = 0, 1, \ldots, 2k$, and $\forall s$ with $m_i \leq s < m_{i+1}$,

$$\sum_{t=m_{i+k}}^{m_{i+k+1}-1} \hat{\lambda}_{st}^e + \sum_{t=m_{i+k+1}}^{m_{i+k+2}-1} \hat{\lambda}_{ts}^e = \nu_{i+k}\Lambda_i^e + \nu_{i+k+1}\Lambda_{i+k+1}^e \leq \hat{c}(e).$$

Actually, the inequality is tight except for $e = f_{i+k+1}$.

To show that $(\hat{\lambda}, \hat{\mu})$ satisfies the second and third types of dual constraints in (11), it suffices to show that the sum of $\hat{\lambda}$-values over edges along the s–t path is the same as the sum of $\hat{\lambda}$-values over edges along the t–s path; i.e., $\sum_{e:s<e\leq t}\hat{\lambda}_{st}^e = \sum_{e:t<e\leq s}\hat{\lambda}_{st}^e$ $\forall s, t \in V$. Clearly we need to show this only for pairs $\{s, t\}$ with $m_i \leq s < m_{i+1}, m_{i+k} \leq t < m_{i+k+1}$ for some i. The claim follows straightforwardly from the fact that the equality

$$\begin{aligned} &(\hat{c}(f_{i+1}) + \cdots + \hat{c}(f_{i+k}))\left(1 + \frac{1}{\tau_i}\frac{1}{\sum_f \hat{c}(f)}\right) \\ &= (\hat{c}(f_{i+k+1}) + \cdots + \hat{c}(f_{i+2k+1}))\left(1 - \frac{1}{\tau_i}\frac{1}{\sum_f \hat{c}(f)}\right) \end{aligned}$$

is equivalent to the equality

$$\begin{aligned} &\frac{1}{\tau_i}\frac{1}{\sum_f \hat{c}(f)}(\hat{c}(f_{i+1}) + \cdots + \hat{c}(f_{i+k}) + \hat{c}(f_{i+k+1}) + \cdots + \hat{c}(f_{i+2k+1})) \\ &= (\hat{c}(f_{i+k+1}) + \cdots + \hat{c}(f_{i+2k+1})) - (\hat{c}(f_{i+1}) + \cdots + \hat{c}(f_{i+k})), \end{aligned}$$

which is evident, since the right-hand side equals $\frac{1}{\tau_i}$ by (12). This completes the feasibility check.

It remains to verify that $\sum_{\{s,t\}\subset V} \hat{\mu}_{st} = C(f_0;\hat{c})$. We use that $C(f_0;\hat{c}) = C(f_i;\hat{c})\ \forall i = 0,1,\dots,2k$ implies that $C(f_0;\hat{c}) = \frac{1}{2k+1}\sum_{i=0}^{2k} C(f_i;\hat{c})$. According to (4), the half-sums that constitute the cost $C(f_i;\hat{c})$ start in nodes $j = m_i,\dots,m_i+n-1$. These starting nodes can be subdivided into $2k+1$ subsets: for $j = i,\dots,i+k$ the half-sums starting in the $\frac{1}{2}\tau_j$ nodes between f_j and $\bar{f}_{j+k+1}$ all have value $\sum_{t=j+1}^{j+k} \hat{c}(f_t)$, whereas for $j = i,\dots,i+k-1$ the half-sums starting in the $\frac{1}{2}\tau_{j+k+1}$ nodes between $\bar{f}_{j+k+1}$ and f_{j+1} all have value $\sum_{t=j+1}^{j+k+1} \hat{c}(f_t)$. Altogether, the total cost is

$$\begin{aligned}\frac{1}{2k+1}\sum_{i=0}^{2k} C(f_i;\hat{c}) &= \frac{1}{2k+1}\sum_{i=0}^{2k}\left(\sum_{j=i}^{i+k-1}\left(\frac{1}{2}\tau_j\sum_{t=j+1}^{j+k}\hat{c}(f_t) + \frac{1}{2}\tau_{j+k+1}\sum_{t=j+1}^{j+k+1}\hat{c}(f_t)\right)\right.\\ &\qquad\qquad\qquad\qquad\left. + \frac{1}{2}\tau_{i+k}\sum_{t=i+k+1}^{i+2k}\hat{c}(f_t)\right)\\ &= \sum_{i=0}^{2k}\left(\frac{1}{2}\tau_i\frac{k+1}{2k+1}\sum_{j=i+1}^{i+k}\hat{c}(f_j) + \frac{1}{2}\tau_{i+k+1}\frac{k}{2k+1}\sum_{j=i+1}^{i+k+1}\hat{c}(f_j)\right). \qquad (14)\end{aligned}$$

In turn, the value of the dual solution can be rewritten as

$$\begin{aligned}\sum_{\{s,t\}\subset V}\hat{\mu}_{st} &= \sum_{i=0}^{2k}\nu_i\nu_{i+k}\sum_{e=f_{i+1},\dots,f_{i+k}}\Lambda_i^e\\ &= \sum_{i=0}^{2k}\frac{1}{2}\tau_i\left(\sum_{j=i+1}^{i+k}\hat{c}(f_j) + \frac{1}{\tau_i}\frac{\sum_{j=i+1}^{i+k}\hat{c}(f_j)}{\sum_f \hat{c}(f)}\right)\\ &= \sum_{i=0}^{2k}\frac{1}{2}\tau_i\left(\sum_{j=i+1}^{i+k}\hat{c}(f_j) + \frac{1}{\tau_i}\frac{k}{2k+1}\right)\\ &= \sum_{i=0}^{2k}\frac{1}{2}\tau_i\left(\sum_{j=i+1}^{i+k}\hat{c}(f_j) + \left(\sum_{j=i+k+1}^{i+2k+1}\hat{c}(f_j) - \sum_{j=i+1}^{i+k}\hat{c}(f_j)\right)\frac{k}{2k+1}\right)\\ &= \sum_{i=0}^{2k}\frac{1}{2}\tau_i\left(\frac{k+1}{2k+1}\sum_{j=i+1}^{i+k}\hat{c}(f_j) + \frac{k}{2k+1}\sum_{j=i+k+1}^{i+2k+1}\hat{c}(f_j)\right),\end{aligned}$$

which equals (14). This settles the proof of Case 3, and thereby the proof of the lemma. □

We are now ready to prove the main result of this section.

THEOREM 3.7. *Let $G = (V,E)$ be an even circuit, let $c : E \to \mathbb{R}_+$, and let $b(i) = 1\ \forall i \in V$. Then the cost of an optimal tree solution equals the value of an optimal dual solution.*

Proof. The proof is by induction on $|\operatorname{supp}(c)|$. The theorem is clearly true if $|\operatorname{supp}(c)| = 1$, when deleting the only edge with positive unit cost yields a tree solution with total cost 0. Setting all dual variables to 0 is feasible and yields value 0.

Now suppose $|\operatorname{supp}(c)| > 1$. Lemma 3.6 (applied to $F := \operatorname{supp}(c)$) tells us that there exist a nonnegative nonzero cost function $\hat{c}$, such that $C(f;\hat{c}) = \min_{e\in E} C(e;\hat{c})$ $\forall f \in \operatorname{supp}(c)$, and a dual solution $(\hat{\lambda},\hat{\mu})$ with respect to $\hat{c}$ with the same objective value. Define cost vector c' as $c' := c - \sigma\hat{c}$, where σ is a scalar chosen such that c' is nonnegative and at least one $f \in \operatorname{supp}(c)$ has $c'(f) = 0$. Such a scalar exists, since $\operatorname{supp}(\hat{c}) \subseteq \operatorname{supp}(c)$ and $\operatorname{supp}(\hat{c}) \neq \emptyset$.

Let $C(e^*; c') = \min_{e \in E} C(e; c')$. By (10), we can assume that $e^* \in \operatorname{supp}(c') \subset \operatorname{supp}(c)$. Since $|\operatorname{supp}(c')| < |\operatorname{supp}(c))|$, the induction hypothesis may be applied to c', giving a feasible dual (λ', μ') with respect to c' of value $\sum \mu'_{vw} = C(e^*; c')$. The solution $(\lambda, \mu) := (\lambda', \mu') + \sigma(\hat{\lambda}, \hat{\mu})$ is feasible with respect to c, as $c = c' + \sigma\hat{c}$. Its value is equal to $C(e^*; c') + \sigma C(e^*; \hat{c}) = \min_{e \in E} C(e; c)$. □

This theorem together with Lemma 3.5 implies our main result, Theorem 1.2.

4. Other cases where the conjecture holds. In this section, we present other classes of instances of the VPN problem on which the Conjecture 2.1 holds, which is equivalent to stating that $OPT(\text{TR}) = OPT(\text{MPR})$ on these instances. We will use the shorthand notation c_{uv} for $c(\{u, v\})$.

We start with the observation that Conjecture 2.1 holds for trees. This is trivial, since it is equivalent to the statement that $OPT(\text{TR}) = OPT(\text{MPR})$, which is trivially true for trees. In fact, it is possible to construct, for any instance (G, b, c) where G is a tree, an explicit dual (λ, μ) with value equal to the cost of the tree.

Indeed, denote the unique path in the tree G between two distinct vertices $i, j \in V$ by P_{ij}. Let K be the cost of the tree G given by (3): $K = \sum_e \min\{b(L_e), b(R_e)\}c(e)$, where L_e and R_e are the two components of $G - e$. Define the dual as follows:

$$\begin{aligned}
\lambda^e_{ij} &:= \frac{b(i)b(j)}{b(L_e)b(R_e)} \min\{b(L_e), b(R_e)\}c(e) && \text{if } |\{i, j\} \cap L_e| = 1,\\
\lambda^e_{ij} &:= 0 && \text{otherwise},\\
\mu_{ij} &:= \sum_{e \in P_{ij}} \lambda^e_{ij} && \text{for } \{i, j\} \subseteq V.
\end{aligned}$$

Then λ, μ is a feasible dual: for $i \in L_e$ we have

$$\begin{aligned}
\sum_{j \neq i} \lambda^e_{ij} &= \sum_{j \in R_e} \frac{b(i)b(j)}{b(L_e)b(R_e)} \min\{b(L_e), b(R_e)\}c(e)\\
&= \frac{b(i)}{b(L_e)} \min\{b(L_e), b(R_e)\}c(e) \leq c(e)b(i),
\end{aligned}$$

and similarly for $i \in R_e$. Since there is only one path between any two vertices i and j, the constraint for μ holds by definition. The value of this dual is, as required,

$$\sum_{i,j} \mu_{ij} = \sum_e \sum_{i,j: e \in P_{ij}} \lambda^e_{ij} = \sum_e \min\{b(L_e), b(R_e)\}c(e) = K.$$

As we gather from the next lemma, for proving Conjecture 2.1, we may assume that the graph G is complete, and that the cost function c is a metric (satisfies the triangle inequality), i.e., $c_{uw} \leq c_{uv} + c_{vw}$ for any three vertices u, v, and w.

An observation that we will use frequently is that for an instance (G, b, c), with G complete and c a metric, an optimal tree solution T, with *balance-point* r, has cost (cf. Proposition 2.2) $\sum_{v \in V} b(v) d^c_T(r, v)$, which is equal to the cost of the star subgraph centered at r. Thus, such instances always have a star as an optimal tree solution.

LEMMA 4.1. *Let $G = (V, E)$, $b : V \to \mathbb{R}_+$, $c : E \to \mathbb{R}_+$ be given. Let $H = (V, F)$ be the complete graph on V. Define $c'(\{u, v\})$ for $\{u, v\} \in F$ as the length of a shortest* *path between u and v with respect to the length function c. If Conjecture* 2.1 *is true for the instance (H, b, c') and the optimal tree solution has value K, then it is also true for the instance (G, b, c) with the same optimal tree solution value K.*

Proof. A dual solution (λ, μ) which is feasible for c' and F is also feasible for c and E, since $c'_e \leq c_e$ $\forall e \in E$. Take an optimal tree solution S for (H, b, c'), having balance-point r and cost $\sum_{v \in V} b(v) d_S^{c'}(r, v)$. As argued above we may assume that S is a star centered at r. Replacing the star S by a shortest path tree T (with respect to c) rooted at r, we obtain $\sum_v b(v) d_T^c(r, v) = \sum_v b(v) d_S^{c'}(r, v)$. □

The next lemma is in itself not very significant, but the result allows for developing further proof tools presented directly hereafter.

LEMMA 4.2. *For any instance $(G = (V, E), b, c)$ such that $r \in V$ exists with $b(r) \geq \sum_{v \neq r} b(v)$, the cost of an optimal tree solution equals the value of an optimal dual solution.*

Proof. By Lemma 4.1, we may assume that G is complete, and that c is metric, whence a star S is an optimal tree solution. The cost of the star centered at r is at most the cost of any other star. Indeed, by Proposition 2.2, the star centered at $s \neq r$ has cost

$$\begin{aligned}
\sum_{v \neq s} b(v) c_{sv} &= b(r) c_{sr} + \sum_{s \neq v \neq r} b(v) c_{sv} \\
&\geq \sum_{v \neq r} b(v) c_{sr} + \sum_{s \neq v \neq r} b(v) c_{sv} \\
&= b(s) c_{sr} + \sum_{s \neq v \neq r} b(v)(c_{sr} + c_{sv}) \\
&\geq b(s) c_{sr} + \sum_{s \neq v \neq r} b(v) c_{rv} \\
&= \sum_{v \neq r} b(v) c_{rv},
\end{aligned}$$

where the first inequality holds by assumption and the second one by the triangle inequality. Therefore, the cost of an optimal tree solution is $\sum_{v \neq r} b(v) c_{rv}$.

Define λ and μ as

$$\begin{aligned}
\lambda_{rv}^e &:= c_e b(v) && \text{for } v \neq r \text{ and } e \in E, \\
\lambda_{uv}^e &:= 0 && \text{for } u \neq r \neq v \text{ and } e \in E, \\
\mu_{rv} &:= c_{rv} b(v) && \text{for } v \neq r, \\
\mu_{uv} &:= 0 && \text{for } u \neq r \neq v.
\end{aligned}$$

Using the assumption of the lemma and the triangle inequality, it is not hard to check that (λ, μ) is feasible for (2) the dual of MPR and has value $\sum_{uv \in F} \mu_{uv} = \sum_{v \neq r} c_{rv} b(v)$. □

The next lemma shows that the property that Conjecture 2.1 holds is preserved under taking 1-sums. A 1-*sum* of two graphs is the graph obtained by identifying a vertex of one graph with a vertex of the other graph. More precisely, let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be disjoint graphs, take any $v_1 \in V_1$ and $v_2 \in V_2$ and identify them, creating a vertex z, which is then the only vertex common to V_1 and V_2, i.e., $V_1 \cap V_2 = \{z\}$. The 1-*sum of* G_1 *and* G_2 *in* z is then the graph $G = (V_1 \cup V_2, E_1 \cup E_2)$.

Lemma 4.3. *Let $G = (V, E)$ be the 1-sum of $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ in a vertex z. Then, for any $b : V \to \mathbb{R}_+$ and $c : E \to \mathbb{R}_+$, the cost of an optimal tree solution equals the value of an optimal dual solution for (G, b, c) if the same holds for (G_1, b_1, c_1) for every $b_1 : V_1 \to \mathbb{R}_+$, and for (G_2, b_2, c_2) for every $b_2 : V_2 \to \mathbb{R}_+$, where c_i denotes the restriction of c to E_i, $i = 1, 2$.*

Proof. Define $B_i := b(V_i)$ for $i = 1, 2$. Then without loss of generality, $B_1 \leq B_2$. Consider the instance (G_1, b_1, c_1) with $b_1(z) = B_2$ and $b_1(v) = b(v), v \in V_1 - z$. For this instance z is a *dominant* vertex in the sense of Lemma 4.2 and therefore the shortest path tree T_1 rooted at z is a tree solution of minimum cost $K_1 := \sum_{v \in V_1} b_1(v) d^{c_1}_{T_1}(z, v)$. Similarly, consider the instance (G_2, b_2, c_2) with $b_2(z) = B_1$ and $b_2(v) = b(v), v \in V_2 - z$. This instance has a tree solution T_2 of minimum cost K_2, which is a shortest path tree with respect to c_2 as length function on the edges, rooted at some $r \in V_2$ (not necessarily $r = z$ this time). Now, remembering (3), it is easy to see that $T := T_1 \cup T_2$ is a tree solution of (G, b, c), rooted at r, with cost $K_1 + K_2$.

Next, we show that a feasible dual solution for (G, b, c) exists with value $K_1 + K_2$. We use the fact that a feasible dual solution λ^i, μ^i of value K_i exists for (G_i, b_i, c_i), $i = 1, 2$. Define

$$\begin{aligned}
\lambda^e_{st} &:= \frac{b(t)}{B_2}\lambda^{1,e}_{sz} && \text{for } s \in V_1 - z,\ t \in V_2, \text{ and } e \in E_1,\\
\lambda^e_{st} &:= \frac{b(s)}{B_1}\lambda^{2,e}_{zt} && \text{for } s \in V_1,\ t \in V_2 - z, \text{ and } e \in E_2,\\
\lambda^e_{st} &:= \lambda^{1,e}_{st} && \text{for } s, t \in V_1 - z,\ e \in E_1,\\
\lambda^e_{st} &:= \lambda^{2,e}_{st} && \text{for } s, t \in V_2 - z,\ e \in E_2,\\
\lambda^e_{st} &:= 0 && \text{otherwise},\\
\mu_{st} &:= \mu^1_{st} && \text{for } s \in V_1 - z,\ t \in V_1 - z,\\
\mu_{st} &:= \mu^2_{st} && \text{for } s \in V_2 - z,\ t \in V_2 - z,\\
\mu_{st} &:= \frac{b(t)}{B_2}\mu^1_{sz} + \frac{b(s)}{B_1}\mu^2_{zt} && \text{for } s \in V_1,\ t \in V_2 \text{ (where } \mu^1_{zz} := 0 =: \mu^2_{zz}).
\end{aligned}$$

We verify feasibility for $s \in V_1 - z$, $e \in E_1$, using that $\sum_{t \in V_2} \frac{b(t)}{B_2} = 1$,

$$\sum_{t \neq s} \lambda^e_{st} = \sum_{t \in V_1 - z - s} \lambda^{1,e}_{st} + \sum_{t \in V_2} \frac{b(t)}{B_2}\lambda^{1,e}_{sz} = \sum_{t \in V_1 - s} \lambda^{1,e}_{st} \leq c_1(e) b_1(s),$$

and similarly for $s \in V_2 - z, e \in E_2$. For $s \in V_1$, $e \in E_2$ we verify

$$\sum_{t \neq s} \lambda^e_{st} = \sum_{t \in V_2 - z} \frac{b(s)}{B_1}\lambda^2_{zt}(e) \leq \frac{b(s)}{B_1} b_2(z) c_2(e) = b(s)c(e),$$

and we do similarly for $s \in V_2$, $e \in E_1$. The definition of μ is exactly such that it satisfies the constraints for μ in (2). Finally, we have

$$\begin{aligned}\sum_{s,t}\mu_{st} &= \sum_{s,t\in V_1-z}\mu^1_{st} + \sum_{s,t\in V_2-z}\mu^2_{st} + \sum_{s\in V_1-z}\sum_{t\in V_2}\frac{b(t)}{B_2}\mu^1_{sz} + \sum_{s\in V_1}\sum_{t\in V_2-z}\frac{b(s)}{B_1}\mu^2_{zt}\\ &= \sum_{s,t\in V_1-z}\mu^1_{st} + \sum_{s,t\in V_2-z}\mu^2_{st} + \sum_{s\in V_1-z}\mu^1_{sz} + \sum_{t\in V_2-z}\mu^2_{zt}\\ &= \sum_{s,t\in V_1}\mu^1_{st} + \sum_{s,t\in V_2}\mu^2_{st} = K_1 + K_2. \qquad \square\end{aligned}$$

As a corollary of this lemma we obtain that Conjecture 2.1 holds for graphs in which each block (defined below) is an edge or a circuit. This gives a slight extension of Theorem 1.2. We use the following definitions. The *connectivity* of a graph $G = (V, E)$ is the minimum size of a subset U of V for which $G - U$ is not connected. If no such U exists (or, equivalently, if G is complete), then the connectivity is ∞. A graph is k-connected if its connectivity is at least k. Now, a k-connected component of a graph $G = (V, E)$ is an inclusion-wise maximal subset U of V for which $G[U]$ (the subgraph of G induced by the vertices in U) is k-connected. A *block* is a 2-connected component U with $|U| \geq 2$. We identify the blocks of a graph with the subgraphs they induce. A connected graph may be obtained from its blocks by taking repeated 1-sums.

THEOREM 4.4. *If G is a graph in which each block is an edge or a circuit, then the cost of an optimal tree solution equals the value of an optimal dual solution for any instance (G, b, c).*

Proof. We obtain the proof directly from Theorem 1.2, from the fact that Conjecture 2.1 holds for trees, and from Lemma 4.3, since G is a 1-sum of edges (a special kind of trees) and circuits. □

Note that the class of graphs described in the above theorem contains those graphs that are often referred to in the network literature as "trees of rings" (see [5]).

The next lemma says that if Conjecture 2.1 holds for an instance, it still holds if we add edges to the graph with cost equal to the length of a shortest path between their end points. It provides a kind of converse to Lemma 4.1.

LEMMA 4.5. *Suppose for the instance $(G = (V, E), b, c)$ a tree solution of cost K and a feasible dual solution of value K exist, and suppose $f := \{s, t\} \notin E$ for $s, t \in V$. Let the instance (G', b, c') be defined by $G' := (V, E \cup \{s, t\})$, $c'(e) = c(e), e \in E$, and $c'(f) = d^c_G(s, t)$. Then (G', b, c') has a tree solution of cost K and a feasible dual solution of value K.*

Proof. A tree solution for (G, b, c) is also a tree solution for (G', b, c'), of the same cost. Moreover, suppose (λ, μ) is a feasible dual solution for (G, b, c) of value $K = \sum_{u,v}\mu_{uv}$. Let P_{st} be a shortest s–t path in G with respect to the length function c. Define, $\forall u, v \in V$, $\hat\lambda^f_{uv} := \sum_{e\in P_{st}}\lambda^e_{uv}$. Set $\hat\lambda^e_{uv} = \lambda^e_{uv}$ $\forall u, v \in V, e \in E$, and $\hat\mu := \mu$. Then $(\hat\lambda, \hat\mu)$ is a feasible dual solution for (G', b, c') of value K. Indeed,

$$\sum_{j\neq i}\hat\lambda^e_{ij} = \sum_{j\neq i}\lambda^e_{ij} \leq c(e)b(i) = c'(e)b(i)\ \forall\ i \in W,\ \forall\ e \in E,$$

$$\sum_{j\neq i}\hat\lambda^f_{ij} = \sum_{j\neq i}\sum_{e\in P_{st}}\lambda^e_{ij} = \sum_{e\in P_{st}}\sum_{j\neq i}\lambda^e_{ij} \leq \sum_{e\in P_{st}} c(e)b(i) = d^c_G(s,t)b(i) = c'(f)b(i). \qquad \square$$

We call c a *circuit metric* if it satisfies the triangle inequality, and every edge outside some Hamilton circuit has cost equal to the length of a shortest path, along the circuit, between its end points. From the above lemma together with Theorem 1.2 it follows that Conjecture 2.1 holds if the cost function c on the graph is a circuit metric.

For the remaining results in this section, we rephrase our conjecture as follows. Let $\Phi(b,c)$ denote the minimum value of the LP-formulation (1) of the MRP VPN problem as a function of b and c. It is easy to see that Φ is concave in c for fixed b, and concave in b for fixed c. For instance, fix b, and let c, $\bar{c}$, and $\hat{c} = \lambda c + (1-\lambda)\bar{c}$ be cost functions for some $\lambda \in (0,1)$. Let (x,y), $(\bar{x},\bar{y})$, and $(\hat{x},\hat{y})$ denote the respective optimal solutions to $\Phi(b,c)$, $\Phi(b,\bar{c})$, and $\Phi(b,\hat{c})$. Then $\Phi(b,\hat{c}) = \sum b_i \hat{c}_e \hat{y}_i^e = \lambda \sum b_i c_e \hat{y}_i^e + (1-\lambda)\sum b_i \bar{c}_e \hat{y}_i^e \geq \lambda \sum b_i c_e y_i^e + (1-\lambda)\sum b_i \bar{c}_e \bar{y}_i^e = \lambda\Phi(b,c) + (1-\lambda)\Phi(b,\bar{c})$. By Lemma 4.1, we may assume that the graph is complete and the cost function c is a metric. As argued before, an optimal star solution is an optimal tree solution for such instances. If it has nonzero cost, then by scaling the cost function c (Lemma 3.2), it is always possible to arrive at an instance with an optimal star solution of cost 1. Therefore, Conjecture 2.1 is true if for every complete graph $G = (V,E)$ the minimum over all b and c of $\Phi(b,c)$ in the following optimization problem is at least 1:

$$\begin{array}{lll} \min_{b,c} & \Phi(b,c) & \\ \text{s.t.} & b_v \geq 0 & \forall\, v \in V, \\ & \displaystyle\sum_{v\neq s} b_v c_{sv} \geq 1 & \forall\, s \in V, \\ & c_e \geq 0 & \forall\, e \in E, \\ & c_{uw} \leq c_{uv} + c_{vw} & \forall\, u,v,w \in V. \end{array} \tag{15}$$

For fixed c, the constraints are linear, and the feasible region for b over which we minimize the concave function $\Phi(b,c)$ is therefore polyhedral. Hence, for fixed c, the minimum is attained in a vertex of the polyhedron determined by (15). Similarly, for fixed b the minimum of $\Phi(b,c)$ is attained in a vertex of the polyhedral feasible region for c.

We will prove that Conjecture 2.1 holds for graphs on at most 4 vertices, by proving that it holds when b is fixed, for all vertices c of (15) for such a graph.

THEOREM 4.6. *For any instance $(G = (V,E), b, c)$ with $|V| \leq 4$, the cost of an optimal tree solution equals the value of an optimal dual solution.*

Proof. We assume that c is a metric. For a graph on 3 or fewer vertices, c is necessarily a *circuit* or *tree metric*, and the conjecture holds by Theorem 1.2 and Lemma 4.5. Since we may assume that the graph is complete, it suffices to prove the theorem for $G = K_4$. Therefore, interpreted as vectors, we have $b \in \mathbb{R}^4_+$ and $c \in \mathbb{R}^6_+$.

We may also assume that $c > 0$, since otherwise applying Lemma 3.3 would bring us back to the case of a graph with 3 vertices. In the case of K_4, there are 4 constraints in (15) saying that every star solution has cost at least 1, 12 triangle inequalities, and 10 nonnegativity constraints. For fixed $b \geq 0$ (chosen such that there exists at least one nonzero metric c such that the optimal star solution has cost 1), the polyhedron of feasible c-vectors has vertices, in which 6 linearly independent constraints for c are tight. Since $c > 0$, in any vertex of this polyhedron at least 2 triangle inequalities are tight. We distinguish 4 cases.

Case 1. There are two tight triangle inequalities on the same triangle.

That is, possibly after renaming the vertices, we have $c_{12} = c_{13} + c_{23}$ and $c_{13} = c_{12} + c_{23}$. Hence, $c_{23} = 0$, contradicting our assumption that $c > 0$.

Thus, triangle inequalities can be tight only on distinct triangles. In K_4, any two distinct triangles intersect in exactly one edge. Suppose from now on that the two triangles with tight triangle inequalities are $\{1, 2, 3\}$ and $\{2, 3, 4\}$, sharing the edge $\{2, 3\}$.

Case 2. The common edge and one of the other edges give the tight inequalities.

That is, possibly after renaming the vertices, we have $c_{12} = c_{13} + c_{23}$ and $c_{23} = c_{24} + c_{34}$. Inserting the latter in the former equality yields $c_{12} = c_{13} + c_{24} + c_{34}$. This together with $c_{12} \leq c_{14} + c_{42} \leq c_{13} + c_{34} + c_{42}$ yields $c_{14} = c_{13} + c_{34}$. This means that c is a *tree metric* completely determined by its value on only the edges $\{1, 3\}$, $\{2, 4\}$, and $\{3, 4\}$. Since Conjecture 2.1 is true for trees, by Lemma 4.5 it is also true for K_4 in this case.

Case 3. Two noncommon edges give the tight inequalities.

That is, possibly after renaming the vertices, we have that $c_{12} = c_{13} + c_{23}$ and $c_{24} = c_{23} + c_{34}$ or $c_{12} = c_{13} + c_{23}$ and $c_{34} = c_{23} + c_{24}$. In the latter case, c is a circuit metric, completely determined by its value on the edges $\{2, 3\}$, $\{1, 3\}$, $\{1, 4\}$, and $\{2, 4\}$; the result follows from Theorem 1.2, using Lemma 4.5. In the former case, c is determined by its value on the edges $\{1, 3\}$, $\{2, 3\}$, $\{1, 4\}$, $\{3, 4\}$, and hence is a *tree plus edge metric*; the result follows from Theorem 4.4, using Lemma 4.5.

Case 4. The common edge gives both tight inequalities.

That is, $c_{23} = c_{12} + c_{13}$ and $c_{23} = c_{24} + c_{34}$. If any of the other triangle inequalities is also tight, we are back in one of the previous cases. So we may assume that the set of six linearly independent tight constraints consists of the above two triangle inequalities together with all four star-inequalities. Hence, the stars centered at the four different vertices all have the same cost of 1, which together with the tight triangle inequalities yields

$$\begin{aligned} b_2c_{12} + b_3c_{13} + b_4c_{14} &= 1, \\ b_1c_{12} + b_3c_{23} + b_4c_{24} &= 1, \\ b_1c_{13} + b_2c_{23} + b_4c_{34} &= 1, \\ b_1c_{14} + b_2c_{24} + b_3c_{34} &= 1, \\ c_{23} - c_{12} - c_{13} &= 0, \\ c_{23} - c_{24} - c_{34} &= 0. \end{aligned}$$

The determinant of the matrix of coefficients of the above set of 6 equations is $2b_1b_4(b_3 - b_2)(b_1 + b_2 + b_3 + b_4)$. As the equations are linearly independent, this determinant is nonzero. This implies that the system has a unique solution for c. Straightforward calculations yield

$$c_{13} + c_{34} - c_{14} = \frac{(b_1 - b_2 + b_3 - b_4)(b_3 - b_1 - b_2 + b_4)(b_1 - b_2 + b_3 + b_4)}{2(b_3 - b_2)b_1b_4(b_1 + b_2 + b_3 + b_4)}$$

and

$$c_{12} + c_{24} - c_{14} = \frac{(b_3 - b_1 - b_2 - b_4)(b_1 - b_2 + b_3 - b_4)(b_3 - b_1 - b_2 + b_4)}{2(b_3 - b_2)b_1b_4(b_1 + b_2 + b_3 + b_4)}.$$

Since both expressions are strictly positive (no more triangle inequalities are tight), their ratio is also positive:

$$\frac{b_1 - b_2 + b_3 + b_4}{b_3 - b_1 - b_2 - b_4} > 0.$$

Thus, either $b_1 - b_2 + b_3 + b_4$ and $b_3 - b_1 - b_2 - b_4$ are both positive or they are both negative. In the former case $b_3 > b_1 + b_2 + b_4$, in the latter case $b_2 > b_1 + b_3 + b_4$. In either case, there is a *dominant* vertex, and the result follows from Lemma 4.2. □

Conjecture 2.1 also holds if G is a complete graph, and the unit cost of all edges is the same.

THEOREM 4.7. *For any instance (G, b, c) with G a complete graph and $c(e) = 1$ $\forall e \in E$, the cost of an optimal tree solution equals the value of an optimal dual solution.*

Proof. Since $c \equiv 1$, c satisfies the triangle inequality, but no triangle inequality is tight. We may assume that b is such that the cost of an optimal tree solution is greater than zero. Then, by scaling b instead of c (Lemma 3.2), it is still possible to arrive at a situation where the star-solution of minimal cost has a cost of 1. Thus, we study the minimization problem (15) for fixed c. The minimum is attained in a vertex of the polyhedron in $\mathbb{R}^{|V|}$ determined by

$$\begin{aligned} b_v &\geq 0 \quad \forall\, v \in V, \\ \sum_{u \neq v} b_u &\geq 1 \quad \forall\, v \in V. \end{aligned} \tag{16}$$

In a vertex of this $|V|$-dimensional polyhedron, $|V|$ independent inequalities are tight. Choose a vertex minimizing $\Phi(b, 1)$ over (16). Suppose that in this vertex some subsets $U \subset V$ and $W \subset V$ with $|U| + |W| = |V|$ correspond to the $|V|$ tight inequalities:

$$\begin{aligned} b_v &= 0 \quad \forall\, v \in U, \\ \sum_{u \in V} b_u - b_v &= 1 \quad \forall\, v \in W. \end{aligned}$$

We will argue that we may assume that $U \cap W = \emptyset$. For any $v \in |U \cap W|$, we have $b_v = 0$ and $\sum_{u \in V} b_u - b_v = 1$, together implying that $\sum_{u \in V} b_u = 1$ and therefore $\sum_{u \in V} b_u - b_w \leq 1$ for any $w \in V$. This enforces $b_w = 0$ $\forall w \in V$, which is infeasible.

Thus, $U \cap W = \emptyset$. Since $\sum_{u \in V} b_u - 1$ is a constant, b_v is a constant, b say, for every $v \in W$. Feasibility for (16) requires that $|W| > 1$. We will explicitly construct a tree solution of minimum cost and a dual solution of the same value.

As argued before, since c satisfies the triangle inequality, there exists a star that is an optimal tree solution. The star centered at a vertex v with $b_v = 0$ has cost $|W|b$, whereas the star centered at a vertex v with $b_v = b$ has cost $(|W| - 1)b$. Thus, any star centered at a vertex of W is an optimal tree solution of cost $(|W| - 1)b$. Define the following:

$$\begin{aligned} \lambda^e_{st} &:= \frac{2b}{|W|} && \text{if } b_s = b_t = b \text{ and } e = \{s, t\}, \\ \lambda^e_{st} &:= \frac{b}{|W|} && \text{if } b_s = b_t = b \text{ and } e \neq \{s, t\}, \\ \lambda^e_{st} &:= 0 && \text{if } b_s = 0 \text{ or } b_t = 0, \\ \mu_{st} &:= \frac{2b}{|W|} && \text{if } b_s = b_t = b, \\ \mu_{st} &:= 0 && \text{if } b_s = 0 \text{ or } b_t = 0. \end{aligned}$$

It is not hard to check that (λ, μ) is feasible and has value

$$\sum_{s,t} \mu_{st} = \frac{|W|(|W|-1)}{2} \cdot \frac{2b}{|W|} = (|W|-1)b. \qquad \square$$

For completeness, we now formulate the most general statement that we can obtain by combining all the results in this section.

THEOREM 4.8. *Suppose $G = (V, E)$ is a connected graph and $c \in \mathbb{R}_+^{|E|}$ is a cost function such that every block $H = (V', E')$ of G endowed with the cost function $c|_{E'}$ is either a circuit, a graph on at most 4 vertices, or a complete graph with uniform edge costs. Then the cost of an optimal tree solution equals the value of an optimal dual for the instance (G, b, c) for any $b \in \mathbb{R}_+^{|V|}$.*

Proof. The theorem follows directly from Theorems 1.2, 4.6, and 4.7 and Lemma 4.3, since G can be obtained from its blocks by taking repeated 1-sums. $\square$

Note that Theorem 4.8 extends Theorem 4.4, since edges are complete graphs for which any cost function is uniform.

REFERENCES

[1] A. Altin, E. Amaldi, P. Belotti, and M. Ç. Pinar, *Virtual private network design under traffic uncertainty*, in Proceedings of CTW04, 2004, pp. 24–27; extended version available online from http://www.elet.polimi.it/upload/belotti/.

[2] C. Chekuri, G. Oriolo, M. G. Scuttellà, and F. B. Shepherd, *Hardness of robust network design*, in Proceedings of the International Network Optimization Conference (INOC) 2005, Lisbon, 2005, pp. 455–461.

[3] N. G. Duffield, P. Goyal, A. Greenberg, P. Mishra, K. K. Ramakrishnan, and J. E. van der Merwe, *A flexible model for resource management in virtual private networks*, ACM SIGCOMM Comput. Comm. Rev., 29 (1999), pp. 95–108.

[4] F. Eisenbrand, F. Grandoni, G. Oriolo, and M. Skutella, *New approaches for virtual private network design*, in Automata, Languages, and Programming, Lecture Notes in Comput. Sci. 3580, Springer, Berlin, 2005, pp. 1151–1162.

[5] T. Erlebach, *Approximation algorithms and complexity results for path problems in trees of rings*, in Mathematical Foundations of Computer Science (Mariánské Lázně), Lecture Notes in Comput. Sci. 2136, Springer, Berlin, 2001, pp. 351–362.

[6] T. Erlebach and M. Rüegg, *Optimal bandwidth reservation in hose-model VPNs with multi-path routing*, in Proceedings of the 23rd INFOCOM Conference of the IEEE Communications Society, Hong Kong, 2004.

[7] J. A. Fingerhut, S. Suri, and J. S. Turner, *Designing least-cost nonblocking broadband networks*, J. Algorithms, 24 (1997), pp. 287–309.

[8] M. Grötschel, L. Lovász, and A. Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 169–197.

[9] A. Gupta, J. Kleinberg, A. Kumar, R. Rastogi, and B. Yener, *Provisioning a virtual private network: A network design problem for multicommodity flow*, in Proceedings of the 33rd Annual ACM Symposium on Theory of Computing (STOC), 2001, pp. 389–398.

[10] A. Gupta, A. Kumar, and T. Roughgarden, *Simpler and better approximation algorithms for network design*, in Proceedings of the 35th Annual ACM Symposium on Theory of Computing (STOC), 2003, pp. 365–372.

[11] C. A. J. Hurkens, J. C. M. Keijsper, and L. Stougie, *Virtual Private Network Design: A Proof of the Tree Routing Conjecture on Ring Networks*, Tech. report, SPOR 2004-15, Department of Mathematics and Computer Science, Technische Universiteit Eindhoven, Eindhoven, The Netherlands, 2004; available online from http://www.win.tue.nl/math/bs/spor/2004-15.pdf.

[12] G. Italiano, S. Leonardi, and G. Oriolo, *Design of networks in the hose model*, in Proceedings of the 3rd Workshop on Approximation and Randomization Algorithms in Communication Networks (ARACNE), Carleton Scientific, 2002, pp. 65–76.

[13] A. Kumar, R. Rastogi, A. Silberschatz, and B. Yener, *Algorithms for provisioning virtual private networks in the hose model*, IEEE/ACM Trans. Networking, 10 (2002), pp. 565–578.

SIAM J. DISCRETE MATH.
Vol. 21, No. 2, pp. 504–522

POLYNOMIAL-TIME ALGORITHMS FOR LINEAR AND CONVEX OPTIMIZATION ON JUMP SYSTEMS*

AKIYOSHI SHIOURA† AND KEN'ICHIRO TANAKA‡

Abstract. The concept of a jump system, introduced by Bouchet and Cunningham [*SIAM J. Discrete Math.*, 8 (1995), pp. 17–32], is a set of integer points with a certain exchange property. In this paper, we discuss several linear and convex optimization problems on jump systems and show that these problems can be solved in polynomial time under the assumption that a membership oracle for a jump system is available. We first present a polynomial-time implementation of the greedy algorithm for the minimization of a linear function. We then consider the minimization of a separable-convex function on a jump system and propose the first polynomial-time algorithm for this problem. The algorithm is based on the domain reduction approach developed in Shioura [*Discrete Appl. Math.*, 84 (1998), pp. 215–220]. We finally consider the concept of M-convex functions on constant-parity jump systems which has been recently proposed by Murota [*SIAM J. Discrete Math.*, 20 (2006), pp. 213–226]. It is shown that the minimization of an M-convex function can be solved in polynomial time by the domain reduction approach.

Key words. jump system, discrete convex function, bisubmodular function, bisubmodular polyhedron, polynomial-time algorithm

AMS subject classifications. 90C10, 90C25, 90C35, 90C27

DOI. 10.1137/060656899

1. Introduction.

The concept of a jump system, introduced by Bouchet and Cunningham [7], is a set of integer points with a certain exchange property (to be described in section 2); see also [9, 13, 18]. It is a generalization of a matroid [17, 24, 27], a delta-matroid [6, 8], and the base polyhedron of an integral polymatroid (or a submodular system) [11, 24, 27]. Jump systems have various examples (see [7, 9, 13, 18]); in particular, the degree sequences of subgraphs of a graph are a fundamental example of jump systems. In this paper, we investigate the following linear and convex optimization problems on jump systems:

(LFMin) minimization of a linear function on a jump system,
(ScFMin) minimization of a separable-convex function on a jump system,
(McFMin) minimization of an M-convex function on a constant-parity jump system.

The main aim of this paper is to show that these problems can be solved in polynomial time under the assumption that a membership oracle for a jump system is available.

1.1. Linear optimization on jump systems.

We discuss the greedy algorithm for the problem (LFMin) in section 3. It is shown [7] (see also [9, 13, 18]) that the problem (LFMin) can be solved by a greedy algorithm. The greedy algorithm finds an optimal solution by iteratively calling a procedure for solving a problem of minimizing (or maximizing) some component of a vector on a jump system. However, the time

*Received by the editors April 10, 2006; accepted for publication (in revised form) December 21, 2006; published electronically June 28, 2007.
http://www.siam.org/journals/sidma/21-2/65689.html

†Graduate School of Information Sciences, Tohoku University, Sendai 980-8579, Japan (shioura@dais.is.tohoku.ac.jp). This work was done while this author was visiting Konrad-Zuse-Zentrum für Informationstechnik Berlin, Berlin, Germany. This author was supported by the Humboldt Research Fellowship of the Alexander von Humboldt Foundation.

‡Graduate School of Information Science and Technology, University of Tokyo, Tokyo 113-8656, Japan (kenitiro@misojiro.t.u-tokyo.ac.jp). This author was supported by the Research Fellowship of the Japan Society for the Promotion of Science for Young Scientists.

complexity of the greedy algorithm is not analyzed in [7, 9, 13, 18], and it is not known so far whether the greedy algorithm runs in polynomial time or not, provided a membership oracle for a jump system is available.

In this paper, we show that the greedy algorithm runs in polynomial time. In particular, we present an implementation of the procedure mentioned above and prove that the procedure runs in polynomial time.

1.2. Convex optimization on jump systems. In section 4, we consider two convex optimization problems (ScFMin) and (McFMin) and propose polynomial-time algorithms for these problems.

We first consider the problem (ScFMin) in section 4.1. A canonical example of this problem arises from the minimization of a separable-convex function on the degree sequences of an undirected graph; a related problem called the minsquare factor problem is discussed in [4, 5]. The problem (ScFMin) is studied in [3], where a local criterion for optimality, as well as a greedy algorithm, is given. Although it is shown that the greedy algorithm runs in pseudopolynomial time, it is not known so far whether the problem (ScFMin) can be solved in polynomial time.

On the other hand, some special cases of (ScFMin) can be solved in polynomial time. One such case is the problem on integral base polyhedra, which is extensively discussed and for which several efficient algorithms have been proposed [14, 15, 20]. Another well-solved case is the problem on integral bisubmodular polyhedra, to which Fujishige [10] applied a min-max theorem for bisubmodular polyhedra and developed a polynomial-time algorithm.

In this paper, we present the first polynomial-time algorithm for the problem (ScFMin). Our algorithm is based on the domain reduction approach [25], which was originally developed for the minimization of a class of discrete convex functions called M-convex functions on base polyhedra [21]. One of the key properties to the domain reduction approach is the "minimizer cut property," which states that a given feasible vector can be easily separated from an optimal solution. We show that the minimizer cut property indeed holds for the problem (ScFMin). By repeatedly applying the minimizer cut property to appropriately chosen feasible vectors, we show that the algorithm finds an optimal solution in polynomial time.

We then discuss in section 4.2 an application of our algorithm to the problem of finding least weakly sub- and supermajorized elements. The concept of (weak) majorization plays a fundamental role in fair resource allocation and related problems (see [19]), and it is shown that any jump system has least weakly sub- and supermajorized elements [1]. By using our algorithm as well as the result in [1], we show that the problem of finding least weakly sub- and supermajorized elements in jump systems can be solved in polynomial time.

We finally consider the problem (McFMin) in section 4.3. The concept of M-convex functions was originally introduced by Murota [21] for functions defined on base polyhedra and recently generalized for functions defined on constant-parity jump systems [22], with a view to providing a general framework for the minsquare factor problem on undirected graphs [4, 5]. Examples of M-convex functions on constant-parity jump systems include a nonseparable-convex function arising from the minimum weight perfect b-matching problem as well as a separable-convex function on the degree sequences of an undirected graph (see section 2). Fundamental properties of M-convex functions on constant-parity jump systems are investigated in [16, 22, 23], where it is shown that a local optimality criterion guarantees global optimality and that a greedy algorithm solves the problem (McFMin) in pseudopolynomial time.

In this paper, we present the first polynomial-time algorithm for the problem (McFMin), which is also based on the domain reduction approach. In fact, the minimizer cut property for (McFMin) is already shown in [23]. By using this fact, we show that a variant of the polynomial-time algorithm for (ScFMin) finds an optimal solution of (McFMin) in polynomial time.

2. Preliminaries on jump systems. Let V be a nonempty finite set and put $n = |V|$. We denote the set of reals and integers by $\mathbb{R}$ and $\mathbb{Z}$, respectively. Also, we denote by $\mathbb{Z}_+$ the set of nonnegative integers. For $x = (x(v) \mid v \in V) \in \mathbb{R}^V$, we define

$$x(Y) = \sum_{v \in Y} x(v) \ (Y \subseteq V), \qquad \|x\|_1 = \sum_{v \in V} |x(v)|, \qquad \mathrm{supp}(x) = \{v \in V \mid x(v) \neq 0\}.$$

We denote by $\mathbf{0}$ the zero vector in $\mathbb{R}^V$. For $u \in V$ we denote by χ_u the characteristic vector of u, with $\chi_u(u) = 1$ and $\chi_u(v) = 0$ for $v \neq u$. We denote by N_1 the set of all integral vectors x with $\|x\|_1 = 1$, i.e., $N_1 = \{+\chi_v, -\chi_v \mid v \in V\}$. For a nonempty finite set $S \subseteq \mathbb{Z}^V$, we define the *size* $\Phi(S)$ of S by

$$\Phi(S) = \max_{v \in V} \bigl\{ \max_{y \in S} y(v) - \min_{y \in S} y(v) \bigr\}.$$

For $x, y \in \mathbb{Z}^V$, a vector $s \in N_1$ is said to be an (x,y)*-increment* if it satisfies $\|(x+s) - y\|_1 = \|x - y\|_1 - 1$. We denote by $\mathrm{inc}(x,y)$ the set of all (x,y)-increments. A nonempty set $J \subseteq \mathbb{Z}^V$ is said to be a *jump system* if it satisfies the exchange axiom

> **(J-EXC$_0$)** For any $x, y \in J$ and for any $s \in \mathrm{inc}(x,y)$, if $x + s \notin J$, then there exists $t \in \mathrm{inc}(x+s, y)$ such that $x + s + t \in J$.

A set $J \subseteq \mathbb{Z}^V$ is said to be a *constant-parity system* if $x(V) - y(V)$ is even for any $x, y \in J$.

We mention here some elementary operations which preserve the property (J-EXC$_0$). Jump systems are closed under reflection.

PROPOSITION 2.1 (see [7]). *Let $J \subseteq \mathbb{Z}^V$ be a jump system and $u \in V$. Then, the set*

$$J^u = \{y \in \mathbb{Z}^V \mid \exists x \in J \textit{ such that } y(u) = -x(u),\ y(v) = x(v)\ (v \in V \setminus \{u\})\}$$

is a jump system.

For any vectors $a, b \in \mathbb{Z}^V$ with $a \leq b$, we define a *box* $[a,b]$ by

$$[a,b] = \{x \in \mathbb{Z}^V \mid a(v) \leq x(v) \leq b(v)\ (v \in V)\}.$$

PROPOSITION 2.2 (cf. [7]). *Let $J \subseteq \mathbb{Z}^V$ be a jump system and $a, b \in \mathbb{Z}^V$ be vectors with $a \leq b$. Then, $J \cap [a,b]$ is a jump system if it is nonempty.*

A univariate function $\varphi : \mathbb{Z} \to \mathbb{R}$ is *convex* if it satisfies

$$2\varphi(\alpha) \leq \varphi(\alpha - 1) + \varphi(\alpha + 1) \quad (\forall \alpha \in \mathbb{Z}).$$

A function $f : \mathbb{Z}^V \to \mathbb{R}$ is said to be *separable-convex* if it is a function of the form $f(x) = \sum_{v \in V} f_v(x(v))$ with univariate convex functions $f_v : \mathbb{Z} \to \mathbb{R}$ $(v \in V)$. The sum of squares $f(x) = \sum_{v \in V} (x(v))^2$ is a special case of a separable-convex function.

Let $J \subseteq \mathbb{Z}^V$ be a constant-parity jump system. A function $f : J \to \mathbb{R}$ is said to be *M-convex* if it satisfies the following property:

> For any $x, y \in J$ and for any $s \in \mathrm{inc}(x, y)$, there exists $t \in \mathrm{inc}(x+s, y)$ such that $x + s + t \in J$, $y - s - t \in J$, and
>
> $$f(x) + f(y) \geq f(x + s + t) + f(y - s - t).$$

We note that the exchange axiom

> **(J-EXC)** For any $x, y \in J$ and for any $s \in \mathrm{inc}(x, y)$, there exists $t \in \mathrm{inc}(x + s, y)$ such that $x + s + t \in J$ and $y - s - t \in J$

characterizes a constant-parity jump system, a fact credited to J. Geelen (see [22] for a proof).

PROPOSITION 2.3. *A nonempty set $J \subseteq \mathbb{Z}^V$ is a constant-parity jump system if and only if it satisfies* (J-EXC).

Examples of jump systems and M-convex functions follow.

Example 2.4. Let $G = (V, E)$ be an undirected graph that may contain loops and parallel edges. For a subgraph $H = (V, F)$ of G, denote its *degree sequence* by $\deg_H = \sum\{\chi_u + \chi_v \mid (u, v) \in F\} \in \mathbb{Z}^V$. It is well known [7, 9, 13, 18] that

$$J = \{\deg_H \mid H \text{ is a subgraph of } G\}$$

forms a constant-parity jump system, called the *degree system* of G. Minimization of a separable-convex function on the degree system J has been investigated in [4, 5].

Given an edge weight function $w : E \to \mathbb{R}$, we define a function $f : J \to \mathbb{R}$ by

$$f(x) = \min\left\{\sum_{e \in F} w(e) \mid H = (V, F) \text{ is a subgraph of } G \text{ with } \deg_H = x\right\},$$

which represents the minimum weight of a subgraph with degree sequence x. Then, f is an M-convex function on a constant-parity jump system [22].

Example 2.5 (see [23]). Let $G = (V, E)$ be an undirected graph that may have loops, but no parallel edges. Let $w : E \to \mathbb{R}$ be an edge weight function and $c : E \to \mathbb{Z}_+$ be an edge capacity function. We define $J \subseteq \mathbb{Z}^V$ as the set of vectors $x \in \mathbb{Z}^V$ such that a c-capacitated perfect x-matching exists in G, i.e., such that there exists $\lambda \in \mathbb{Z}^E$ satisfying

$$\sum\{\lambda(e) \mid \text{edge } e \text{ is incident to } v\} = x(v) \ (\forall v \in V), \quad 0 \leq \lambda(e) \leq c(e) \ (\forall e \in E).$$

Then, J is a constant-parity jump system.

We then define a function $f : J \to \mathbb{R}$ as the minimum weight of a c-capacitated perfect x-matching, i.e.,

$$f(x) = \min\left\{\sum_{e \in E} \lambda(e) w(e) \;\middle|\; \begin{array}{l} \sum\{\lambda(e) \mid \text{edge } e \text{ is incident to } v\} = x(v) \ (\forall v \in V), \\ \lambda(e) \in \mathbb{Z},\ 0 \leq \lambda(e) \leq c(e) \ (\forall e \in E) \end{array}\right\}.$$

Then, f is an M-convex function on a constant-parity jump system. Moreover, the function $\tilde{f} : J \to \mathbb{R}$ given as

$$\tilde{f}(x) = f(x) + \sum_{v \in V} f_v(x(v)),$$

where $f_v : \mathbb{Z} \to \mathbb{R}$ $(v \in V)$ is a family of univariate convex functions, is also M-convex.

3. Polynomiality of the greedy algorithm for linear optimization on jump systems. In this section, we consider the problem of minimizing a linear function on a jump system:

$$\text{(LFMin)} \quad \text{Minimize} \quad w^T x \qquad \text{subject to} \quad x \in J,$$

where $w \in \mathbb{R}^V$ and J is a finite jump system. We show that the greedy algorithm for the problem (LFMin) runs in polynomial time. We assume that a membership oracle for the jump system J is available and that a vector in J is given.

3.1. Greedy algorithm. It is shown [7, 18] that the problem (LFMin) can be solved by the following greedy algorithm.

```
Algorithm Greedy.
Step 0: Let $x_0$ be any vector in $J$. Put $J_0 = J$. Compute an integer $k$ and an ordering of the elements in $V = \{v_1, v_2, \ldots, v_n\}$ such that
        $$|w(v_1)| \geq \cdots \geq |w(v_k)| > |w(v_{k+1})| = \cdots = |w(v_n)| = 0.$$
Step 1: For $i = 1, 2, \ldots, k$, do the following:
  Step 1-1: Compute the value $\alpha_i \in \mathbb{Z}$ given by
        $$\alpha_i = \begin{cases} \min\{y(v_i) \mid x \in J_{i-1}\} & (\text{if } w_i > 0), \\ \max\{y(v_i) \mid x \in J_{i-1}\} & (\text{if } w_i < 0). \end{cases}$$
  Step 1-2: Let $x_i$ be any vector in $J_{i-1}$ with $x_i(v_i) = \alpha_i$. Put
        $$J_i = \{y \in J_{i-1} \mid y(v_i) = \alpha_i\}.$$
Step 2: Output $x_k$.
```

Theorem 3.1 (see [7, 18]). *The algorithm* Greedy *outputs an optimal solution of* (LFMin).

We show that the algorithm Greedy runs in polynomial time.

Theorem 3.2. *The algorithm* Greedy *finds an optimal solution of* (LFMin) *in* $O(n^2 \log \Phi(J))$ *time, provided a vector in J is given.*

Proof. The most time-consuming part is the computation of α_i in Step 1-1, which can be done in $O(n \log \Phi(J))$ time by using the vector x_{i-1}, as shown later in section 3.2. Hence, the algorithm Greedy runs in $O(n^2 \log \Phi(J))$ time. □

In the next section, we explain in detail how to compute α_i in $O(n \log \Phi(J))$ time.

3.2. Computation of upper and lower bounds of jump systems. We present two procedures to compute the values $\max\{y(u) \mid y \in J\}$ and $\min\{y(u) \mid y \in J\}$ in polynomial time for a finite jump system $J \subseteq \mathbb{Z}^V$ and an element $u \in V$.

```
Procedure Upper_Bound$(J, u)$.
Step 0: Let $x := x_0$ be an initial vector in $J$.
Step 1: Put $x := x + \bar{\alpha}\chi_u$, where $\bar{\alpha} = \max\{\alpha \in \mathbb{Z}_+ \mid x + \alpha\chi_u \in J\}$.
Step 2: For each $v \in V \setminus \{u\}$, do the following:
  Step 2-1: Put $x := x + \bar{\beta}_v(\chi_u + \chi_v)$, where
        $$\bar{\beta}_v = \max\{\beta \in \mathbb{Z}_+ \mid x + \beta(\chi_u + \chi_v) \in J\}.$$
  Step 2-2: Put $x := x + \bar{\gamma}_v(\chi_u - \chi_v)$, where
        $$\bar{\gamma}_v = \max\{\gamma \in \mathbb{Z}_+ \mid x + \gamma(\chi_u - \chi_v) \in J\}.$$
Step 3: Output $x$.
```

```
PROCEDURE LOWER_BOUND($J, u$).
Step 0: Let $x := x_0$ be an initial vector in $J$.
Step 1: Put $x := x - \bar{\alpha}\chi_u$, where $\bar{\alpha} = \max\{\alpha \in \mathbb{Z}_+ \mid x - \alpha\chi_u \in J\}$.
Step 2: For each $v \in V \setminus \{u\}$, do the following:
  Step 2-1: Put $x := x + \bar{\beta}_v(-\chi_u + \chi_v)$, where
      $\bar{\beta}_v = \max\{\beta \in \mathbb{Z}_+ \mid x + \beta(-\chi_u + \chi_v) \in J\}$.
  Step 2-2: Put $x := x + \bar{\gamma}_v(-\chi_u - \chi_v)$, where
      $\bar{\gamma}_v = \max\{\gamma \in \mathbb{Z}_+ \mid x + \gamma(-\chi_u - \chi_v) \in J\}$.
Step 3: Output $x$.
```

THEOREM 3.3. *For a finite jump system J and $u \in V$, the procedure* UPPER_BOUND(J, u) *(resp.,* LOWER_BOUND(J, u)*) finds a vector $x \in J$ satisfying $x(u) = \max\{y(u) \mid y \in J\}$ (resp., $x(u) = \min\{y(u) \mid y \in J\}$) in $\mathrm{O}(n \log \Phi(J))$ time, provided a vector $x_0 \in J$ is given.*

The proof of Theorem 3.3 is given in sections 3.2.1 and 3.2.2.

COROLLARY 3.4. *Suppose that J is a jump system given as the intersection $J = \tilde{J} \cap [a, b]$ of another jump system $\tilde{J}$ and a box $[a, b]$, and that a membership oracle for $\tilde{J}$ is available. For any $u \in V$, we can find vectors $x, x' \in J$ with $x(u) = \max\{y(u) \mid y \in J\}$ and $x'(u) = \min\{y(u) \mid y \in J\}$ in $\mathrm{O}(n \log \Phi(J))$ time, provided a vector in J is given.*

Proof. Although it takes $\mathrm{O}(n)$ time to check whether a given vector is contained in $\tilde{J} \cap [a, b]$, we can implement the procedures UPPER_BOUND(J, u) and LOWER_BOUND(J, u) so that they run in $\mathrm{O}(n \log \Phi(J))$ time.

When the procedures need to check whether $x \in \tilde{J} \cap [a, b]$, the vector x is of the form $x = y + \alpha(\chi_u \pm \chi_v)$ with $y \in \tilde{J} \cap [a, b]$, $\alpha \in \mathbb{Z}_+$, and $v \in V$. Hence, we have $x \in \tilde{J} \cap [a, b]$ if and only if $x \in \tilde{J}$, $a(u) \leq x(u) \leq b(u)$, and $a(v) \leq x(v) \leq b(v)$, which can be checked in constant time. This shows that the procedures run in $\mathrm{O}(n \log \Phi(J))$ time for the jump system $J = \tilde{J} \cap [a, b]$ as well. □

3.2.1. Validity. We show the validity of the procedure UPPER_BOUND(J, u). The validity of LOWER_BOUND(J, u) can be shown similarly and therefore omitted.

LEMMA 3.5. *Let $x \in J$ and $u \in V$. Suppose that $x + \chi_u + t \notin J$ holds for all $t \in (N_1 \cup \{\mathbf{0}\}) \setminus \{-\chi_u\}$. Then, we have $x(u) = \max\{y(u) \mid y \in J\}$.*

Proof. Assume, to the contrary, that there exists some $x' \in J$ with $x'(u) > x(u)$. Since $x + \chi_u \notin J$ by assumption, (J-EXC$_0$) implies that there exists some $t \in \mathrm{inc}(x + \chi_u, x')$ such that $x + \chi_u + t \in J$, which is a contradiction since $t \in N_1 \setminus \{-\chi_u\}$. □

LEMMA 3.6. *Let $u \in V$ and $x, y \in J$ be vectors such that*

$$y(u) - x(u) \geq \sum_{v \in V \setminus \{u\}} |x(v) - y(v)| + 1.$$

Then, we have $\{x + \chi_u,\ x + 2\chi_u\} \cap J \neq \emptyset$.

Proof. We prove the claim by induction on the value $\|x - y\|_1$.

We first consider the case where $x(v) = y(v)$ for all $v \in V \setminus \{u\}$, which contains the base case where $\|x - y\|_1 = 1$. Then, (J-EXC$_0$) for x and y implies $\{x + \chi_u, x + 2\chi_u\} \cap J \neq \emptyset$.

We then assume that $x(w) \neq y(w)$ for some $w \in V \setminus \{u\}$, where it may be assumed that $x(w) < y(w)$. Since $-\chi_w \in \mathrm{inc}(y, x)$, (J-EXC$_0$) for y and x implies that there exists $t \in \mathrm{inc}(y - \chi_w, x) \cup \{\mathbf{0}\}$ such that $y' = y - \chi_w + t \in J$. The vector y' satisfies

$$y'(u) - x(u) \geq y(u) - x(u) - 1 \geq \sum_{v \in V \setminus \{u\}} |x(v) - y(v)| \geq \sum_{v \in V \setminus \{u\}} |x(v) - y'(v)| + 1$$

and $\|y' - x\|_1 < \|y - x\|_1$. Hence, the induction hypothesis implies $\{x + \chi_u,\, x + 2\chi_u\} \cap J \neq \emptyset$. □

Lemma 3.7. *Let $u \in V$ and $x, y \in J$ be vectors such that*

$$y(u) - x(u) \geq \sum_{v \in V \setminus \{u\}} |x(v) - y(v)|.$$

If $\{x + \chi_u,\, x + 2\chi_u\} \cap J = \emptyset$, then $\{y + \chi_u,\, y + 2\chi_u\} \cap J = \emptyset$.

Proof. Suppose, to the contrary, that $\{y + \chi_u,\, y + 2\chi_u\} \cap J \neq \emptyset$ and let $y' \in \{y + \chi_u,\, y + 2\chi_u\} \cap J$. Then, we have $y'(u) - x(u) \geq \sum_{v \in V \setminus \{u\}} |x(v) - y'(v)| + 1$, and therefore $\{x + \chi_u,\, x + 2\chi_u\} \cap J \neq \emptyset$ by Lemma 3.6, a contradiction. □

Lemma 3.8. *Let $u, w \in V$ be distinct elements and $x, y \in J$ be vectors such that*

$$y(u) - x(u) \geq \sum_{v \in V \setminus \{u\}} |x(v) - y(v)|, \qquad |x(w) - y(w)| \geq 1.$$

(i) *If $x(w) < y(w)$, then $\{x + \chi_u,\, x + 2\chi_u,\, x + \chi_u + \chi_w\} \cap J \neq \emptyset$.*

(ii) *If $x(w) > y(w)$, then $\{x + \chi_u,\, x + 2\chi_u,\, x + \chi_u - \chi_w\} \cap J \neq \emptyset$.*

Proof. We prove (i) by induction on the value $\|x - y\|_1$. The claim (ii) can be shown similarly.

We first consider the case where $x(v) = y(v)$ holds for all $v \in V \setminus \{u, w\}$, which contains the base case where $\|x - y\|_1 = 2$. Then, $y = x + \alpha\chi_u + \beta\chi_w$ for some positive integers α and β with $\alpha \geq \beta$. Since $+\chi_u \in \mathrm{inc}(x, y)$, (J-EXC$_0$) for x and y implies $\{x + \chi_u,\, x + 2\chi_u,\, x + \chi_u + \chi_w\} \cap J \neq \emptyset$.

We then assume $x(v') \neq y(v')$ for some $v' \in V \setminus \{u, w\}$, where we may assume $x(v') < y(v')$. Since $-\chi_{v'} \in \mathrm{inc}(y, x)$, (J-EXC$_0$) for y and x implies $y' = y - \chi_{v'} + t \in J$ for $t \in \mathrm{inc}(y - \chi_{v'}, x) \cup \{\mathbf{0}\}$.

Case 1 ($t \neq -\chi_u$). y' satisfies

$$y'(u) - x(u) = y(u) - x(u) \geq \sum_{v \in V \setminus \{u\}} |x(v) - y(v)| \geq \sum_{v \in V \setminus \{u\}} |x(v) - y'(v)| + 1.$$

Hence, we have $\{x + \chi_u,\, x + 2\chi_u\} \cap J \neq \emptyset$ by Lemma 3.6.

Case 2 ($t = -\chi_u$). We have

$$y'(u) - x(u) = y(u) - x(u) - 1 \geq \sum_{v \in V \setminus \{u\}} |x(v) - y(v)| - 1 = \sum_{v \in V \setminus \{u\}} |x(v) - y'(v)|$$

and $y'(w) = y(w) > x(w)$. Since $\|y' - x\|_1 = \|y - x\|_1 - 2$, the induction hypothesis implies $\{x + \chi_u,\, x + 2\chi_u,\, x + \chi_u + \chi_w\} \cap J \neq \emptyset$. □

Lemma 3.9. *Let $u, w \in V$ be distinct elements and $x, y \in J$ be vectors such that*

$$y(u) - x(u) \geq \sum_{v \in V \setminus \{u\}} |x(v) - y(v)|.$$

Then, we have the following:

(i) *If* $\{x+\chi_u,\, x+2\chi_u,\, x+\chi_u+\chi_w\} \cap J = \emptyset$, *then* $y+\chi_u+\chi_w \notin J$.

(ii) *If* $\{x+\chi_u,\, x+2\chi_u,\, x+\chi_u-\chi_w\} \cap J = \emptyset$, *then* $y+\chi_u-\chi_w \notin J$.

Proof. We prove (i) only. Assume, to the contrary, that $y' = y+\chi_u+\chi_w \in J$. We first consider the case where $y(w) \geq x(w)$. Then,

$$\begin{aligned} y'(u)-x(u) = y(u)-x(u)+1 &\geq \sum_{v\in V\setminus\{u\}} |x(v)-y(v)|+1 \\ &= \sum_{v\in V\setminus\{u\}} |x(v)-y'(v)| \end{aligned}$$

and $y'(w)-x(w) = y(w)-x(w)+1 > 0$. Hence, Lemma 3.8 implies $\{x+\chi_u,\, x+2\chi_u,\, x+\chi_u+\chi_w\} \cap J \neq \emptyset$, a contradiction.

We then consider the case where $y(w) < x(w)$. Then,

$$\begin{aligned} y'(u)-x(u) = y(u)-x(u)+1 &\geq \sum_{v\in V\setminus\{u\}} |x(v)-y(v)|+1 \\ &= \sum_{v\in V\setminus\{u\}} |x(v)-y'(v)|+2. \end{aligned}$$

Hence, Lemma 3.6 implies $\{x+\chi_u,\, x+2\chi_u\} \cap J \neq \emptyset$, a contradiction. □

LEMMA 3.10. *The procedure* UPPER_BOUND(J,u) *finds a vector* $x \in J$ *satisfying* $x(u) = \max\{y(u) \mid y \in J\}$.

Proof. By the definition of $\bar{\alpha}$, we have $\{x+\chi_u,\, x+2\chi_u\} \cap J = \emptyset$ immediately after Step 1. Therefore, $\{x+\chi_u,\, x+2\chi_u\} \cap J = \emptyset$ holds during the iterations in Step 2 by Lemma 3.7. Similarly, we have $x+\chi_u+\chi_w \notin J$ (resp., $x+\chi_u-\chi_w \notin J$) immediately after Step 2-1 (resp., Step 2-2) with $v = w$ is performed, and therefore $x+\chi_u+\chi_w \notin J$ (resp., $x+\chi_u-\chi_w \notin J$) holds in the following iterations in Step 2 by Lemma 3.9. At the end of the procedure, the vector x satisfies $x+\chi_u+t \notin J$ for all $t \in (N_1 \cup \{\mathbf{0}\}) \setminus \{-\chi_u\}$. Hence, Lemma 3.5 implies $x(u) = \max\{y(u) \mid y \in J\}$. □

3.2.2. Time complexity. We then analyze the time complexity of the procedure UPPER_BOUND(J,u). The analysis of LOWER_BOUND(J,u) is similar and therefore omitted.

LEMMA 3.11. *Let* $x \in J$ *and* $u \in V$, *and put* $\bar{\alpha} = \max\{\alpha \in \mathbb{Z}_+ \mid x+\alpha\chi_u \in J\}$. *Then, we have* $\{x+\alpha\chi_u,\, x+(\alpha+1)\chi_u\} \cap J \neq \emptyset$ *for any* $\alpha \in \mathbb{Z}$ *with* $0 \leq \alpha < \bar{\alpha}$.

Proof. The claim follows immediately from (J-EXC$_0$). □

LEMMA 3.12. *Let* $x \in J$ *and* $u, w \in V$ *be distinct elements. Suppose that* $\{x+\chi_u,\, x+2\chi_u\} \cap J = \emptyset$.

(i) *Let* $\bar{\beta}_w = \max\{\beta \in \mathbb{Z}_+ \mid x+\beta(\chi_u+\chi_w) \in J\}$. *Then,* $x+\beta(\chi_u+\chi_w) \in J$ *for all* $\beta \in \mathbb{Z}$ *with* $0 \leq \beta \leq \bar{\beta}_w$.

(ii) *Let* $\bar{\gamma}_w = \max\{\gamma \in \mathbb{Z}_+ \mid x+\gamma(\chi_u-\chi_w) \in J\}$. *Then,* $x+\gamma(\chi_u-\chi_w) \in J$ *for all* $\gamma \in \mathbb{Z}$ *with* $0 \leq \gamma \leq \bar{\gamma}_w$.

Proof. We prove (i) only. It suffices to show that for any positive integer β with $\beta \geq 2$ and $x+\beta(\chi_u+\chi_w) \in J$, it holds that $x+(\beta-1)(\chi_u+\chi_w) \in J$. By (J-EXC$_0$) applied to $y = x+\beta(\chi_u+\chi_w)$ and x, we have $y-\chi_w+t \in J$ for some $t \in \{\mathbf{0}, -\chi_u, -\chi_w\}$. Since $\{x+\chi_u,\, x+2\chi_u\} \cap J = \emptyset$, it follows from Lemma 3.6 that $\{y-\chi_w,\, y-2\chi_w\} \cap J = \emptyset$. Therefore, we have $y-\chi_w-\chi_u = x+(\beta-1)(\chi_u+\chi_w) \in J$. □

LEMMA 3.13. *For any* $u \in V$, *the procedure* UPPER_BOUND(J,u) *runs in* $\mathrm{O}(n \log \Phi(J))$ *time, provided a vector* $x_0 \in J$ *is given.*

Proof. By Lemma 3.11, the value $\bar{\alpha}$ in Step 1 can be computed in $\mathrm{O}(\log \Phi(J))$ time by a variant of binary search. Similarly, we can compute $\bar{\beta}_v$ and $\bar{\gamma}_v$ $(v \in V \setminus \{u\})$ by binary search in $\mathrm{O}(\log \Phi(J))$ time by Lemma 3.12. Hence, the claim follows. □

This concludes the proof of Theorem 3.3.

4. Polynomial-time algorithms for convex optimization on jump systems. In this section, we consider the following two convex optimization problems on jump systems. The first problem is the minimization of a separable-convex function on a jump system:

$$\text{(ScFMin)} \quad \text{Minimize} \quad f(x) \equiv \sum_{v \in V} f_v(x(v)) \qquad \text{subject to} \quad x \in J,$$

where $f_v : \mathbb{Z} \to \mathbb{R}$ $(v \in V)$ is a family of univariate convex functions and J is a finite jump system. The second problem is the minimization of an M-convex function on a constant-parity jump system:

$$\text{(McFMin)} \quad \text{Minimize} \quad f(x) \qquad \text{subject to} \quad x \in J,$$

where $J \subseteq \mathbb{Z}^V$ is a finite constant-parity jump system and $f : J \to \mathbb{R}$ is an M-convex function. For both of the problems, we assume that a membership oracle for J and an oracle for evaluating the function value of f are available and that a vector in J is given. We present polynomial-time algorithms for the two problems.

4.1. A polynomial-time algorithm for minimizing a separable-convex function on a jump system. We first show some properties for optimal solutions of the problem (ScFMin). The global optimality of the problem (ScFMin) is characterized by a local optimality.

THEOREM 4.1 (see [3, Corollary 4.2]). *A vector $x \in J$ is an optimal solution of* (ScFMin) *if and only if $f(x) \leq f(x+s+t)$ for all $s, t \in N_1 \cup \{\mathbf{0}\}$ with $x+s+t \in J$.*

The next property shows that a given nonoptimal vector in J can be easily separated from an optimal solution.

THEOREM 4.2 (minimizer cut property for (ScFMin)). *Let $x \in J$ be a vector which is not an optimal solution of* (ScFMin). *Suppose that $s_* \in N_1$ satisfies*

$$\begin{aligned} s_* \in \arg\min\{f(x+s) \mid s \in N_1,\ \exists t \in N_1 \cup \{\mathbf{0}\} \\ \textit{such that } x+s+t \in J \textit{ and } f(x+s+t) < f(x)\}. \end{aligned} \tag{4.1}$$

Then, there exists an optimal solution x_* of (ScFMin) *satisfying*

$$\begin{cases} x_*(u) \leq x(u) - 1 & (\textit{if } s_* = -\chi_u), \\ x_*(u) \geq x(u) + 1 & (\textit{if } s_* = +\chi_u). \end{cases}$$

The proof of Theorem 4.2 will be given in section 4.4.1.

Our algorithm maintains a box $[a, b]$ containing an optimal solution of (ScFMin). Note that $J \cap [a, b]$ is a jump system by Proposition 2.2. The box $[a, b]$ is reduced iteratively by using the minimizer cut property (Theorem 4.2), and finally, an optimal solution is found.

Given a finite set $J \subseteq \mathbb{Z}^V$, we define a set $J^\circ \subseteq \mathbb{Z}^V$ by

$$J^\circ = J \cap [a_J^\circ, b_J^\circ], \tag{4.2}$$

where

$$a_J(v) = \min\{y(v) \mid y \in J\}, \quad b_J(v) = \max\{y(v) \mid y \in J\} \qquad (v \in V),$$
$$a^\circ_J(v) = \left\lfloor \left(1 - \frac{1}{n}\right) a_J(v) + \frac{1}{n} b_J(v) \right\rfloor \qquad (v \in V),$$
$$b^\circ_J(v) = \left\lceil \frac{1}{n} a_J(v) + \left(1 - \frac{1}{n}\right) b_J(v) \right\rceil \qquad (v \in V).$$

The set J° is intended to represent a set of vectors in J lying away from the boundary.

THEOREM 4.3. *Let J be a finite jump system.*

(i) *J° is nonempty and hence a jump system.*

(ii) *A vector in J° can be found in $\mathrm{O}(n^2 \log \Phi(J))$ time, provided a vector in J is given.*

Proof. The proof is given in sections 4.4.2 and 4.4.3. □

ALGORITHM DOMAIN_REDUCTION_SCFMIN.

Step 0: Set $a(v) := a_J(v)$ and $b(v) := b_J(v)$ for $v \in V$.

Step 1: Find a vector $x \in (J \cap [a, b])^\circ$.

Step 2: If $f(x) \leq f(x + s + t)$ for all $s, t \in N_1 \cup \{\mathbf{0}\}$ with $x + s + t \in J$, then stop (x is optimal).

Step 3: Find $s_* \in N_1$ satisfying (4.1).

Step 4: Put $\{u\} = \mathrm{supp}(s_*)$. Modify a or b as follows:

$$\begin{cases} b(u) := x(u) - 1 & (\text{if } s_* = -\chi_u), \\ a(u) := x(u) + 1 & (\text{if } s_* = +\chi_u). \end{cases}$$

Go to Step 1.

We analyze the number of iterations of the algorithm. Denote by a_i, b_i the vectors a, b at the beginning of the ith iteration. It is clear that $b_i(v) - a_i(v)$ is nonincreasing w.r.t. i. Furthermore, we have the following property.

LEMMA 4.4. *Let $u \in V$ be the element with $\{u\} = \mathrm{supp}(s_*)$, where s_* is the vector chosen in Step 2 of the ith iteration. Then, we have*

$$b_{i+1}(u) - a_{i+1}(u) < \left(1 - \frac{1}{n}\right)\{b_i(u) - a_i(u)\}.$$

Proof. We show the inequality in the case $s_* = -\chi_u$ only. Let $x \in (J \cap [a_i, b_i])^\circ$ be the vector chosen in Step 1 of the ith iteration. Then,

$$\begin{aligned} b_{i+1}(u) - a_{i+1}(u) &= x(u) - 1 - a_i(u) \\ &\leq \left\lceil \frac{1}{n} a_i(u) + \left(1 - \frac{1}{n}\right) b_i(u) \right\rceil - 1 - a_i(u) \\ &< \left(1 - \frac{1}{n}\right) \{b_i(u) - a_i(u)\}. \qquad \square \end{aligned}$$

We have $b_0(v) - a_0(v) \leq \Phi(J)$ for all $v \in V$ at the beginning of the algorithm, and if $b_i(v) - a_i(v) < 1$ for all $v \in V$, then we obtain an optimal solution immediately. Hence, it follows from Lemma 4.4 that the algorithm DOMAIN_REDUCTION_SCFMIN terminates in $\mathrm{O}(n^2 \log \Phi(J))$ iterations.

By Theorem 4.3, a vector in $(J \cap [a, b])^\circ$ can be found in $\mathrm{O}(n^2 \log \Phi(J))$ time. Steps 2, 3, and 4 can be done in $\mathrm{O}(n^2)$ time.

THEOREM 4.5. *The algorithm* DOMAIN_REDUCTION_SCFMIN *finds an optimal solution of the problem* (ScFMin) *in $\mathrm{O}(n^4 (\log \Phi(J))^2)$ time, provided a vector in J is given.*

4.2. Application to weak majorized elements in jump systems. We explain an application of our algorithm to the problem of finding least weakly sub- and supermajorized elements in jump systems discussed in [1] (see also [26]).

For a vector $x \in \mathbb{R}^V$, let $x_{[1]} \geq x_{[2]} \geq \cdots \geq x_{[n]}$ denote the components of x in decreasing order. For two vectors $x, y \in \mathbb{R}^V$, the vector x is said to be *weakly submajorized by* y if

$$\sum_{i=1}^{j} x_{[i]} \leq \sum_{i=1}^{j} y_{[i]} \qquad (j = 1, 2, \ldots, n).$$

For a nonempty subset S of $\mathbb{R}^V$, a vector $x \in S$ is said to be a *least weakly submajorized element* of S if x is weakly submajorized by y for all $y \in S$.

The concept of weak supermajorization is similarly defined. For a vector $x \in \mathbb{R}^V$, let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ denote the components of x in increasing order. For two vectors $x, y \in \mathbb{R}^V$, the vector x is said to be *weakly supermajorized by* y if

$$\sum_{i=1}^{j} x_{(i)} \geq \sum_{i=1}^{j} y_{(i)} \qquad (j = 1, 2, \ldots, n).$$

It is easy to see that x is weakly supermajorized by y if and only if $-x$ is weakly submajorized by $-y$. For a nonempty subset S of $\mathbb{R}^V$, a vector $x \in S$ is said to be a *least weakly supermajorized element* of S if x is weakly supermajorized by y for all $y \in S$.

The following statement conjectured by Tamir [26] is proven by Ando [1].

THEOREM 4.6 (see [1]). *Any finite jump system has a least weakly sub- and supermajorized elements.*

The proof of Theorem 4.6 in [1] shows that the problem of finding a least weakly submajorized element of a jump system J can be reduced to the following convex quadratic optimization problem:

$$\text{Minimize} \quad \sum_{v \in V} (x(v) + M)^2 \qquad \text{subject to} \quad x \in J,$$

where M is a nonnegative real number such that $x(v) + M \geq 0$ for all $x \in J$ and $v \in V$. Such M is given by

$$M = \begin{cases} 0 & (\text{if } J \subseteq \mathbb{Z}_+^V), \\ -\min_{v \in V}\{\min\{y(v) \mid y \in J\}\} & (\text{otherwise}) \end{cases}$$

and can be computed in $\mathrm{O}(n^2 \log \Phi(J))$ time by Theorem 3.3. Then, the convex quadratic optimization problem above can be solved in $\mathrm{O}(n^4 (\log \Phi(J))^2)$ time by using the algorithm DOMAIN_REDUCTION_SCFMIN.

THEOREM 4.7. *Least weakly sub- and supermajorized elements of a finite jump system J can be computed in* $\mathrm{O}(n^4(\log \Phi(J))^2)$ *time, provided a vector in J is given.*

4.3. A polynomial-time algorithm for minimization of an M-convex function on a constant-parity jump system. The problem (McFMin) can be solved in polynomial time in a similar way as the problem (ScFMin), due to the following useful properties. The global optimality of the problem (McFMin) is characterized by a local optimality.

THEOREM 4.8 (see [22, Theorem 3.3]). *A vector $x \in J$ is an optimal solution of* (McFMin) *if and only if $f(x) \le f(x+s+t)$ holds for all $s, t \in N_1$ with $x+s+t \in J$.*

The minimizer cut property holds for the problem (McFMin) as well.

THEOREM 4.9 (minimizer cut property for (McFMin) [23, Theorem 4.1]). *Let $x \in J$ be a vector which is not an optimal solution of* (McFMin)*, and $s_*, t_* \in N_1$ satisfy*

$$f(x+s_*+t_*) = \min\{f(x+s+t) \mid s,t \in N_1\}.$$

Put $\{u\} = \mathrm{supp}(s_*)$ and $\{w\} = \mathrm{supp}(t_*)$. Then, there exists $x_* \in \arg\min f$ such that

$$x_*(u) \begin{cases} \le x(u)-1 & (\textit{if } s_* = -\chi_u), \\ \ge x(u)+1 & (\textit{if } s_* = +\chi_u), \end{cases} \qquad x_*(w) \begin{cases} \le x(w)-1 & (\textit{if } t_* = -\chi_w), \\ \ge x(w)+1 & (\textit{if } t_* = +\chi_w). \end{cases}$$

Based on Theorems 4.8 and 4.9, we consider a variant of the algorithm DOMAIN_REDUCTION_SCFMIN in section 4.1.

ALGORITHM DOMAIN_REDUCTION_MCFMIN.

Step 0: Set $a(v) := a_J(v)$ and $b(v) := b_J(v)$ for $v \in V$.

Step 1: Find a vector $x \in (J \cap [a,b])^\circ$.

Step 2: If $f(x) \le f(x+s+t)$ for all $s,t \in N_1$ with $x+s+t \in J$, then stop (x is optimal).

Step 3: Find $s_*, t_* \in N_1$ satisfying $f(x+s_*+t_*) = \min\{f(x+s+t) \mid s,t \in N_1\}$.

Step 4: Put $\{u\} = \mathrm{supp}(s_*)$ and $\{w\} = \mathrm{supp}(t_*)$. Modify a and b as follows:

$$\begin{cases} b(u) := x(u)-1 & (\text{if } s_* = -\chi_u), \\ a(u) := x(u)+1 & (\text{if } s_* = +\chi_u), \end{cases} \qquad \begin{cases} b(w) := x(w)-1 & (\text{if } t_* = -\chi_w), \\ a(w) := x(w)+1 & (\text{if } t_* = +\chi_w). \end{cases}$$

Go to Step 1.

We can show the following result, where the proof is quite similar to that for Theorem 4.5 and therefore omitted.

THEOREM 4.10. *The algorithm* DOMAIN_REDUCTION_MCFMIN *solves the problem* (McFMin) *in* $\mathrm{O}(n^4(\log \Phi(J))^2)$ *time, provided a vector in J is given.*

4.4. Proofs.

4.4.1. Proof of the minimizer cut property for (ScFMin). In this section, we prove Theorem 4.2. A proof of Theorem 4.2 is given for a special case where J is a convex jump system [2, Theorem 5.2]. A jump system J is said to be *convex* if every integral point in the convex hull of J is contained in J. In the following, we give a proof for the general case.

The proof uses the following fundamental properties of separable-convex functions.

PROPOSITION 4.11. *Let $f : \mathbb{Z}^V \to \mathbb{R}$ be a separable-convex function.*

(i) *For any $x, y \in \mathbb{Z}^V$ and any $s \in \mathrm{inc}(x,y)$, we have*

$$f(x) + f(y) \ge f(x+s) + f(y-s).$$

(ii) *For any $x \in \mathbb{Z}^V$ and any $s, t \in N_1$ with* $\mathrm{supp}(s) \ne \mathrm{supp}(t)$*, we have*

$$f(x+s+t) - f(x) = \{f(x+s) - f(x)\} + \{f(x+t) - f(x)\}.$$

Let $x \in J$ be a vector which is not an optimal solution of (ScFMin) and $s_* \in N_1$ be a vector satisfying (4.1). We assume, without loss of generality, that $s_* = +\chi_u$

for some $u \in V$. Let x_* be an optimal solution of (ScFMin) maximizing the value $x_*(u)$, and assume that x_* minimizes $\|x_* - x\|_1$ among all such x_*. We assume, to the contrary, that $x_*(u) \leq x(u)$ and derive a contradiction.

By the definition of s_*, there exists $t_* \in N_1 \cup \{\mathbf{0}\}$ such that

$$x + s_* + t_* \in J, \qquad f(x + s_* + t_*) < f(x). \tag{4.3}$$

LEMMA 4.12. $f(x + s_*) < f(x)$.

Proof. We assume $t_* \neq \mathbf{0}$ since otherwise the claim is obvious from (4.3). If $t_* = s_*$, then the separable convexity of f and (4.3) imply $\{f(x + s_*) - f(x)\} \leq \{f(x + 2s_*) - f(x)\}/2 < 0$. If $t_* \neq s_*$, then (4.1) and Proposition 4.11(ii) imply

$$\begin{aligned} 2\{f(x + s_*) - f(x)\} &\leq \{f(x + s_*) - f(x)\} + \{f(x + t_*) - f(x)\} \\ &= f(x + s_* + t_*) - f(x), \end{aligned}$$

which, together with (4.3), yields $f(x + s_*) < f(x)$. □

LEMMA 4.13. *There exists $p \in \mathrm{inc}(x_*, x)$ such that $f(x_* + p) > f(x_*)$ and $f(x - p) < f(x + s_*)$.*

Proof. Since $s_* \in \mathrm{inc}(x_*, x + s_*)$, Proposition 4.11(i) and Lemma 4.12 imply

$$f(x_* + s_*) - f(x_*) \leq f(x + s_*) - f(x) < 0, \tag{4.4}$$

which, together with the optimality of x_*, yields $x_* + s_* \notin J$. Since $s_* \in \mathrm{inc}(x_*, x + s_* + t_*)$, (J-EXC$_0$) implies that there exists $p \in \mathrm{inc}(x_* + s_*, x + s_* + t_*)$ such that $x_* + s_* + p \in J$. By the optimality of x_*, we have

$$f(x_* + s_* + p) > f(x_*). \tag{4.5}$$

Claim 1. $p \neq s_*$.

Proof of claim. Assume, to the contrary, that $p = s_*$. We consider the following two cases and derive a contradiction.

Case 1 ($s_* = t_*$). Separable convexity of f, the inequality $x_*(u) \leq x(u)$, and (4.3) imply

$$f(x_* + 2s_*) - f(x_*) \leq f(x + 2s_*) - f(x) < 0,$$

which contradicts the inequality (4.5).

Case 2 ($s_* \neq t_*$). Inequality (4.5) implies $f(x_* + 2s_*) > f(x_*)$, from which follows

$$f(x_* + 2s_*) - f(x_* + s_*) \geq (1/2)\{f(x_* + 2s_*) - f(x_*)\} > 0. \tag{4.6}$$

Since $s_* = p \in \mathrm{inc}(x_* + s_*, x + s_* + t_*)$, Proposition 4.11(i) implies

$$f(x + s_* + t_*) - f(x + t_*) \geq f(x_* + 2s_*) - f(x_* + s_*). \tag{4.7}$$

Since $f(x + s_* + t_*) - f(x + t_*) = f(x + s_*) - f(x)$ by Proposition 4.11(ii), it follows from (4.6) and (4.7) that $f(x + s_*) > f(x)$, a contradiction to Lemma 4.12. □

We first show that $p \in \mathrm{inc}(x_*, x)$. Assume, to the contrary, that $p \notin \mathrm{inc}(x_*, x)$. Since $p \in \mathrm{inc}(x_* + s_*, x + s_* + t_*) = \mathrm{inc}(x_*, x + t_*)$, we have $p = t_*$. Then, $t_* \neq s_*$ by Claim 1. Therefore, Proposition 4.11(ii) implies

$$\begin{aligned} f(x_* + s_* + p) - f(x_*) &= f(x_* + s_* + t_*) - f(x_*) \\ &= \{f(x_* + s_*) - f(x_*)\} + \{f(x_* + t_*) - f(x_*)\} \\ &\leq \{f(x + s_*) - f(x)\} + \{f(x + t_*) - f(x)\} \\ &= f(x + s_* + t_*) - f(x) \ < \ 0, \end{aligned}$$

where the first inequality is by $s_* \in \mathrm{inc}(x_*, x + s_*)$ and $t_* = p \in \mathrm{inc}(x_*, x + t_*)$, and the second inequality by (4.3). This, however, is a contradiction to (4.5).

We then show that $f(x_* + p) > f(x_*)$ and $f(x - p) < f(x + s_*)$. By (4.5), Proposition 4.11(ii), and Claim 1, we have

$$0 < f(x_* + s_* + p) - f(x_*) = \{f(x_* + s_*) - f(x_*)\} + \{f(x_* + p) - f(x_*)\}. \tag{4.8}$$

Therefore, it holds that

$$\begin{aligned} f(x - p) - f(x) &\le f(x_*) - f(x_* + p) \\ &< f(x_* + s_*) - f(x_*) \\ &\le f(x + s_*) - f(x) \ < \ 0, \end{aligned}$$

where the first inequality is by Proposition 4.11(i) and $p \in \mathrm{inc}(x_*, x)$, the second is by (4.8), and the last two inequalities are by (4.4). This implies $f(x_* + p) > f(x_*)$ and $f(x - p) < f(x + s_*)$. □

Let $p_1 \in \mathrm{inc}(x_*, x)$ be a vector with $f(x_* + p_1) > f(x_*)$ minimizing the value $f(x - p_1)$ among all such vectors. It follows from Lemmas 4.12 and 4.13 that

$$f(x - p_1) < f(x + s_*) < f(x), \tag{4.9}$$

which implies $x - p_1 \notin J$ by (4.1). Hence, (J-EXC$_0$) implies that there exists $q \in \mathrm{inc}(x - p_1, x_*)$ such that $x - p_1 + q \in J$. By (4.1) and (4.9), we have

$$f(x - p_1 + q) \ge f(x). \tag{4.10}$$

Lemma 4.14. $q \ne -p_1$.

Proof. Assume, to the contrary, that $q = -p_1$. Since $-p_1 = q \in \mathrm{inc}(x - p_1, x_*)$, Proposition 4.11(i) implies

$$f(x_*) - f(x_* + p_1) \ge f(x - 2p_1) - f(x - p_1). \tag{4.11}$$

By (4.9) and (4.10), we have

$$f(x - 2p_1) - f(x - p_1) \ge f(x) - f(x - p_1) > 0. \tag{4.12}$$

It follows from (4.11) and (4.12) that $f(x_*) > f(x_* + p_1)$, a contradiction to the choice of p_1. □

Since $q \in \mathrm{inc}(x - p_1, x_*) \subseteq \mathrm{inc}(x, x_*)$, it follows from Proposition 4.11(i) that

$$f(x_*) - f(x_* - q) \ge f(x + q) - f(x). \tag{4.13}$$

By Proposition 4.11(ii), (4.10), and Lemma 4.14, we have

$$f(x + q) - f(x) \ge f(x) - f(x - p_1). \tag{4.14}$$

It follows from (4.9), (4.13), and (4.14) that

$$f(x_*) - f(x_* - q) \ge f(x) - f(x - p_1) > 0. \tag{4.15}$$

From this inequality we have $x_* - q \notin J$ since x_* is an optimal solution. Hence, (J-EXC$_0$) implies that there exists $p_2 \in \mathrm{inc}(x_* - q, x)$ such that $x_* - q + p_2 \in J$. We note that $(x_* - q + p_2)(u) \geq x_*(u)$ since $-s_* \notin \{-q, p_2\}$ and that $\|(x_* - q + p_2) - x\|_1 < \|x_* - x\|_1$. Therefore, we have

$$f(x_* - q + p_2) > f(x_*) \tag{4.16}$$

by the choice of x_*.

LEMMA 4.15. $p_2 \neq -q$.

Proof. Assume, to the contrary, that $p_2 = -q$. Since $-q = p_2 \in \mathrm{inc}(x_* - q, x)$, Proposition 4.11(i) implies

$$f(x) - f(x + q) \geq f(x_* - 2q) - f(x_* - q) > 0, \tag{4.17}$$

where the last inequality is by (4.15) and (4.16). On the other hand, Proposition 4.11(ii), (4.10), and Lemma 4.14 imply

$$f(x + q) - f(x) \geq f(x) - f(x - p_1) > 0,$$

where the last inequality is by (4.9). This inequality, however, is a contradiction to (4.17). □

By Proposition 4.11(ii), (4.16), and Lemma 4.15, we have

$$f(x_* + p_2) - f(x_*) > f(x_*) - f(x_* - q). \tag{4.18}$$

Since $p_2 \in \mathrm{inc}(x_* - q, x) \subseteq \mathrm{inc}(x_*, x)$, it follows from Proposition 4.11(i) that

$$f(x) - f(x - p_2) \geq f(x_* + p_2) - f(x_*),$$

which, together with (4.15) and (4.18), implies $f(x_* + p_2) > f(x_*)$ and $f(x - p_2) < f(x - p_1)$, a contradiction to the choice of p_1.

This concludes the proof of Theorem 4.2.

4.4.2. Nonemptiness of J°. We prove Theorem 4.3(i), the nonemptiness of the set $J^\circ = J \cap [a_J^\circ, b_J^\circ]$ defined by (4.2).

We first show that the intersection of the convex hull $\mathrm{conv}(J)$ of J and the box $[a_J^\circ, b_J^\circ]$ is nonempty.

We define

$$3^V = \{(X, Y) \mid X, Y \subseteq V,\ X \cap Y = \emptyset\}.$$

Given a function $\rho : 3^V \to \mathbb{R}$, we define a polyhedron $P_*(\rho)$ as

$$P_*(\rho) = \{x \in \mathbb{R}^V \mid x(X) - x(Y) \leq \rho(X, Y)\ (\forall (X, Y) \in 3^V)\}.$$

A function $\rho : 3^V \to \mathbb{R}$ is called a *bisubmodular function* if it satisfies the following inequality for all $(X_1, Y_1), (X_2, Y_2) \in 3^V$:

$$\begin{aligned}&\rho(X_1, Y_1) + \rho(X_2, Y_2)\\ &\geq \rho(X_1 \cap X_2,\, Y_1 \cap Y_2) + \rho((X_1 \cup X_2) \setminus (Y_1 \cup Y_2),\, (Y_1 \cup Y_2) \setminus (X_1 \cup X_2)).\end{aligned}$$

THEOREM 4.16 (see [7]). *Let $J \subseteq \mathbb{Z}^V$ be a jump system. Then, there exists an integer-valued bisubmodular function $\rho_J : 3^V \to \mathbb{Z} \cup \{+\infty\}$ such that $\rho_J(\emptyset, \emptyset) = 0$ and* $\mathrm{conv}(J) = P_*(\rho_J)$. *Moreover, such ρ_J is uniquely determined by*

$$\rho_J(X, Y) = \sup\{x(X) - x(Y) \mid x \in J\} \qquad ((X, Y) \in 3^V). \tag{4.19}$$

THEOREM 4.17 (see [12]). *Let $\rho : 3^V \to \mathbb{R}$ be a bisubmodular function with $\rho(\emptyset, \emptyset) = 0$ and $a, b \in \mathbb{R}^V$ be vectors with $a \leq b$. Then, the set $P_*(\rho) \cap [a, b]$ is nonempty if and only if*

$$a(X) - b(Y) \leq \rho(X, Y) \qquad (\forall (X, Y) \in 3^V). \tag{4.20}$$

LEMMA 4.18. *For a finite jump system $J \subseteq \mathbb{Z}^V$, it holds that* $\mathrm{conv}(J) \cap [a_J^\circ, b_J^\circ] \neq \emptyset$.

Proof. Let $\rho = \rho_J$ be a function defined by (4.19). It follows from Theorem 4.16 that ρ is an integer-valued bisubmodular function satisfying $\rho(\emptyset, \emptyset) = 0$ and $\mathrm{conv}(J) = P_*(\rho_J)$. Moreover, we have

$$a_J^\circ(v) = \left\lfloor -\left(1 - \frac{1}{n}\right) \rho(\emptyset, \{v\}) + \frac{1}{n} \rho(\{v\}, \emptyset) \right\rfloor \qquad (v \in V), \tag{4.21}$$

$$b_J^\circ(v) = \left\lceil -\frac{1}{n} \rho(\emptyset, \{v\}) + \left(1 - \frac{1}{n}\right) \rho(\{v\}, \emptyset) \right\rceil \qquad (v \in V) \tag{4.22}$$

since $\rho(\emptyset, \{v\}) = -a_J(v)$ and $\rho(\{v\}, \emptyset) = b_J(v)$ hold. To prove $\mathrm{conv}(J) \cap [a_J^\circ, b_J^\circ] \neq \emptyset$, it suffices to show that $a_J^\circ(X) - b_J^\circ(Y) \leq \rho(X, Y)$ for all $(X, Y) \in 3^V$ by Theorem 4.17.

Let $(X, Y) \in 3^V$ and put $k = |X| + |Y|$. We claim that

$$\begin{aligned} &k\rho(X, Y) + k \sum_{v \in Y} \rho(\{v\}, \emptyset) + k \sum_{v \in X} \rho(\emptyset, \{v\}) \\ &\geq \sum_{v \in Y} \{\rho(\{v\}, \emptyset) + \rho(\emptyset, \{v\})\} + \sum_{v \in X} \{\rho(\{v\}, \emptyset) + \rho(\emptyset, \{v\})\}. \end{aligned} \tag{4.23}$$

Indeed, the bisubmodularity of ρ implies

$$\begin{aligned} \text{LHS of (4.23)} &= \sum_{w \in Y} \left\{ \rho(X, Y) + \sum_{v \in Y \setminus \{w\}} \rho(\{v\}, \emptyset) + \sum_{v \in X} \rho(\emptyset, \{v\}) \right\} \\ &\quad + \sum_{w \in X} \left\{ \rho(X, Y) + \sum_{v \in Y} \rho(\{v\}, \emptyset) + \sum_{v \in X \setminus \{w\}} \rho(\emptyset, \{v\}) \right\} \\ &\quad + \sum_{v \in Y} \rho(\{v\}, \emptyset) + \sum_{v \in X} \rho(\emptyset, \{v\}) \\ &\geq \sum_{w \in Y} \left\{ \rho(X, Y) + \rho(Y \setminus \{w\}, \emptyset) + \rho(\emptyset, X) \right\} \\ &\quad + \sum_{w \in X} \left\{ \rho(X, Y) + \rho(Y, \emptyset) + \rho(\emptyset, X \setminus \{w\}) \right\} \\ &\quad + \sum_{v \in Y} \rho(\{v\}, \emptyset) + \sum_{v \in X} \rho(\emptyset, \{v\}) \\ &\geq \text{RHS of (4.23)}. \end{aligned}$$

Since the LHS of (4.23) is nonnegative and $k \leq n$, the integer k in (4.23) can be replaced with n. Thus,

$$\begin{aligned}\rho(X,Y) &\geq \sum_{v\in X}\left\{-\left(1-\frac{1}{n}\right)\rho(\emptyset,\{v\})+\frac{1}{n}\rho(\{v\},\emptyset)\right\}\\ &\quad-\sum_{v\in Y}\left\{-\frac{1}{n}\rho(\emptyset,\{v\})+\left(1-\frac{1}{n}\right)\rho(\{v\},\emptyset)\right\}\\ &\geq a^{\circ}_J(X)-b^{\circ}_J(Y),\end{aligned}$$

where the last inequality follows from (4.21) and (4.22). □

We prove the nonemptiness of J° by using the following theorem.

THEOREM 4.19 (see [18, Theorem 5.1]). *Let J be a finite jump system and $a, b \in \mathbb{Z}^V$ be vectors with $a(v) < b(v)$ for all $v \in V$. Then, there exists a vector $w \in \{-1, 0, +1\}^V$ such that*

$$\min\{\|x-y\|_1 \mid x \in [a,b],\ y \in J\} = \min\{w^T x \mid x \in [a,b]\} - \max\{w^T y \mid y \in J\}.$$

LEMMA 4.20. *For a finite jump system J, the set J° defined by* (4.2) *is nonempty.*

Proof. Let $V' = \{v \in V \mid a^{\circ}_J(v) < b^{\circ}_J(v)\}$. We denote by $J' \subseteq \mathbb{Z}^{V'}$ the orthogonal projection of J onto $\mathbb{Z}^{V'}$, i.e.,

$$J' = \{y \in \mathbb{Z}^{V'} \mid \exists x \in J \text{ such that } y(v) = x(v)\ (v \in V')\}.$$

For $v \in V \setminus V'$, we have $a^{\circ}_J(v) = b^{\circ}_J(v) = a_J(v) = b_J(v)$, implying that $y(v) = a^{\circ}_J(v)\ (= b^{\circ}_J(v))$ for all $y \in J$. Therefore, $J \cap [a^{\circ}_J, b^{\circ}_J] \neq \emptyset$ if and only if

$$J' \cap \{x \in \mathbb{Z}^{V'} \mid a^{\circ}_J(v) \leq x(v) \leq b^{\circ}_J(v)\ (v \in V')\} \neq \emptyset,$$

where it is noted that $a^{\circ}_{J'}(v) = a^{\circ}_J(v)$ and $b^{\circ}_{J'}(v) = b^{\circ}_J(v)$ for $v \in V'$. Hence, it suffices to consider the case where $a^{\circ}_J(v) < b^{\circ}_J(v)$ for all $v \in V$.

By Theorem 4.19, there exists some $w \in \{-1, 0, +1\}^V$ such that

$$\min\{\|x-y\|_1 \mid x \in [a^{\circ}_J, b^{\circ}_J],\ y \in J\} = \min\{w^T x \mid x \in [a^{\circ}_J, b^{\circ}_J]\} - \max\{w^T y \mid y \in J\}. \tag{4.24}$$

Since $\mathrm{conv}(J) \cap [a^{\circ}_J, b^{\circ}_J] \neq \emptyset$ by Lemma 4.18, we have

$$\begin{aligned}&\min\{w^T x \mid x \in [a^{\circ}_J, b^{\circ}_J]\} - \max\{w^T y \mid y \in J\}\\ &= \min\{w^T x \mid x \in [a^{\circ}_J, b^{\circ}_J]\} - \max\{w^T y \mid y \in \mathrm{conv}(J)\} \ \leq\ 0.\end{aligned} \tag{4.25}$$

It follows from (4.24) and (4.25) that $\min\{\|x-y\|_1 \mid x \in [a^{\circ}_J, b^{\circ}_J],\ y \in J\} = 0$, implying that $J^{\circ} = J \cap [a^{\circ}_J, b^{\circ}_J] \neq \emptyset$. □

This concludes the proof of Theorem 4.3(i).

4.4.3. Finding a vector in J°. We prove Theorem 4.3(ii) by providing an algorithm to find a vector in J°. More generally, we consider how to find a vector in the (nonempty) intersection of a jump system J and a box $[a, b]$.

Our algorithm is based on the following simple observation.

LEMMA 4.21. *Let J be a jump system, $u \in V$, and α, β be integers such that $\alpha \leq \beta$ and $J \cap \{y \in \mathbb{Z}^V \mid \alpha \leq y(u) \leq \beta\} \neq \emptyset$. Then, we have*

$$\max\{y(u) \mid y \in J,\ y(u) \leq \beta\} \geq \alpha, \quad \min\{y(u) \mid y \in J,\ y(u) \geq \alpha\} \leq \beta.$$

Proof. Let x be any vector in $J \cap \{y \in \mathbb{Z}^V \mid \alpha \leq y(u) \leq \beta\}$. Then, we have

$$\begin{aligned}\max\{y(u) \mid y \in J,\ y(u) \leq \beta\} \geq x(u) \geq \alpha,\\ \min\{y(u) \mid y \in J,\ y(u) \geq \alpha\} \leq x(u) \leq \beta.\end{aligned}$$ □

Given a jump system J and vectors $a, b \in \mathbb{Z}^V$ with $a \leq b$ and $J \cap [a, b] \neq \emptyset$, the following algorithm finds a vector in $J \cap [a, b]$, provided a vector in J is given.

```
ALGORITHM FIND_VECTOR_IN_$J \cap [a, b]$.
Step 0: Let $x := x_0$ be an initial vector in $J$.
Step 1: While $\{v \in V \mid x(v) < a(v)\} \neq \emptyset$, do the following steps:
  Step 1-1: Choose an element $u \in V$ with $x(u) < a(u)$.
  Step 1-2: Find a vector $x_*$ in $J'$ maximizing the value $x_*(u)$, where
      $J' = J \cap \{y \in \mathbb{Z}^V \mid y(u) \leq b(u),$
            $\min(x(v), a(v)) \leq y(v) \leq \max(x(v), b(v)) \ (v \in V \setminus \{u\})\}.$
  Step 1-3: Put $x := x_*$.
Step 2: While $\{v \in V \mid x(v) > b(v)\} \neq \emptyset$, do the following steps:
  Step 2-1: Choose an element $u \in V$ with $x(u) > b(u)$.
  Step 2-2: Find a vector $x_*$ in $J'$ minimizing the value $x_*(u)$, where
      $J' = J \cap \{y \in \mathbb{Z}^V \mid y(u) \geq a(u),$
            $\min(x(v), a(v)) \leq y(v) \leq \max(x(v), b(v)) \ (v \in V \setminus \{u\})\}.$
  Step 2-3: Put $x := x_*$.
Step 3: Output $x$.
```

We observe that if the inequality $a(v) \leq x(v) \leq b(v)$ for some $v \in V$ is once satisfied, then it is kept until termination of the algorithm. Note that the set J' defined in Step 1-1 is a jump system by Proposition 2.2. This, together with Lemma 4.21, implies that the vector x in Step 1-3 satisfies $x \in J'$ and $a(u) \leq x(u) \leq b(u)$. Similarly, for each $u \in V$ with $x(u) > b(u)$, the inequality $a(u) \leq x(u) \leq b(u)$ is satisfied in Step 2-3. Thus, the vector x satisfies $x \in J \cap [a, b]$ at the end of the algorithm.

Each iteration of Steps 1 and 2 requires $\mathrm{O}(n \log \Phi(J'))$ time by Corollary 3.4, and we have $\Phi(J') \leq \Phi(J)$ since $J' \subseteq J$. Hence, the algorithm runs in $\mathrm{O}(n^2 \log \Phi(J))$ time.

This concludes the proof of Theorem 4.3(ii).

REFERENCES

[1] K. Ando, *Weak Majorization of Finite Jump Systems*, manuscript, 1996. Available from http://coconut.sys.eng.shizuoka.ac.jp/ando/maj/maj11.dvi.

[2] K. Ando, S. Fujishige, and T. Naitoh, *A greedy algorithm for minimizing a separable convex function over an integral bisubmodular polyhedron*, J. Oper. Res. Soc. Japan, 37 (1994), pp. 188–196.

[3] K. Ando, S. Fujishige, and T. Naitoh, *A greedy algorithm for minimizing a separable convex function over a finite jump system*, J. Oper. Res. Soc. Japan, 38 (1995), pp. 362–375.

[4] N. Apollonio and A. Sebő, *Minsquare factors and maxfix covers of graphs*, in Integer Programming and Combinatorial Optimization, D. Bienstock and G. Nemhauser, eds., Lecture Notes in Comput. Sci. 3064, Springer, Berlin, 2004, pp. 388–400.

[5] N. Apollonio and A. Sebő, *Minconvex Factors of Prescribed Size in Graphs*, Leibniz-IMAG preprint 145, Laboratoire Leibniz, Grenoble, France, 2006.

[6] A. Bouchet, *Greedy algorithm and symmetric matroids*, Math. Programming, 38 (1987), pp. 147–159.

[7] A. Bouchet and W. H. Cunningham, *Delta-matroids, jump systems, and bisubmodular polyhedra*, SIAM J. Discrete Math., 8 (1995), pp. 17–32.

[8] R. Chandrasekaran and S. N. Kabadi, *Pseudomatroids*, Discrete Math., 71 (1988), pp. 205–217.

[9] W. H. Cunningham, *Matching, matroids, and extensions*, Math. Program., 91 (2002), pp. 515–542.

[10] S. Fujishige, *A min-max theorem for bisubmodular polyhedra*, SIAM J. Discrete Math., 10 (1997), pp. 294–308.

[11] S. Fujishige, *Submodular Functions and Optimization*, 2nd ed., Ann. Discrete Math. 58, Elsevier, Amsterdam, 2005.

[12] S. Fujishige and S. B. Patkar, *The Box Convolution and the Dilworth Truncation of Bisubmodular Functions*, Report 94823, Forschungsinstitut für Diskrete Mathematik, Universität Bonn, Bonn, Germany, 1994.

[13] J. F. Geelen, *Lectures on Jump System*, manuscript, 1996. Available from http://www.math.uwaterloo.ca/~jfgeelen/publications/js.ps.

[14] H. Groenevelt, *Two algorithms for maximizing a separable concave function over a polymatroid feasible region*, European J. Oper. Res., 54 (1991), pp. 227–236.

[15] D. S. Hochbaum, *Lower and upper bounds for the allocation problem and other nonlinear optimization problems*, Math. Oper. Res., 19 (1994), pp. 390–409.

[16] Y. Kobayashi, K. Murota, and K. Tanaka, *Operations on M-convex functions on jump systems*, SIAM J. Discrete Math., 21 (2007), pp. 107–129.

[17] E. L. Lawler, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart, and Winston, New York, 1976, reprinted by Dover Publications, Mineola, NY, 2001.

[18] L. Lovász, *The membership problem in jump systems*, J. Combin. Theory Ser. B, 70 (1997), pp. 45–66.

[19] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.

[20] S. Moriguchi and A. Shioura, *On Hochbaum's proximity-scaling algorithm for the general resource allocation problem*, Math. Oper. Res., 29 (2004), pp. 394–397.

[21] K. Murota, *Discrete Convex Analysis*, SIAM Monogr. Discrete Math. Appl. 10, SIAM, Philadelphia, 2003.

[22] K. Murota, *M-convex functions on jump systems: A general framework for minsquare graph factor problem*, SIAM J. Discrete Math., 20 (2006), pp. 213–226.

[23] K. Murota and K. Tanaka, *A steepest descent algorithm for M-convex functions on jump systems*, IEICE Trans. Fundamentals, E89-A (2006), pp. 1160–1165.

[24] A. Schrijver, *Combinatorial Optimization: Polyhedra and Efficiency*, Springer-Verlag, Berlin, 2003.

[25] A. Shioura, *Minimization of an M-convex function*, Discrete Appl. Math., 84 (1998), pp. 215–220.

[26] A. Tamir, *Least majorized elements and generalized polymatroids*, Math. Oper. Res., 20 (1995), pp. 583–589.

[27] D. Welsh, *Matroid Theory*, Academic Press, New York, 1976.

SIAM J. DISCRETE MATH.
Vol. 21, No. 2, pp. 523–531

FURTHER RESULTS ON BAR *k*-VISIBILITY GRAPHS[*]

STEPHEN G. HARTKE[†], JENNIFER VANDENBUSSCHE[†], AND PAUL WENGER[†]

Abstract. A bar visibility representation of a graph G is a collection of horizontal bars in the plane corresponding to the vertices of G such that two vertices are adjacent if and only if the corresponding bars can be joined by an unobstructed vertical line segment. In a bar k-visibility graph, two vertices are adjacent if and only if the corresponding bars can be joined by a vertical line segment that intersects at most k other bars. Bar k-visibility graphs were introduced by Dean et al. [*J. Graph Algorithms Appl.*, 11 (2007), pp. 45–59]. In this paper, we present sharp upper bounds on the maximum number of edges in a bar k-visibility graph on n vertices and the largest order of a complete bar k-visibility graph. We also discuss regular bar k-visibility graphs and forbidden induced subgraphs of bar k-visibility graphs.

Key words. visibility graph, bar visibility graph, bar k-visibility graph, visibility representation of a graph

AMS subject classification. 05C62

DOI. 10.1137/050644240

1. Introduction.

The idea of representing a graph using a visibility relation has received much attention due to its applications to circuit layout (see [10] and additional references in [1]). Define a *bar visibility representation* of a graph G to be a set of disjoint horizontal closed line segments (or *bars*) in the plane in one-to-one correspondence with the vertices of G such that v and w are adjacent in G if and only if a vertical line segment can be drawn joining their associated bars that does not intersect any other bar. We say that G is a bar visibility graph if it has a bar visibility representation. (In the literature, these graphs have also been referred to as "visibility graphs" or "strong visibility graphs.") This particular model was first introduced by [7]. They observed that bar visibility graphs must be planar, and later provided a characterization of bar visibility graphs with the restriction that bars have distinct x coordinates as endpoints [8]. Later, [11] gave a somewhat complicated characterization for the general case and obtained some results concerning connectivity and bar visibility representations. [1] showed that determining whether a given graph is a bar visibility graph is an NP-complete problem.

Recently, Dean et al. introduced in [3, 4] the following generalization of bar visibility graphs. A *bar k-visibility representation* of a graph G is a bar representation in which v and w are adjacent in G if and only if a vertical line segment can be drawn joining their associated bars and which intersects *at most* k other bars. We denote the family of bar k-visibility graphs as $\mathcal{F}_k$. Notice that $\mathcal{F}_0$ is the family of bar visibility graphs defined above. Dean et al. obtain a bound on the number of edges for any graph in $\mathcal{F}_k$: If $n(G) \geq 2k+2$, then G has at most $(k+1)(3n - \frac{7}{2}k - 5) - 1$ edges. They give a construction showing that the complete graph on $4k+4$ vertices K_{4k+4} is in $\mathcal{F}_k$ and conjecture an improved edge bound of $(k+1)(3n-4k-6)$, which

[*] Received by the editors November 3, 2005; accepted for publication (in revised form) January 5, 2007; published electronically June 28, 2007. The work of the second author was partially supported by National Science Foundation VIGRE grant DMS-9983160. The work of the third author was partially supported by National Science Foundation VIGRE grant DMS-0083160.

http://www.siam.org/journals/sidma/21-2/64424.html

[†] Department of Mathematics, University of Illinois, Urbana, IL 61801 (hartke@math.uiuc.edu, jarobin1@math.uiuc.edu, pwenger2@math.uiuc.edu).

is attained by K_{4k+4}. They prove this conjecture for $k = 0, 1$ and use their edge bound to establish $K_{5k+5} \notin \mathcal{F}_k$. Other results include bounds on the thickness and the chromatic number of bar k-visibility graphs.

In this paper, we prove that the upper bound on the number of edges conjectured in [3] is correct, yielding K_{4k+4} as the largest complete bar k-visibility graph. We also prove that for each k, $\mathcal{F}_{k-1}$ and $\mathcal{F}_k$ are incomparable under set inclusion. We restrict the graphs that are in $\mathcal{F}_k$ for some k by proving that triangle-free graphs which are nonplanar are forbidden as induced subgraphs in bar k-visibility graphs. Finally, inspired by the result that the only regular interval graphs are complete graphs, we prove that if G is regular of degree $d < 2k+2$ and $G \in \mathcal{F}_k$, then G is a complete graph. However, we have constructions of $(2k+2)$-regular noncomplete graphs with bar k-visibility representations for $k \in \{0, 1, 2, 3, 4\}$.

Bar k-visibility graphs can be seen as a generalization of interval graphs. In particular, as k approaches infinity, it is easy to see that $\mathcal{F}_k$ approaches the family of interval graphs; hence results on interval graphs inspired many of our investigations into bar k-visibility graphs. A variation of interval graphs that has been studied is the idea of t-interval graphs, where each vertex of G is allotted t distinct intervals in its interval representation. This idea was extended to bar visibility graphs in [2], where each vertex of G is permitted t bars in its representation, and vertices are adjacent if there is a direct line of sight between any of their t bars. Another well-studied variation of bar visibility graphs is ϵ-visibility graphs, introduced by Melnikov [9]. These graphs are defined just as bar k-visibility graphs, except bars are replaced with arbitrary intervals which may not contain their endpoints. [12] and, independently, [11] gave a very simple characterization of ϵ-visibility graphs.

Throughout this paper, all graphs are simple graphs with no loops. A *bar* refers to a closed interval of the real line with an associated height. A *bar k-representation* of a graph G, $k \geq 0$, is a one-to-one correspondence between the vertices of G and a set of bars such that vertices of G are adjacent if and only if there is a vertical line segment joining their associated bars that intersects at most k other bars. (Note that in contrast to some other visibility models, a sight line of zero width is sufficient in this model.) When we are referring to a particular bar k-visibility representation of a graph G, $B(v)$ will refer to the bar associated with v, and $I(v)$ will refer to the projection of $B(v)$ onto the x-axis. We also use $u \leftrightarrow v$ to denote that u and v are adjacent and $u \not\leftrightarrow v$ to denote that u and v are nonadjacent.

2. An upper bound on the number of edges.

[3] gave an upper bound of $(k+1)(3n - \frac{7}{2}k - 5) - 1$ on the number of edges in an n-vertex bar k-visibility graph with $n \geq 2k+2$. They also conjectured an upper bound of $(k+1)(3n-4k-6)$, which would be sharp by a construction given in [3]. We prove their conjectured upper bound on the number of edges by refining their edge-counting technique.

THEOREM 1. *If G is a bar k-visibility graph with more than $2k+2$ vertices, G has at most $(k+1)(3n-4k-6)$ edges.*

Proof. Consider a bar k-visibility representation of G with vertices $v_1, v_2, \ldots, v_n$. We may assume that no two bars are at the same height, and hence we index the bars in order such that v_1 is the topmost bar and v_n is the bottommost bar. As noted in [3], we may also assume that the left and right endpoints of the associated intervals are distinct. If the endpoints are not distinct, then perturbing the endpoints slightly cannot decrease the number of edges in the resulting graph.

We sweep a vertical line from left to right over the representation, counting the number of edges that are created as we encounter sightlines. When the left endpoint of a bar $B(v)$ is encountered, the number of visibility-blocking bars is only increased. Hence the only new visibilities involve the new bar. If the left endpoint of $B(v)$ is the ith left endpoint encountered, then when $i \leq 2k+2$, $B(v)$ can see at most $i-1$ other bars. When $i > 2k+2$, $B(v)$ can possibly see $k+1$ bars above and $k+1$ bars below, for a total of $2k+2$ new edges. Thus the maximum number of edges counted by encountering left endpoints is

$$\begin{aligned}\sum_{i=1}^{2k+2}(i-1) + \sum_{i=2k+3}^{n}(2k+2) &= \frac{(2k+2)(2k+1)}{2} + (n-2k-2)(2k+2)\\ &= (k+1)(2n-2k-3).\end{aligned}$$

When the right endpoint of a bar $B(w)$ is encountered, the number of visibility-blocking bars between other bars is decreased and new visibilities may be created. If the right endpoint of $B(w)$ is the ith right endpoint encountered, then when $i \leq n-2k-2$, up to $k+1$ bars above $B(w)$ have the potential to each see one new bar below $B(w)$. Hence, at most $k+1$ new visibilities may be created. Once there are only $2k+2$ bars remaining, each time a bar ends the potential number of new edges decreases by one. Hence when $n-2k-2 < i < n-k-1$, there are at most $n-k-1-i$ new visibilities created. When $i \geq n-k-1$, no new sightlines are created since every bar has already seen every other remaining bar. Thus the maximum number of edges counted by encountering right endpoints is

$$\begin{aligned}\sum_{i=1}^{n-2k-2}(k+1) + \sum_{i=n-2k-1}^{n-k-2}(n-k-1-i) + \sum_{i=n-k-1}^{n} 0 &= (n-2k-2)(k+1) + \frac{k(k+1)}{2}\\ &= (k+1)\left(n - \frac{3}{2}k - 2\right).\end{aligned}$$

Hence, as was shown in [3], we have an upper bound of $(k+1)(3n-(7/2)k-5)$ for the number of edges in G.

Notice, however, that this bound is attained only if the top $k+1$ and the bottom $k+1$ bars are among the first $2k+2$ left endpoints (as we must begin a bar with at least $k+1$ bars above and $k+1$ bars below as soon as we are able to do so) and the last $2k+2$ right endpoints (as we must also end bars with $k+1$ bars above and $k+1$ bars below as long as we are able to do so). This implies, however, that we have twice counted the $(k+1)^2$ edges between these two sets of vertices, once when we encountered their left endpoints and once when bars between them ended. If, on the other hand, a bar from the top $k+1$ begins after at least $2k+2$ other bars have arrived, then we do not gain $2k+2$ new edges when it begins; when $B(v_i)$ begins, we gain at most $k+1+i-1$ new edges, $k+1$ from the bars below and $i-1$ from the bars above. Similarly, if a bar $B(v_i)$ for $i \leq k+1$ ends among the first $n-(2k-2)$, then there are not $k+1$ bars above it that can gain visibility. Instead, we gain at most $i-1$ new visibilities. The same holds for bars among the bottom $k+1$. We use these two facts to improve our edge bound.

Let ℓ be the number of bars of the top $k+1$ whose left endpoint is among the first $2k+2$ left endpoints and whose right endpoint is among the last $2k+2$ right endpoints, and m the similar number of bars in the bottom $k+1$. We observe first that each of the ℓm edges between these two sets of vertices is counted both when these bars begin (as a left-endpoint edge) and when visibility increases between them sufficiently as bars end (as a right-endpoint edge). For the remaining $k+1-\ell$ bars of the top $k+1$, they either begin among the last $n-(2k+2)$ or end among the first $n-(2k+2)$, or perhaps both. Either the left endpoint of these bars or the right endpoint of these bars, then, does not contribute the maximum possible number of edges. When $B(v_i)$ begins late or ends early for $i \leq k+1$, we overcount by $i-1$; thus we have overcounted the fewest number of edges when these $k+1-\ell$ bars are $B(v_{k+1}), B(v_k), \ldots, B(v_{k+1-\ell})$. In this case, our edge bound has overcounted at least

$$1+2+\cdots+(k+1-\ell)=\frac{1}{2}(k+2-\ell)(k+1-\ell)$$

edges. Similarly, we obtained at least an extra $\frac{1}{2}(k+2-m)(k+1-m)$ in our edge count by assuming the $k+1-m$ bars from the bottom $k+1$ yielded $2k+2$ edges when they began and $k+1$ edges when they ended. Therefore our graph has at most

$$\left(3n-\frac{7}{2}k-5\right)(k+1)-\left[\ell m+\frac{1}{2}(k+2-\ell)(k+1-\ell)+\frac{1}{2}(k+2-m)(k+1-m)\right]$$

edges; we seek to minimize the function

$$\begin{aligned} f(\ell,m) &= \ell m+\frac{1}{2}(k+2-\ell)(k+1-\ell)+\frac{1}{2}(k+2-m)(k+1-m) \\ &= \frac{1}{2}(\ell+m)^2-\frac{2k+3}{2}(\ell+m)+(k+1)(k+2). \end{aligned}$$

As this is a quadratic in $(\ell+m)$, we find that local extrema occur when $\ell+m=\frac{2k+3}{2}$, yielding a minimum objective value of $\frac{8k^2+24k+14}{16}$. Hence $f(\ell,m)\geq \frac{1}{2}k^2+\frac{3}{2}k+\frac{7}{8}$, and therefore our graph has at most

$$\left(3n-\frac{7}{2}k-5\right)(k+1)-\left(\frac{1}{2}k^2+\frac{3}{2}k+\frac{7}{8}\right)=3nk+3n-4k^2-10k-\frac{47}{8}$$

edges; since the number of edges must be integer-valued, we get an upper bound of

$$3nk+3n-4k^2-10k-6=(k+1)(3n-4k-6). \qquad \square$$

COROLLARY 2. *If K_n is a bar k-visibility graph, then $n \leq 4k+4$.*

Proof. If $n=4k+4+m$, then K_n has

$$(4k+4+m)(4k+4+m-1)\frac{1}{2}=8k^2+14k+4mk+6+\frac{7}{2}m+\frac{1}{2}m^2$$

edges. Theorem 1 gives an upper bound of

$$(k+1)(3(4k+4+m)-4k-6)=8k^2+14k+3mk+6+3m$$

edges in a k-visibility graph with $4k+4+m$ vertices. Hence

$$\begin{aligned} 8k^2+14k+4mk+6+\frac{7}{2}m+\frac{1}{2}m^2 &\leq 8k^2+14k+3mk+6+3m, \\ 4mk+\frac{7}{2}m+\frac{1}{2}m^2 &\leq 3mk+3m, \\ mk+\frac{1}{2}(m+m^2) &\leq 0, \end{aligned}$$

and therefore $m \leq 0$. Hence $n \leq 4k+4$. $\square$

[3] gave a construction achieving this edge bound for all $n \geq 4k+4$ (see Figure 1). Notice when $n=4k+4$, the construction is the complete graph K_{4k+4}. When $n<4k+4$, the k-visibility graph with the most edges is a complete graph on n vertices, obtained by leaving any $4k+4-n$ bars out of the K_{4k+4} representation.

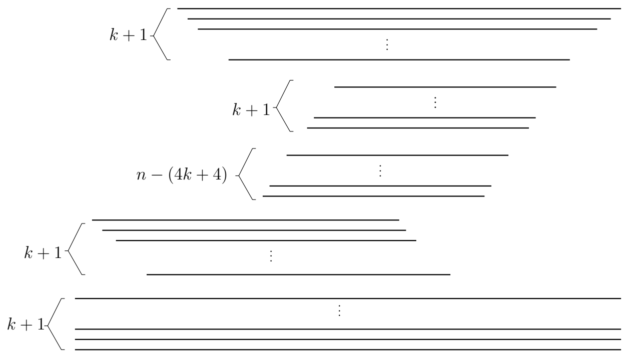


FIG. 1. *A bar k-visibility representation with $(k+1)(3n-4k-6)$ edges.*

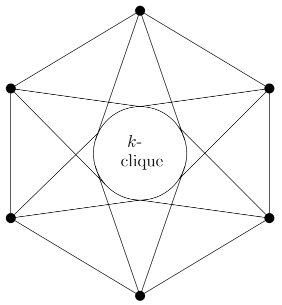


FIG. 2. *Example: W_6^k.*

3. Comparing the families $\mathcal{F}_{k-1}$ and $\mathcal{F}_k$. Corollary 2 shows that $K_{4k+4} \in \mathcal{F}_k$ but not $\mathcal{F}_{k-1}$. A natural question is whether $\mathcal{F}_{k-1}$ is contained in $\mathcal{F}_k$. In order to answer this question, we will need the following lemma.

LEMMA 3. *Suppose in some bar k-visibility representation of a graph G that $I(v) \cap I(w) \neq \emptyset$ but $v \not\leftrightarrow w$. Then for any vertical line ℓ intersecting $I(v) \cap I(w)$, if ℓ crosses $B(x)$, x is contained in a $(k+2)$-clique whose intervals also intersect ℓ.*

Proof. If $v \not\leftrightarrow w$, then there must be at least $k+1$ bars blocking $B(v)$ from $B(w)$. Any consecutive $k+2$ bars along ℓ, including $B(x)$, can all see each other, and hence their associated vertices must form a $(k+2)$-clique. □

Define a k-wheel W_n^k to be the graph obtained by joining every vertex of a k-clique with every vertex of an n-cycle (see Figure 2).

PROPOSITION 4. *For $n \geq 5$, W_n^k is not a bar k-visibility graph.*

Proof. Since W_n^k contains an induced long cycle C_n, it is not an interval graph. Therefore there must be two vertices v and w such that $I(v) \cap I(w) \neq \emptyset$ but $v \not\leftrightarrow w$. Let $I(v) \cap I(w) = [a, b]$. As the only $(k+2)$-clique containing v or w contains the middle k-clique, then by Lemma 3 any vertical line intersecting $[a, b]$ must intersect all k bars of the k-clique. The k-clique is not sufficient to obstruct $B(v)$'s view of $B(w)$; hence there must be another bar located between them. Let $B(x)$ be the first bar *not* associated with the middle k-clique that is intersected by a vertical line drawn from v to w. Note that x is one of v's two neighbors in C_n. Let $I(x) = [a', b']$. Let v' be v's other neighbor in C_n; note that $v' \neq w$ and $v' \not\leftrightarrow x$. Now, $n \geq 5$ implies that v' and x have no common neighbor on the cycle, so there can be no $(k+1)$-clique between $B(x)$ and $B(v')$. Therefore we must have $I(v') \cap I(x) = \emptyset$. Since v and v'

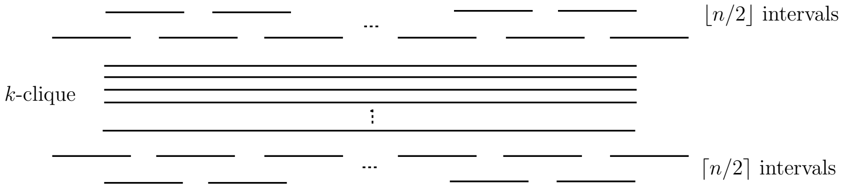

$\lfloor n/2 \rfloor$ intervals
...
k-clique
⋮
...
$\lceil n/2 \rceil$ intervals

FIG. 3. *A bar $(k-1)$-visibility representation of W_n^k.*

must be adjacent, $I(v) \cap I(v') \neq \emptyset$; assume by symmetry that $I(v) \cap I(v') = [c, d]$ for some $d < a'$. Since deleting v, x, and the k-clique from W_n^k leaves a connected graph, there must be some interval $I(z)$ that intersects $[d, a']$, where z is not v, x, or a vertex of the k-clique. The bar $B(z)$ closest to $B(v)$ with $I(z)$ intersecting $[d, a']$ must be visible to $B(v)$, since only the bars corresponding to the vertices of the k-clique could be located between $B(v)$ and $B(z)$. Since v has no other neighbors, this is a contradiction. □

Note that an easy construction shows that W_4^k is a bar k-visibility graph, so the result is sharp.

The results above combine to give the following.

THEOREM 5. *For all k, the families $\mathcal{F}_k$ and $\mathcal{F}_{k-1}$ are incomparable under inclusion.*

Proof. $\mathcal{F}_k \nsubseteq \mathcal{F}_{k-1}$ follows from Corollary 2. For the reverse inclusion, we observe that $W_n^k \in \mathcal{F}_{k-1}$; Figure 3 gives a bar $(k-1)$-visibility representation. By Proposition 4, $W_n^k \notin \mathcal{F}_k$, and hence $\mathcal{F}_{k-1} \nsubseteq \mathcal{F}_k$. □

4. Induced subgraphs.

We have already observed that as k increases, $\mathcal{F}_k$ approaches the family of interval graphs. It is known that all interval graphs are chordal graphs and interval graphs do not contain any induced subdivided complete bipartite graph $K_{1,3}$ [6]. We have shown already that W_n^k is a $(k-1)$-visibility graph and hence k-visibility graphs may contain induced long cycles. One may wonder whether an induced subdivision of $K_{1,3}$ prevents a graph from being in $\mathcal{F}_k$ for any k. The following proposition answers this question.

PROPOSITION 6. *For every tree T and every $k \geq 0$, there exists a graph G such that G contains T as an induced subgraph and G is a bar k-visibility graph.*

Proof. Choose a vertex r of T to be the root, and fix some integer d. We define the placement of $V(T)$'s bars in a k-visibility representation inductively. Assign the root the bar $B(r)$. Having assigned bars to all vertices at distance ℓ from the root, we place the bars for the vertices at distance $\ell + 1$ as follows: For a vertex v at level ℓ, find its children $v_1, \ldots, v_m$. Divide $B(v)$ into $2m-1$ closed segments, and assign v_i the $(2i-1)$st segment. Translate this segment down a distance of d to obtain $B(v_i)$.

Having assigned bars to all vertices of T in this way, if v and w are adjacent in T, then $I(v) \cap I(w) \neq \emptyset$. By placing a k-clique between each level, we ensure that only vertices at adjacent levels can see each other. The graph induced by this bar k-visibility representation is the desired graph G. □

Figure 4 gives an example of when T is an induced subdivided $K_{1,3}$.

We prove instead that certain nonplanar subgraphs are forbidden as induced subgraphs.

PROPOSITION 7. *Suppose that a graph G contains a triangle-free nonplanar induced subgraph. Then G is not a bar k-visibility graph for any k.*

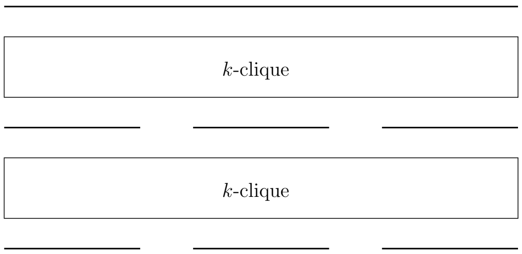


FIG. 4. *A bar k-visibility representation of a graph with an induced subdivided $K_{1,3}$.*

Proof. Suppose that G is a bar k-visibility graph for some k and has a triangle-free nonplanar induced subgraph H. Fix a bar k-visibility representation of G. Any two adjacent vertices of H are also adjacent in G, and thus their associated intervals intersect in the bar k-visibility representation of G. A pair of adjacent vertices u and v must exist in H such that any vertical line segment joining $B(u)$ and $B(v)$ intersects the bar of at least one other vertex w in H. Otherwise, if we restrict the bar k-visibility representation of G to the vertices in H, we would obtain a planar representation of H. However, by assumption, H is nonplanar. Thus, when $B(u)$ sees $B(v)$ in the bar k-visibility representation of G, the line of sight intersects $B(w)$. Therefore u, v, and w form a triangle in G which will also be in H. This contradicts the assumption that H is a triangle-free induced subgraph of G, and hence G cannot be a bar k-visibility graph for any k. ☐

5. Regular bar *k*-visibility graphs. It is easy to show that the only connected regular interval graphs are complete graphs. For small degrees, this fact remains true for bar k-visibility graphs.

PROPOSITION 8. *If G is a connected d-regular bar k-visibility graph with $d \leq 2k+1$, then G is a complete graph.*

Proof. Let v be a vertex whose bar begins last; that is, no vertex has a bar whose left endpoint is farther right than v's. Let $I(v) = [a, b]$, and let $v_1, v_2, \dots, v_m$ be the vertices whose intervals contain the point a, where the vertices are ordered from top to bottom by the height of their corresponding bars. Note that $v = v_i$ for some i.

All of v's neighbors are among $v_1, \dots, v_m$, so $\deg(v) \leq m-1$. Since $v_{\lfloor m/2 \rfloor}$ can see $k+1$ bars above it and $k+1$ bars below it if enough bars are present, $\deg(v_{\lfloor m/2 \rfloor}) \geq m-1$. Hence, $d = \deg(v) = \deg(v_{\lfloor m/2 \rfloor}) = m-1$.

If G is not a complete graph, there exists at least one vertex whose bar ends before v's bar begins. Among all such vertices, let c be the maximum value of a right endpoint of the associated intervals. Note that $c < a$. Let $z_1, z_2, \dots, z_p$ be all the vertices whose intervals' right endpoints are c. As G is connected, some v_i must be adjacent to some z_j. Among the bars $B(v_1), \dots, B(v_m)$ seeing some $B(z_j)$ at the point c, choose i to minimize $|\lfloor m/2 \rfloor - i|$.

We claim that $\deg(v_i) \geq m$. We know that z_j is adjacent to v_i; suppose some v_ℓ is not in v_i's neighborhood, $i \neq \ell$. But then $c \in I(v_\ell)$, since any bar that begins after c must see $B(v_i)$ in order to have degree $m-1$. Consider the point $c+\epsilon$, where ϵ is chosen to be small enough such that no interval begins in $[c, c+\epsilon]$. As v_i was chosen to be the "most central" bar extending left to the point c, there cannot be k intervals containing the point $c+\epsilon$ blocking $B(v_i)$ from $B(v_\ell)$. Therefore $v_i \leftrightarrow v_j$, and hence $\deg(v_i) \geq m$, contradicting the assumption that G is regular. ☐

When $d = 2k+2$ and $k \in \{0, 1, 2, 3, 4\}$, there exist d-regular noncliques in $\mathcal{F}_k$.

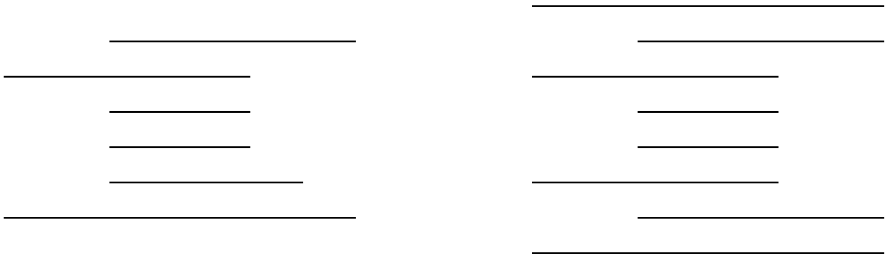


FIG. 5. *On the left is a bar* 1*-visibility representation of the* 4*-regular graph formed by removing a perfect matching from* K_6*. On the right is a bar* 2*-visibility representation of the* 6*-regular graph formed by removing a perfect matching from* K_8*.*

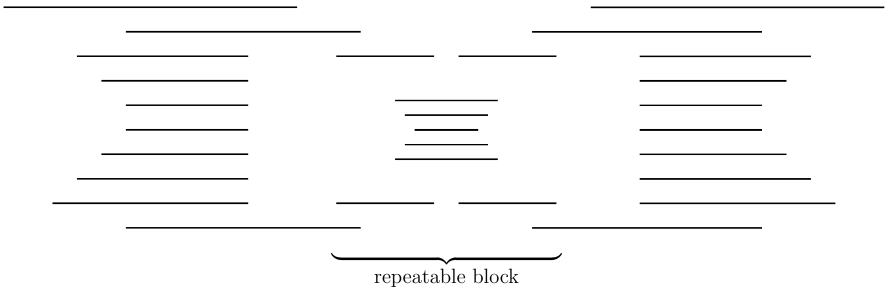


FIG. 6. *A bar* 3*-visibility representation of an* 8*-regular graph. The* 9 *bars in the repeatable block can be repeated horizontally as many times as desired, or omitted entirely. Consecutive blocks may need to be perturbed vertically a small amount so that the top and bottom bars can see the top and bottom bars from the next block, but are still disjoint from those bars.*

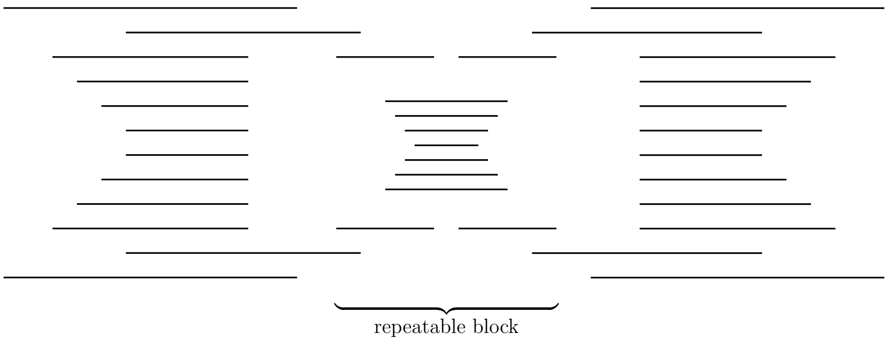


FIG. 7. *A bar* 4*-visibility representation of a* 10*-regular graph. The* 11 *bars in the repeatable block can be repeated horizontally as many times as desired, or omitted entirely.*

For $k = 0$, every cycle C_n with $n \geq 4$ is a 2-regular bar 0-visibility graph. Figure 5 shows noncliques that are $2k+2$ regular when $k = 1$ and $k = 2$. Figures 6 and 7 show constructions for an infinite number of regular graphs of degree $2k + 2$ when $k = 3$ and $k = 4$, respectively.

The question remains open for larger values of d and k.

6. Conclusion. There are many open questions remaining about bar k-visibility graphs, with the primary goal being a complete characterization of bar k-visibility graphs. There are also several other interesting questions that may serve as intermediate steps toward this goal.

1. Are there forbidden induced subgraphs for bar k-visibility graphs besides triangle-free nonplanar graphs?
2. Does every graph that is not a bar k-visibility graph for any k contain an induced triangle-free nonplanar subgraph?
3. Are there $(2k+2)$-regular bar k-visibility graphs for $k \geq 5$?
4. Are there d-regular bar k-visibility graphs with $d \geq 2k+3$?

Dean et al. [3] also present several open questions regarding the chromatic number, genus, and thickness of bar k-visibility graphs. [5] further investigates the thickness of bar 1-visibility graphs.

It is worth noting that while we now have a sharp edge bound, it does not improve the upper bound of $6k+6$ in [3] for the chromatic number of bar k-visibility graphs. We feel that this bound can be lowered, possibly through a deeper exploration of the structural aspects of bar k-visibility graphs and their connection to minimum degree and degeneracy.

Acknowledgments. The authors thank Douglas B. West for a helpful suggestion that simplified the statement of Proposition 7, and Mareike Massow for comments that improved the readability of the paper.

REFERENCES

[1] T. Andreae, *Some results on visibility graphs*, Discrete Appl. Math., 40 (1992), pp. 5–17.

[2] Y.-W. Chang, J. P. Hutchinson, M. S. Jacobson, J. Lehel, and D. B. West, *The bar visibility number of a graph*, SIAM J. Discrete Math., 18 (2004), pp. 462–471.

[3] A. Dean, W. Evans, E. Gethner, J. Laison, M. A. Safari, and W. Trotter, *Bar k-visibility graphs*, J. Graph Algorithms Appl., 11 (2007), pp. 45–59.

[4] A. M. Dean, W. Evans, E. Gethner, J. D. Laison, M. A. Safari, and W. Trotter, *Bar k-visibility graphs: Bounds on the number of edges, chromatic number, and thickness*, in Graph Drawing, Lecture Notes in Comput. Sci. 3843, Springer, Berlin, 2006, pp. 73–82.

[5] S. Felsner and M. Massow, *Thickness of Bar 1-Visibility Graphs*, preprint.

[6] C. G. Lekkerkerker and J. C. Boland, *Representation of a finite graph by a set of intervals on the real line*, Fund. Math., 51 (1962/1963), pp. 45–64.

[7] F. Luccio, S. Mazzone, and C. K. Wong, *Visibility Graphs*, University of Pisa, Prog. Naz. Teoria degli Algoritmi, Report 9, Pisa, Italy, 1983.

[8] F. Luccio, S. Mazzone, and C. K. Wong, *A note on visibility graphs*, Discrete Math., 64 (1987), pp. 209–219.

[9] L. S. Melnikov, *Problem at the 6th Hungarian Colloquium on Combinatorics*, 1981.

[10] M. Schlag, F. Luccio, P. Maestrini, D. T. Lee, and C. K. Wong, *A visibility problem in VLSI layout compaction*, in Advances in Computing Research, Vol. 2, JAI Press, Greenwich, CT, 1985, pp. 259–282.

[11] R. Tamassia and I. G. Tollis, *A unified approach to visibility representations of planar graphs*, Discrete Comput. Geom., 1 (1986), pp. 321–341.

[12] S. K. Wismath, *Characterizing bar line-of-sight graphs*, in Proceedings of the ACM Symposium on Computational Geometry, Baltimore, MD, 1985, pp. 147–152.

SIAM J. DISCRETE MATH.
Vol. 21, No. 2, pp. 532–547

OPTIMAL TREE STRUCTURES FOR GROUP KEY MANAGEMENT WITH BATCH UPDATES[*]

RONALD L. GRAHAM[†], MINMING LI[‡], AND FRANCES F. YAO[‡]

Abstract. We investigate the key management problem for broadcasting applications. Previous work showed that batch rekeying can be more cost-effective than individual rekeying. Under the assumption that every user has probability p of being replaced by a new user during a batch rekeying period, we study the structure of the optimal key trees. Constant bounds on both the branching degree and the subtree size at any internal node are established for the optimal tree. These limits are then utilized to give an $O(n)$ dynamic programming algorithm for constructing the optimal tree for n users and any fixed value of p. In particular, we show that when $p > 1 - 3^{-1/3} \approx 0.307$, the optimal tree is an n-star, and when $p \leq 1 - 3^{-1/3}$, each nonroot internal node has a branching degree of at most 4. We also study the case when p tends to 0 and show that the optimal tree resembles a balanced ternary tree to varying degrees depending on certain number-theoretical properties of n.

Key words. key trees, group keys, batch updates, optimality

AMS subject classifications. 05C05, 49K99

DOI. 10.1137/06064929X

1. Introduction.

In the group broadcast problem, we have n subscribers and a group controller (GC) that periodically broadcasts messages (e.g., a video clip) to all the subscribers over an insecure channel. To guarantee that only the authorized users can decode the contents of the messages, the GC will dynamically update the group key for the whole group. Whenever some user leaves or joins, the GC will generate a new group key and in some way notify the remaining users in the group. A recent survey of the key management problem for groups of low-state devices can be found in [1]. Here, we consider the key tree model [4] for the key management problem. We describe this model briefly as follows (precise formulation is given in section 2). Every leaf node of the key tree represents a user and stores his individual key. Every internal node stores a key shared by all leaf descendants of the internal node. Every user possesses all the keys along the path from the leaf node (representing the user) to the root node. To prevent revoked users from knowing future message contents and also to prevent new users from knowing past message contents, the GC updates a set of keys, whenever a new user joins or a current user leaves, as follows. So long as there is a user change among the leaf descendants of an internal node v, the GC will: (1) replace the old key stored at v with a new key, and (2) broadcast (to all users) the new key encrypted with the key stored at each child node of v. Note that only users represented by leaf descendants of v can get useful information from the broadcast. Furthermore, this procedure must be done in a bottom-up fashion (i.e., starting with

[*] Received by the editors January 6, 2006; accepted for publication (in revised form) January 8, 2007; published electronically June 28, 2007. The work described in this paper was supported in part by the National Natural Science Foundation of China (60135010, 60321002, 60223004), the Chinese National Key Foundation Research & Development Plan (2004CB318108), the U.S. National Science Foundation grant CCR-0310991, and a grant from the Research Grants Council of Hong Kong under project CityU 1165/04E.

http://www.siam.org/journals/sidma/21-2/64929.html

[†] Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA 92037 (graham@ucsd.edu).

[‡] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, People's Republic of China (minmli@cs.cityu.edu.hk, csfyao@cityu.edu.hk).

the lowest v whose key must be updated—see section 2 for details) to guarantee that a revoked user will not know the new keys. The cost of the above procedure counts the number of encryptions used in step 2 (or equivalently, the number of broadcasts made by the GC).

When users change frequently, this method for updating the group keys whenever a user leaves or joins may be too costly. Thus, a batch rekeying strategy was proposed by Li et al. in [2] whereby rekeying was done only periodically instead of immediately after each member change. They designed a marking algorithm for processing the batch updates. It was shown by simulation that, using their algorithm, among the totally balanced key trees (where internal nodes of the tree have branching degree 2^i), a degree of 4 is the best when the number of requests (leave/join) within a batch is not large. For a large number of requests, using a star (a tree of depth 1) to organize the users outperforms all the balanced key trees mentioned above in their simulation. Further analysis of the batch rekeying model was done by Zhu, Chan, and Noubir in [5]. They introduced a new model to investigate the special case when the number of joins always equals the number of leaves during a batch period. Thus, they assumed that during a certain period, every user has a probability p of being replaced by a new user. They then studied the optimal tree under two assumptions: (A) the tree is totally balanced, and (B) every node on level i has 2^{k_i} children. Under these assumptions, characterizations of the optimal key tree were given together with an algorithm for computing it. In particular, it was shown that every node on an intermediate level of the tree should have a degree of 4 (that is, $i = 2$).

In this paper, we carry out the first theoretical investigation for the optimal tree as modelled by [5] but without assumptions (A) and (B). In the following discussion, an "optimal tree" always refers to a true minimum-cost key tree without any a priori restrictions on its structure; i.e., we allow the degree of each node in the tree to be arbitrary and independent of each other. Denote such a tree by a (p, n)-*optimal tree* where n is the number of users (i.e., number of leaves of the tree), and p is the probability that a leaf gets updated. The main results of the paper are as follows. We identify for p a range $1 \geq p \geq 1 - 3^{-1/3} \approx 0.307$, where a star (a tree of depth 1) is the (p, n)-optimal tree for all n. We also prove, for all (p, n)-optimal trees, a constant upper bound 4 to the branching degree of any node v other than the root, and an upper bound (as a function of p) to the size of the subtree rooted at v. These characterizations enable us to design a dynamic programming algorithm to compute the optimal tree in time $O(n)$. We further study the case when $p \to 0$ and show that the optimal tree is as close to a balanced ternary tree with n leaves as possible, subject to number-theoretical properties of n (see Theorem 4 for a precise statement).

The rest of the paper is organized as follows. In section 2, we describe the batch update model in detail. In section 3, we find the range of p when the (p, n)-optimal tree is a star. We derive properties of the general optimal trees in section 4 and give a linear-time dynamic programming algorithm for constructing them in section 5. Finally, we carry out the analysis for the case $p \to 0$ in section 6 and show that, in most cases, a branching degree of 3 is employed by the optimal tree.

2. Preliminaries. Before giving a precise formulation of the tree optimization problem to be considered, we briefly discuss its motivation and review the basic key tree model for group key management. This model is referred to in the literature either as key tree [4] or LKH (logical key hierarchy) [3].

In the key tree model, there is a GC, represented by the root, and n subscribers (or users) represented by the n leaves of the tree. The tree structure is used by the

GC for key management purposes. Associated with every node of the tree (whether internal node or leaf) is an encryption key. The key associated with the root is called the traffic encryption key (TEK), which is used for accessing encrypted service contents by the subscribers. The key k_v associated with each nonroot node v is called a key encryption key (KEK) which is used for updating the TEK when necessary. Each subscriber possesses all the keys along the path from the leaf representing the subscriber to the root.

In the batch update model to be considered, only simultaneous join/leave is allowed, that is, whenever there is a revoked user, a new user will be assigned to that vacant position. This assumption is justified since, in a steady state, the number of joins and departures would be roughly equal during a batch processing period. To guarantee forward and backward security, a new user assigned to a position (leaf) will be given a new key by the GC and, furthermore, all the keys associated with the ancestors of the leaf must be updated by the GC. The updates are performed from the lowest ancestor upward for security reasons. We will explain the updating procedure together with the updating cost in what follows.

The GC first communicates with each new subscriber separately to assign a new key to the corresponding leaf. After that, the GC will broadcast certain encrypted messages to all subscribers in such a way so that each valid subscriber will know all the new keys associated with its leaf-to-root path while the revoked subscribers will not know any of the new keys. The GC accomplishes this task by broadcasting the new keys, in encrypted form, from the lowest level upward recursively as follows. Let v be an internal node at the lowest level whose key needs to be (but has not yet been) updated. For each child u of v, the GC broadcasts a message containing $E_{k_u^{new}}(k_v^{new})$, which means the encryption of k_v^{new} with the key k_u^{new}. Thus the GC sends out d_v broadcast messages for updating k_v if v has d_v children. Updating this way ensures that the revoked subscribers will not know any information about the new keys while current subscribers can use one of their KEKs to decrypt the useful $E_{k_u^{new}}(k_v^{new})$ sequentially until they get the new TEK.

We adopt the probabilistic model introduced in [5] that each of the n positions has the same probability p to independently experience subscriber change during a batch rekeying period. Under this model, an internal node v with N_v leaf descendants will have a probability of $1 - q^{N_v}$ for which its associated key k_v requires updating, where $q = 1 - p$. The updating incurs $d_v \cdot (1 - q^{N_v})$ expected broadcast messages by the procedure described above. We thus define the expected updating cost $C(T)$ of a key tree T by $C(T) = \sum_v d_v \cdot (1 - q^{N_v})$, where the sum is taken over all the internal nodes v of T. It is more convenient to remove the factor d_v from the formula by associating the cost $1 - q^{N_v}$ with each of v's children. This way we express $C(T)$ as a node weight summation: for each nonroot tree node u, its node weight is defined to be $1 - q^{N_v}$, where v is u's parent. The optimization problem we are interested in can now be formulated as follows.

Optimal key tree for batch updates. We are given two parameters $0 \leq p \leq 1$ and $n > 0$. Let $q = 1-p$. For a rooted tree T with n leaves and node set V (including internal nodes and leaves), define a weight function $c(u)$ on V as follows. Let $c(r) = 0$ for root r. For every nonroot node u, let $c(u) = 1 - q^{N_v}$, where v is u's parent. Define the cost of T as $C(T) = \sum_{u \in V} c(u)$. Find a T for which $C(T)$ is minimized. We say that such a tree is (p, n)-*optimal* and denote its cost by $OPT(p, n)$.

We will first study the case when p is any fixed constant and later the case for $p \to 0$. Notice that $C(T) \geq \sum_v d_v(1 - q) \geq pn$ implies $OPT(p, n) \geq pn$. On the other hand, we have $OPT(p,n) \leq (1-q^n)n$ by considering a tree where all leaves are attached directly to the root (i.e., a star). Thus we know asymptotically $OPT(p,n) = \Omega(n)$ when p is a constant. However, it is still interesting to identify the *exact* optimal tree which can achieve significantly better cost than a star, especially when p is a small constant.

3. The star optimal bound.

We start with some basic definitions about rooted trees. We say a tree is of *depth k* if the longest leaf-root path consists of k edges. A tree of depth 2 is also referred to as a *two-level tree.* A tree of depth 1 is called a *k-star* if it has k leaves. A tree edge (u,v), where u is a child of v, is said to be at *depth k* if the path from u to the root consists of k edges. The *branching degree* of a node v is the number of children of v; the *subtree size* of v, denoted by N_v, refers to the number of leaf descendants of v.

LEMMA 1. *If the n-*star *can achieve $OPT(p,n)$, then the $(n-1)$-*star *can also achieve $OPT(p,n-1)$.*

Proof. We prove the lemma by contradiction. Suppose the $(n-1)$-star cannot achieve $OPT(p,n-1)$. Let the degree of the root in a $(p,n-1)$-optimal tree be k, where $k < n-1$. Write the optimal cost as $OPT(p,n-1) = k(1-q^{n-1}) + C$, where C represents the contribution to the cost by edges at depth ≥ 2. Thus we have $(n-1)(1-q^{n-1}) > k(1-q^{n-1}) + C$, which implies $n(1-q^n) > (k+1)(1-q^n) + C$. This means we can reduce the cost of $OPT(p,n)$ by adopting the same structure of the $(p,n-1)$-optimal tree but with root degree $k+1$, a contradiction. □

LEMMA 2. *When $0 \leq q \leq 3^{-1/3}$, the n-*star *is strictly better than any two-level tree.*

Proof. A two-level tree can be obtained from a star by successively grouping certain nodes together to form a subtree of the root. To prove the lemma, we need only show that the above operation always increases the cost of the tree; i.e., for any grouping size k, where $1 < k < n$, we will show that $1 - q^n < \frac{1}{k}(1-q^n) + 1 - q^k$. This is trivially true when $k = 1$ or $q = 0$, so we assume that $k \geq 2$ and $q > 0$.

With fixed $q > 0$, define $f(k)$ for integer k by $f(k) = k \log_k(1/q)$. We observe that for any fixed q, where $0 < q < 1$, the value of $f(k)$ is minimized when $k = 3$. Thus, when $0 < q \leq 3^{-1/3}$, we have $k \log_k(1/q) \geq 1$ which implies $kq^k \leq 1$. Hence, for $0 < q \leq 3^{-1/3}$, we have

$$\begin{aligned} 1 - q^n - \left(\frac{1}{k}(1-q^n) + 1 - q^k\right) &= \frac{1}{k}(kq^k - 1 - (k-1)q^n) \\ &< \frac{1}{k}(kq^k - 1) \\ &\leq 0. \qquad □ \end{aligned}$$

LEMMA 3. *Let p and n be given. If the n-*star *is strictly better than any two-level tree, then the n-*star *is the (p,n)-optimal tree.*

Proof. If the n-star is not a (p,n)-optimal tree, then we can transform the (p,n)-optimal tree from the bottom up, every time combining two levels into a star. By Lemma 1, the m-star is strictly better than any two-level tree for $1 < m < n$ and p. Since one level is always better than two levels, we can eventually transform the optimal tree into the n-star without increasing the cost. □

THEOREM 1. *When $1 \geq p \geq 1 - 3^{-1/3}$, the n-*star *is the (p,n)-optimal tree for any n. For $2 \leq n \leq 4$, the n-*star *is the (p,n)-optimal tree for any $p > 0$.*

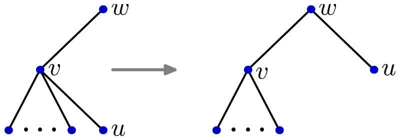


FIG. 1. *Tree transformation* 1.

Proof. The first part of the theorem follows from Lemmas 2 and 3. The cases of $2 \le n \le 4$ can be verified easily. □

4. Properties of an optimal tree. By Theorem 1, the structure of a (p, n)-optimal tree is uniquely determined for $0 \le q \le 3^{-1/3} \approx 0.693$ (or $1 \ge p \ge 1-3^{-1/3} \approx 0.307$). We now derive some properties of the optimal trees which will be used for constructing a (p, n)-optimal tree in the remaining range $1 \ge q > 3^{-1/3}$. Note that Lemmas 4 and 5 as well as Theorem 2 are true for all (p, n)-optimal trees where $0 \le p \le 1$ and $n > 0$.

For a tree T, we associate a value $t_v = q^{N_v}$ with every node v (thus $t_v = q$ if v is a leaf). The subtree rooted at u is denoted by T_u. We say T_u is a subtree of v if u is a child of v.

LEMMA 4. *For a nonroot internal node v with a branching degree of k in a (p, n)-optimal tree, every child u of v satisfies $t_u \ge \frac{k-1}{k}$.*

Proof. Assume $t_u < \frac{k-1}{k}$, we can then move u up to become a sibling of v, as shown in Figure 1. In this way, we increase the total cost of the tree by

$$\begin{aligned}\Delta C &= (1-q^{N_w}) + (k-1)(1-q^{N_v-N_u}) - k(1-q^{N_v})\\ &< 1 + (k-1)(1-q^{N_v-N_u}) - k(1-q^{N_v-N_u}t_u)\\ &= q^{N_v-N_u}(kt_u - (k-1))\\ &< 0,\end{aligned}$$

where w is the parent of v and N_v represents the value before the transformation. This contradicts the cost optimality of the original tree. □

LEMMA 5. *Every nonroot internal node in a (p, n)-optimal tree has a branching degree of* ≤ 5.

Proof. By Lemma 4, if a nonroot internal node v in the optimal tree has a branching degree of $k \ge 6$, then for any child u of v we have $t_u \ge \frac{k-1}{k}$. We can group together two children of v, with the largest and the second largest t_u values, to form a single subtree of v as shown in Figure 2. By this transformation, we increase the total cost by

$$\begin{aligned}\Delta C &= 2(1-t_{u_1}t_{u_2}) - (1-t_v)\\ &= t_v - 2t_{u_1}t_{u_2} + 1.\end{aligned}$$

Note that t_v is the product of t_u over all children u of v. Thus, we have $t_v < (t_{u_1}t_{u_2})^{k/2}$, which implies $\Delta C < z^k - 2z^2 + 1$, where $\frac{k-1}{k} \le z \le 1$.

It is easy to verify that $z^6 - 2z^2 + 1 < 0$ for $5/6 \le z \le 1$. Because the value of $z^k - 2z^2 + 1$ decreases with k for fixed z, we see that $\Delta C < 0$ for any $k \ge 6$ and $\frac{k-1}{k} \le z \le 1$, which proves the lemma. □

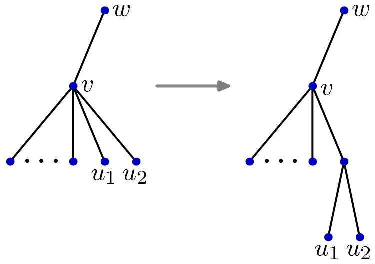


FIG. 2. *Tree transformation* 2.

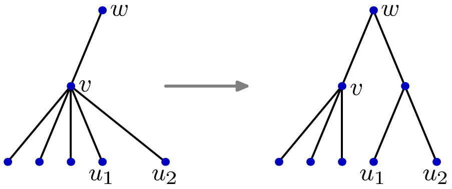


FIG. 3. *Tree transformation* 3.

THEOREM 2. *In a (p, n)-optimal tree,*

(1) *any internal node other than the root must have a branching degree of ≤ 4;*

(2) *the size of any subtree T_v, where v is a child of the root, is upper bounded by* $\max\{4(\log q^{-1})^{-1}, 1\}$.

Proof. By using Lemma 5, we can complete the proof of (1) by showing that the optimal tree does not have any internal node with a branching degree of 5.

Assume there is a nonroot internal node v with five children $u_1, \ldots, u_5$. For simplicity, we write t_{u_i} as t_i and assume $t_1 \geq \cdots \geq t_5$. First, observe that $z^5-2z^2+1 < 0$ when $z \geq 0.86$. According to Lemma 4 and the proof of Lemma 5, it must be the case that both conditions $t_1t_2 < (0.86)^2 = 0.7396$ and $0.8 \leq t_i < 0.86$ for $2 \leq i \leq 5$ hold. We now prove that under these conditions, another tree transformation will reduce the total cost which contradicts the tree's optimality. We transform the optimal tree into tree T' as shown in Figure 3. By doing so, we increase the total cost by

$$\begin{aligned}\Delta C &= 3(1 - t_3t_4t_5) + 2(1 - t_1t_2) + (1 - t_w) - 5(1 - t_v)\\ &< -2t_1t_2 - 3t_3t_4t_5 + 5t_v + 1\\ &= (5t_3t_4t_5 - 2)t_1t_2 - (3t_3t_4t_5 - 1),\end{aligned}$$

where t_v represents the value before transformation. By using the fact that $t_i \geq 0.8$ for $1 \leq i \leq 5$ and $t_1t_2 < 0.7396$, it can be verified that $\Delta C < 0$. This completes the proof of property (1).

For property (2), as we have shown in Lemma 4, any child u of v satisfies $q(u) = q^{N_u} > 1/2$. Thus, $N_u < (\log q^{-1})^{-1}$. Since v has a branching degree of at most 4 by property (1) and also $N_v \geq 1$, we have $N_v \leq \max\{4(\log q^{-1})^{-1}, 1\}$. This completes the proof of the theorem. □

5. Algorithm for constructing the optimal tree. We will construct a (p, n)-optimal tree by assembling a forest of suitable subtrees. We first generalize the cost function from trees to forests as follows.

Definition 1. *For $L \leq n$, we define a (p,n,L)-forest to be a forest of key trees with L leaves in total. The cost of the tree edges in the forest are defined as before, while the cost of the forest is the sum of individual tree costs plus $k \cdot (1-q^n)$, where k is the number of trees in the forest. We refer to the (p,n,L)-forest with minimum cost as the optimal (p,n,L)-forest and denote that minimum cost by $F(p,n,L)$.*

Theorem 3. *For any fixed p, Algorithm 1 computes the (p,n)-optimal tree cost in $O(n)$ time.*

Proof. Based on Theorem 2, in a (p,n)-optimal tree, any subtree T_v, where v is a child of the root, satisfies (1) its size is at most $\max\{4(\log q^{-1})^{-1}, 1\}$ and (2) the branching degree of any internal node in T_v is at most 4. For fixed q, we view $4(\log q^{-1})^{-1}$ as a constant and denote it by K. For each i, where $2 \leq i \leq K$, we consider the minimum cost $R(i)$ of any tree T_v with size i and subject to the degree restriction stated in (2). The value of $R(i)$ can be computed in constant time as follows. First, define (k_1,k_2,k_3,k_4) to be an *i-quadruple* if $k_1+k_2+k_3+k_4 = i$, $0 \leq k_1,k_2,k_3,k_4 \leq i$, and $k^* \geq 2$, where k^* is the number of positive elements in (k_1,k_2,k_3,k_4). Then $R(i)$ is the minimum value of $R(k_1)+R(k_2)+R(k_3)+R(k_4)+(1-q^i)\cdot k^*$ over all i-quadruples (k_1,k_2,k_3,k_4), and it can be computed in $O(K^3)$ time using dynamic programming. Now, we can obtain the true (p,n)-optimal tree also by dynamic programming, by computing optimal (p,n,L)-forests as subproblems of size L for $1 \leq L \leq n$ as given in Algorithm 1. Thus, the total running time of the algorithm is $O(n \cdot K + K^4)$ which is $O(n)$ for fixed p. Algorithm 1 focuses on computing the optimal tree cost; the tree structure can be obtained by keeping track of the optimal branching at every dynamic programming iteration. ∎

Algorithm 1 *Computing optimal key tree*

```
Input: $n$ and $p$ $(0 \leq p \leq 1)$
Output: Optimal tree cost $OPT(n,p)$
  $q = 1-p$
  $K = \min\{4(\log q^{-1})^{-1}, n\}$
  if $q \leq 3^{-1/3}$ or $2 \leq n \leq 4$ then
    $OPT(p,n) \leftarrow n \cdot (1-q^n)$
    Return $OPT(p,n)$
  end if
  $R(1) = 0$
  for $i = 2$ to 4 do
    $R(i) = i * (1-q^i)$
  end for
  $i = 5$
  while $i < K$ do
    Compute $R(i)$, cost of the restricted $(p,i)$-optimal tree.
    $i = i+1$
  end while
  $F(p,n,0) \leftarrow 0$
  for $L = 1$ to $n$ do
    $F(p,n,L) \leftarrow \min(R(j)+1-q^n+F(p,n,L-j))$ over all $j$, $1 \leq j \leq \min\{K,L\}$.
  end for
  $OPT(n,p) \leftarrow F(p,n,n)$
  Return $OPT(n,p)$
```

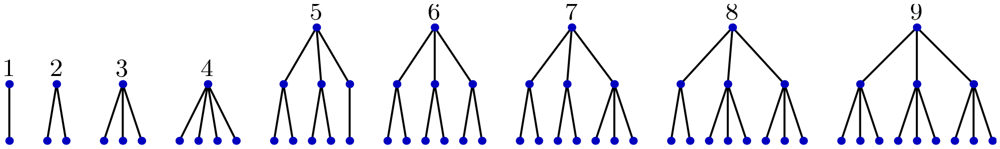


Fig. 4. *Optimal small trees.*

6. Optimal trees as $p \to 0$. Algorithm 1 has running time $O(n \cdot K + K^4)$, where K is upper bounded by $4(\log q^{-1})^{-1}$ (and also by n). We can regard $O(K^4)$ as a constant term for a fixed value of p, but its values get large as $p \to 0$. Therefore, in this final section we will study the structure of (p, n)-optimal trees as $p \to 0$. It turns out the structure of (p, n)-optimal trees depends rather critically on certain number-theoretic properties of n. (Some of the detailed computations will be suppressed.) Suppose $T = T(n)$ denotes a rooted tree with n leaves and edge set E.

For convenience, we let $L(e) = N_v$ if $e = (u, v)$ and u is a child of v. We express the cost of T as

$$P_T(p) = C(T) = \sum_{e \in E} (1 - (1-p)^{L(e)}).$$

Of course, $P_T(0) = 0$ and $P_T(1) = |E|$. The optimal trees as $p \to 0$ are those with $P_T(p)$ having the smallest slope at $p = 0$. Any such optimal tree will remain optimal for an interval $[0, c]$ for some (small) $c > 0$. The slope of $P_T(p)$ at $p = 0$ is denoted by λ_T and can be expressed as

$$\lambda_T = \sum_{e \in E} L(e).$$

We let $\lambda^*(n)$ denote the smallest possible value of λ_T over all trees $T = T(n)$ having n leaves. Our first task will be to determine the exact value of $\lambda^*(n)$ for all values of n.

To begin with, it is easy to check by hand that the trees shown in Figure 4 are optimal for the values $1 \le n \le 9$. This implies the corresponding values of $\lambda^*(n)$ shown below:

n	1	2	3	4	5	6	7	8	9
$\lambda^*(n)$	0	4	9	16	23	30	38	44	54

For integers $t \ge 0$, define the sets $I_t = \{3^t, 3^t + 1, \ldots, 2 \cdot 3^t\}$ and $J_t = \{2 \cdot 3^t, 2 \cdot 3^t + 1, \ldots, 3^{t+1}\}$. Notice that $\lambda^*(n)$ is *linear* on I_0, I_1, J_0, and J_1. We will show that this holds in general for all I_t and J_t.

First, let us extend $\lambda^*(n)$ to a real function $\lambda^*(x)$ for all $x \ge 1$ by linear interpolation (see Figure 5).

The basic recurrence that $\lambda^*(n)$ satisfies is

$$\lambda^*(n) = \min_{2 \le r \le n} \left\{ rn + \sum \lambda^*(i_k) \right\}, \tag{6.1}$$

where the sum is taken over all $i_k \ge 1$ such that $i_1 + \cdots + i_r = n$ and r denotes the degree of the root ρ.
What we will prove is in the following lemma.

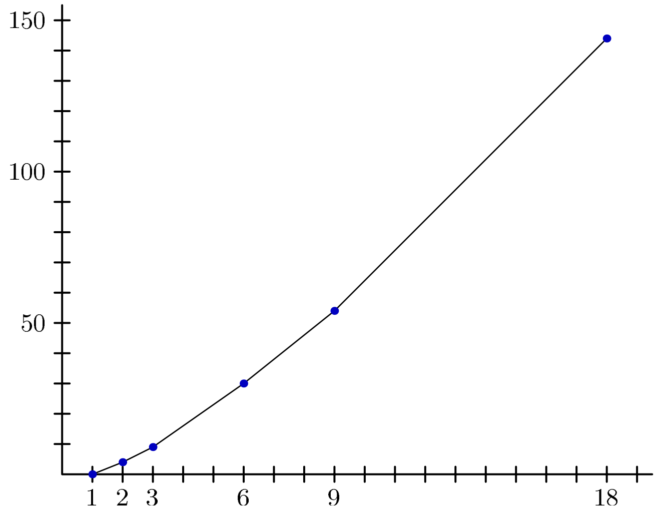


FIG. 5. *Graph of* $\lambda^*(n)$.

LEMMA 6.

$$\lambda^*(x) = \begin{cases} (3t+4)x - 4\cdot 3^t & \text{if} \quad x \in I_t, \\ (3t+5)x - 6\cdot 3^t & \text{if} \quad x \in J_t. \end{cases} \tag{6.2}$$

Proof. It is easy to check that this holds for $t = 0$ and 1. Assume that this holds for some value $t \geq 1$. We also assume that $\lambda^*(x)$ is linear on I_s and J_s, $s \leq t$, and strictly convex between intervals, (i.e., the slopes are strictly increasing on successive intervals). Thus, for each fixed value of r, the sum $\sum \lambda^*(i_k)$ is minimized by taking all the i_k to be as equal as possible. In fact, we can take all $i_k = \frac{n}{r}$, by the assumption of linearity of $\lambda^*(x)$ on the I_s and J_s, $s \leq t$. (Here we need the elementary fact that if $\frac{n}{r} \in [a, a+1]$ for some integer a, then n can be expressed as the sum of u a's and $r-u$ $a+1$'s, for some u with $0 \leq u \leq r$.) Thus, we can write

$$\lambda^*(n) = \min_{2\leq r\leq n} \left\{ rn + r\lambda^*\left(\frac{n}{r}\right) \right\}.$$

Our next job is to eliminate values of r as candidates for achieving the minimum. For example, let us show that

$$5n + 5\lambda^*\left(\frac{n}{5}\right) > 4n + 4\lambda^*\left(\frac{n}{4}\right).$$

Take $n \in I_t$ and consider $\frac{n}{4}$ and $\frac{n}{5}$. Since their ratio is $\frac{5}{4}$, then there are only four possibilities (for some $s < t$):

(i) $\frac{n}{5}, \frac{n}{4} \in \bar{I}_s$;
(ii) $\frac{n}{5}, \frac{n}{4} \in \bar{J}_s$;
(iii) $\frac{n}{5} \in \bar{I}_s$, $\frac{n}{4} \in \bar{J}_s$;
(iv) $\frac{n}{5} \in \bar{J}_s$, $\frac{n}{4} \in \bar{I}_{s+1}$,

where we use $\bar{I}_s$ and $\bar{J}_s$ to denote the intervals $[3^s, 2\cdot 3^s]$ and $[2\cdot 3^s, 3^{s+1}]$, respectively.

In case (i), we have

$$5\lambda^*\left(\frac{n}{5}\right) = 5(3s+4)\frac{n}{5} - 5\cdot 4\cdot 3^s = (3s+4)n - 20\cdot 3^s$$

and

$$4\lambda^*\left(\frac{n}{4}\right) = 4(3s+4)\frac{n}{4} - 4 \cdot 4 \cdot 3^s = (3s+4)n - 16 \cdot 3^s.$$

Thus,

$$5n + 5\lambda^*\left(\frac{n}{5}\right) - 4n - 4\lambda^*\left(\frac{n}{4}\right) = n - 4 \cdot 3^s \geq (5-4)3^s = 3^s > 0$$

since $n \geq 5 \cdot 3^s$ (because $\frac{n}{5} \in \bar{I}_s$).

Similarly, for case (ii) we have

$$5\lambda^*\left(\frac{n}{5}\right) = (3s+5)n - 30 \cdot 3^s$$

and

$$4\lambda^*\left(\frac{n}{4}\right) = (3s+5)n - 24 \cdot 3^s.$$

Thus,

$$5n + 5\lambda^*\left(\frac{n}{5}\right) - 4n - 4\lambda^*\left(\frac{n}{4}\right) = n - 6 \cdot 3^s \geq (10-6)3^s = 4 \cdot 3^s > 0$$

since $n \geq 10 \cdot 3^s$ (because $\frac{n}{5} \in \bar{J}_s$).

For case (iii), we have

$$5\lambda^*\left(\frac{n}{5}\right) = (3s+4)n - 20 \cdot 3^s$$

and

$$4\lambda^*\left(\frac{n}{4}\right) = (3s+5)n - 24 \cdot 3^s.$$

Thus,

$$5n + 5\lambda^*\left(\frac{n}{5}\right) - 4n - 4\lambda^*\left(\frac{n}{4}\right) = 4 \cdot 3^s > 0.$$

Finally, for case (iv) we have

$$5\lambda^*\left(\frac{n}{5}\right) = (3s+5)n - 30 \cdot 3^s$$

and

$$4\lambda^*\left(\frac{n}{4}\right) = (3s+7)n - 48 \cdot 3^s.$$

Thus,

$$5n + 5\lambda^*\left(\frac{n}{5}\right) - 4n - 4\lambda^*\left(\frac{n}{4}\right) = -n + 18 \cdot 3^s \geq 3 \cdot 3^s > 0$$

since $\frac{n}{5} \in \bar{J}_s \Longrightarrow n \leq 15 \cdot 3^s$.

Hence, in all cases it is better to use $r = 4$ than $r = 5$; i.e., the value of r which determines $\lambda^*(n)$ for $n \in I_t$ cannot be 5. A similar argument rules out *any* value of $r \geq 5$ (we omit the calculations which are very similar to the case we have just done). Also, it is easy to see that the same arguments apply if we initially assumed that $n \in J_{t+1}$, where now we can assume that the induction hypothesis holds for all $s \leq t$, and for I_{t+1}, as well.

Thus, we are left with the possibilities that $r = 2, 3$, or 4. Here, things become a little more interesting! When we apply the preceding argument to compare $3n + \lambda^*(n/3)$ and $4n + \lambda^*(n/4)$, we find that the difference is positive in cases (ii) and (iii), but can be 0 in cases (i) and (iv) exactly when $n = 4 \cdot 3^{t-1}$.

For $r = 2$, the same arguments show that there is a *whole interval* of values for n where the difference $2n + 2\lambda^*(\frac{n}{2}) - 3n - 3\lambda^*(\frac{n}{3})$ can be 0, namely when $4 \cdot 3^t \leq n \leq 6 \cdot 3^t$ (it can never be negative).

With this information, we can now compute the values of $\lambda^*(n)$ for $n \in I_{t+1} \cup J_{t+1}$. When we do this and extend to the real function $\lambda^*(x)$, we find that (6.2) holds for $t+1$. With this, the induction step is completed, and we have shown that (6.2) holds for all t. □

In particular, we have

$$\lambda^*(3^t) = t \cdot 3^{t+1}, \quad \lambda^*(2 \cdot 3^t) = (6t+4) \cdot 3^t, \quad \lambda^*(4 \cdot 3^t) = (12t+16) \cdot 3^t. \tag{6.3}$$

Let $T^*(n)$ denote an optimal tree with n leaves as $p \to 0$. Although we now know the *slope* $\lambda^*(n)$ of $P_{T^*(n)}$, we don't yet know the degree of the root of $P_{T^*(n)}$ in the case that n is in the "ambiguous" range $4 \cdot 3^t \leq n \leq 6 \cdot 3^t$. This will depend on the second derivative of $P_T(p)$ evaluated at $p = 0$. This is just

$$-\sum_{e \in E} \binom{L(e)}{2}.$$

Since this is negative, we want to make the sum $\sum_{e \in E} \binom{L(e)}{2}$ as large as possible in order to find the optimal tree $T^*(n)$. Let $\mu^*(n)$ denote the largest possible value of this sum over all trees $T(n)$ for which $\lambda_{T(n)} = \lambda^*(n)$. We know, in general, that an optimal tree $T^*(n)$ with root degree r has on each of its root edges an optimal subtree $T^*(i_k)$, where

$$\sum_{k=1}^{r} i_k = n$$

and *all the i_k must lie in the same interval I_t (or J_t)*, since otherwise the optimal value $\lambda^*(n)$ would not be achieved.

First, observe that we know the optimal tree $T^*(3^t)$ since it must have a root degree of 3, so by induction we can deduce that $\mu^*(3^t) = \frac{1}{4}(3^{2t+2} - (2t+3)3^{t+1})$. Now, for $T^*(2 \cdot 3^t)$, either the root has degree 2, in which case the two subtrees must both be $T^*(3^t)$'s, or the root has degree 3, in which case the three subtrees are all $T^*(2 \cdot 3^{t-1})$'s (since in both cases, the subtree sizes are endpoints of an I or J interval). By induction, we can conclude that degree 3 wins here and that $\mu^*(2 \cdot 3^t) = 3^{2t+2} - (3t+7)3^t$. Finally, for $T^*(4 \cdot 3^t)$ (where there are three possible choices for the root degree), we find that degree 4 wins in this case, and we have $\mu^*(4 \cdot 3^t) = 41 \cdot 3^{2t} - (6t+17)3^t$. These are the only values of n for which a root degree of 4 is optimal.

The remaining problem is to eliminate the possibility of a root degree of 2 for the values $4 \cdot 3^t < n \leq 6 \cdot 3^t$. It should be noted that to obtain the largest possible $\mu^*(n)$, the subtree sizes will tend to be as far apart as possible (consistent with staying in

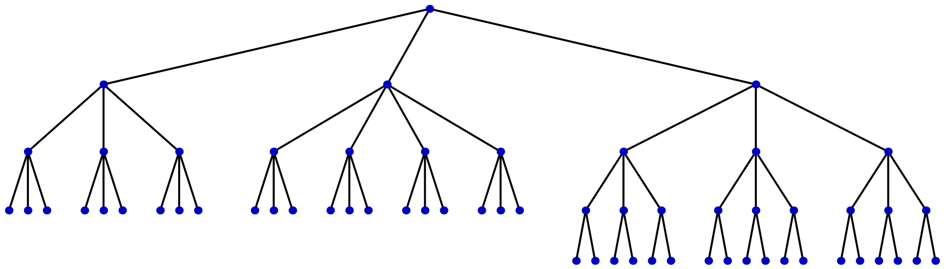


FIG. 6. *An optimal* $T^*(39)$.

the same I or J interval), again because of the tendency for $\mu^*(n)$ to be convex. It isn't convex everywhere (or even monotone), however, because of the unusually large values at $4 \cdot 3^t$. After all, since

$$\mu^*(n) = \max_{2 \le r \le 4} \left(r\binom{n}{2} + \sum_{i_1 \ldots i_r} \mu^*(i_k) \right),$$

then when $r = 4$, we get an especially large contribution from the term $r\binom{n}{2}$.

First, let us deal with $n \in J_t$, i.e., $2 \cdot 3^t \le n \le 3^{t+1}$. In this case, $T^*(n)$ has a root degree of 3 and so all three (optimal) subtrees have sizes $i_k \in J_{t-1}$. We put the proof of the following lemma in the appendix.

LEMMA 7. *Let* $n = 2 \cdot 3^t + r$ *with* $0 \le r \le 3^t$. $\max\{\mu^*(i_1) + \mu^*(i_2) + \mu^*(i_3) : i_1 + i_2 + i_3 = n, \text{all}\, i_k \in J_{t-1}\}$ *occurs when the* i_k *are "maximally spread," i.e., when at least two of the* i_k *are equal to the endpoint values* $2 \cdot 3^{t-1}$ *and* 3^t *of* J_{t-1}.

We derive in the following lemma that the optimal tree as $p \to 0$ cannot have a root degree of 2. Its proof is also put in the appendix.

LEMMA 8. $T^*(n)$ *cannot have a root degree of* 2 *for any* $n > 2$.

We summarize what we have shown in the following result.

THEOREM 4. *As* $p \to 0$, *the* (p, n)*-optimal tree* $T^*(n)$ *always has a root degree of* 3 *except for* n *of the form* $4 \cdot 3^t$, *in which case* $T^*(4 \cdot 3^t)$ *has a root degree of* 4, *and for* $n = 2$, *when* $T^*(2)$ *has a root degree of* 2. *When* $2 \cdot 3^t \le n \le 3^{t+1}$, *then* $T^*(n)$ *is as close to a balanced ternary tree with n leaves as possible. Namely, all subtrees (as well as* $T^*(n)$ *itself) have root degrees of* 3, *except for the very bottom level, where subtrees of size* 2 *can occur. However, when* $3^t \le n < 2 \cdot 3^t$, $T^*(n)$ *can deviate substantially from a balanced ternary tree.*

As an example for the situation $3^t \le n \le 2 \cdot 3^t$, note that $T^*(39)$ has three subtrees of sizes $9, 12$, and 18 (see Figure 6). In general, it seems to be difficult to predict the sizes of the subtrees in the optimal tree $T^*(n)$ for certain values of n. For example, for $n = 1252$, the sizes are $280, 486$, and 486, while for $n = 1253$, the sizes are $324, 443$, and 486. This is an example of the effect of a number of the form $324 = 4 \cdot 3^4$ having an especially large value of μ^*, thereby causing a preferential bias towards using it. In this case, $\mu^*(324) = 265680$ while $\mu^*(323) = 240997, \mu^*(325) = 244014$. In fact, we already see this happening at $n = 11$, where the optimal tree $T^*(11)$ has subtree sizes $3, 4$, and 4, which are not maximally spread in I_1.

Appendix.

Proof of Lemma 7. This is true by inspection for $t = 0, 1$. Suppose it holds for some value of $t \geq 1$ and let $n = 2 \cdot 3^{t+1} + r \in J_{t+1}$. Now, by (1),

$$\mu^*(n) = 3\binom{n}{2} + \max\{\mu^*(i_1) + \mu^*(i_2) + \mu^*(i_3) : i_1 + i_2 + i_3 = n, \text{ all } i_k \in J_t\}. \tag{A.1}$$

By induction, the maximum is achieved when the i_k are maximally spread. In particular, this means that, assuming $i_1 \leq i_2 \leq i_3$:

(a) if $0 \leq r \leq 3^t$, then $i_1 = 2 \cdot 3^t$, $i_2 = 2 \cdot 3^t$, $i_3 = 2 \cdot 3^t + r'$;
(b) if $3^t \leq r \leq 2 \cdot 3^t$, then $i_1 = 2 \cdot 3^t$, $i_2 = 2 \cdot 3^t + r'$, $i_3 = 3^{t+1}$;
(c) if $2 \cdot 3^t \leq r \leq 3^{t+1}$, then $i_1 = 2 \cdot 3^t + r'$, $i_2 = 3^{t+1}$, $i_3 = 3^{t+1}$.

Let $\Delta(m)$ denote the difference $\mu^*(m+1) - \mu^*(m)$. Then this implies, for $0 \leq r < 3^{t+1}$, the fundamental equation

$$\Delta(2 \cdot 3^{t+1} + r) = \Delta(2 \cdot 3^t + r') + 3n, \tag{A.2}$$

where $0 \leq r' < 3^t$ and $r \equiv r' \pmod{3^t}$. The term $3n$ comes from the difference $3\binom{n+1}{2} - 3\binom{n}{2} = 3n$. From this it follows (by induction) that a sum of k consecutive values

$$\Delta(2 \cdot 3^{t+1} + u) + \Delta(2 \cdot 3^{t+1} + (u+1)) + \cdots + \Delta(2 \cdot 3^{t+1} + (u+k-1))$$

for $0 \leq u \leq 3^{t+1} - k + 1$ is *minimized* by taking $u = 0$ since the sum is a monotone function of u. From this, the claim now follows, since any choice of the i_k which isn't maximally spread can be replaced by a maximally spread choice which can only increase the value of $\mu^*(i_1) + \mu^*(i_2) + \mu^*(i_3)$ (the difference in the two values being equal to the difference of two interval sums of the Δ's).

The next step is to obtain an explicit expression for the value $\Delta(n)$ for $n = 2 \cdot 3^t + r$ with $0 \leq r < 3^t$. We do this by iterating (5). The result is

$$\Delta(2 \cdot 3^t + r) = 3^{t+2} - 2 + 3R_t(r), \tag{A.3}$$

where $R_t(r)$ is defined as follows. Write r in its base 3 expansion as $r = r_{t-1}r_{t-2}r_{t-3}\ldots r_2r_1r_0$, where each $r_k \in \{0, 1, 2\}$. Then $R_t(r)$ is the sum of the t numbers corresponding to the t rows (read in base 3) of the array

$$\begin{array}{ccccccc}
r_{t-1} & r_{t-2} & r_{t-3} & \ldots & r_2 & r_1 & r_0 \\
0 & r_{t-2} & r_{t-3} & \ldots & r_2 & r_1 & r_0 \\
0 & 0 & r_{t-3} & \ldots & r_2 & r_1 & r_0 \\
0 & 0 & 0 & \ldots & r_2 & r_1 & r_0 \\
 & & & \vdots & & & \\
0 & 0 & 0 & \ldots & r_2 & r_1 & r_0 \\
0 & 0 & 0 & \ldots & 0 & r_1 & r_0 \\
0 & 0 & 0 & \ldots & 0 & 0 & r_0.
\end{array}$$

Thus, $R_t(0) = 0$, $R_t(1) = 3t$, $R_t(14) = 14t - 21$, etc.

From this we can write down an "explicit" expression for $\mu^*(n)$ for $n = 2 \cdot 3^t + r \in J_t$, which is

$$\begin{aligned}
\mu^*(2 \cdot 3^t + r) &= \mu^*(2 \cdot 3^t) + \Delta(2 \cdot 3^t) + \Delta(2 \cdot 3^t + 1) + \cdots + \Delta(2 \cdot 3^t + r - 1) \\
&= 3^{2t+2} - (3t+7)3^t + r \cdot (3^{t+2} - 2) + 3\sum_{k=0}^{r-1} R_t(k). \qquad \square
\end{aligned} \tag{A.4}$$

Proof of Lemma 8. As we have seen, this is only possible when $4 \cdot 3^t \le n \le 6 \cdot 3^t$. For such n, $2 \cdot 3^t \le \frac{n}{2} \le 3^{t+1}$. This implies that if $T^*(n)$ has a root degree of 2, then the two subtrees $T^*(i_1)$ and $T^*(i_2)$ with $i_1 + i_2 = n$ must have i_1 and i_2 both in J_t and be maximally spread (by the same inductive argument as before), which in this case means that at least one of the i_k must be one of the endpoint values $2 \cdot 3^t$ or 3^{t+1} of J_t. More precisely, if $T^{(2)}(n)$ denotes the best tree on n leaves with a root degree of 2 (and, of course, achieving the optimal value of $\lambda^*(n)$), and we assume that $i_1 \le i_2$, then

(a) if $4 \cdot 3^t \le n = 4 \cdot 3^t + r_1 \le 5 \cdot 3^t$, then $i_1 = 2 \cdot 3^t$, $i_2 = 2 \cdot 3^t + r_1$;

(b) if $5 \cdot 3^t \le n = 5 \cdot 3^t + r_2 \le 6 \cdot 3^t$, then $i_1 = 2 \cdot 3^t + r_2$, $i_2 = 3^{t+1}$.

In either case, we have

$$\mu_{T^{(2)}(n)} = 2\binom{n}{2} + \mu^*(i_1) + \mu^*(i_2), \tag{A.5}$$

where we know exactly the values of $\mu^*(i_1)$ and $\mu^*(i_2)$, since both arguments are in J_t. What we would like to do is compare this with the best tree $T'(n)$ with a root degree of 3 and show that $\mu_{T'(n)} > \mu_{T^{(2)}(n)}$. Unfortunately, we don't know the *best* tree $T'(n)$ with a root degree of 3. However, we know a pretty good one. This is the tree for which all of its subtrees also have root degrees of 3, recursively, until down to the very last level, in which any subtree with only 2 leaves must have a root degree of 2. For this family of trees $T'(n)$, we let $\mu'(n)$ denote the corresponding value of the second derivative (with its sign changed), and we let $\Delta'(m) = \mu'(m+1) - \mu'(m)$. Thus, $\Delta(n)$ and $\Delta'(n)$ agree for $n = 2, 3, 4, 5, 7, 8, 9$ while we have $\Delta(6) = 24$ and $\Delta'(6) = 20$.

More important is that $\Delta'(n)$ satisfies the same recurrence (5) that $\Delta(n)$ does, namely,

$$\Delta'(3^{t+1} + r) = \Delta'(3^t + r') + 3n, \tag{A.6}$$

where $n = 3^{t+1} + r$, $0 \le r < 3^{t+1}, 0 \le r' < 3^t$, and $r \equiv r' \pmod{3^t}$. This follows from the same arguments which established (5). Iterating this, we obtain the analogue of (6):

$$\Delta'(3^t + r) = \frac{3^{t+2} - 5}{2} + R_t(r). \tag{A.7}$$

Finally, summing this over r and using the fact that $\mu(3^t) = \frac{1}{4}(3^{2t+2} - (6t+9)3^t)$, we obtain the following analogue of (7):

$$\mu'(3^t + r) = \frac{1}{4}\big(3^{2t+2} - (6t+9)3^t\big) + \frac{r(3^{t+2} - 5)}{2} + 3\sum_{k=0}^{r-1} R_t(k). \tag{A.8}$$

We are now in a position to complete the argument. Let $n = 4 \cdot 3^t + r$. There are two cases.

Case 1. $0 \le r \le 3^t$. In this case

$$\mu'(n) = \mu^*(3^t) + \mu^*(2 \cdot 3^t) + \mu'(3^t + r) + 3\binom{n}{2}$$

and

$$\mu_{T^{(2)}(n)} = \mu^*(2 \cdot 3^t) + \mu^*(2 \cdot 3^t + r) + 2\binom{n}{2}.$$

Hence,

$$\mu'(n) > \mu_{T^{(2)}(n)}$$

if and only if

$$\mu^*(3^t) + \mu^*(2\cdot 3^t) + \mu'(3^t + r) + 3\binom{n}{2} > \mu^*(2\cdot 3^t) + \mu^*(2\cdot 3^t + r) + 2\binom{n}{2}$$

if and only if

$$\frac{3^{2t+2} - (6t+9)3^t}{2} + \frac{r(3^{t+2}-5)}{2} + 3\sum_{k=0}^{r-1} R_t(k) + \binom{n}{2}$$

$$> 3^{2t+2} - (3t+7)3^t + r(3^{t+2}-2) + 3\sum_{k=0}^{r-1} R_t(k)$$

if and only if

$$2\binom{4\cdot 3^t + r}{2} > 3^{2t+2} + r3^{t+2} - 5\cdot 3^t + r,$$

which is easily verified. This finishes Case 1.

Case 2. $3^t \le r \le 2\cdot 3^t$. Let $s = r - 3^t$. In this case

$$\mu'(n) = \mu'(3^t + s) + \mu^*(2\cdot 3^t) + \mu^*(2\cdot 3^t) + 3\binom{n}{2}$$

and

$$\mu_{T^{(2)}(n)} = \mu^*(2\cdot 3^t + s) + \mu^*(3^{t+1}) + 2\binom{n}{2}.$$

Hence,

$$\mu'(n) > \mu_{T^{(2)}(n)}$$

if and only if

$$\mu'(3^t + s) + \mu^*(2\cdot 3^t) + \mu^*(2\cdot 3^t) + \binom{n}{2} > \mu^*(2\cdot 3^t + s) + \mu^*(3^{t+1})$$

if and only if

$$\frac{3^{2t+2} - (6t+9)3^t}{4} + \frac{s(3^{t+2}-5)}{2} + 3\sum_{k=0}^{r-1} R_t(k) + 2(3^{2t+2} - (3t+7)3^t) + \binom{5\cdot 3^t + s}{2}$$

$$> \; 3^{2t+2} - (3t+7)3^t + s(3^{t+2}-2) + 3\sum_{k=0}^{r-1} R_t(k) + \frac{3^{2t+4} - (6t+15)3^{t+1}}{4}$$

if and only if

$$2\binom{5\cdot 3^t + s}{2} > 18\cdot 3^{2t} - 4\cdot 3^t + s(3^{t+2}+1).$$

As before, it is easy to check that this inequality holds, and Case 2 is completed.

Thus, we have shown that in the ambiguous range $4 \cdot 3^t \leq n < 6 \cdot 3^t$, there is always a (ternary) tree $T'(n)$ which dominates the best tree $T^{(2)}(n)$ with a degree 2 root (where all trees under consideration must achieve the optimal value of $\lambda^*(n)$). Consequently, the *optimal* tree $T^*(n)$ also dominates this $T^{(2)}(n)$ as well. $\square$

REFERENCES

[1] M. T. Goodrich, J. Z. Sun, and R. Tamassia, *Efficient tree-based revocation in groups of low-state devices*, in Advances in Cryptology—CRYPTO, Lecture Notes in Comput. Sci. 3152, Springer, Berlin, 2004, pp. 511–527.

[2] X. S. Li, Y. R. Yang, M. G. Gouda, and S. S. Lam, *Batch re-keying for secure group communications*, in Proceedings of the Tenth International Conference on the World Wide Web, WWW10, Hong Kong, 2001, pp. 525–534.

[3] D. M. Wallner, E. G. Harder, and R. C. Agee, *Key management for multicast: Issues and architectures*, in internet draft *draft-waller-key-arch*-01*.txt*, Sept. 1998

[4] C. K. Wong, M. G. Gouda, and S. S. Lam, *Secure group communications using key graphs*, IEEE/ACM Trans. Networking, 8 (2003), pp. 16–30.

[5] F. Zhu, A. Chan, and G. Noubir, *Optimal tree structure for key management of simultaneous join/leave in secure multicast*, in Proceedings of the IEEE Military Communication Conference (MILCOM), Boston, 2003, pp. 773–778.

SIAM J. Discrete Math.
Vol. 21, No. 2, pp. 548–550

ERRATUM: ENUMERATING TYPICAL CIRCULANT COVERING PROJECTIONS ONTO A CIRCULANT GRAPH*

RONGQUAN FENG†, JIN HO KWAK‡, AND YOUNG SOO KWON§

Abstract. This paper consists of an erratum to the previously published *Enumerating Typical Circulant Covering Projections onto a Circulant Graph*.

DOI. 10.1137/060653020

In Lemma 4 of [1], the authors incorrectly rephrased the characterization theorem of two isomorphic graph coverings given in [3] for a regular covering case. This should be corrected as follows.

Lemma 4′. *Let ϕ and ψ be typical voltage assignments in $C^1(G;\mathbb{Z}_p)$. Then, two typical circulant p-fold coverings $p_\phi : G^\phi \to G$ and $p_\psi : G^\psi \to G$ are isomorphic if and only if there exists a function $g : V(G) \to S_p$ such that $\psi(uv) = g(v)\phi(uv)g(u)^{-1}$ in S_p for each $uv \in D(G)$, where S_p is the symmetric group on the elements of $\mathbb{Z}_p$ and $\mathbb{Z}_p$ is considered as the left regular subgroup of S_p.*

However, if the voltage assignments ϕ and ψ in $C^1(G;\mathbb{Z}_p)$ are assumed to be trivial on a spanning tree of a graph G, two coverings $p_\phi : G^\phi \to G$ and $p_\psi : G^\psi \to G$ are isomorphic if and only if there exists an automorphism $\sigma \in \mathrm{Aut}\,(\mathbb{Z}_p)$ such that $\phi(uv)^\sigma = \psi(uv)$ for every arc uv of G (see [2, 4]).

Because of the error in Lemma 4, Lemma 5 of [1] is also incorrect. Instead of these two lemmas, we use a minor extension of Lemma 10 for our enumeration, which can be stated as follows.

Lemma 10′. *Let $G = \mathrm{Cay}(\mathbb{Z}_n, Y)$ be a connected circulant graph. For any natural number ℓ (not necessarily prime), let $f, g : \mathbb{Z}_{\ell n} \to \mathbb{Z}_n$ be two group epimorphisms. Then two connected typical coverings $f_* : \mathrm{Cay}(\mathbb{Z}_{\ell n}, X_1) \to G$ and $g_* : \mathrm{Cay}(\mathbb{Z}_{\ell n}, X_2) \to G$ are isomorphic if and only if there exists an automorphism $\Phi \in \mathrm{Aut}\,(\mathbb{Z}_{\ell n})$ such that $g \circ \Phi = f$ and $\Phi(X_1) = X_2$.*

The necessity of Lemma 10′ is proved in [1] and the sufficiency is clear.

Based on Lemmas 2, 3, and 7 in [1] and Lemma 10′, Theorem 8 in [1], which counts the connected typical circulant prime-fold coverings, should be corrected as follows.

Theorem 8′. *For any odd prime p, the number of isomorphism classes of connected typical circulant p-fold coverings of $G = \mathrm{Cay}(\mathbb{Z}_n, Y)$ is $\frac{1}{p-1}(p^{\lfloor \frac{d}{2} \rfloor} - 1)$ if $(p, n) = 1$ and is $p^{\lfloor \frac{d}{2} \rfloor - 1}$ otherwise, where d is the valency of G.*

Proof. Let $Y = \{\pm i_1, \pm i_2, \ldots, \pm i_k\}$ or $\{\pm i_1, \pm i_2, \ldots, \pm i_k, \frac{n}{2}\}$ according to whether the valency d of G is even $2k$ or odd $2k+1$, where $0 < i_1, i_2, \ldots, i_k < \lfloor \frac{n+1}{2} \rfloor$. Then, by Lemma 3, any typical circulant p-fold covering of G can be derived from a typical voltage assignment, or from a k-tuple $(\delta_1, \delta_2, \ldots, \delta_k) \in \mathbb{Z}_p^k$, with the assumption that any typical covering projection sends 1 in $\mathbb{Z}_{pn}$ to 1 in $\mathbb{Z}_n$ by Lemma 2. Let Δ

*Received by the editors February 26, 2006; accepted for publication (in revised form) September 22, 2006; published electronically June 28, 2007.
http://www.siam.org/journals/sidma/21-2/65302.html

†LMAM, School of Mathematical Sciences, Peking University, Beijing 100871, People's Republic of China (fengrq@math.pku.edu.cn).

‡Department of Mathematics, Pohang University of Science and Technology, Pohang, 790-784 Korea (jinkwak@postech.ac.kr).

§Department of Mathematics, Yeungnam University, Kyeongsan, 712-749 Korea (ysookwon@ynu.ac.kr).

denote the set of k-tuples $(\delta_1, \delta_2, \ldots, \delta_k) \in \mathbb{Z}_p^k$ which induce connected p-fold coverings of G. Then $|\Delta| = p^{\lfloor \frac{d}{2} \rfloor} - 1$ if $(p, n) = 1$ and $|\Delta| = p^{\lfloor \frac{d}{2} \rfloor}$ if $p|n$ by Lemma 7. Furthermore, by Lemma 10′, any two k-tuples $(\delta_1, \delta_2, \ldots, \delta_k)$ and $(\delta'_1, \delta'_2, \ldots, \delta'_k)$ in Δ induce isomorphic coverings if and only if there exists a $\Phi \in \mathrm{Aut}\,(\mathbb{Z}_{pn})$ such that $\Phi(i_j + \delta_j n) = i_j + \delta'_j n$ for every $j = 1, 2, \ldots, k$. In this case, Φ becomes a covering isomorphism between induced coverings and $\Phi(1) = 1 + an$ for some $a = 0, 1, \ldots, p-1$. Note that a map Φ defined by $\Phi(1) = 1 + an$ for some $a = 0, 1, \ldots, p-1$ is an automorphism of $\mathbb{Z}_{pn}$ if and only if $(pn, 1 + an) = 1$. Let $\mathcal{S} = \{\Phi \in \mathrm{Aut}\,(\mathbb{Z}_{pn}) \mid \Phi(1) \equiv 1(\mathrm{mod}\ n)\}$. Then, $\mathcal{S}$ is a subgroup of $\mathrm{Aut}\,(\mathbb{Z}_{pn})$. Define an $\mathcal{S}$-action on Δ by $\Phi(\delta_1, \delta_2, \ldots, \delta_k) = (\delta'_1, \delta'_2, \ldots, \delta'_k)$ for any $\Phi \in \mathcal{S}$ and $(\delta_1, \delta_2, \ldots, \delta_k) \in \Delta$, where δ'_j is uniquely determined by the relation $\Phi(i_j + \delta_j n) = i_j + \delta'_j n$ for every $j = 1, 2, \ldots, k$. This action is well defined and the number of isomorphism classes of connected typical circulant p-fold coverings of G is the number of orbits under the $\mathcal{S}$-action on Δ. Let an arbitrary $\delta = (\delta_1, \delta_2, \ldots, \delta_k) \in \Delta$ be given. Since no $\Phi \in \mathcal{S}$ fixes the k-tuple $(i_1 + \delta_1 n, i_2 + \delta_2 n, \ldots, i_k + \delta_k n)$ except the identity Φ, the orbit size of δ equals the cardinality $|\mathcal{S}|$, that is, the number of automorphisms Φ of $\mathbb{Z}_{pn}$ such that $\Phi(1) = 1 + an$ for some $a = 0, 1, \ldots, p-1$. As the first case, let $(p, n) = 1$. Then, $\gcd(pn, 1 + an) = 1$ except exactly one of $a = 0, 1, \ldots p-1$. Hence, the orbit size of δ is $p-1$. Since $\delta \in \Delta$ is given arbitrarily, it gives the proof of the case $(p, n) = 1$. As the remaining case, let $p|n$. Then, for each $a = 0, 1, \ldots p-1$, we get $\gcd(1 + an, pn) = 1$. Hence, the orbit size of δ is p. This completes the proof. □

Comparing with the old enumeration in Theorem 8 in [1], the number of isomorphism classes of connected typical circulant p-fold coverings of $G = \mathrm{Cay}(\mathbb{Z}_n, Y)$ is corrected as the multiple of the old value by $\frac{1}{p-1}$ when $(p, n) = 1$ in Theorem 8′. The following corrections will be listed as the last part of this manuscript.

Since Lemma 7 in [1] counts the number of disconnected typical circulant p-fold coverings of $G = \mathrm{Cay}(\mathbb{Z}_n, Y)$, one can get the number of isomorphism classes of (connected or not) typical circulant p-fold coverings of G with the help of Theorem 8′. This provides a correct version of Theorem 6 in [1].

Now, Theorem 13 in [1], which counts the connected typical circulant ℓ-fold coverings for any composite number ℓ, can be revised as follows.

Theorem 13′. *Let $\ell = p_1^{r_1} p_2^{r_2} \cdots p_s^{r_s}$ be the prime factorization of a positive integer ℓ and let G be a connected circulant graph of order n and valency d. Then the number N of isomorphism classes of connected typical circulant ℓ-fold coverings of G is*

$$N = \begin{cases} 0 & \textit{if } \ell \textit{ is even and } d \textit{ is odd}, \\ \displaystyle\prod_{i=1}^{s} N_i & \textit{otherwise}, \end{cases}$$

where

$$N_i = \begin{cases} p_i^{r_i(\lfloor \frac{d}{2} \rfloor - 1)} & \textit{if } p_i | n, \\ p_i^{(r_i - 1)(\lfloor \frac{d}{2} \rfloor - 1)} \left(p_i^{\lfloor \frac{d}{2} \rfloor} - 1 \right) / (p_i - 1) & \textit{if } (p_i, n) = 1. \end{cases}$$

Corollaries 14, 15, 16, and 18 and Table 1 should be revised as follows.

Corollary 14′. *Let G be a connected circulant graph of order n and valency d. For any prime p and any natural number r, the number of isomorphism classes*

Table 0.1

The number of isomorphism classes of connected typical circulant ℓ-fold coverings of the complete graph K_n for small ℓ and small n

n	3	4	5	6	7	8	9	10	11	12	13	14	$\cdots$
$\ell = 2$	1	0	3	0	7	0	15	0	31	0	63	0	$\cdots$
$\ell = 3$	1	1	4	3	13	13	27	40	121	81	364	364	
$\ell = 4$	1	0	6	0	28	0	120	0	496	0	2016	0	
$\ell = 5$	1	1	5	6	31	31	131	125	321	321	3906	3906	
$\ell = 6$	1	0	12	0	91	0	405	0	3751	0	22932	0	$\cdots$
$\vdots$	$\vdots$											$\vdots$	

of connected typical circulant p^r-fold coverings of G is 0 when $p = 2$ and d is odd. Otherwise, this number is $p^{r(\lfloor \frac{d}{2} \rfloor - 1)}$ if $p|n$, and is $p^{(r-1)(\lfloor \frac{d}{2} \rfloor - 1)}(p^{\lfloor \frac{d}{2} \rfloor} - 1)/(p-1)$ if $(p, n) = 1$.

Corollary 15′. *Let G be a connected circulant graph of order n and valency d. For any two distinct primes p and q, the number N of isomorphism classes of connected typical circulant pq-fold coverings of G is 0 when d is odd and one of p and q is 2. Otherwise, the number N is*

$$N = \begin{cases} p^{\lfloor \frac{d}{2} \rfloor - 1} q^{\lfloor \frac{d}{2} \rfloor - 1} & \text{if } pq|n, \\ p^{\lfloor \frac{d}{2} \rfloor - 1}(q^{\lfloor \frac{d}{2} \rfloor} - 1)/(q-1) & \text{if } p|n \text{ but } (q, n) = 1, \\ (p^{\lfloor \frac{d}{2} \rfloor} - 1) q^{\lfloor \frac{d}{2} \rfloor - 1}/(p-1) & \text{if } q|n \text{ but } (p, n) = 1, \\ (p^{\lfloor \frac{d}{2} \rfloor} - 1)(q^{\lfloor \frac{d}{2} \rfloor} - 1)/((p-1)(q-1)) & \text{if } (pq, n) = 1. \end{cases}$$

Corollary 16′. *Let $\ell = p_1^{r_1} p_2^{r_2} \cdots p_s^{r_s}$ be the prime factorization of a positive integer ℓ. Then no connected typical circulant ℓ-fold covering of K_n exists when both ℓ and n are even. Otherwise the number of isomorphism classes of connected typical circulant ℓ-fold coverings of K_n is $\prod_{i=1}^{s} N_i$, where*

$$N_i = \begin{cases} p_i^{r_i(\lfloor \frac{n-1}{2} \rfloor - 1)} & \text{if } p_i|n, \\ p_i^{(r_i - 1)(\lfloor \frac{n-1}{2} \rfloor - 1)} \left(p_i^{\lfloor \frac{n-1}{2} \rfloor} - 1 \right)/(p_i - 1) & \text{if } (p_i, n) = 1. \end{cases}$$

Corollary 18′. *Let G be a connected circulant trivalent graph of order n but $G \neq K_4$ or $K_{3,3}$. If ℓ is even, then G has no connected circulant ℓ-fold coverings. If ℓ is odd, then G has only one connected circulant ℓ-fold covering up to isomorphism.*

All statements in [1] that were not mentioned remain valid.

REFERENCES

[1] R. Feng, J. H. Kwak, and Y. S. Kwon, *Enumerating typical circulant covering projections onto a circulant graph*, SIAM J. Discrete Math., 19 (2005), pp. 196–207.

[2] S. Hong, J. H. Kwak, and J. Lee, *Regular graph coverings whose covering transformation groups have the isomorphism extension property*, Discrete Math., 148 (1996), pp. 85–105.

[3] J. H. Kwak and J. Lee, *Isomorphism classes of graph bundles*, Canad. J. Math., 42 (1990), pp. 747–761.

[4] M. Škoviera, *A contribution to the theory of voltage graphs*, Discrete Math., 61 (1986), pp. 281–292.

SIAM J. Discrete Math.
Vol. 21, No. 3, pp. 551–572

THE INTEGER KNAPSACK COVER POLYHEDRON*

HANDE YAMAN†

Abstract. We study the integer knapsack cover polyhedron which is the convex hull of the set of vectors $x \in \mathbb{Z}^n_+$ that satisfy $C^T x \geq b$, with $C \in \mathbb{Z}^n_{++}$ and $b \in \mathbb{Z}_{++}$. We present some general results about the nontrivial facet-defining inequalities. Then we derive specific families of valid inequalities, namely, rounding, residual capacity, and lifted rounding inequalities, and identify cases where they define facets. We also study some known families of valid inequalities called 2-partition inequalities and improve them using sequence-independent lifting.

Key words. integer knapsack cover polyhedron, valid inequalities, facets, sequence-independent lifting

AMS subject classifications. 90C10, 90C57

DOI. 10.1137/050639624

1. Introduction.

The purpose of this paper is to study the integer knapsack cover polyhedron. Let $N = \{1, 2, \dots, n\}$. Item $i \in N$ has capacity c_i. We would like to cover a demand of b using integer amounts of items in N. We assume that b and c_i for $i \in N$ are positive integers.

We are interested in the integer knapsack cover set

$$X = \left\{ x \in \mathbb{Z}^n_+ : \sum_{i \in N} c_i x_i \geq b \right\} \tag{1}$$

and its convex hull $PX = conv(X)$. The constraint $\sum_{i \in N} c_i x_i \geq b$ is called the *cover constraint.*

Set X is a relaxation of the feasible sets of many optimization problems involving demands that may be covered with different types of items. Pochet and Wolsey [15] study a special case to derive valid inequalities for a network design problem. Mazur [11] uses the polyhedral results on PX to generate strong valid inequalities for the multifacility location problem. Yaman [18] uses the same relaxation to strengthen formulations for the heterogeneous vehicle routing problem, which generalizes the well-known capacitated vehicle routing problem by introducing the choice between different vehicle types. Yaman and Sen [19] arrive at the same relaxation in the context of the manufacturer's mixed pallet design problem, where each customer can buy integer numbers of pallets with different configurations to satisfy its demand. Knowledge about polyhedral properties of PX can be used in deriving strong formulations for these problems. For recent work in understanding the structure of simple mixed integer and integer sets, see, e.g., [3, 7, 12, 13, 15].

There has been a lot of work on the polytope of the 0/1 knapsack problem (e.g., [5, 8, 9, 16, 17, 20]). The situation is different for the integer knapsack cover polyhedron. Despite the many application areas where set X may appear as a relaxation, the literature on the polyhedral properties of its convex hull is quite limited.

Pochet and Wolsey [15] study the special case where c_{i+1} is an integer multiple of c_i for all $i = 1, 2, \dots, n-1$. They derive the partition inequalities and show that

*Received by the editors September 6, 2005; accepted for publication (in revised form) March 28, 2007; published electronically July 11, 2007.

http://www.siam.org/journals/sidma/21-3/63962.html

†Bilkent University, Department of Industrial Engineering, Bilkent 06800 Ankara, Turkey (hyaman@bilkent.edu.tr).

these inequalities define the convex hull together with the nonnegativity constraints. They derive conditions under which these inequalities are valid in the general case.

Mazur [11] and Mazur and Hall [12] study the general case. They show that $\dim(PX) = n$, $x_i \geq 0$ defines a facet of PX for $i \in N$, and if $\sum_{i \in N} \alpha_i x_i \geq \alpha_0$ is a nontrivial facet-defining inequality of PX, then $\alpha_i > 0$ for all $i \in N$ and $\alpha_0 > 0$. Let $c_1^{'}, \ldots, c_m^{'}$ be the distinct c_i values that are less than b. An important result by Mazur [11] is that, if one knows the description of $conv(\{x \in \mathbb{Z}_+^m : \sum_{i=1}^m c_i^{'} x_i \geq b\})$, it is trivial to obtain the description of PX. The inequality $\sum_{i \in N} \alpha_i x_i \geq \alpha_0$ is a nontrivial facet-defining inequality for PX if and only if $\alpha_i = \alpha_j$ for all $i, j \in N$ with $c_i = c_j$, $\alpha_i = \alpha_0$ for all $i \in N$ with $c_i \geq b$, and $\sum_{i=1}^m \alpha_i^{'} x_i \geq \alpha_0$ is a nontrivial facet-defining inequality for $conv(\{x \in \mathbb{Z}_+^m : \sum_{i=1}^m c_i^{'} x_i \geq b\})$, where $\alpha_i^{'} = \alpha_j$ if $c_i^{'} = c_j$ for $i = 1, \cdots, m$ and $j \in N$. So interesting instances satisfy $c_1 < c_2 < \cdots < c_n < b$.

Mazur and Hall [12] also study the integer capacity cover polyhedron defined as the convex hull of the set $\{(y, x) \in \{0,1\}^q \times \mathbb{Z}_+^n : \sum_{i \in N} c_i x_i \geq \sum_{i=1}^q y_i\}$. They use simultaneous lifting to derive facet-defining inequalities for this polyhedron using those of the integer knapsack cover polyhedron. They remark that little is known about the polyhedral properties of the latter polyhedron, and it is difficult to identify its facets.

Atamturk [1] presents a family of facet-defining inequalities and lifting results for the polytope $conv(X \cap \{x \in \mathbb{Z}^n : x \leq u\})$ for $u \in \mathbb{Z}_{++}^n$.

In this paper, we derive several families of valid inequalities and discuss when they define facets of PX. We investigate the domination relations between these families of valid inequalities. Most of our results on facet-defining inequalities are for the special case where $c_1 = 1$.

This work is motivated by the results of Mazur and Hall [12], where valid inequalities for the integer knapsack cover polyhedron are lifted to valid inequalities for a more complicated polyhedron, the integer capacity cover polyhedron. We are also motivated by the positive results in [18, 19], which demonstrate the use of simple valid inequalities based on the integer knapsack cover relaxation in closing the duality gap for complicated mixed integer programming problems studied in these papers.

The paper is organized as follows. In section 2, we give the general properties of nontrivial facet-defining inequalities of PX. In sections 3–6, we introduce four families of valid inequalities, namely, rounding, residual capacity, lifted rounding, and lifted 2-partition inequalities. We compare their relative strengths and give conditions under which they define facets of PX. In section 7, we investigate the use of lifted rounding and lifted 2-partition inequalities in solving the manufacturer's mixed pallet design problem introduced by Yaman and Sen [19]. We conclude in section 8.

2. General results on facet-defining inequalities.

In this section, we derive general properties of nontrivial facet-defining inequalities of PX.

In the sequel, we assume that $c_1, \ldots, c_n$ and b are positive integers and that they satisfy $c_1 < c_2 < \cdots < c_n < b$ (this assumption is made without loss of generality due to the result of Mazur [11] mentioned above). Let c be the greatest common divisor of c_i's. We replace c_i with $\frac{c_i}{c}$ for each $i \in N$ and b with $\left\lceil \frac{b}{c} \right\rceil$. This does not change the set X but strengthens the cover constraint. Let e_i denote the n-dimensional unit vector with 1 at the ith place and 0 elsewhere.

Proposition 1. *Let $\sum_{i \in N} \alpha_i x_i \geq \alpha_0$ be a nontrivial facet-defining inequality for PX. Then*

$$0 < \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n \leq \alpha_0 \leq \min_{i \in N} \alpha_i \left\lceil \frac{b}{c_i} \right\rceil.$$

Proof. Suppose that $\sum_{i \in N} \alpha_i x_i \geq \alpha_0$ is a nontrivial facet-defining inequality for PX. The fact that $\alpha_i > 0$ for $i = 0, 1, \ldots, n$ is proved in [11, 12].

Let j and l be such that $j < l$ and $x \in PX$ be such that $\sum_{i \in N} \alpha_i x_i = \alpha_0$, with $x_j \geq 1$. Consider $x^{'} = x - e_j + e_l$. As $c_l > c_j$, $x^{'} \in PX$. Then $\sum_{i \in N} \alpha_i x_i^{'} \geq \alpha_0$, implying that $\alpha_l \geq \alpha_j$. So $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$.

Let $x \in PX$ be such that $\sum_{i \in N} \alpha_i x_i = \alpha_0$, with $x_n \geq 1$. Then $\alpha_n x_n \leq \alpha_0$ and, as $x_n \geq 1$, $\alpha_n \leq \alpha_0$.

For $i \in N$, $x = \lceil \frac{b}{c_i} \rceil e_i$ is in PX, and so $\alpha_i \lceil \frac{b}{c_i} \rceil \geq \alpha_0$. Thus $\alpha_0 \leq \min_{i \in N} \alpha_i \lceil \frac{b}{c_i} \rceil$. □

We have a necessary condition for a nontrivial inequality to be facet-defining.

THEOREM 1. *Let $\sum_{i \in N} \alpha_i x_i \geq \alpha_0$ be a nontrivial facet-defining inequality for PX. Let $j \in \arg\max_{i \in N} \frac{c_i}{\alpha_i}$. Then $(\alpha_0 - \alpha_i) \frac{c_j}{\alpha_j} + c_i \geq b$ for all $i \in N \setminus \{j\}$.*

Proof. Assume that there exists $l \in N \setminus \{j\}$ such that $(\alpha_0 - \alpha_l) \frac{c_j}{\alpha_j} + c_l < b$. Let $x \in X$ be such that $\sum_{i \in N} \alpha_i x_i = \alpha_0$. Then $x_j = \frac{\alpha_0 - \sum_{i \in N \setminus \{j\}} \alpha_i x_i}{\alpha_j}$. The left-hand side of the cover constraint evaluated at x is $\sum_{i \in N} c_i x_i = \sum_{i \in N \setminus \{j\}} (c_i - \frac{c_j}{\alpha_j} \alpha_i) x_i + \frac{c_j}{\alpha_j} \alpha_0$. This is less than or equal to $(c_l - \frac{c_j}{\alpha_j} \alpha_l) x_l + \frac{c_j}{\alpha_j} \alpha_0$, since $c_i - \frac{c_j}{\alpha_j} \alpha_i \leq 0$ for all $i \in N \setminus \{j\}$. Now as $(\alpha_0 - \alpha_l) \frac{c_j}{\alpha_j} + c_l < b$ and $c_l - \frac{c_j}{\alpha_j} \alpha_l \leq 0$, whenever $x_l \geq 1$, $(c_l - \frac{c_j}{\alpha_j} \alpha_l) x_l + \frac{c_j}{\alpha_j} \alpha_0 < b$. This proves that, for any $x \in X$ such that $\sum_{i \in N} \alpha_i x_i = \alpha_0$, we have $x_l = 0$. □

Next, we give necessary and sufficient conditions for some inequalities to be facet-defining. Later, we use this result to identify specific families of facet-defining inequalities.

THEOREM 2. *Let $\sum_{i \in N} \alpha_i x_i \geq \alpha_0$ be a valid inequality for PX, with $\alpha_i > 0$ and integer for all $i \in N \cup \{0\}$ and $\alpha_1 = 1$. Let j be the largest index, with $\alpha_j = 1$. If $\alpha_i \geq \frac{c_i}{c_j}$ for all $i = j+1, \ldots, n$, then the inequality $\sum_{i \in N} \alpha_i x_i \geq \alpha_0$ is facet-defining for PX if and only if $(\alpha_0 - \alpha_i) c_j + c_i \geq b$ for $i = j+1, \ldots, n$ and $(\alpha_0 - 1) c_j + c_1 \geq b$.*

Proof. If the conditions of the theorem are satisfied, then $\alpha_0 e_j$, $(\alpha_0 - 1) e_j + e_i$ for $i = 1, \ldots, j-1$, and $(\alpha_0 - \alpha_i) e_j + e_i$ for $i = j+1, \ldots, n$ are in PX; they satisfy $\sum_{i \in N} \alpha_i x_i = \alpha_0$ and are affinely independent. This proves that the inequality $\sum_{i \in N} \alpha_i x_i \geq \alpha_0$ is facet-defining for PX.

The necessity of the conditions are implied by Theorem 1. □

To conclude this section, we investigate when the cover constraint is facet-defining for PX. If c_j divides b for all $j \in N$, then the nonnegativity constraints and the cover constraint describe the polyhedron PX, i.e., $PX = \{x \in \mathbb{R}^n_+ : \sum_{j \in N} c_j x_j \geq b\}$.

Using Theorem 2, we identify another case where the cover constraint is facet-defining.

COROLLARY 1. *If $c_1 = 1$, then the cover constraint is facet-defining for PX.*

The conclusion of Theorem 1 is trivially satisfied for the cover constraint. But the cover constraint is not necessarily facet-defining for PX. The following simple example proves this statement.

Example 1. Let $X^1 = \{x \in \mathbb{Z}^2_+ : 3x_1 + 4x_2 \geq 14\}$. The polyhedron $conv(X^1) = \{(x_1, x_2) \in \mathbb{R}^2_+ : x_1 + x_2 \geq 4, 2x_1 + 3x_2 \geq 10\}$.

3. Rounding inequalities. In this section, we derive a family of valid inequalities, called the *rounding inequalities*, and identify some cases where they are facet-defining for PX.

For $\lambda > 0$, the rounding inequality

$$\sum_{i \in N} \left\lceil \frac{c_i}{\lambda} \right\rceil x_i \geq \left\lceil \frac{b}{\lambda} \right\rceil \tag{2}$$

is a valid inequality for PX. It is obtained using the well-known Chvatal–Gomory procedure (see, e.g., Nemhauser and Wolsey [14]). These inequalities have been used by Yaman [18]. Here we investigate under which conditions these inequalities are facet-defining for PX. The inequality for $\lambda = c_n$ is $\sum_{i \in N} x_i \geq \lceil \frac{b}{c_n} \rceil$. Mazur [11] proves that this inequality is facet-defining for PX if and only if $b \leq \big(\lceil \frac{b}{c_n} \rceil - 1\big)c_n + c_1$. Inequality (2) for any $\lambda > c_n$ is dominated by the corresponding inequality for c_n. So we are interested in $\lambda < c_n$.

The result below is a corollary to Theorem 2.

COROLLARY 2. *Let λ be such that $c_j \leq \lambda < c_{j+1}$ for some $j \in \{1, \dots, n-1\}$. If $\lceil \frac{c_i}{\lambda} \rceil \geq \frac{c_i}{c_j}$ for all $i = j+1, \dots, n$, then inequality* (2) *is facet-defining if and only if $\big(\lceil \frac{b}{\lambda} \rceil - 1\big)c_j + c_1 \geq b$ and $\big(\lceil \frac{b}{\lambda} \rceil - \lceil \frac{c_i}{\lambda} \rceil\big)c_j + c_i \geq b$ for all $i = j+1, \dots, n$.*

Proof. As $\lceil \frac{c_i}{\lambda} \rceil$ for $i \in N$ and $\lceil \frac{b}{\lambda} \rceil$ are positive integers, $\lceil \frac{c_1}{\lambda} \rceil = 1$, j is the largest index with coefficient 1 in inequality (2), and $\lceil \frac{c_i}{\lambda} \rceil \geq \frac{c_i}{c_j}$ for all $i = j+1, \dots, n$, Theorem 2 applies. □

We have a necessary condition as a corollary to Theorem 1.

COROLLARY 3. *Let $\lambda > 0$. If there exists $j \in N$ such that c_j is divisible by λ and if inequality* (2) *is facet-defining for PX, then $(\lceil \frac{b}{\lambda} \rceil - \lceil \frac{c_i}{\lambda} \rceil)\lambda + c_i \geq b$ for all $i \in N \setminus \{j\}$.*

Proof. For $i \in N$, $\frac{c_i}{\lceil \frac{c_i}{\lambda} \rceil} \leq \lambda$. So, if $j \in N$ is such that λ divides c_j, $j \in \arg\max_{i \in N} \frac{c_i}{\lceil \frac{c_i}{\lambda} \rceil}$, and we can apply Theorem 1. □

We consider the subset of inequalities (2) defined by λ equal to $c_1, \dots, c_n$. In the following corollary, we generalize the result by Mazur [11].

COROLLARY 4. *For $j \in N$, the inequality*

$$\sum_{i \in N} \left\lceil \frac{c_i}{c_j} \right\rceil x_i \geq \left\lceil \frac{b}{c_j} \right\rceil \tag{3}$$

is facet-defining for PX if and only if $\big(\lceil \frac{b}{c_j} \rceil - 1\big)c_j + c_1 \geq b$ and $\big(\lceil \frac{b}{c_j} \rceil - \lceil \frac{c_i}{c_j} \rceil\big)c_j + c_i \geq b$ for all $i = j+1, \dots, n$.

Proof. Take $\lambda = c_j$. As $\lceil \frac{c_i}{c_j} \rceil \geq \frac{c_i}{c_j}$ for all $i = j+1, \dots, n$, we apply Corollary 2 to obtain the result. □

Atamturk [1] studies the polytope $conv(X \cap \{x \in \mathbb{Z}^n : x \leq u\})$ for $u \in \mathbb{Z}^n_{++}$ and proves that inequality (3) for $j \in N$ such that $u_j c_j \geq b$ is facet-defining if and only if the conditions of Corollary 4 are satisfied.

We go back to Example 1 and see if rounding inequalities are facet-defining.

Example 2. Consider set X^1 defined in Example 1. The rounding inequality for $\lambda = c_1$ is not facet-defining since $\big(\lceil \frac{14}{3} \rceil - \lceil \frac{4}{3} \rceil\big)3 + 4 = 13 < 14 = b$. The inequality is $x_1 + 2x_2 \geq 5$ and is dominated by $2x_1 + 3x_2 \geq 10$. We can obtain the latter inequality by lifting inequality $x_1 \geq 5$, which is a rounding inequality when $x_2 = 0$ with variable x_2 (see section 5).

The rounding inequality for $\lambda = c_2$ is facet-defining since $\big(\lceil \frac{14}{4} \rceil - 1\big)4 + 3 = 15 \geq 14 = b$. This is the inequality $x_1 + x_2 \geq 4$.

The convex hull of X^1 is described by the nonnegativity constraints, a rounding inequality ($x_1 + x_2 \geq 4$), and a lifted rounding inequality ($2x_1 + 3x_2 \geq 10$).

In the next example, we see two sets that are defined by parameters which differ only in the right-hand side of the cover constraint. The rounding inequalities for $\lambda = c_2, c_3, \dots, c_n$ are facet-defining for the polyhedron when the right-hand side is b, and none are facet-defining when the right-hand side is $b+1$.

Example 3. Consider the set $X^2 = \{x \in \mathbb{Z}_+^4 : x_1 + 4x_2 + 5x_3 + 6x_4 \geq 61\}$. The convex hull of X^2 is described by the nonnegativity constraints and the following inequalities (these results are obtained using PORTA [6]):

$$x_1 + 4x_2 + 5x_3 + 6x_4 \geq 61, \tag{4}$$

$$x_1 + 2x_2 + 3x_3 + 3x_4 \geq 31, \tag{5}$$

$$x_1 + x_2 + 2x_3 + 2x_4 \geq 16, \tag{6}$$

$$x_1 + x_2 + x_3 + 2x_4 \geq 13, \tag{7}$$

$$x_1 + x_2 + x_3 + x_4 \geq 11. \tag{8}$$

Inequality (4) is the cover constraint. By Corollary 1, as $c_1 = 1$, we know that the cover constraint is facet-defining. Inequalities (6)–(8) are rounding inequalities. It is easy to verify that the conditions of Corollary 4 are satisfied. Note that inequality (5) is the rounding inequality for $\lambda = 2$, and the conditions of Corollary 3 are satisfied.

Now consider the set $X^3 = \{x \in \mathbb{Z}_+^4 : x_1 + 4x_2 + 5x_3 + 6x_4 \geq 62\}$. The following inequalities together with the nonnegativity constraints describe the convex hull of X^3:

$$x_1 + 4x_2 + 5x_3 + 6x_4 \geq 62, \tag{9}$$

$$x_1 + 2x_2 + 3x_3 + 4x_4 \geq 32, \tag{10}$$

$$x_1 + 2x_2 + 2x_3 + 3x_4 \geq 26, \tag{11}$$

$$x_1 + 2x_2 + 2x_3 + 2x_4 \geq 22. \tag{12}$$

The cover constraint (9) is facet-defining, but the rounding inequalities for $\lambda = c_2, c_3, c_4$ do not define facets. Inequality (10) dominates the rounding inequality for $\lambda = c_2$, which is $x_1 + x_2 + 2x_3 + 2x_4 \geq 16$, (11) dominates inequality $x_1 + x_2 + x_3 + 2x_4 \geq 13$, which is the rounding inequality for $\lambda = c_3$, and (12) dominates $x_1 + x_2 + x_3 + x_4 \geq 11$, which is the rounding inequality for $\lambda = c_4$. In the following section, we will identify these inequalities (10)–(12).

4. Residual capacity inequalities.

4. Residual capacity inequalities. Residual capacity inequalities are introduced by Magnanti, Mirchandani, and Vachani [10] for the single arc design problem. Here we present inequalities that are based on a similar idea.

Assume that the demand b is covered using some item $j \in N$. Then at least $\lceil \frac{b}{c_j} \rceil$ units of item j need to be used. If $\lceil \frac{b}{c_j} \rceil - 1$ units are used to full capacity, then the capacity of the last unit to be used is $r_j = b - (\lceil \frac{b}{c_j} \rceil - 1)c_j$. If only $\lceil \frac{b}{c_j} \rceil - 1$ units of item j are used, then the remaining items should cover a demand equal to r_j. This is expressed in the following valid inequality.

For $j \in N$, define $N_j = \{1, 2, \ldots, j\}$ and $N_j^{'} = \{i \in N_j : c_i \geq r_j\}$. For $N^0 \subset N$ and $N^1 = N \setminus N^0$, let $X_h(N^1) = \{x \in \mathbb{Z}_+^n : \sum_{i \in N} c_i x_i \geq h, x_i = 0 \text{ for all } i \in N^0\}$.

THEOREM 3. *For $j \in N$, the inequality*

$$\sum_{i=1}^{j} \min\{c_i, r_j\} x_i + \sum_{i=j+1}^{n} c_i x_i \geq r_j \left\lceil \frac{b}{c_j} \right\rceil \tag{13}$$

is valid for PX.

Proof. If $\sum_{i \in N_j'} x_i = \lceil \frac{b}{c_j} \rceil$, then the inequality is satisfied. If $\sum_{i \in N_j'} x_i = \lceil \frac{b}{c_j} \rceil - p$ for some $p \geq 1$, then the feasibility of x implies $\sum_{i \in N_j \setminus N_j'} c_i x_i + \sum_{i=j+1}^{n} c_i x_i \geq$

$b-\sum_{i\in N_j'} c_i x_i \geq b - c_j \sum_{i\in N_j'} x_i = r_j + (p-1)c_j$. As $r_j+(p-1)c_j \geq r_j p$, inequality (13) is satisfied. □

For $j \in N$, if $r_j = c_j$, then b is divisible by c_j and inequality (13) is the same as the cover constraint.

Theorem 4. *If $c_1 = 1$ for $j \in N$, the inequality*

$$\sum_{i=1}^{j} \min\{c_i, r_j\} x_i \geq r_j \left\lceil \frac{b}{c_j} \right\rceil \tag{14}$$

is facet-defining for conv$(X_b(N_j))$.

Proof. Let $F = \{x \in X_b(N_j) : \sum_{i=1}^{j} \min\{c_i, r_j\} x_i = r_j \lceil \frac{b}{c_j} \rceil\}$. Assume that all $x \in F$ satisfy $\sum_{i=1}^{j} \alpha_i x_i = \alpha_0$. As $\lceil \frac{b}{c_j} \rceil e_j \in F$, we need $\alpha_0 = \lceil \frac{b}{c_j} \rceil \alpha_j$. For $i \in N_j'$, $(\lceil \frac{b}{c_j} \rceil - 1) e_j + e_i \in F$, implying that $\alpha_i = \alpha_j$. As $c_1 = 1$, we have $(\lceil \frac{b}{c_j} \rceil - 1) e_j + r_j e_1 \in F$. So $\alpha_1 = \frac{\alpha_j}{r_j}$. Finally, for $i \in N_j \setminus (N_j' \cup \{1\})$, $(\lceil \frac{b}{c_j} \rceil - 1) e_j + e_i + (r_j - c_i) e_1 \in F$. Hence, $\alpha_i = \frac{\alpha_j c_i}{r_j}$. Then $\sum_{i=1}^{j} \alpha_i x_i = \alpha_0$ is a $\frac{\alpha_j}{r_j}$ multiple of $\sum_{i=1}^{j} \min\{c_i, r_j\} x_i = r_j \lceil \frac{b}{c_j} \rceil$. □

For $j \in N$, if $r_j = 1$, then inequality (14) is $\sum_{i=1}^{j} x_i \geq \lceil \frac{b}{c_j} \rceil$ and is the same as the rounding inequality for $\lambda = c_j$ for *conv*$(X_b(N_j))$. By Corollary 4, it is facet-defining since $\left(\lceil \frac{b}{c_j} \rceil - 1\right) c_j + c_1 = b - r_j + c_1 \geq b$.

For $j = n$, *conv*$(X_b(N_n)) = PX$, and the following result can be deduced from Theorem 4.

Corollary 5. *If $c_1 = 1$, inequality (13) for $j = n$ is facet-defining for PX.*

Example 4. Consider the set X^3 given in Example 3. For item 2, $r_2 = 2$ and $\lceil \frac{b}{c_2} \rceil = 16$. Inequality (13) for item 2 is $x_1 + 2x_2 + 5x_3 + 6x_4 \geq 32$ and is dominated by inequality (10). For item 3, $r_3 = 2$ and $\lceil \frac{b}{c_3} \rceil = 13$. The corresponding inequality (13) is $x_1 + 2x_2 + 2x_3 + 6x_4 \geq 26$ and is dominated by inequality (11). For item 4, $r_4 = 2$ and $\lceil \frac{b}{c_4} \rceil = 11$. Inequality (13) is $x_1 + 2x_2 + 2x_3 + 2x_4 \geq 22$ and is the same as inequality (12). In the remaining of this section, we will try to identify inequalities (10) and (11).

We can generalize inequality (13) as follows.

Theorem 5. *For $j \in N$, let $\mu \geq 0$ be such that $\lceil \frac{r_j(r_j+\mu)}{c_j} + \mu \rceil \geq r_j$ and $r_j + \mu \leq c_j$. The inequality*

$$\sum_{i=1}^{j} \min\{c_i, r_j\} x_i + \sum_{i=j+1}^{n} \left\lceil \frac{c_i(r_j+\mu)}{c_j} \right\rceil x_i \geq r_j \left\lceil \frac{b}{c_j} \right\rceil \tag{15}$$

is valid for PX.

Proof. If $\sum_{i\in N_j'} x_i = \lceil \frac{b}{c_j} \rceil$, then the inequality is satisfied. If $\sum_{i\in N_j'} x_i = \lceil \frac{b}{c_j} \rceil - 1$, then inequality (15) simplifies to $\sum_{i\in N_j \setminus N_j'} c_i x_i + \sum_{i=j+1}^{n} \lceil \frac{c_i(r_j+\mu)}{c_j} \rceil x_i \geq r_j$. By feasibility, we need to have $\sum_{i\in N_j \setminus N_j'} c_i x_i + \sum_{i=j+1}^{n} c_i x_i \geq r_j$. Using coefficient reduction, we obtain $\sum_{i\in N_j \setminus N_j'} c_i x_i + \sum_{i=j+1}^{n} r_j x_i \geq r_j$. As $\lceil \frac{c_i(r_j+\mu)}{c_j} \rceil \geq r_j$ for all $i = j+1, \ldots, n$, inequality (15) is satisfied.

If $\sum_{i\in N_j'} x_i = \lceil \frac{b}{c_j} \rceil - p$ for some $p \geq 2$, then inequality (15) simplifies to $\sum_{i\in N_j \setminus N_j'} c_i x_i + \sum_{i=j+1}^{n} \lceil \frac{c_i(r_j+\mu)}{c_j} \rceil x_i \geq r_j p$. The feasibility of x implies that $\sum_{i\in N_j \setminus N_j'} c_i x_i + \sum_{i=j+1}^{n} c_i x_i \geq r_j + (p-1)c_j$. We multiply this inequality with $\frac{r_j+\mu}{c_j}$

and obtain $\sum_{i\in N_j\setminus N_j'} c_i\frac{r_j+\mu}{c_j}x_i+\sum_{i=j+1}^{n} c_i\frac{r_j+\mu}{c_j}x_i \geq \frac{r_j(r_j+\mu)}{c_j}+(p-1)(r_j+\mu)$. Now, as $r_j+\mu\leq c_j$ and so $\sum_{i\in N_j\setminus N_j'} c_ix_i+\sum_{i=j+1}^{n}\lceil\frac{c_i(r_j+\mu)}{c_j}\rceil x_i\geq\sum_{i\in N_j\setminus N_j'} c_i\frac{(r_j+\mu)}{c_j}x_i+\sum_{i=j+1}^{n} c_i\frac{(r_j+\mu)}{c_j}x_i$, we have $\sum_{i\in N_j\setminus N_j'} c_ix_i+\sum_{i=j+1}^{n}\lceil\frac{c_i(r_j+\mu)}{c_j}\rceil x_i \geq \frac{r_j(r_j+\mu)}{c_j}+(p-1)(r_j+\mu)$. Since the left-hand side is always an integer, we round up the right-hand side and get $\lceil\frac{r_j(r_j+\mu)}{c_j}+(p-1)\mu\rceil+(p-1)r_j$. As $\lceil\frac{r_j(r_j+\mu)}{c_j}+\mu\rceil\geq r_j$, $\mu\geq 0$, and $p\geq 2$, we obtain $\sum_{i\in N_j\setminus N_j'} c_ix_i+\sum_{i=j+1}^{n}\lceil\frac{c_i(r_j+\mu)}{c_j}\rceil x_i\geq r_jp$. So x satisfies inequality (15). $\square$

For $\mu=c_j-r_j$, inequality (15) is the same as inequality (13).

As μ increases, inequality (15) gets weaker. So for given $j\in N$, we are interested in inequality (15) defined by the smallest μ that satisfies the condition $\lceil\frac{r_j(r_j+\mu)}{c_j}+\mu\rceil\geq r_j$. Let $\epsilon>0$ be very small. We take $\mu_j=\frac{c_j(r_j-1)-r_j^2}{r_j+c_j}+\epsilon$, if $\lceil\frac{r_j^2}{c_j}\rceil<r_j$, and $\mu_j=0$, otherwise.

Observe that nondominated residual capacity inequalities (15) are defined per item, so there are $O(n)$ of them.

Example 5. Consider again the set X^3 of Example 3. For item 2, $r_2=2$. As $\lceil\frac{r_2^2}{c_2}\rceil=1<2=r_2$, $\mu_2=\frac{4(2-1)-4}{2+4}+\epsilon=\epsilon$. The corresponding inequality (15) is $x_1+2x_2+3x_3+4x_4\geq 32$ and is the same as inequality (10). For item 3, $r_3=2$. As $\lceil\frac{r_3^2}{c_3}\rceil=1<2=r_3$, $\mu_3=\frac{5(2-1)-4}{2+5}+\epsilon=\frac{1}{7}+\epsilon$. The corresponding inequality (15) is $x_1+2x_2+2x_3+3x_4\geq 26$ and is the same as inequality (11).

If $r_j=1$, then $\mu_j=0$ and inequality (15) is the same as the rounding inequality (3) for $\lambda=c_j$.

If $r_j=c_j$, then again $\mu_j=0$. This time inequality (15) is the same as the cover constraint.

We have a necessary condition for inequality (15) to be facet-defining.

COROLLARY 6. *For $j\in N$, if inequality* (15) *is facet-defining for PX and $r_j<c_j$, then $c_i+\lceil\frac{b}{c_j}\rceil c_j-\frac{c_j}{r_j}\lceil\frac{c_i(r_j+\mu_j)}{c_j}\rceil\geq b$ for all $i=j+1,\dots,n$.*

Proof. As $c_i-\frac{c_j}{r_j}\min\{c_i,r_j\}\leq 0$ for all $i=1,\dots,j-1$ and $\big(c_i-\frac{c_j}{r_j}\lceil\frac{c_i(r_j+\mu_j)}{c_j}\rceil\big)\leq 0$ for all $i=j+1,\dots,n$, we apply Theorem 1. So, if inequality (15) is facet-defining for PX, then $\lceil\frac{b}{c_j}\rceil c_j-\min\{c_i,r_j\}\frac{c_j}{r_j}+c_i\geq b$ for $i=1,\dots,j-1$ and $\lceil\frac{b}{c_j}\rceil c_j-\frac{c_j}{r_j}\lceil\frac{c_i(r_j+\mu_j)}{c_j}\rceil+c_i\geq b$ for all $i=j+1,\dots,n$.

For $i\in N_j^{'}$, the condition is $\lceil\frac{b}{c_j}\rceil c_j-c_j+c_i\geq b$. The left-hand side is equal to $\lfloor\frac{b}{c_j}\rfloor c_j+c_i\geq\lfloor\frac{b}{c_j}\rfloor c_j+r_j=b$. For $i\in N_j\setminus N_j^{'}$, the condition is $\lceil\frac{b}{c_j}\rceil c_j-c_i\frac{c_j}{r_j}+c_i\geq b$. The left-hand side is equal to $b-r_j+c_j-c_i\frac{c_j-r_j}{r_j}=b+(c_j-r_j)\frac{(r_j-c_i)}{r_j}\geq b$ since $c_j\geq r_j$ and $r_j\geq c_i$. So the conditions of Theorem 1 are always satisfied for $i\in N_j$. $\square$

5. Lifted rounding inequalities. In this section, we derive valid inequalities using lifting. For $N^0\subset N$ and $N^1=N\setminus N^0$, let $\sum_{i\in N^1}\alpha_ix_i\geq\alpha_0$ be a valid inequality for $X_b(N^1)$.

Suppose we lift inequality $\sum_{i\in N^1}\alpha_ix_i\geq\alpha_0$, with x_l with $l\in N^0$. The optimal lifting coefficient of x_l is

$$\begin{aligned}\alpha_l=\max&\frac{\alpha_0-\sum_{i\in N^1}\alpha_ix_i}{x_l}\\ \text{s.t. }& x_l\geq 1\\ & x\in X_b(N^1\cup\{l\}).\end{aligned}$$

Consider the case where $\alpha_i = 1$ for all $i \in N^1$, $j = \arg\max_{i \in N^1} c_i$, and $\alpha_0 = \lceil \frac{b}{c_j} \rceil$. For $l \in N^0$, the nonlinear lifting problem simplifies to

$$\alpha_l = \max_{x_l \in \mathbb{Z}_{++}} \frac{\lceil \frac{b}{c_j} \rceil - \lceil \frac{(b - c_l x_l)^+}{c_j} \rceil}{x_l}.$$

Clearly, a maximizing x_l cannot be larger than $\lceil \frac{b}{c_l} \rceil$. Hence, we obtain

$$\alpha_l = \max_{x_l \in \{1,2,\ldots,\lceil \frac{b}{c_l} \rceil\}} \frac{\lceil \frac{b}{c_j} \rceil - \lceil \frac{(b - c_l x_l)^+}{c_j} \rceil}{x_l},$$

and we can compute α_l by enumeration.

Example 6. Consider the set X^1 defined in Example 1. Inequality $x_1 \geq 5$ is facet-defining for *conv*$(X^1 \cap \{x \in \mathbb{Z}^2_+ : x_2 = 0\})$. We lift inequality $x_1 \geq 5$ with variable x_2. The optimal lifting coefficient $\alpha_2 = \max_{x_2 \in \{1,2,3,4\}} \frac{5 - \lceil \frac{(14 - 4x_2)^+}{3} \rceil}{x_2} = \max\{1, \frac{3}{2}, \frac{4}{3}, \frac{5}{4}\} = \frac{3}{2}$. The corresponding inequality is $2x_1 + 3x_2 \geq 10$ and is facet-defining for *conv*(X^1).

Computation of the optimal lifting coefficients of variables that are lifted in later in the sequence may become harder. So we are interested in sequence-independent lifting.

Atamturk [4] studies sequence-independent lifting for mixed integer programming. The following can be derived from his results. Consider the lifting function $\Phi(a) = \alpha_0 - \min_{x \in X_{b-a}(N^1)} \sum_{i \in N^1} \alpha_i x_i$. If this function is subadditive, i.e., if $\Phi(a) + \Phi(d) \geq \Phi(a + d)$ for all $a, d \in \mathbb{R}$, then the lifting is sequence-independent. In this case, the inequality $\sum_{i \in N^1} \alpha_i x_i + \sum_{i \in N^0} \Phi(c_i) x_i \geq \alpha_0$ is a valid inequality for PX. In the general case, let Θ be a subadditive function, with $\Theta \geq \Phi$. Then the inequality $\sum_{i \in N^1} \alpha_i x_i + \sum_{i \in N^0} \Theta(c_i) x_i \geq \alpha_0$ is a valid inequality for PX. If the inequality $\sum_{i \in N^1} \alpha_i x_i \geq \alpha_0$ is facet-defining for *conv*$(X_b(N^1))$ and $\Theta(c_i) = \Phi(c_i)$ for all $i \in N^0$, then inequality $\sum_{i \in N^1} \alpha_i x_i + \sum_{i \in N^0} \Theta(c_i) x_i \geq \alpha_0$ is facet-defining for PX.

THEOREM 6. *Let $N^1 \subset N$ and $\sum_{i \in N^1} \alpha_i x_i \geq \alpha_0$ be a valid inequality for $X_b(N^1)$. If there exists $j \in N^1$ such that $\alpha_i \geq \alpha_j \lceil \frac{c_i}{c_j} \rceil$ for all $i \in N^1 \setminus \{j\}$, then the lifting function is*

$$\Phi(a) = \alpha_0 - \alpha_j \left\lceil \frac{(b - a)^+}{c_j} \right\rceil.$$

Proof. Suppose there exists $j \in N^1$ such that $\alpha_i \geq \alpha_j \lceil \frac{c_i}{c_j} \rceil$ for all $i \in N^1 \setminus \{j\}$. The lifting function is $\Phi(a) = \alpha_0 - \min_{x \in X_{b-a}(N^1)} \sum_{i \in N^1} \alpha_i x_i$. Let x be an optimal solution to the minimization problem. Consider $\overline{x} = x - \sum_{i \in N^1 \setminus \{j\}} x_i e_i + \lceil \frac{\sum_{i \in N^1 \setminus \{j\}} c_i x_i}{c_j} \rceil e_j$. Clearly, $\overline{x} \in X_{b-a}(N^1)$. The objective function evaluated at $\overline{x}$ is equal to

$$\begin{aligned}\sum_{i \in N^1} \alpha_i \overline{x}_i &= \sum_{i \in N^1} \alpha_i x_i - \sum_{i \in N^1 \setminus \{j\}} \alpha_i x_i + \alpha_j \left\lceil \frac{\sum_{i \in N^1 \setminus \{j\}} c_i x_i}{c_j} \right\rceil \\ &\leq \sum_{i \in N^1} \alpha_i x_i - \sum_{i \in N^1 \setminus \{j\}} \alpha_i x_i + \alpha_j \sum_{i \in N^1 \setminus \{j\}} \left\lceil \frac{c_i}{c_j} \right\rceil x_i.\end{aligned}$$

As $\alpha_i \geq \alpha_j \lceil \frac{c_i}{c_j} \rceil$ for all $i \in N^1 \setminus \{j\}$, $\sum_{i \in N^1} \alpha_i \overline{x}_i \leq \sum_{i \in N^1} \alpha_i x_i$, and so $\overline{x}$ is also optimal. Hence $\lceil \frac{(b-a)^+}{c_j} \rceil e_j$ is also optimal and the optimal value is $\alpha_j \lceil \frac{(b-a)^+}{c_j} \rceil$. □

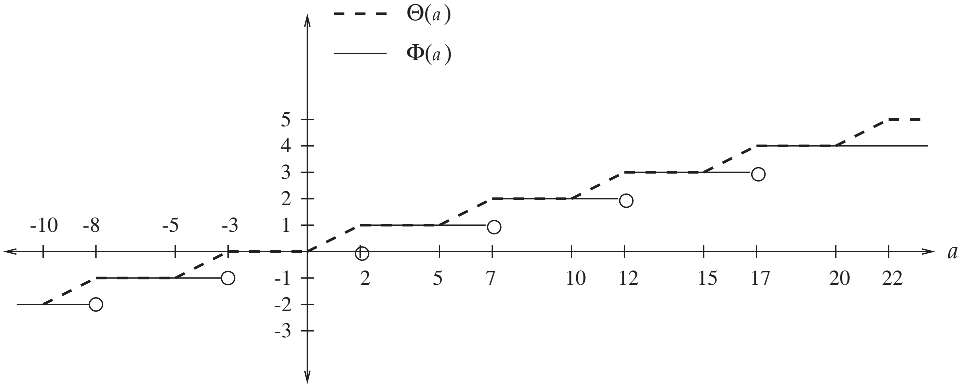


FIG. 1. *Lifting function Φ and subadditive function Θ for $b = 17$ and $c_j = 5$.*

Suppose there exists $j \in N^1$ such that $\alpha_i \geq \alpha_j \lceil \frac{c_i}{c_j} \rceil$ for all $i \in N^1 \setminus \{j\}$, $\alpha_j = 1$, and $\alpha_0 = \lceil \frac{b}{c_j} \rceil$. The lifting function for the inequality $\sum_{i \in N^1} \alpha_i x_i \geq \lceil \frac{b}{c_j} \rceil$ is $\Phi(a) = \lceil \frac{b}{c_j} \rceil - \lceil \frac{(b-a)^+}{c_j} \rceil$. The function Φ is not subadditive. An example where $b = 17$ and $c_j = 5$ is depicted in Figure 1. Here for $a = 11$ and $d = 6$, we have $\lceil \frac{b}{c_j} \rceil - \lceil \frac{b-a}{c_j} \rceil + \lceil \frac{b}{c_j} \rceil - \lceil \frac{b-d}{c_j} \rceil = 4 - 2 + 4 - 3 = 3 < \lceil \frac{b}{c_j} \rceil - \lceil \frac{b-a-d}{c_j} \rceil = 4 - 0 = 4$.

For $j \in N$ and $a \in \mathbb{R}$, define

$$\rho_j(a) = a - \left\lfloor \frac{a}{c_j} \right\rfloor c_j.$$

LEMMA 1. *For $j \in N$, if $\rho_j(b) > 0$, the function $\Theta(a) = \lfloor \frac{a}{c_j} \rfloor + \min\{\frac{\rho_j(a)}{\rho_j(b)}, 1\}$ (see Figure 1) is subadditive.*

Proof. Let $a, d \in \mathbb{R}$. Then $\Theta(a) + \Theta(d) = \lfloor \frac{a}{c_j} \rfloor + \min\{\frac{\rho_j(a)}{\rho_j(b)}, 1\} + \lfloor \frac{d}{c_j} \rfloor + \min\{\frac{\rho_j(d)}{\rho_j(b)}, 1\}$. There are two cases: (i) $\rho_j(a) + \rho_j(d) = \rho_j(a + d)$ and (ii) $\rho_j(a) + \rho_j(d) = \rho_j(a + d) + c_j$. In case (i), since $\rho_j(a) + \rho_j(d) = \rho_j(a + d)$, we have $\lfloor \frac{a}{c_j} \rfloor + \lfloor \frac{d}{c_j} \rfloor = \lfloor \frac{a+d}{c_j} \rfloor$. If $\min\{\frac{\rho_j(a)}{\rho_j(b)}, 1\} = 1$ or $\min\{\frac{\rho_j(d)}{\rho_j(b)}, 1\} = 1$, then $\Theta(a) + \Theta(d) \geq \lfloor \frac{a+d}{c_j} \rfloor + 1 \geq \Theta(a + d)$. Otherwise, $\min\{\frac{\rho_j(a)}{\rho_j(b)}, 1\} = \frac{\rho_j(a)}{\rho_j(b)}$ and $\min\{\frac{\rho_j(a)}{\rho_j(b)}, 1\} = \frac{\rho_j(a)}{\rho_j(b)}$. Then $\Theta(a) + \Theta(d) = \lfloor \frac{a+d}{c_j} \rfloor + \frac{\rho_j(a)}{\rho_j(b)} + \frac{\rho_j(d)}{\rho_j(b)} = \lfloor \frac{a+d}{c_j} \rfloor + \frac{\rho_j(a+d)}{\rho_j(b)} \geq \Theta(a + d)$. In case (ii), as $\rho_j(a) + \rho_j(d) = \rho_j(a + d) + c_j$, $\lfloor \frac{a}{c_j} \rfloor + \lfloor \frac{d}{c_j} \rfloor = \lfloor \frac{a+d}{c_j} \rfloor - 1$. If $\min\{\frac{\rho_j(a)}{\rho_j(b)}, 1\} = 1$ and $\min\{\frac{\rho_j(d)}{\rho_j(b)}, 1\} = 1$, then $\Theta(a) + \Theta(d) = \lfloor \frac{a+d}{c_j} \rfloor + 1 \geq \Theta(a + d)$. If $\min\{\frac{\rho_j(a)}{\rho_j(b)}, 1\} = \frac{\rho_j(a)}{\rho_j(b)}$ and $\min\{\frac{\rho_j(d)}{\rho_j(b)}, 1\} = 1$, then $\Theta(a) + \Theta(d) = \lfloor \frac{a+d}{c_j} \rfloor + \frac{\rho_j(a)}{\rho_j(b)}$. Since $\rho_j(d) \leq c_j$, $\rho_j(a) \geq \rho_j(a+d)$. So $\lfloor \frac{a+d}{c_j} \rfloor + \frac{\rho_j(a)}{\rho_j(b)} \geq \Theta(a+d)$. The case where $\min\{\frac{\rho_j(a)}{\rho_j(b)}, 1\} = 1$ and $\min\{\frac{\rho_j(d)}{\rho_j(b)}, 1\} = \frac{\rho_j(d)}{\rho_j(b)}$ is similar. Finally, if $\min\{\frac{\rho_j(a)}{\rho_j(b)}, 1\} = \frac{\rho_j(a)}{\rho_j(b)}$ and $\min\{\frac{\rho_j(a)}{\rho_j(b)}, 1\} = \frac{\rho_j(a)}{\rho_j(b)}$, $\Theta(a) + \Theta(d) = \lfloor \frac{a+d}{c_j} \rfloor - 1 + \frac{\rho_j(a)}{\rho_j(b)} + \frac{\rho_j(d)}{\rho_j(b)} = \lfloor \frac{a+d}{c_j} \rfloor - 1 + \frac{\rho_j(a+d)}{\rho_j(b)} + \frac{c_j}{\rho_j(b)}$. Since $c_j \geq \rho_j(b)$, $\lfloor \frac{a+d}{c_j} \rfloor - 1 + \frac{\rho_j(a+d)}{\rho_j(b)} + \frac{c_j}{\rho_j(b)} \geq \lfloor \frac{a+d}{c_j} \rfloor + \frac{\rho_j(a+d)}{\rho_j(b)} \geq \Theta(a + d)$. This proves that Θ is subadditive. □

Now we will lift the inequality $\sum_{i \in N^1} \alpha_i x_i \geq \lceil \frac{b}{c_j} \rceil$ using the function Θ.

THEOREM 7. *Let $N^0 \subset N$, $N^1 = N \setminus N^0$, and $\sum_{i \in N^1} \alpha_i x_i \geq \alpha_0$ be a valid inequality for $X_b(N^1)$. If there exists $j \in N^1$ such that $\alpha_j = 1$, $\alpha_i \geq \lceil \frac{c_i}{c_j} \rceil$ for all*

$i \in N^1 \setminus \{j\}$, $\alpha_0 = \lceil \frac{b}{c_j} \rceil$, and $\rho_j(b) > 0$, then the inequality

$$(16) \qquad \sum_{i \in N^1} \rho_j(b) \alpha_i x_i + \sum_{i \in N^0} \left(\rho_j(b) \left\lfloor \frac{c_i}{c_j} \right\rfloor + \min\{\rho_j(c_i), \rho_j(b)\} \right) x_i \geq \rho_j(b) \left\lceil \frac{b}{c_j} \right\rceil$$

is a valid inequality for PX.

Proof. The inequality $\sum_{i \in N^1} \alpha_i x_i \geq \lceil \frac{b}{c_j} \rceil$ is valid for $X_b(N_1)$. Consider the subadditive function $\Theta(a) = \lfloor \frac{a}{c_j} \rfloor + \min\{\frac{\rho_j(a)}{\rho_j(b)}, 1\}$ given in Lemma 1. We will show that $\Theta \geq \Phi$. If $a < b$ and $\rho_j(a) < \rho_j(b)$, then $\rho_j(b-a) = \rho_j(b) - \rho_j(a) > 0$. So $\Phi(a) = \lceil \frac{b}{c_j} \rceil - \lceil \frac{b-a}{c_j} \rceil = \frac{b - \rho_j(b) + c_j}{c_j} - \frac{b - a - \rho_j(b) + \rho_j(a) + c_j}{c_j} = \frac{a - \rho_j(a)}{c_j} = \lfloor \frac{a}{c_j} \rfloor \leq \Theta(a)$. If $a < b$ and $\rho_j(a) \geq \rho_j(b)$, then $\Theta(a) = \lceil \frac{a}{c_j} \rceil \geq \lceil \frac{b}{c_j} \rceil - \lceil \frac{b-a}{c_j} \rceil = \Phi(a)$. If $a \geq b$, then $\Phi(a) = \lceil \frac{b}{c_j} \rceil$. If $\lceil \frac{a}{c_j} \rceil = \lceil \frac{b}{c_j} \rceil$, then $\rho_j(a) \geq \rho_j(b)$. So $\Theta(a) = \lceil \frac{a}{c_j} \rceil = \Phi(a)$. If $\lceil \frac{a}{c_j} \rceil \geq \lceil \frac{b}{c_j} \rceil + 1$, then $\Theta(a) \geq \lfloor \frac{a}{c_j} \rfloor \geq \lceil \frac{b}{c_j} \rceil = \Phi(a)$. So the inequality $\sum_{i \in N^1} \alpha_i x_i + \sum_{i \in N^0} \big(\lfloor \frac{c_i}{c_j} \rfloor + \min\{\frac{\rho_j(c_i)}{\rho_j(b)}, 1\} \big) x_i \geq \lceil \frac{b}{c_j} \rceil$ is a valid inequality for PX. Multiplying both sides with $\rho_j(b)$, we obtain inequality (16). □

Some of the inequalities (16) are dominated by others. Indeed, as given in the following proposition, the number of nondominated inequalities (16) is polynomial.

Proposition 2. *For $j \in N$ with $\rho_j(b) > 0$, the inequality*

$$(17) \quad \sum_{i=1}^{j} \min\{c_i, \rho_j(b)\} x_i + \sum_{i=j+1}^{n} \left(\rho_j(b) \left\lfloor \frac{c_i}{c_j} \right\rfloor + \min\{\rho_j(c_i), \rho_j(b)\} \right) x_i \geq \rho_j(b) \left\lceil \frac{b}{c_j} \right\rceil$$

is valid and dominates inequality (16) *for $N^0 \subset N$, $N^1 = N \setminus N^0$ such that $j \in N^1$, $\alpha_j = 1$, $\alpha_i \geq \lceil \frac{c_i}{c_j} \rceil$ for all $i \in N^1 \setminus \{j\}$ and $\alpha_0 = \lceil \frac{b}{c_j} \rceil$.*

Proof. Inequality (17) is valid since it is the same as inequality (16) for $N^1 = \{j\}$.

Let $N^0 \subset N$, $N^1 = N \setminus N^0$ such that $j \in N^1$, $\alpha_j = 1$, $\alpha_i \geq \lceil \frac{c_i}{c_j} \rceil$ for all $i \in N^1 \setminus \{j\}$, and $\alpha_0 = \lceil \frac{b}{c_j} \rceil$. For $i \in N^1$, $\rho_j(b) \lfloor \frac{c_i}{c_j} \rfloor + \min\{\rho_j(c_i), \rho_j(b)\} \leq \rho_j(b) \lceil \frac{c_i}{c_j} \rceil \leq \rho_j(b) \alpha_i$. So the coefficient of x_i in (17) is less than or equal to its coefficient in (16). The coefficients of x_i for $i \in N^0$ and the right-hand sides are the same in both inequalities. Hence inequality (17) dominates inequality (16). □

We call inequalities (17) *lifted rounding inequalities.* The number of lifted rounding inequalities that are not dominated is $O(n)$.

It is interesting to note that even though inequalities (16) are not, inequalities (17) are special cases of the multifacility cut-set inequalities derived by Atamturk [2] for the single commodity-multifacility network design problem.

For $j \in N$ such that $\rho_j(b) > 0$, consider the inequality $x_j \geq \lceil \frac{b}{c_j} \rceil$, which is facet-defining for $conv(X_b(\{j\}))$. If $c_1 \geq \rho_j(b)$, then, for $i < j$, $c_i \geq \rho_j(b)$. So $\Phi(c_i) = \Theta(c_i) = 1$. For $i > j$, if $\rho_j(c_i) = 0$ or $\rho_j(c_i) \geq \rho_j(b)$, then $\Phi(c_i) = \Theta(c_i) = \lceil \frac{c_i}{c_j} \rceil$. By Theorem 5 in Atamturk [4], the resulting inequality

$$(18) \qquad \sum_{i=1}^{j} x_i + \sum_{i=j+1}^{n} \left\lceil \frac{c_i}{c_j} \right\rceil x_i \geq \left\lceil \frac{b}{c_j} \right\rceil$$

is facet-defining for PX. Notice that this is the same inequality as the rounding inequality (2) for $\lambda = c_j$. The condition $c_1 \geq \rho_j(b)$ implies that $\big(\lceil \frac{b}{c_j} \rceil - 1 \big) c_j + c_1 \geq b$. For $i < j$, if $\rho_j(c_i) = 0$, then $\big(\lceil \frac{b}{c_j} \rceil - \lceil \frac{c_i}{c_j} \rceil \big) c_j + c_i = \lceil \frac{b}{c_j} \rceil c_j \geq b$. If $\rho_j(c_i) \geq \rho_j(b)$, then $\big(\lceil \frac{b}{c_j} \rceil - \lceil \frac{c_i}{c_j} \rceil\big)c_j + c_i = \big(\frac{b+c_j-\rho_j(b)}{c_j} - \frac{c_i+c_j-\rho_j(c_i)}{c_j}\big)c_j + c_i = b - \rho_j(b) + \rho_j(c_i) \geq b$. As a result, the conditions stated above are the same as the conditions of Corollary 4. However, Corollary 4 is a stronger result, since it states that these conditions are both necessary and sufficient.

Now we compare inequalities (17) and (3). The two following propositions are easy to prove.

PROPOSITION 3. *For $j \in N$ with $\rho_j(b) = 1$, inequalities* (17) *and* (3) *are the same.*

PROPOSITION 4. *For $j \in N$ with $\rho_j(b) \geq 2$, inequality* (17) *dominates inequality* (3).

If, for $j \in N$, $\rho_j(b) > 0$ (or, equivalently, $r_j < c_j$), then $\rho_j(b) = r_j$. So residual capacity inequalities (15) and inequalities (17) look very similar. Coefficients of variables x_i, with $i \in \{1, \dots, j\}$, are the same in both inequalities. The right-hand sides are also the same. Only coefficients of variables x_i, with $i \in \{j+1, \dots, n\}$, may be different.

PROPOSITION 5. *For $j \in N$, if $r_j < c_j$ and $\lceil \frac{r_j^2}{c_j} \rceil \geq r_j$, then inequality* (15) *for $\mu = 0$ and inequality* (17) *are the same.*

Proof. If $\lceil \frac{r_j^2}{c_j} \rceil \geq r_j$, then the coefficient of x_i, with $i \in \{j+1, \dots, n\}$, is $\lceil \frac{c_i r_j}{c_j} \rceil$ in inequality (15) with $\mu = 0$. This is equal to

$$\left\lceil \frac{(\lfloor \frac{c_i}{c_j} \rfloor c_j + \rho_j(c_i))r_j}{c_j} \right\rceil = \left\lfloor \frac{c_i}{c_j} \right\rfloor r_j + \left\lceil \frac{\rho_j(c_i)r_j}{c_j} \right\rceil.$$

Since $\rho_j(c_i) \leq c_j$ and $r_j \leq c_j$, $\lceil \frac{\rho_j(c_i)r_j}{c_j} \rceil \leq \min\{\rho_j(c_i), r_j\}$. So the coefficient of x_i in (15) is less than or equal to its coefficient in (17).

If $\rho_j(c_i) \geq r_j$, then $\lceil \frac{\rho_j(c_i)r_j}{c_j} \rceil \geq \lceil \frac{r_j^2}{c_j} \rceil \geq r_j$. Now assume that $\rho_j(c_i) < r_j$ and $\lceil \frac{\rho_j(c_i)r_j}{c_j} \rceil < \rho_j(c_i)$. Then $\rho_j(c_i)r_j \leq (\rho_j(c_i) - 1)c_j$. This is equivalent to $c_j \leq (c_j - r_j)\rho_j(c_i)$. Since $\lceil \frac{r_j^2}{c_j} \rceil \geq r_j$, $r_j^2 > (r_j - 1)c_j$. So $c_j > (c_j - r_j)r_j > (c_j - r_j)\rho_j(c_i)$. This contradicts $c_j \leq (c_j - r_j)\rho_j(c_i)$. Hence if $\rho_j(c_i) < r_j$, then $\lceil \frac{\rho_j(c_i)r_j}{c_j} \rceil \geq \rho_j(c_i)$. So the coefficients of variable x_i in inequalities (15) and (17) are the same. □

PROPOSITION 6. *For $j \in N$, if $\lceil \frac{r_j^2}{c_j} \rceil < r_j$, then inequality* (17) *dominates inequality* (15) *for $\mu = \mu_j$.*

Proof. If $\lceil \frac{r_j^2}{c_j} \rceil < r_j$, then the coefficient of x_i, with $i > j$, in (15) for $\mu = \mu_j$ is $\lceil \frac{c_i(r_j+\mu_j)}{c_j} \rceil = \lfloor \frac{c_i}{c_j} \rfloor r_j + \lceil \lfloor \frac{c_i}{c_j} \rfloor \mu_j + \frac{\rho_j(c_i)(r_j+\mu_j)}{c_j} \rceil$. If $\rho_j(c_i) \geq r_j$, then $\lceil \lfloor \frac{c_i}{c_j} \rfloor \mu_j + \frac{\rho_j(c_i)(r_j+\mu_j)}{c_j} \rceil \geq \lceil \lfloor \frac{c_i}{c_j} \rfloor \mu_j + \frac{r_j(r_j+\mu_j)}{c_j} \rceil$. Since $c_i \geq c_j$, $\lceil \lfloor \frac{c_i}{c_j} \rfloor \mu_j + \frac{r_j(r_j+\mu_j)}{c_j} \rceil \geq \lceil \mu_j + \frac{r_j(r_j+\mu_j)}{c_j} \rceil \geq r_j$.

Assume that $\rho_j(c_i) < r_j$ and $\lceil \lfloor \frac{c_i}{c_j} \rfloor \mu_j + \frac{\rho_j(c_i)(r_j+\mu_j)}{c_j} \rceil < \rho_j(c_i)$. Then $\rho_j(c_i)(r_j + \mu_j) \leq c_j(\rho_j(c_i) - 1 - \lfloor \frac{c_i}{c_j} \rfloor \mu_j)$ or, equivalently, $c_j \leq \rho_j(c_i)(c_j - r_j) - \mu_j c_i$. Since $\frac{r_j(r_j+\mu_j)}{c_j} + \mu_j > r_j - 1$, we have that $c_j > r_j(c_j - r_j - \mu_j) - \mu_j c_j$, and now, since $r_j > \rho_j(c_i)$, $c_j > \rho_j(c_i)(c_j - r_j - \mu_j) - \mu_j c_j$. Putting together with $c_j \leq \rho_j(c_i)(c_j - r_j) - \mu_j c_i$, we obtain $\rho_j(c_i)(c_j - r_j) - \mu_j c_i > \rho_j(c_i)(c_j - r_j - \mu_j) - \mu_j c_j$. This is equivalent to $\rho_j(c_i) + c_j > c_i$ since $\mu_j > 0$. But this is impossible. So if $\rho_j(c_i) < r_j$, then $\lceil \lfloor \frac{c_i}{c_j} \rfloor \mu_j + \frac{\rho_j(c_i)(r_j+\mu_j)}{c_j} \rceil \geq \rho_j(c_i)$. This proves that the coefficient of x_i in (15) is greater than or equal to its coefficient in (17). □

These four propositions show that, for $j \in N$ with $\rho_j(b) > 0$, the lifted rounding inequality (17) dominates the rounding inequality (2) for $\lambda = c_j$ and the residual capacity inequality (15) for $\mu = \mu_j$. For a special case, these inequalities (17) are facet-defining for PX.

THEOREM 8. *For $j \in N$ such that $\rho_j(b) > 0$, if $c_1 = 1$, then inequality* (17) *is facet-defining for PX.*

Proof. Suppose that $\rho_j(b) > 0$ and $c_1 = 1$. Assume that all points in X which satisfy inequality (17) at equality also satisfy $\sum_{i=1}^n \alpha_i x_i = \alpha_0$. The point $\lceil \frac{b}{c_j} \rceil e_j$ is in X and satisfies inequality (17) at equality. So $\alpha_0 = \alpha_j \lceil \frac{b}{c_j} \rceil$.

Notice that, if we remove one item j, the remaining demand to be covered is $\rho_j(b)$. For $i < j$, if $c_i > \rho_j(b)$, then consider the point $e_i + (\lceil \frac{b}{c_j} \rceil - 1)e_j$. It is easy to verify that this point is also in X and that inequality (17) is tight at this point. Then we have $\alpha_i = \alpha_j$.

For $i < j$, if $c_i \le \rho_j(b)$, then the point $e_i + (\lceil \frac{b}{c_j} \rceil - 1)e_j + (\rho_j(b) - c_i)e_1$ is in X and inequality (17) is tight at this point. So $\alpha_i = \alpha_j - (\rho_j(b) - c_i)\alpha_1$. Since $c_1 = 1 \le \rho_j(b)$, we obtain $\alpha_1 = \frac{\alpha_j}{\rho_j(b)}$. Then $\alpha_i = c_i \frac{\alpha_j}{\rho_j(b)}$. Hence for $i < j$, $\alpha_i = \min\{c_i, \rho_j(b)\} \frac{\alpha_j}{\rho_j(b)}$.

For $i > j$, if $\rho_j(c_i) = 0$, consider point $e_i + (\lceil \frac{b}{c_j} \rceil - \frac{c_i}{c_j})e_j$. The left-hand side of inequality (17) at this point is equal to $\rho_j(b)\frac{c_i}{c_j} + (\lceil \frac{b}{c_j} \rceil - \frac{c_i}{c_j})\rho_j(b) = \lceil \frac{b}{c_j} \rceil \rho_j(b)$. So inequality (17) is tight. The left-hand side of the cover constraint is equal to $c_i + (\lceil \frac{b}{c_j} \rceil - \frac{c_i}{c_j})c_j = \lceil \frac{b}{c_j} \rceil c_j \ge b$. Thus this point is in X. Then we have $\alpha_i = \alpha_j \frac{c_i}{c_j}$.

Finally, for $i > j$, with $\rho_j(c_i) > 0$, consider $e_i + (\lceil \frac{b}{c_j} \rceil - \lceil \frac{c_i}{c_j} \rceil)e_j + (\rho_j(b) - \rho_j(c_i))^+ e_1$. The left-hand side of inequality (17) evaluated at this point is equal to $\rho_j(b)\lfloor \frac{c_i}{c_j} \rfloor + \min\{\rho_j(c_i), \rho_j(b)\} + (\lceil \frac{b}{c_j} \rceil - \lceil \frac{c_i}{c_j} \rceil)\rho_j(b) + (\rho_j(b) - \rho_j(c_i))^+ = \rho_j(b)\lfloor \frac{c_i}{c_j} \rfloor + \rho_j(b) + (\lceil \frac{b}{c_j} \rceil - \lceil \frac{c_i}{c_j} \rceil)\rho_j(b)$. Since $\rho_j(c_i) > 0$, this is equal to $\rho_j(b) + (\lceil \frac{b}{c_j} \rceil - 1)\rho_j(b) = \lceil \frac{b}{c_j} \rceil \rho_j(b)$, showing that inequality (17) is tight at this point. The left-hand side of the cover constraint is equal to

$$c_i + \left(\left\lceil \frac{b}{c_j} \right\rceil - \left\lceil \frac{c_i}{c_j} \right\rceil \right) c_j + (\rho_j(b) - \rho_j(c_i))^+. \tag{19}$$

If $\rho_j(c_i) > \rho_j(b)$, then (19) is equal to $c_i + (\lceil \frac{b}{c_j} \rceil - \lceil \frac{c_i}{c_j} \rceil)c_j = c_i + (\lceil \frac{b}{c_j} \rceil - \lfloor \frac{c_i}{c_j} \rfloor - 1)c_j = \rho_j(c_i) + (\lceil \frac{b}{c_j} \rceil - 1)c_j > \rho_j(b) + (\lceil \frac{b}{c_j} \rceil - 1)c_j = b$. If $\rho_j(c_i) \le \rho_j(b)$, then (19) is equal to $c_i + (\lceil \frac{b}{c_j} \rceil - \lceil \frac{c_i}{c_j} \rceil)c_j + \rho_j(b) - \rho_j(c_i) = c_i + (\lfloor \frac{b}{c_j} \rfloor - \lfloor \frac{c_i}{c_j} \rfloor)c_j + \rho_j(b) - \rho_j(c_i) = b$. So this point is in X. This proves that $\alpha_i = \alpha_j \lceil \frac{c_i}{c_j} \rceil - (\rho_j(b) - \rho_j(c_i))^+ \alpha_1 = \alpha_j \lceil \frac{c_i}{c_j} \rceil - (\rho_j(b) - \rho_j(c_i))^+ \frac{\alpha_j}{\rho_j(b)}$. If $\rho_j(b) \le \rho_j(c_i)$, then $\alpha_i = \alpha_j \lceil \frac{c_i}{c_j} \rceil = \alpha_j(\lfloor \frac{c_i}{c_j} \rfloor + 1)$. If $\rho_j(b) > \rho_j(c_i)$, then $\alpha_i = \alpha_j \lceil \frac{c_i}{c_j} \rceil - (\rho_j(b) - \rho_j(c_i))\frac{\alpha_j}{\rho_j(b)} = \alpha_j(\lfloor \frac{c_i}{c_j} \rfloor + 1) - \alpha_j + \rho_j(c_i)\frac{\alpha_j}{\rho_j(b)} = \alpha_j(\lfloor \frac{c_i}{c_j} \rfloor + \frac{\rho_j(c_i)}{\rho_j(b)})$. So, for $i < j$, $\alpha_i = \alpha_j(\lfloor \frac{c_i}{c_j} \rfloor + \min\{\frac{\rho_j(c_i)}{\rho_j(b)}, 1\}) = \frac{\alpha_j}{\rho_j(b)}(\lfloor \frac{c_i}{c_j} \rfloor \rho_j(b) + \min\{\rho_j(c_i), \rho_j(b)\})$.

Hence $\sum_{i=1}^n \alpha_i x_i = \alpha_0$ has the form

$$\sum_{i=1}^{j-1} \min\{c_i, \rho_j(b)\} \frac{\alpha_j}{\rho_j(b)} x_i + \alpha_j x_j + \sum_{j+1}^{n} \frac{\alpha_j}{\rho_j(b)} \left(\left\lfloor \frac{c_i}{c_j} \right\rfloor \rho_j(b) + \min\{\rho_j(c_i), \rho_j(b)\} \right) x_i = \alpha_j \left\lceil \frac{b}{c_j} \right\rceil.$$

This is $\frac{\alpha_j}{\rho_j(b)}$ times $\sum_{i=1}^{j} \min\{c_i, \rho_j(b)\} x_i + \sum_{i=j+1}^{n} \left(\rho_j(b)\lfloor \frac{c_i}{c_j} \rfloor + \min\{\rho_j(c_i), \rho_j(b)\}\right) x_i = \rho_j(b) \lceil \frac{b}{c_j} \rceil$. $\square$

Example 7. Consider the set $X^4 = \{x \in \mathbb{Z}_+^7 : x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 + 6x_6 + 7x_7 \geq 38\}$. The convex hull of X^4 is described by the nonnegativity constraints and the following inequalities (obtained using PORTA [6]):

$$x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 + 6x_6 + 7x_7 \geq 38, \tag{20}$$

$$2x_1 + 2x_2 + 4x_3 + 4x_4 + 5x_5 + 6x_6 + 6x_7 \geq 34, \tag{21}$$

$$x_1 + 2x_2 + 3x_3 + 3x_4 + 4x_5 + 5x_6 + 5x_7 \geq 28, \tag{22}$$

$$x_1 + 2x_2 + 2x_3 + 3x_4 + 4x_5 + 4x_6 + 5x_7 \geq 26, \tag{23}$$

$$x_1 + 2x_2 + 3x_3 + 3x_4 + 3x_5 + 4x_6 + 5x_7 \geq 24, \tag{24}$$

$$x_1 + 2x_2 + 2x_3 + 2x_4 + 3x_5 + 4x_6 + 4x_7 \geq 20, \tag{25}$$

$$x_1 + 2x_2 + 3x_3 + 3x_4 + 3x_5 + 3x_6 + 3x_7 \geq 18, \tag{26}$$

$$x_1 + x_2 + 2x_3 + 2x_4 + 2x_5 + 3x_6 + 3x_7 \geq 16, \tag{27}$$

$$x_1 + 2x_2 + 2x_3 + 2x_4 + 2x_5 + 2x_6 + 3x_7 \geq 14, \tag{28}$$

$$x_1 + x_2 + 2x_3 + 2x_4 + 2x_5 + 2x_6 + 2x_7 \geq 12. \tag{29}$$

As $c_1 = 1$, the cover constraint (20) is facet-defining for $conv(X^4)$. None of the rounding inequalities for items $\lambda = c_2, \ldots, c_7$ is facet-defining for $conv(X^4)$. For item 2, $\rho_2(38) = 0$. For item 3, $\rho_3(38) = 2$. Inequality (17) for 3, $x_1+2x_2+2x_3+3x_4+4x_5+4x_6+5x_7 \geq 26$, is a valid inequality and is facet-defining since $c_1 = 1$ and $\rho_3(38) > 0$. Indeed, it is the same as inequality (23). For item 4, $\rho_4(38) = 2$. Inequality (17) reads $x_1+2x_2+2x_3+2x_4+3x_5+4x_6+4x_7 \geq 20$ and is a valid inequality. This is the same as inequality (25) and is facet-defining. Note here that $\mu_4 = \epsilon$ and inequality (15) for item 4, $x_1 + 2x_2 + 2x_3 + 3x_4 + 3x_5 + 4x_6 + 4x_7 \geq 20$, is dominated by inequality (25). For item 5, $\rho_5(38) = 3$. Inequality (17), $x_1 + 2x_2 + 3x_3 + 3x_4 + 3x_5 + 4x_6 + 5x_7 \geq 24$, is the same as inequality (24). For item 6, $\rho_6(38) = 2$. The corresponding inequality (17) is $x_1+2x_2+2x_3+2x_4+2x_5+2x_6+3x_7 \geq 14$ and is the same as inequality (28). For item 7, $\rho_7(38) = 3$. The inequality $x_1 + 2x_2 + 3x_3 + 3x_4 + 3x_5 + 3x_6 + 3x_7 \geq 18$ is valid and facet-defining for $conv(X^4)$. This is the same as inequality (26).

6. Lifted 2-partition inequalities. Pochet and Wolsey [15] derive partition inequalities for PX where c_i divides c_{i+1} for all $i = 1, \ldots, n-1$. Then they prove that these inequalities are valid for PX in general under some conditions. Let $(i_1, \ldots, j_1), \ldots, (i_p, \ldots, j_p)$ be a partition of N such that $i_1 = 1$, $j_p = n$, and $i_t = j_{t-1} + 1$ for all $t = 2, \ldots, p$. Let $\beta_p = b$. For $t = p, \ldots, 1$, compute $\kappa_t = \lceil \frac{\beta_t}{c_{i_t}} \rceil$ and $\beta_{t-1} = \beta_t - (\kappa_t - 1)c_{i_t}$. The inequality

$$\sum_{t=1}^{p} \left(\prod_{s=1}^{t-1} \kappa_s \right) \sum_{j=i_t}^{j_t} \min \left\{ \left\lceil \frac{c_j}{c_{i_t}} \right\rceil, \kappa_t \right\} x_j \geq \prod_{s=1}^{p} \kappa_s \tag{30}$$

is called the *partition inequality*. Pochet and Wolsey [15] prove that the partition inequality is valid for PX if $\kappa_{t-1} \leq \lfloor \frac{c_{i_t}}{c_{i_{t-1}}} \rfloor$ for all $t = 2, \ldots, p$. If c_i divides c_{i+1} for all $i = 1, \ldots, n-1$, then the partition inequalities are valid without any condition, and they describe PX together with nonnegativity constraints.

Consider the case where $i_1 = 1$ and $j_1 = n$. Then inequality (30) reduces to the inequality $\sum_{j=1}^{n} \min \{ \lceil \frac{c_j}{c_1} \rceil, \kappa_1 \} x_j \geq \kappa_1$. This is the same as the rounding inequality (2) for $\lambda = c_1$ since $\kappa_1 = \lceil \frac{b}{c_1} \rceil$ and $c_j < b$ for all $j \in N$.

The next special case is when $i_1 = 1$, $j_1 = j - 1$, $i_2 = j$, and $j_2 = n$. Then $\kappa_2 = \lceil \frac{b}{c_j} \rceil$, $\beta_1 = b - (\lceil \frac{b}{c_j} \rceil - 1)c_j$. Notice that $\beta_1 = r_j$. Finally, $\kappa_1 = \lceil \frac{r_j}{c_1} \rceil$. Inequality

(30) becomes

$$\sum_{i=1}^{j-1}\min\left\{\left\lceil\frac{c_i}{c_1}\right\rceil,\left\lceil\frac{r_j}{c_1}\right\rceil\right\}x_i+\left\lceil\frac{r_j}{c_1}\right\rceil\sum_{i=j}^{n}\left\lceil\frac{c_i}{c_j}\right\rceil x_i\geq\left\lceil\frac{r_j}{c_1}\right\rceil\left\lceil\frac{b}{c_j}\right\rceil \tag{31}$$

and is valid if $\lceil\frac{r_j}{c_1}\rceil\leq\lfloor\frac{c_j}{c_1}\rfloor$. We refer to these inequalities as 2-*partition inequalities.*

PROPOSITION 7. *For $j\in N$, if $c_1=1$, inequality* (31) *is dominated by the cover constraint or inequality* (17).

Proof. If $c_1=1$, then the inequality simplifies to

$$\sum_{i=1}^{j}\min\{c_i,r_j\}x_i+r_j\sum_{i=j+1}^{n}\left\lceil\frac{c_i}{c_j}\right\rceil x_i\geq r_j\left\lceil\frac{b}{c_j}\right\rceil \tag{32}$$

and is always valid. If, moreover, $r_j=c_j$, then the inequality becomes $\sum_{i=1}^{j}c_ix_i+\sum_{i=j+1}^{n}c_j\lceil\frac{c_i}{c_j}\rceil x_i\geq b$ and is dominated by the cover constraint. If $r_j<c_j$, then $r_j=\rho_j(b)$ and $\rho_j(b)>0$. For $i>j$, if c_i is divisible by c_j, then $r_j\lceil\frac{c_i}{c_j}\rceil=\rho_j(b)\lfloor\frac{c_i}{c_j}\rfloor+\rho_j(c_i)$ since $\rho_j(c_i)=0$. If c_i is not divisible by c_j, then $r_j\lceil\frac{c_i}{c_j}\rceil=\rho_j(b)\lfloor\frac{c_i}{c_j}\rfloor+\rho_j(b)$. So the coefficient of x_i in (32) is greater than or equal to its coefficient in inequality (17). For $i\leq j$, the variable x_i has the same coefficient in (32) and (17). Also, the right-hand sides of (32) and (17) are the same. Hence if $c_1=1$ and $r_j<c_j$, inequality (17) dominates inequality (32). □

If $\lceil\frac{c_i}{c_1}\rceil\geq\lceil\frac{r_j}{c_1}\rceil$ for all $i<j$, then inequality (31) simplifies to $\sum_{i=1}^{j}x_i+\sum_{i=j+1}^{n}\lceil\frac{c_i}{c_j}\rceil x_i\geq\lceil\frac{b}{c_j}\rceil$, which is the rounding inequality (2) for $\lambda=c_j$.

Now we will improve the 2-partition inequalities (31) using lifting. Let $N^0\subset N$, $N^1=N\setminus N^0$, $j_{min}=\arg\min_{i\in N^1}c_i$, and $j\in N^1$, with $j_{min}\neq j$. The 2-partition inequality for the partition $N^-=\{i\in N^1:i<j\}$ and $N^+=\{i\in N^1:i\geq j\}$ is

$$\sum_{i\in N^-}\min\left\{\left\lceil\frac{c_i}{c_{j_{min}}}\right\rceil,\left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\right\}x_i+\left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\sum_{i\in N^+}\left\lceil\frac{c_i}{c_j}\right\rceil x_i\geq\left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\left\lceil\frac{b}{c_j}\right\rceil \tag{33}$$

and is valid when $x_i=0$ for all $i\in N^0$ if $\lceil\frac{r_j}{c_{j_{min}}}\rceil\leq\lfloor\frac{c_j}{c_{j_{min}}}\rfloor$.

The lifting function for inequality (33) is

$$\begin{aligned}\beta(a)=&\left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\left\lceil\frac{b}{c_j}\right\rceil\\&-\min_{x\in X_{b-a}(N^1)}\left(\sum_{i\in N^-}\min\left\{\left\lceil\frac{c_i}{c_{j_{min}}}\right\rceil,\left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\right\}x_i+\left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\sum_{i\in N^+}\left\lceil\frac{c_i}{c_j}\right\rceil x_i\right).\end{aligned}$$

LEMMA 2. *If $r_j\leq c_j-1$ and $\lceil\frac{r_j}{c_{j_{min}}}\rceil\leq\lfloor\frac{c_j}{c_{j_{min}}}\rfloor$, for $a\in\mathbb{R}$,*

$$\beta(a)=\begin{cases}\lceil\frac{r_j}{c_{j_{min}}}\rceil\lceil\frac{a}{c_j}\rceil-\lceil\frac{\rho_j(b-a)}{c_{j_{min}}}\rceil & \textit{if } a<b \textit{ and } 0<\rho_j(a)<r_j,\\ \lceil\frac{r_j}{c_{j_{min}}}\rceil\lceil\frac{a}{c_j}\rceil & \textit{if } a<b \textit{ and } \rho_j(a)\geq r_j \textit{ or } \rho_j(a)=0,\\ \lceil\frac{r_j}{c_{j_{min}}}\rceil\lceil\frac{b}{c_j}\rceil & \textit{if } a\geq b.\end{cases}$$

Proof. For $d\in\mathbb{R}$, let

$$z(d)=\min_{x\in X_d(N^1)}\left(\sum_{i\in N^-}\min\left\{\left\lceil\frac{c_i}{c_{j_{min}}}\right\rceil,\left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\right\}x_i+\left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\sum_{i\in N^+}\left\lceil\frac{c_i}{c_j}\right\rceil x_i\right).$$

If $d \le 0$, then $z(d) = 0$. If $d > 0$, Pochet and Wolsey [15] prove that there exists an optimal solution where $x_i = 0$, for $i \neq j_{min}$ and $i \neq j$, and $x_{j_{min}} \le \lceil \frac{r_j}{c_{j_{min}}} \rceil - 1$. Consider such optimal solutions. If $d < c_j$, then e_j or $\lceil \frac{d}{c_{j_{min}}} \rceil e_{j_{min}}$ is optimal. Hence $z(d) = \min\{\lceil \frac{r_j}{c_{j_{min}}} \rceil, \lceil \frac{d}{c_{j_{min}}} \rceil\}$. If $d \ge c_j$, then $x_j \ge \lfloor \frac{d}{c_j} \rfloor$ since otherwise $x_{j_{min}} \ge \lceil \frac{c_j}{c_{j_{min}}} \rceil$. So $\lfloor \frac{d}{c_j} \rfloor e_j + \lceil \frac{\rho_j(d)}{c_{j_{min}}} \rceil e_{j_{min}}$ or $\lceil \frac{d}{c_j} \rceil e_j$ is optimal, and $z(d) = \min\{\lceil \frac{r_j}{c_{j_{min}}} \rceil \lfloor \frac{d}{c_j} \rfloor + \lceil \frac{\rho_j(d)}{c_{j_{min}}} \rceil, \lceil \frac{r_j}{c_{j_{min}}} \rceil \lceil \frac{d}{c_j} \rceil\}$. So if $a < b$, then

$$\beta(a) = \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left\lceil \frac{b}{c_j} \right\rceil - \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left\lfloor \frac{b-a}{c_j} \right\rfloor - \min\left\{ \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil, \left\lceil \frac{\rho_j(b-a)}{c_{j_{min}}} \right\rceil \right\}.$$

Consider $a < b$. If $\rho_j(b-a) = \rho_j(b) - \rho_j(a)$ and $\rho_j(a) > 0$, then

$$\begin{aligned}
\beta(a) &= \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left(\left\lceil \frac{b}{c_j} \right\rceil - \left\lfloor \frac{b-a}{c_j} \right\rfloor \right) - \left\lceil \frac{\rho_j(b-a)}{c_{j_{min}}} \right\rceil \\
&= \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left(\frac{b-\rho_j(b)+c_j}{c_j} - \frac{b-a-\rho_j(b-a)}{c_j} \right) - \left\lceil \frac{\rho_j(b-a)}{c_{j_{min}}} \right\rceil \\
&= \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left(\frac{b-\rho_j(b)+c_j}{c_j} - \frac{b-a-\rho_j(b)+\rho_j(a)}{c_j} \right) - \left\lceil \frac{\rho_j(b-a)}{c_{j_{min}}} \right\rceil \\
&= \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left(\frac{a-\rho_j(a)+c_j}{c_j} \right) - \left\lceil \frac{\rho_j(b-a)}{c_{j_{min}}} \right\rceil \\
&= \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left\lceil \frac{a}{c_j} \right\rceil - \left\lceil \frac{\rho_j(b-a)}{c_{j_{min}}} \right\rceil.
\end{aligned}$$

If $\rho_j(b-a) = \rho_j(b) - \rho_j(a) + c_j$, then

$$\begin{aligned}
\beta(a) &= \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left(\left\lceil \frac{b}{c_j} \right\rceil - \left\lfloor \frac{b-a}{c_j} \right\rfloor \right) - \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \\
&= \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left(\frac{b-\rho_j(b)+c_j}{c_j} - \frac{b-a-\rho_j(b-a)}{c_j} - 1 \right) \\
&= \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left(\frac{b-\rho_j(b)+c_j}{c_j} - \frac{b-a-\rho_j(b)+\rho_j(a)-c_j}{c_j} - 1 \right) \\
&= \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \frac{a-\rho_j(a)+c_j}{c_j} \\
&= \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left\lceil \frac{a}{c_j} \right\rceil.
\end{aligned}$$

If $\rho_j(a) = 0$, then

$$\begin{aligned}
\beta(a) &= \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left(\left\lceil \frac{b}{c_j} \right\rceil - \left\lfloor \frac{b-a}{c_j} \right\rfloor \right) - \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \\
&= \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left(\frac{b-\rho_j(b)+c_j}{c_j} - \frac{b-a-\rho_j(b)}{c_j} - 1 \right) \\
&= \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \frac{a}{c_j}. \qquad \square
\end{aligned}$$

Function β is not subadditive in general. Consider $b = 18$, $c_j = 5$, and $c_{j_{min}} = 2$. Let $a = 2.5$ and $b = 5.5$. Then $\beta(2.5) = 1$, $\beta(5.5) = 2$, and $\beta(8) = 4$. So, $\beta(2.5) +$

$\beta(5.5) < \beta(8)$. So, to do lifting, we need a subadditive function which is greater than or equal to β. We first study the case where $c_{j_{min}}$ divides r_j. Notice that, in this case, $\lceil \frac{r_j}{c_{j_{min}}} \rceil \leq \lfloor \frac{c_j}{c_{j_{min}}} \rfloor$ is always satisfied.

THEOREM 9. *Let $N^0 \subset N$, $N^1 = N \setminus N^0$, $j_{min} = \arg\min_{i \in N^1} c_i$, $j \in N^1$, with $j_{min} < j$, $r_j \leq c_j - 1$, and $\rho_{j_{min}}(r_j) = 0$, $N^- = \{i \in N^1 : i < j\}$, and $N^+ = \{i \in N^1 : i \geq j\}$. The inequality*

$$\sum_{i \in N^-} \min\left\{\left\lceil \frac{c_i}{c_{j_{min}}} \right\rceil, \frac{r_j}{c_{j_{min}}}\right\} x_i + \frac{r_j}{c_{j_{min}}} \sum_{i \in N^+} \left\lceil \frac{c_i}{c_j} \right\rceil x_i$$
$$+ \sum_{i \in N^0} \left(\frac{r_j}{c_{j_{min}}} \left\lfloor \frac{c_i}{c_j} \right\rfloor + \min\left\{ \frac{\rho_j(c_i)}{c_{j_{min}}}, \frac{r_j}{c_{j_{min}}} \right\} \right) x_i \geq \frac{r_j}{c_{j_{min}}} \left\lceil \frac{b}{c_j} \right\rceil \tag{34}$$

is valid for PX.

Proof. Consider the function $\sigma(a) = \frac{r_j}{c_{j_{min}}} \lfloor \frac{a}{c_j} \rfloor + \min\{\frac{\rho_j(a)}{c_{j_{min}}}, \frac{r_j}{c_{j_{min}}}\}$. Notice that $\sigma(a) = \frac{r_j}{c_{j_{min}}} \Theta(a)$ for all $a \in \mathbb{R}$. Since Θ is subadditive (see Lemma 1) and $\frac{r_j}{c_{j_{min}}} > 0$, σ is subadditive. So, to prove the validity of (34), we need to show that $\sigma(a) \geq \beta(a)$ for all $a \in \mathbb{R}$.

If $a \geq b$ and $\lceil \frac{a}{c_j} \rceil = \lceil \frac{b}{c_j} \rceil$, then $\rho_j(a) \geq \rho_j(b)$. So $\sigma(a) = \frac{r_j}{c_{j_{min}}} \lceil \frac{a}{c_j} \rceil = \beta(a)$. If $a > b$ and $\lceil \frac{a}{c_j} \rceil \geq \lceil \frac{b}{c_j} \rceil + 1$, then $\sigma(a) \geq \frac{r_j}{c_{j_{min}}} \lfloor \frac{a}{c_j} \rfloor \geq \beta(a)$. If $a < b$ and $0 < \rho_j(a) < r_j$, then $\sigma(a) = \frac{r_j}{c_{j_{min}}} \lfloor \frac{a}{c_j} \rfloor + \frac{\rho_j(a)}{c_{j_{min}}}$ and $\beta(a) = \frac{r_j}{c_{j_{min}}} \lceil \frac{a}{c_j} \rceil - \lceil \frac{\rho_j(b-a)}{c_{j_{min}}} \rceil = \frac{r_j}{c_{j_{min}}} \lceil \frac{a}{c_j} \rceil - \frac{r_j}{c_{j_{min}}} - \lceil \frac{-\rho_j(a)}{c_{j_{min}}} \rceil = \frac{r_j}{c_{j_{min}}} \lfloor \frac{a}{c_j} \rfloor + \lfloor \frac{\rho_j(a)}{c_{j_{min}}} \rfloor \leq \sigma(a)$. If $a < b$ and $\rho_j(a) \geq r_j$ or $\rho_j(a) = 0$, then $\sigma(a) = \beta(a)$. Hence $\sigma(a) \geq \beta(a)$ for all $a \in \mathbb{R}$. □

These inequalities are not useful as they are dominated by the lifted rounding inequalities.

PROPOSITION 8. *For $j \in N$ with $r_j \leq c_j - 1$, inequality* (17) *dominates inequality* (34) *for all choices of $N^0 \subset N$, $N^1 = N \setminus N^0$, with $j \in N^1$, $j_{min} = \arg\min_{i \in N^1} c_i$, $j_{min} \neq j$, and $\rho_{j_{min}}(r_j) = 0$.*

Proof. Let $N^0 \subset N$, $N^1 = N \setminus N^0$, with $j \in N^1$, $j_{min} = \arg\min_{i \in N^1} c_i$, $j_{min} \neq j$, and $\rho_{j_{min}}(r_j) = 0$. If we divide inequality (17) by $c_{j_{min}}$, we obtain

$$\sum_{i=1}^{j} \min\left\{ \frac{c_i}{c_{j_{min}}}, \frac{r_j}{c_{j_{min}}} \right\} x_i + \sum_{i=j+1}^{n} \left(\frac{r_j}{c_{j_{min}}} \left\lfloor \frac{c_i}{c_j} \right\rfloor + \min\left\{ \frac{\rho_j(c_i)}{c_{j_{min}}}, \frac{r_j}{c_{j_{min}}} \right\} \right) x_i$$
$$\geq \frac{r_j}{c_{j_{min}}} \left\lceil \frac{b}{c_j} \right\rceil. \tag{35}$$

In inequality (34), variable x_i has the coefficient $\min\{\lceil \frac{c_i}{c_{j_{min}}} \rceil, \frac{r_j}{c_{j_{min}}}\} \geq \min\{\frac{c_i}{c_{j_{min}}}, \frac{r_j}{c_{j_{min}}}\}$ if $i \in N^-$. For $i \in N^+$, the variable x_i has the coefficient $\frac{r_j}{c_{j_{min}}} \lceil \frac{c_i}{c_j} \rceil \geq \frac{r_j}{c_{j_{min}}} \lfloor \frac{c_i}{c_j} \rfloor + \min\{\frac{\rho_j(c_i)}{c_{j_{min}}}, \frac{r_j}{c_{j_{min}}}\}$. The coefficient of x_i for $i \in N^0$ and the right-hand sides are equal in inequalities (17) and (34). □

Now we are interested in cases where $c_{j_{min}}$ does not divide r_j.

LEMMA 3. *If $r_j \leq c_j - 1$, $\lceil \frac{r_j}{c_{j_{min}}} \rceil \leq \lfloor \frac{c_j}{c_{j_{min}}} \rfloor$, and $\rho_{j_{min}}(r_j) > 0$, then the function*

$$\gamma(a) = \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left\lfloor \frac{a}{c_j} \right\rfloor + \min\left\{ \frac{\rho_j(a)}{\rho_{j_{min}}(r_j)}, \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \right\}$$

for $a \in \mathbb{R}$ is subadditive.

Proof. For $a, d \in \mathbb{R}$, if $\rho_j(a) + \rho_j(d) = \rho_j(a+d)$, then $\lfloor \frac{a}{c_j} \rfloor + \lfloor \frac{d}{c_j} \rfloor = \lfloor \frac{a+d}{c_j} \rfloor$. If $\min\{\frac{\rho_j(a)}{\rho_{j_{min}}(r_j)}, \lceil \frac{r_j}{c_{j_{min}}} \rceil\} = \lceil \frac{r_j}{c_{j_{min}}} \rceil$ or $\min\{\frac{\rho_j(d)}{\rho_{j_{min}}(r_j)}, \lceil \frac{r_j}{c_{j_{min}}} \rceil\} = \lceil \frac{r_j}{c_{j_{min}}} \rceil$, then $\gamma(a) + \gamma(d) \geq \lceil \frac{r_j}{c_{j_{min}}} \rceil \lfloor \frac{a+d}{c_j} \rfloor + \lceil \frac{r_j}{c_{j_{min}}} \rceil \geq \gamma(a+d)$. Otherwise, $\gamma(a) + \gamma(d) = \lceil \frac{r_j}{c_{j_{min}}} \rceil \lfloor \frac{a+d}{c_j} \rfloor + \frac{\rho_j(a+d)}{\rho_{j_{min}}(r_j)} \geq \gamma(a+d)$. If $\rho_j(a) + \rho_j(d) = \rho_j(a+d) + c_j$, then $\lfloor \frac{a}{c_j} \rfloor + \lfloor \frac{d}{c_j} \rfloor = \lfloor \frac{a+d}{c_j} \rfloor - 1$. If $\min\{\frac{\rho_j(a)}{\rho_{j_{min}}(r_j)}, \lceil \frac{r_j}{c_{j_{min}}} \rceil\} = \lceil \frac{r_j}{c_{j_{min}}} \rceil$ and $\min\{\frac{\rho_j(d)}{\rho_{j_{min}}(r_j)}, \lceil \frac{r_j}{c_{j_{min}}} \rceil\} = \lceil \frac{r_j}{c_{j_{min}}} \rceil$, then $\gamma(a) + \gamma(d) = \lceil \frac{r_j}{c_{j_{min}}} \rceil \lfloor \frac{a+d}{c_j} \rfloor + \lceil \frac{r_j}{c_{j_{min}}} \rceil \geq \gamma(a+d)$. If $\min\{\frac{\rho_j(a)}{\rho_{j_{min}}(r_j)}, \lceil \frac{r_j}{c_{j_{min}}} \rceil\} = \frac{\rho_j(a)}{\rho_{j_{min}}(r_j)}$ and $\min\{\frac{\rho_j(d)}{\rho_{j_{min}}(r_j)}, \lceil \frac{r_j}{c_{j_{min}}} \rceil\} = \lceil \frac{r_j}{c_{j_{min}}} \rceil$, then $\gamma(a) + \gamma(d) = \lceil \frac{r_j}{c_{j_{min}}} \rceil \lfloor \frac{a+d}{c_j} \rfloor + \frac{\rho_j(a)}{\rho_{j_{min}}(r_j)} \geq \lceil \frac{r_j}{c_{j_{min}}} \rceil \lfloor \frac{a+d}{c_j} \rfloor + \frac{\rho_j(a+d)}{\rho_{j_{min}}(r_j)} \geq \gamma(a+d)$. The case where $\min\{\frac{\rho_j(a)}{\rho_{j_{min}}(r_j)}, \lceil \frac{r_j}{c_{j_{min}}} \rceil\} = \lceil \frac{r_j}{c_{j_{min}}} \rceil$ and $\min\{\frac{\rho_j(d)}{\rho_{j_{min}}(r_j)}, \lceil \frac{r_j}{c_{j_{min}}} \rceil\} = \frac{\rho_j(d)}{\rho_{j_{min}}(r_j)}$ is similar. Finally, if we have $\min\{\frac{\rho_j(a)}{\rho_{j_{min}}(r_j)}, \lceil \frac{r_j}{c_{j_{min}}} \rceil\} = \frac{\rho_j(a)}{\rho_{j_{min}}(r_j)}$ and $\min\{\frac{\rho_j(d)}{\rho_{j_{min}}(r_j)}, \lceil \frac{r_j}{c_{j_{min}}} \rceil\} = \frac{\rho_j(d)}{\rho_{j_{min}}(r_j)}$, then $\gamma(a) + \gamma(d) = \lceil \frac{r_j}{c_{j_{min}}} \rceil (\lfloor \frac{a+d}{c_j} \rfloor - 1) + \frac{\rho_j(a+d)+c_j}{\rho_{j_{min}}(r_j)}$. Since $\lceil \frac{r_j}{c_{j_{min}}} \rceil \leq \lfloor \frac{c_j}{c_{j_{min}}} \rfloor$ and $\frac{c_j}{\rho_{j_{min}}(r_j)} \geq \frac{c_j}{c_{j_{min}}} \geq \lfloor \frac{c_j}{c_{j_{min}}} \rfloor$, $\gamma(a) + \gamma(d) \geq \lceil \frac{r_j}{c_{j_{min}}} \rceil \lfloor \frac{a+d}{c_j} \rfloor + \frac{\rho_j(a+d)}{\rho_{j_{min}}(r_j)} \geq \gamma(a+d)$. So γ is subadditive. □

Using function γ, we will lift inequality (33).

THEOREM 10. *Let $N^0 \subset N$, $N^1 = N \setminus N^0$, $j_{min} = \arg\min_{i \in N^1} c_i$, $j \in N^1$, with $j_{min} < j$, $r_j \leq c_j - 1$, $\rho_{j_{min}}(r_j) > 0$, and $\lceil \frac{r_j}{c_{j_{min}}} \rceil \leq \lfloor \frac{c_j}{c_{j_{min}}} \rfloor$, $N^- = \{i \in N^1 : i < j\}$, and $N^+ = \{i \in N^1 : i \geq j\}$. The lifted 2-partition inequality*

$$\sum_{i \in N^-} \min\left\{\left\lceil \frac{c_i}{c_{j_{min}}} \right\rceil, \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil\right\} x_i + \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \sum_{i \in N^+} \left\lceil \frac{c_i}{c_j} \right\rceil x_i$$
$$+ \sum_{i \in N^0} \left(\left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left\lfloor \frac{c_i}{c_j} \right\rfloor + \min\left\{ \frac{\rho_j(c_i)}{\rho_{j_{min}}(r_j)}, \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \right\} \right) x_i \geq \left\lceil \frac{r_j}{c_{j_{min}}} \right\rceil \left\lceil \frac{b}{c_j} \right\rceil \tag{36}$$

is valid for PX.

Proof. To prove the validity of (36), we need to show that $\gamma(a) \geq \beta(a)$ for all $a \in \mathbb{R}$. For $a < b$, with $0 < \rho_j(a) < r_j$, if $\min\{\frac{\rho_j(a)}{\rho_{j_{min}}(r_j)}, \lceil \frac{r_j}{c_{j_{min}}} \rceil\} = \frac{\rho_j(a)}{\rho_{j_{min}}(r_j)}$, then

$$\begin{aligned}
\gamma(a) - \beta(a) &= \lceil \frac{r_j}{c_{j_{min}}} \rceil \lfloor \frac{a}{c_j} \rfloor + \frac{\rho_j(a)}{\rho_{j_{min}}(r_j)} - \lceil \frac{r_j}{c_{j_{min}}} \rceil \lceil \frac{a}{c_j} \rceil + \lceil \frac{\rho_j(b-a)}{c_{j_{min}}} \rceil \\
&= \frac{\rho_j(a)}{\rho_{j_{min}}(r_j)} - \lceil \frac{r_j}{c_{j_{min}}} \rceil + \lceil \frac{\rho_j(b-a)}{c_{j_{min}}} \rceil \\
&= \frac{\rho_j(a)}{\rho_{j_{min}}(r_j)} - \lceil \frac{r_j}{c_{j_{min}}} \rceil + \lceil \frac{r_j - \rho_j(a)}{c_{j_{min}}} \rceil \\
&= \frac{\rho_j(a)}{\rho_{j_{min}}(r_j)} - \lceil \frac{r_j}{c_{j_{min}}} \rceil \\
&\qquad + \lceil \frac{r_j - \rho_{j_{min}}(r_j) + c_{j_{min}} - \rho_j(a) + \rho_{j_{min}}(r_j) - c_{j_{min}}}{c_{j_{min}}} \rceil \\
&= \frac{\rho_j(a)}{\rho_{j_{min}}(r_j)} + \lceil \frac{-\rho_j(a) + \rho_{j_{min}}(r_j) - c_{j_{min}}}{c_{j_{min}}} \rceil.
\end{aligned}$$

If $\rho_j(a) < \rho_{j_{min}}(r_j)$, then $\lceil \frac{-\rho_j(a)+\rho_{j_{min}}(r_j)-c_{j_{min}}}{c_{j_{min}}} \rceil = 0$ and $\gamma(a) - \beta(a) = \frac{\rho_j(a)}{\rho_{j_{min}}(r_j)} \geq 0$. If $\rho_j(a) \geq \rho_{j_{min}}(r_j)$, then $\gamma(a) - \beta(a) = \frac{\rho_j(a)}{\rho_{j_{min}}(r_j)} - 1 + \lceil \frac{-\rho_j(a)+\rho_{j_{min}}(r_j)}{c_{j_{min}}} \rceil \geq \frac{\rho_j(a)-\rho_{j_{min}}(r_j)}{c_{j_{min}}} - \lfloor \frac{\rho_j(a)-\rho_{j_{min}}(r_j)}{c_{j_{min}}} \rfloor \geq 0$.

If $\min\left\{\frac{\rho_j(a)}{\rho_{j_{min}}(r_j)}, \left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\right\} = \left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil$, $\gamma(a) = \left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\left\lceil\frac{a}{c_j}\right\rceil \geq \beta(a)$. For $a < b$, with $\rho_j(a) = 0$, $\gamma(a) = \left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\left\lfloor\frac{a}{c_j}\right\rfloor = \beta(a)$. For $a < b$, with $\rho_j(a) \geq r_j$,

$$\begin{aligned}
\frac{\rho_j(a)}{\rho_{j_{min}}(r_j)} - \left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil &= \frac{\rho_j(a)}{\rho_{j_{min}}(r_j)} - \frac{r_j - \rho_{j_{min}}(r_j) + c_{j_{min}}}{c_{j_{min}}} \\
&= \frac{\rho_j(a)c_{j_{min}} - \rho_{j_{min}}(r_j)(r_j - \rho_{j_{min}}(r_j) + c_{j_{min}})}{\rho_{j_{min}}(r_j)c_{j_{min}}} \\
&\geq \frac{r_j c_{j_{min}} - \rho_{j_{min}}(r_j)(r_j - \rho_{j_{min}}(r_j) + c_{j_{min}})}{\rho_{j_{min}}(r_j)c_{j_{min}}} \\
&= \frac{r_j(c_{j_{min}} - \rho_{j_{min}}(r_j)) - \rho_{j_{min}}(r_j)(-\rho_{j_{min}}(r_j) + c_{j_{min}})}{\rho_{j_{min}}(r_j)c_{j_{min}}} \\
&= \frac{(r_j - \rho_{j_{min}}(r_j))(c_{j_{min}} - \rho_{j_{min}}(r_j))}{\rho_{j_{min}}(r_j)c_{j_{min}}} \geq 0.
\end{aligned}$$

So $\gamma(a) = \left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\left\lceil\frac{a}{c_j}\right\rceil = \beta(a)$. For $a \geq b$, if $\left\lceil\frac{a}{c_j}\right\rceil = \left\lceil\frac{b}{c_j}\right\rceil$, then $\rho_j(a) \geq r_j$ and $\gamma(a) = \left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\left\lceil\frac{a}{c_j}\right\rceil = \beta(a)$. Otherwise, $\left\lceil\frac{a}{c_j}\right\rceil = \left\lceil\frac{b}{c_j}\right\rceil + 1$, and so $\gamma(a) \geq \beta(a)$. Hence $\gamma(a) \geq \beta(a)$ for all $a \in \mathbb{R}$. $\square$

As in the case of lifted rounding inequalities, the lifted 2-partition inequalities are also dominated by a subset of them which is polynomial in size.

PROPOSITION 9. *Let* $\{j_{min}, j\} \subseteq N$, *with* $j_{min} < j$, $r_j \leq c_j - 1$, $\rho_{j_{min}}(r_j) > 0$, *and* $\left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil \leq \left\lfloor\frac{c_j}{c_{j_{min}}}\right\rfloor$. *The inequality*

$$\begin{aligned}
&\sum_{i=1}^{j_{min}-1} \min\left\{\frac{c_i}{\rho_{j_{min}}(r_j)}, \left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\right\}x_i + \sum_{i=j_{min}}^{j-1} \min\left\{\left\lceil\frac{c_i}{c_{j_{min}}}\right\rceil, \frac{c_i}{\rho_{j_{min}}(r_j)}, \left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\right\}x_i \\
&\quad + \sum_{i=j}^{n}\left(\left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\left\lfloor\frac{c_i}{c_j}\right\rfloor + \min\left\{\frac{\rho_j(c_i)}{\rho_{j_{min}}(r_j)}, \left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\right\}\right)x_i \geq \left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\left\lceil\frac{b}{c_j}\right\rceil \qquad (37)
\end{aligned}$$

is valid and dominates inequality (36) *for* $N^0 \subset N$, $N^1 = N \backslash N^0$, *with* $\{j_{min}, j\} \subset N^1$ *and* $j_{min} = \arg\min_{i \in N^1} c_i$.

Proof. Let $\{j_{min}, j\} \subseteq N$, with $j_{min} < j$, $r_j \leq c_j - 1$, $\rho_{j_{min}}(r_j) > 0$, and $\left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil \leq \left\lfloor\frac{c_j}{c_{j_{min}}}\right\rfloor$. Consider $N^- = \{j_{min} \leq i < j : \left\lceil\frac{c_i}{c_{j_{min}}}\right\rceil \leq \frac{c_i}{\rho_{j_{min}}(r_j)}\}$, $N^+ = \{j\}$, $N^1 = N^- \cup N^+$, and $N^0 = N \setminus N^1$. For this choice of subsets, inequality (36) is the same as inequality (37).

Let $N^1 \subset N$, with $\{j_{min}, j\} \subset N^1$ and $j_{min} = \arg\min_{i \in N^1} c_i$. In inequality (36), for $i \in N^1$, if $i < j$, then x_i has the coefficient $\min\left\{\left\lceil\frac{c_i}{c_{j_{min}}}\right\rceil, \left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\right\}$, and if $i \in N^0$, then it has the coefficient $\min\left\{\frac{c_i}{\rho_{j_{min}}(r_j)}, \left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\right\}$. In both cases, its coefficient in inequality (36) is greater than or equal to its coefficient in inequality (37). If $i > j$ and $i \in N^1$, then the coefficient of x_i in inequality (36) is $\left\lceil\frac{r_j}{c_{j_{min}}}\right\rceil\left\lceil\frac{c_i}{c_j}\right\rceil$ and is greater than or equal to its coefficient in inequality (37). Other variables have the same coefficients in both inequalities. As the right-hand sides are also the same, we can conclude that inequality (37) dominates inequality (36). $\square$

The number of lifted 2-partition inequalities that are not dominated is $O(n^2)$.

7. Preliminary computational results. We mentioned in the introduction that the inequalities presented in this paper could be used to solve some hard mixed integer programming problems such as the *heterogeneous vehicle routing problem* (see

[18]) and the *manufacturer's mixed pallet design problem* (MPD) (see [19]). Some preliminary results with the rounding inequalities and the lifted rounding inequalities are presented in [18] and [19], respectively.

In this section, we investigate the effect of the lifted rounding inequalities and the lifted 2-partition inequalities in solving the MPD instances. The rounding inequalities for $\lambda = c_j$ for some $j \in N$ and the residual capacity inequalities are not included in this study as they are the same as or dominated by the lifted rounding inequalities.

We first give a brief definition of the MPD. For details, we refer the reader to [19]. Let C be the set of customers, N be the set of products, and $T = \{1, 2, \ldots, \tau\}$ be the set of periods. Each customer $k \in C$ has a demand of d_{kit} units for product $i \in N$ in period $t \in T$. Products are of identical dimensions and are sold in pallets. Each pallet has Q_1 rows, and, in each row, there are Q_2 units of a product. A pallet which contains more than one product type is called a mixed pallet. Let P denote the set of potential mixed pallet designs and q_{ij} denote the number of rows of product $i \in N$ in pallet design $j \in P$. The manufacturer also offers full pallets for each product $i \in N$, which consists of $Q_1 Q_2$ units of product i. We denote by h_{kit} and π_{kit} the unit inventory holding cost and the unit backlogging cost, respectively, for product $i \in N$ and customer $k \in C$ at the end of period $t \in T$. No backlogging is permitted at the end of period τ. The problem is to select at most m mixed pallet designs from set P to minimize the sum of customers' inventory holding and backlogging costs in periods $1, 2, \ldots, \tau$.

Let p_j be 1, if mixed pallet design $j \in P$ is offered, and 0, otherwise. Let P_k denote the set of mixed pallets that customer $k \in C$ can buy. Define y_{kjt} to be the number of pallets of type $j \in P_k$ that customer $k \in C$ buys in period $t \in T$ and f_{kit} to be the number of full pallets of product type $i \in N$ that customer $k \in C$ buys in period $t \in T$. In addition, define I_{kit} and B_{kit} to be the amount of product $i \in N$ that remains in inventory and that is backlogged at the end of period $t \in T$ for customer $k \in C$, respectively. Let M be a very large number. The MPD is formulated as follows in [19]:

$$\min \sum_{k \in C} \sum_{i \in N} \sum_{t \in T} (\pi_{kit} B_{kit} + h_{kit} I_{kit}) \tag{38}$$

$$\text{s.t.} \quad \sum_{j \in P} p_j \leq m, \tag{39}$$

$$I_{kit-1} - B_{kit-1} + Q_1 Q_2 f_{kit} + \sum_{j \in P_k} Q_2 q_{ij} y_{kjt} = d_{kit} + I_{kit} - B_{kit} \qquad \forall k \in C, i \in N, t \in T, \tag{40}$$

$$y_{kjt} \leq M p_j \qquad \forall k \in C, j \in P_k, t \in T, \tag{41}$$

$$I_{ki0} = B_{ki0} = B_{ki\tau} = 0 \qquad \forall k \in C, i \in N, \tag{42}$$

$$I_{kit}, B_{kit} \geq 0 \qquad \forall k \in C, i \in N, t \in T, \tag{43}$$

$$f_{kit} \geq 0 \text{ and integer} \qquad \forall k \in C, i \in N, t \in T, \tag{44}$$

$$y_{kjt} \geq 0 \text{ and integer} \qquad \forall k \in C, j \in P_k, t \in T, \tag{45}$$

$$p_j \in \{0, 1\} \qquad \forall j \in P. \tag{46}$$

The objective function (38) is the sum of inventory holding and backlogging costs over all periods. At most m mixed pallet designs can be offered due to constraint (39). Constraints (40) are the balance equations. Constraints (41) ensure that customers do

Table 1
Results with and without valid inequalities.

	Model1		Model2				Model3		
Problem	Nodes	CPU	(17)	% gap	Nodes	CPU	(37)	Nodes	CPU
1	1040094	168.38	33	96.57	84039	15.06	4	49348	10.05
2	3158201	662.17	39	97.18	189635	40.64	53	67257	16.24
3	29531186	6774.68	43	97.53	621224	159.62	59	248578	68.50
4	25242255	5800.38	48	95.43	600664	152.93	65	266693	76.80
5	2008508	1535.85	54	97.96	42476	34.12	77	87575	75.31
6	7650540	6310.95	58	98.30	395894	329.14	83	175031	150.45
7	110344292	7751.75	63	96.65	148494	121.37	89	48285	45.20

not buy mixed pallets that are not offered. Constraints (42) are beginning and ending conditions. Constraints (43)–(46) are nonnegativity and integrality constraints.

Yaman and Sen prove that the optimal value of the linear programming relaxation of MPD is zero. As a result it is important to derive strong valid inequalities for this problem to be able to improve the linear programming-based lower bounds.

For $k \in C$ and $i \in N$, let $D_{ki} = \left\lceil \sum_{t\in T} d_{kit}/Q_2 \right\rceil$. The inequality

$$\sum_{t\in T}\left(\min\{Q_1, D_{ki}\}f_{kit} + \sum_{j\in P_k}\min\{q_{ij}, D_{ki}\}y_{kjt}\right) \geq D_{ki} \tag{47}$$

is satisfied by all feasible solutions of MPD. Remark that the set of nonnegative integer solutions satisfying inequality (47) is an integer knapsack cover set. Hence we can generate valid lifted rounding and lifted 2-partition inequalities for the MPD based on inequalities (47).

We test the use of these valid inequalities on seven problem instances. We start with two base instances. In the first instance the number of products is two, and in the second instance the number of products is three. In both base instances, the number of periods is three, and the maximum number of mixed pallet designs to be offered is one. Using the first base instance, we generated four problems where the number of customers takes values 4, 5, 6, and 7. Using the second base instance, we generated three problems with 5, 6, and 7 customers.

For each problem instance, we first solve the model without valid inequalities. We call this *Model*1. We report the number of nodes in the branch and bound tree (in column *node*) and the CPU time in seconds (in column *CPU*). Then we form *Model*2 by adding the nondominated lifted rounding inequalities (17) to Model1. For Model2, we report the number of inequalities (17) added (in column (17)), the percentage duality gap (in column *%gap*, where $\%gap = \frac{opt-lp}{opt} * 100$, *opt* is the optimal value, and *lp* is the lower bound obtained from the linear programming relaxation), the number of nodes in the branch and bound tree, and the CPU time in seconds. Finally, we form *Model*3 by adding the nondominated lifted 2-partition inequalities (37) to Model2. We report here the number of inequalities (37) added (in column (37)), the number of nodes in the branch and bound tree, and the CPU time in seconds. The percentage duality gaps remained the same as the ones of Model2 and so are not reported. We solve the models using the mixed integer programming (MIP) solver of CPLEX 8.1 on an AMD Opteron 252 processor (2.6 GHz) with 2 GB of RAM. The results are given in Table 1.

The results show that both families of valid inequalities have been useful in decreasing the number of nodes in the branch and bound tree and the solution times for these instances. The solution time for Model3 is larger than the one of Model2 for instance five, but still it is about twenty times less than the one of Model1. The averages of percentage improvements obtained in the number of nodes and CPU time with the addition of inequalities (17) are 96.29% and 95.85%, respectively. The averages of percentage improvements obtained in the number of nodes and CPU time compared to Model2 with the addition of inequalities (37) are 34.07% and 28.07%, respectively.

8. Conclusion.

We studied the polyhedral properties of the convex hull of the integer knapsack cover set which appears as a relaxation of many optimization problems that concern covering a given demand using integer numbers of different types of items. We derived four families of valid inequalities, investigated when they dominate each other, and gave some conditions under which some are facet-defining. We used sequence-independent lifting to derive that last two families of valid inequalities. These inequalities can be used to solve problems such as those investigated in [11, 18, 19].

Except the rounding inequalities for arbitrary λ values, the valid inequalities derived in this paper share some common features. There exists always an item $j \in N$ such that the right-hand side of the inequality is equal to the coefficient of x_j times $\left\lceil \frac{b}{c_j} \right\rceil$. We know that this is an upper bound on the value of the right-hand side (see Proposition 1). Clearly, there are facet-defining inequalities which do not follow this rule. For instance, the cover constraint is facet-defining for $conv(\{x \in \mathbb{Z}_+^3 : 3x_1 + 4x_2 + 5x_3 \geq 13\})$.

Again excluding rounding inequalities, another common feature is that the number of inequalities that are nondominated within a family is polynomial even when the family has an exponential number of inequalities. These inequalities can be further lifted or modified to define larger families of valid inequalities for more complicated problems in consideration. For instance, an exponential number of valid inequalities can be derived for the integer capacity cover polyhedron using the inequalities of this paper and the lifting results of Mazur and Hall [12].

REFERENCES

[1] A. Atamturk, *Cover and pack inequalities for (mixed) integer programming*, Ann. Oper. Res., 139 (2005), pp. 21–38.

[2] A. Atamturk, *On capacitated network design cut-set polyhedra*, Math. Program., 92 (2002), pp. 425–437.

[3] A. Atamturk, *On the facets of the mixed-integer knapsack polyhedron*, Math. Program., 98 (2003), pp. 145–175.

[4] A. Atamturk, *Sequence independent lifting for mixed-integer programming*, Oper. Res., 52 (2004), pp. 487–490.

[5] E. Balas, *Facets of the knapsack polytope*, Math. Program., 8 (1975), pp. 146–164.

[6] T. Christof, *PORTA - a POlyhedron Representation Transformation Algorithm*, Version 1.3.2, 1999, http://www.iwr.uni-heidelberg.de/groups/comopt/software/PORTA/.

[7] S. Dash and O. Gunluk, *Valid inequalities based on simple mixed-integer sets*, Math. Program., 105 (2006), pp. 29–53.

[8] Z. Gu, G. L. Nemhauser, and M. W. P. Savelsbergh, *Cover inequalities for* 0-1 *linear programs: Computation*, INFORMS J. Comput., 10 (1998), pp. 427–437.

[9] P. L. Hammer, E. L. Johnson, and U. N. Peled, *Facets of regular* 0-1 *polytopes*, Math. Program., 8 (1975), pp. 179–206.

[10] T. L. Magnanti, P. Mirchandani, and R. Vachani, *The convex hull of two core capacitated network design problems*, Math. Program., 60 (1993), pp. 233–250.

[11] D. R. Mazur, *Integer Programming Approaches to a Multi-Facility Location Problem*, Ph.D. Thesis, John Hopkins University, Baltimore, MD, 1999.

[12] D. R. Mazur and L. A. Hall, *Facets of a Polyhedron Closely Related to the Integer Knapsack-Cover Problem*, Technical report, 2002, http://www.optimization-online.org/.

[13] A. J. Miller and L. A. Wolsey, *Tight formulations for some simple mixed integer programs and convex objective integer programs*, Math. Program., 98 (2003), pp. 73–88.

[14] G. L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*, Wiley, New York, 1988.

[15] Y. Pochet and L. A. Wolsey, *Integer knapsack and flow covers with divisible coefficients: Polyhedra, optimization and separation*, Discrete Appl. Math., 59 (1995), pp. 57–74.

[16] R. Weismantel, *On the 0/1 knapsack polytope*, Math. Program., 77 (1997), pp. 49–68.

[17] L. Wolsey, *Faces for a linear inequality in 0-1 variables*, Math. Program., 8 (1975), pp. 165–178.

[18] H. Yaman, *Formulations and valid inequalities for the heterogeneous vehicle routing problem*, Math. Program., 106 (2006), pp. 365–390.

[19] H. Yaman and A. Sen, *Manufacturer's Mixed Pallet Design Problem*, European Journal of Operational Research, to appear.

[20] E. Zemel, *Easily computable facets of the knapsack polytope*, Math. Oper. Res., 14 (1989), pp. 760–764.

SIAM J. Discrete Math.
Vol. 21, No. 3, pp. 573–591

RECOGNIZING CHORDAL PROBE GRAPHS AND CYCLE-BICOLORABLE GRAPHS[*]

ANNE BERRY[†], MARTIN CHARLES GOLUMBIC[‡], AND MARINA LIPSHTEYN[‡]

Abstract. A graph $G = (V, E)$ is a chordal probe graph if its vertices can be partitioned into two sets, P (probes) and N (non-probes), where N is a stable set and such that G can be extended to a chordal graph by adding edges between non-probes. We give several characterizations of chordal probe graphs, first, in the case of a fixed given partition of the vertices into probes and non-probes, and second, in the more general case where no partition is given. In both of these cases, our results are obtained by introducing new classes, namely, N-triangulatable graphs and cycle-bicolorable graphs. We give polynomial time recognition algorithms for each class. N-triangulatable graphs have properties similar to chordal graphs, and we characterize them using graph separators and using a vertex elimination ordering. For cycle-bicolorable graphs, which are shown to be perfect, we prove that any cycle-bicoloring of a graph renders it N-triangulatable. The corresponding recognition complexity for chordal probe graphs, given a partition of the vertices into probes and non-probes, is $O(|P||E|)$, thus also providing an interesting tractable subcase of the chordal graph sandwich problem. If no partition is given in advance, the complexity of our recognition algorithm is $O(|E|^2)$.

Key words. chordal graph, probe graph, triangulation, perfect graph, elimination scheme, bicoloring

AMS subject classifications. 05C17, 05C75

DOI. 10.1137/050637091

1. Introduction.

1.1. Motivation: Interval probe graphs.

The study of chordal probe graphs [11, 12] was originally motivated as a generalization of the interval probe graphs which occur in applications involving physical mapping of DNA. Interval probe graphs were introduced by Zhang [20, 24] to model problems in physical mapping of DNA when the intervals are either probes or non-probes, and the information on the overlaps between non-probes is missing. As a result, Zhang defined a graph to be interval probe if its vertex set can be partitioned into probes and non-probes in such a fashion that it can be completed into an interval graph by adding only edges between non-probes. This shows two different facets of the problem: either the partition is given in advance, or a partition has to be proposed as part of the solution.

Recently, for partitioned interval probe graphs, an $O(n^2)$ time recognition algorithm was first reported in [15] which uses PQtrees. Another method, given in [18], uses modular decomposition and has complexity $O(n + m \log n)$ for a graph with n vertices and m edges. In the case of trees, Sheng [22] gives characterizations by a family of forbidden subgraphs for both the partitioned and non-partitioned case, thus ensuring polynomial time recognition of trees which are interval probe graphs (see also [13]). The polynomial time complexity of recognizing interval probe graphs (when no partition is given) has been given in [5].

[*] Received by the editors July 29, 2005; accepted for publication (in revised form) January 11, 2007; published electronically July 18, 2007.
http://www.siam.org/journals/sidma/21-3/63709.html

[†] LIMOS UMR CNRS 6158, Ensemble Scientifique des Cézeaux, Bat ISIMA D123, Université Blaise Pascal, F-63 173 Aubière, France (berry@isima.fr).

[‡] Caesarea Rothschild Institute and Department of Computer Science, University of Haifa, Haifa, Israel (golumbic@cs.haifa.ac.il, marina_lipstein@atricia.com).

1.2. Chordal probe graphs and their generalizations. Generalizing interval probe graphs, Golumbic and Lipshteyn [11, 12] introduced chordal probe graphs as a new class of perfect graphs: A graph G is *chordal probe* if its vertex set can be partitioned into a set P of probes and a stable set N of non-probes in such a fashion that G can be completed into a chordal graph by adding only edges between non-probes. They gave $O(m^2)$ algorithms to recognize chordal probe graphs which are also even-chordal, which is exactly the subfamily of chordal probe graphs which have no even hole; this class includes the interval probe graphs and is also weakly chordal [14].

Among the results in this paper, we solve the general problem for chordal probe graphs, by giving polynomial time recognition algorithms for the partitioned as well as the non-partitioned case. In doing so, we introduce two new graph superclasses, the N-triangulatable graphs and cycle-bicolorable graphs, proving interesting properties for both of them.

In Part I, we examine the partitioned case. In fact, we solve a broader problem in which the set N is not assumed to be a stable set, which defines the class of N-triangulatable graphs. We investigate the structural properties of this class and show that several properties of chordal graphs can be extended to this class, namely we characterize them using graph separators and using a vertex elimination ordering. These results enable us to propose a recognition algorithm with a complexity of $O(|P|m)$. Section 3 deals with N-triangulatable graphs and section 4 discusses the subcase of partitioned chordal probe graphs.

In Part II, we discuss the case where no partition is given in advance. Our approach uses a lemma from [12], which remarks that in a partitioned chordal probe graph, probes and non-probes must alternate on every chordless cycle. Thus, in section 6, we again solve a broader problem by introducing the class of cycle-bicolorable graphs, a superclass of chordal probe graphs. In section 7, we characterize chordal probe graphs as cycle-bicolorable graphs in which one color defines a stable set and give a corresponding $O(n^2m)$ time recognition algorithm. These results are based on a new graph decomposition, introduced in section 5, which groups together the cycles of the graph into so-called C-components. The polynomial time complexity relies on the theory of graph separators.

2. Background and previous results.

2.1. General definitions and properties of chordal graphs. The graphs in this work are undirected and finite. A graph is denoted $G = (V, E)$, with $n = |V|$ and $m = |E|$. We let G_A denote the subgraph induced by a vertex set $A \subset V$; similarly, $\overline{G}_A$ denotes the subgraph induced by A in the complement $\overline{G}$ of G. A *clique* in a graph is a set of pairwise adjacent vertices, and a *stable set* in a graph is a set of pairwise non-adjacent vertices. We say that we *saturate* a set of vertices if we add all the edges necessary to make it a clique. In this paper, a *connected component* is a vertex set which induces a maximal connected subgraph.

The (open) *neighborhood* of a vertex x in graph G is the set $\mathscr{N}_G(x) = \{y \neq x \mid xy \in E\}$; we will say that a vertex x *sees* another vertex y if $xy \in E$. The *closed neighborhood* of x is $\mathscr{N}_G[x] = \mathscr{N}_G(x) \cup \{x\}$. We extend the notion of neighborhood to a set of vertices A by defining $\mathscr{N}_G(A) = \cup_{x \in A} \mathscr{N}_G(x) - A$ and $\mathscr{N}_G[A] = \mathscr{N}_G(A) \cup \{A\}$. When there is no ambiguity as to which graph is referred to, the subscript will be omitted, i.e., $\mathscr{N}_G(x)$ will be written simply $\mathscr{N}(x)$. The degree of vertex x will be denoted by $deg(x) = |\mathscr{N}(x)|$.

A chordless cycle of length k is denoted by C_k and we always assume $k \geq 4$. A *hole* is a chordless cycle of length at least five; a hole is called *odd* or *even* depending on the parity of its length. An *antihole* is the complement of a hole. A graph G is called *perfect* if every induced subgraph G_A satisfies the equality $\omega(G_A) = \chi(G_A)$, where ω denotes the size of the largest clique and χ is the chromatic number. The *Strong Perfect Graph Theorem* [6], originally conjectured by Berge, states that *a graph is perfect if and only if it contains neither an odd hole nor an odd antihole.*

Chordal graphs are a well known family of perfect graphs. A graph is defined to be *chordal* if every cycle of length 4 or greater has a chord, that is, an edge joining two non-consecutive vertices of the cycle. Chordal graphs have important application areas including acyclic relational database schemes, facility location problems, statistical analysis, and the problem of tracing genetic mutations over an evolutionary period by constructing phylogenetic trees (see [9, 19]).

It is often the case in such applications, however, that an input graph G has edges missing due to incomplete data. This gives rise to the problem of adding additional edges F in order to complete it into a chordal supergraph $G^{'} = (V, E + F)$ of G. The edge set F is said to be *minimal* if no proper subset defines a chordal graph when added; the resulting chordal graph is then called a *minimal triangulation.* When $|F|$ is required to be smallest possible, it is called a *minimum triangulation.*

The *chordal graph sandwich problem* (see [4, 10, 23]) is another variation where a specified set of optional edges E_0 is given with the input, and the triangulation F (not necessarily minimum) must satisfy $F \subseteq E_0$. Both the minimum triangulation problem and the chordal graph sandwich problem are NP-complete.

An undirected graph $G = (V, E)$ is a *chordal probe graph* if its vertex set can be partitioned into two subsets, P (probes) and N (non-probes), where N is a stable set and there exists a completion $F \subseteq N \times N$ such that $H = (V, E+F)$ is a chordal graph. The class of chordal probe graphs was introduced in [12] as a generalization of interval probe graphs. Interval probe graphs are defined similarly, where the completed graph H must be an interval graph.

A vertex is *simplicial* if its neighborhood is a clique. The notion of *simplicial vertex* was introduced independently by Dirac in 1961 [7] and by Lekkerkerker and Boland in 1962 [17] as an extension of the notion of a leaf in a tree and is the basis for the following theorem by Dirac.

THEOREM 2.1 (see [7]). *A non-clique chordal graph has at least two non-adjacent simplicial vertices.*

This led Fulkerson and Gross [8] to characterize chordal graphs in an algorithmic manner as follows.

CHARACTERIZATION 2.2 (see [8]). *A graph is chordal if and only if one can repeatedly find a simplicial vertex and delete it from the graph until no vertex is left.*

This defines an ordering on the vertices called a *perfect elimination ordering (peo).*

One of the earliest ways which was used to compute a triangulation was to force the graph into respecting this characterization by using an ordering α on the vertices and repeatedly choosing the next vertex in this ordering, forcing its neighborhood into a clique by the addition of any missing edges and removing the vertex; we refer to this process as the *elimination game on* (G, α). Each of the successive graphs obtained is called a *transitory elimination graph* and is denoted by G_i. At the end of the process, the set F of added edges define a triangulation $G^+_\alpha = (V, E + F)$ of the input graph (V, E) (see [3]).

The following property is well known.

PROPERTY 2.3. *If* G' *is a triangulation of* G *and* α *is a peo of* G', *then* $G' = G^+_\alpha$.

2.2. Minimal separators and minimal triangulations. Minimal separators were introduced by Dirac [7]. A subset S of vertices of a connected graph G is called a *separator* if G_{V-S} is not connected. A separator S is called an *ab-separator* if a and b are in different connected components of G_{V-S}, a separator S is a *minimal ab-separator* if S is an ab-separator and no proper subset of S is an ab-separator; finally, a separator S is a *minimal separator* if there is some pair $\{a, b\}$ such that S is a minimal ab-separator. Equivalently, a separator S is minimal if there exist two distinct components C_1 and C_2 in G_{V-S} such that $\mathscr{N}(C_1) = \mathscr{N}(C_2) = S$; such components are called *full components of S*.

Minimal separators turn out to be a very useful tool for computing a minimal triangulation.

Definition 2.4 (Kloks, Kratsch, and Spinrad [16]). *Let S and T be two minimal separators of G. Then S* crosses *T if there exist two components X_1, X_2 of $G_{(V-T)}$, $X_1 \neq X_2$, such that $S \cap X_1 \neq \emptyset$ and $S \cap X_2 \neq \emptyset$.*

It is shown in [21] that the crossing relation is symmetric, i.e., S *crosses* T if and only if T *crosses* S.

Property 2.5. *Let S and T be two minimal separators of G. Then S* crosses *T if and only if T has a vertex in each full component of S.*

Theorem 2.6 (see [21]). *When a minimal separator S is saturated, creating graph G':*

1. *All the minimal separators which cross S disappear.*
2. *All the minimal separators which do not cross S remain.*
3. *No new minimal separator appears.*
4. *Any minimal triangulation of G' is a minimal triangulation of G.*

Thus, computing a minimal triangulation of a graph G is equivalent to saturating a maximal set of pairwise non-crossing minimal separators of G (see [21]).

The following is a consequence of Theorem 2.6.

Property 2.7. *S and T are two crossing minimal separators of a graph G if and only if S contains two non-adjacent vertices x and y such that T is a minimal xy-separator of G.*

Lekkerkerker and Boland in [17] introduced the following notion which will be fundamental to this paper.

Definition 2.8. *A* substar *S of x is a subset of $\mathscr{N}(x)$ such that for some connected component U of $G_{V-\mathrm{N}[x]}$, $S = \mathscr{N}(U)$,* i.e., all the vertices of a substar see some common connected component of $G_{V-\mathrm{N}[x]}$.

Note that the substars of x are exactly the minimal separators included in the neighborhood of x.

Example 2.9. In Figure 1, the substars of j are $\{b, c\}$ and $\{e, i, k\}$; j is on a chordless cycle with e, i, and k, because substar $\{e, i, k\}$ is an independent set, but j is *not* on a cycle with b nor c, because substar $\{b, c\}$ is a clique.

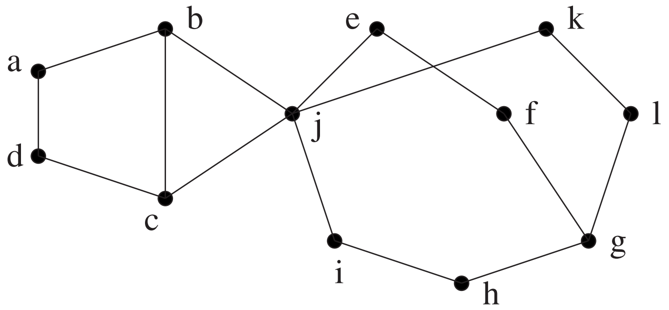


Fig. 1. *A graph. The substars of j are $\{b, c\}$ and $\{e, i, k\}$.*

Additional properties which are useful are the following.

PROPERTY 2.10 (see [1]). *For a vertex x, the substars of x are pairwise non-crossing.*

PROPERTY 2.11 (see [1]). *Let x, y be two non-adjacent vertices of a graph G, then no substar of x can cross a substar of y.*

PROPERTY 2.12 (see [17]). *A vertex x is on a chordless cycle if and only if at least one of its substars is not a clique. More precisely, if X is a connected component of $G_{V-N[x]}$ such that $S = N(X)$ contains the non-edge yz, then x is on a chordless cycle C on which it sees y and z, and all the other vertices of C are in X.*

For example, in Figure 1, vertex j is on a chordless cycle with e, i and k, because substar $\{e, i, k\}$ is an independent set, but j is *not* on a cycle with b nor c, because substar $\{b, c\}$ is a clique.

This leads us to the following definition.

DEFINITION 2.13. *We say that a vertex x is* LB-simplicial *if all the substars of x are cliques.*

Finally, we recall the following characterization of chordal graphs which does not appear to be well known.

CHARACTERIZATION 2.14 (see [17]). *A graph is chordal if and only if every vertex is LB-simplicial.*

Part I. The partitioned case. The original motivation for this work has been the recognition of chordal probe graphs, in the non-partitioned as well as in the partitioned case. We will first address the partitioned case. In order to do this, we solve a more general problem.

3. *N*-triangulatable partitioned graphs.

We introduce a new problem, namely, triangulating a graph whose vertex set is bipartitioned into "probes" and "non-probes" by adding only edges between non-probes. The corresponding class, which we call N-triangulatable graphs, is studied in this section.

One of the interesting developments is that N-triangulatable graphs turn out to be very similar to chordal graphs: We will show that several properties and characterizations of chordal graphs can be very profitably extended to N-triangulatable graphs, and that they yield the tools we need to handle this class efficiently.

DEFINITION 3.1. *We will say that a graph $G = (P + N, E)$ is* N-triangulatable (N-T) *if a triangulation of G can be obtained by adding only edges whose endpoints are non-probes. We will call such a triangulation an* N-triangulation *of G.*

Remark 3.2.

1. If G is N-T, then G_P is a chordal graph.
2. An induced subgraph of an N-T graph is an N-T graph.
3. In the case where $P = \emptyset$, the graph becomes an arbitrary graph, and it is always N-T.
4. In the case where N is a stable set, G is N-T if and only if G is chordal probe with respect to this partition.
5. Recognizing N-T graphs is a special case of the chordal graph sandwich problem, where the optional edges E_0 consist of all non-edges between non-probes.

We will now see that both Lekkerkerker and Boland's Characterization 2.14 and Fulkerson and Gross' Characterization 2.2 can be extended to recognize this class. These will be studied, respectively, in sections 3.1 and 3.3.

3.1. Quasi LB-simpliciality of *N*-T graphs.

In this section, we extend Characterization 2.14 of chordal graphs due to Lekkerkerker and Boland to N-T graphs.

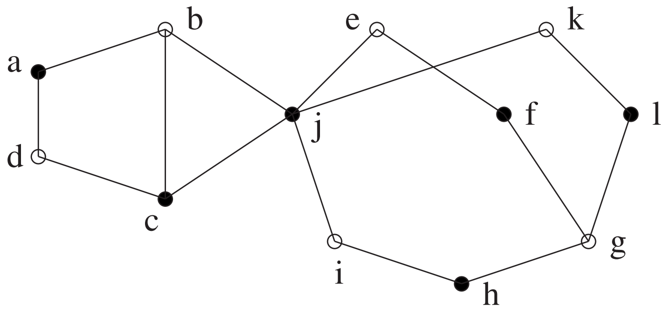


FIG. 2. *An N-T graph with white non-probes and black probes.*

This will also enable us to give a recognition algorithm for N-T graphs using separators.

DEFINITION 3.3. *We will say that a vertex x is* quasi-LB-simplicial *if all the non-edges of all the substars of x have both endpoints that are non-probes.*

Example 3.4. In Figure 2, if black vertices are probes and white are non-probes, c is quasi-LB-simplicial, as its substars are $\{b, d\}$ and $\{j\}$.

We will see that examining the substars of the probes of the graph is sufficient to characterize N-T graphs.

DEFINITION 3.5. *We will say that the substars of a probe are P-*substars.

THEOREM 3.6. *The following conditions are equivalent for a graph $G = (P + N, E)$:*

1. *G is an N-T graph.*
2. *All probes of G are quasi-LB-simplicial.*
3. *G contains no chordless cycle with two adjacent probes.*

Proof. (1) ⇒ (3): Let $G = (P + N, E)$ be an N-T graph, let $V = P + N$, and let G' be an N-triangulation of G. Suppose by contradiction that in G there is a chordless cycle $(p_1, p_2, v_3, \ldots, v_k, p_1)$, where p_1 and p_2 are probes. In G', p_1 sees v_k and p_2; $v_3, v_4, \ldots, v_{k-1}$ belong to the same connected component X of $G'_{V-\mathrm{N}\,[p_1]}$. $\mathscr{N}(X)$ is a substar of G', but it fails to be a clique, as it contains v_k and p_2, which are non-adjacent. This contradicts Characterization 2.14 for chordal graphs.

(3) ⇒ (2): Assume in $G = (P + N, E)$ there is no chordless cycle with two adjacent probes. Suppose by contradiction that there exists a probe x which fails to be quasi-LB-simplicial: x has two non-adjacent neighbors, y and z, one of which is a probe; w.l.o.g. y is a probe. According to Property 2.12, there exists a chordless cycle which contains y, x, and z consecutively, a contradiction.

(2) ⇒ (1): Let $G = (P + N, E)$ be a graph such that all probes are quasi-LB-simplicial; we will prove that G is an N-T graph.

Let us use the minimal triangulation algorithm LB-TRIANG described in [1], which repeatedly chooses a vertex x, saturates its substars, and removes x. Regardless of the order in which the vertices are processed, LB-TRIANG computes a minimal triangulation of the input graph; we will run it by first choosing all the probes.

We claim that no new P-substar can appear. Because of Theorem 2.6, the only way a P-substar can be created is by adding edges which will cause a previous minimal separator S, which was not a P-substar, to be in the neighborhood of a probe. However, no new edge can be added incident to a probe, so this cannot happen.

After all the P-substars have been processed and eliminated, only vertices from N are left in the graph. When we finish the execution, we will have computed a minimal triangulation of G which has added only edges between two non-probes, which is thus an N-triangulation of G. By definition, G is an N-T graph. □

Complexity. The recognition algorithm based on Theorem 3.6 runs in $O(|P|m)$ time: The implementation of Algorithm LB-TRIANG proposed in [1], as in the proof of Theorem 3.6, uses a data structure inspired from clique trees and requires only $O(m)$ time per processed vertex; a global $O(m)$ time is then used to check that only edges between pairs of non-probes have been added.

Computing a minimal triangulation of an N-T graph costs $O(nm)$ time, which is the same as computing a minimal triangulation of any graph. However, in order to recognize N-T graphs, it is not necessary to actually compute an N-triangulation. Therefore, unless P is of order n, it is cheaper to recognize the class than to exhibit an N-triangulation for it.

3.2. Properties of N-T graphs.

THEOREM 3.7. *Let $G = (P + N, E)$ be an N-T graph. The P-substars of G are pairwise non-crossing.*

Proof. Let $G = (P + N, E)$ be an N-T graph, and let $V = P + N$. Let us assume by contradiction that there are two crossing P-substars S_1 and S_2. By Property 2.11, S_1 and S_2 must be substars of two adjacent vertices p_1 and p_2. By Property 2.7, there must be two non-adjacent vertices x and y in S_1, such that S_2 is a minimal xy-separator.

Let us first suppose that p_2 belongs to S_1. Since p_1 is quasi-LB-simplicial, p_2 must see x and y. Therefore, every minimal xy-separator must contain p_2, which contradicts the fact that S_2 is a minimal xy-separator.

Let us now examine the case where p_2 is not in S_1. Let X be the connected component of $G_{V-\mathrm{N}\,[p_1]}$ such that $S_1 = \mathscr{N}(X)$. According to Property 2.12, x and y belong to some chordless cycle C on which x, p_1 and y are consecutive, with all other vertices in X. Suppose p_2 sees some of these intermediate vertices $C \cap (V - \mathscr{N}[p_1])$. Then p_2 would belong to $\mathscr{N}(X)$ and thus to S_1, which is impossible. Therefore, S_2 has no vertex in X, which is a full component of S_1; by Property 2.5, S_1 and S_2 are non-crossing. □

COROLLARY 3.8. *The number of P-substars in an N-T graph is less than n.*

Proof. This follows from the simple observation that, since the P-substars are non-crossing minimal separators, they can all be chosen to be saturated and preserved in some minimal triangulation of G, which, as all chordal graphs, has less than n minimal separators. □

Recall that in the proof of Theorem 3.6, we ran LB-TRIANG by using all the probes in a first phase and then the non-probes in a second phase. Since the minimal separators which are chosen as substars in the first phase are pairwise non-crossing, the resulting set F_P of edges which is added is the same, regardless of the order in which the probes are processed; the edges of F_P are mandatory, and we will use them to define G^* below. The set of edges computed by the second phase, however, depends on the order in which the non-probes are processed.

DEFINITION 3.9. *We define the* enhanced graph G^* *of G to be the graph obtained from G by saturating all the P-substars of G.*

Example 3.10. Figure 3 gives the enhanced graph of the graph of Figure 2.

THEOREM 3.11. *Any minimal triangulation of G^* is a minimal triangulation of G and an N-triangulation of G.*

Proof. By Theorem 3.7, the P-substars of G are pairwise non-crossing. Therefore, G^* is obtained by saturating a set of pairwise non-crossing minimal separators of G.

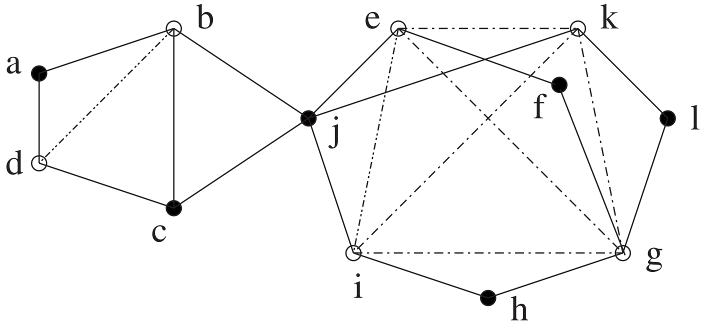


Fig. 3. *The corresponding enhanced graph G^*, where the probe substars have been saturated.*

By Theorem 2.6, any minimal triangulation of G^* is a minimal triangulation of G. If we run Algorithm LB-TRIANG as in the proof of Theorem 3.6 by first choosing the probes, making these simplicial will only add edges between two non-probes. After the probes are processed and eliminated, only non-probes are left in the graph, and the chosen subsequent triangulation will also add only edges between two non-probes; the minimal triangulation thus obtained is an N-triangulation. $\square$

3.3. Quasi-perfect elimination in N-T graphs. We will now go on to show that Fulkerson and Gross' Characterization 2.2 can also be extended to an N-T graph and that, as is the case with chordal graphs with respect to perfect elimination orderings (peos), a greedy approach to playing the quasi-peo elimination game will successfully recognize N-T graphs.

Definition 3.12. *Let $G = (P + N, E)$ be an N-T graph. We will say that a vertex v of G is* quasi-simplicial *if every non-edge of $\mathscr{N}(v)$ has both endpoints which are non-probes.*

Definition 3.13. *We will say that an ordering α on the vertices of G is a* quasi-perfect elimination ordering (qpeo) *if at each step i of the elimination game on (G, α), vertex $\alpha(i)$ is quasi-simplicial in the transitory elimination graph G_i.*

Example 3.14. In Figure 2, if black vertices are probes and white are non-probes, a is quasi-simplicial and d is not. However, if a is chosen first in a qpeo, saturating $\mathscr{N}(a)$ and removing a will make d quasi-simplicial in the transitory graph; $\alpha = (a, d, c, b, j, h, l, f, e, k, g, i)$ is a qpeo.

Lemma 3.15. *Let $G = (P+N, E)$ be an N-T graph, and let v be a quasi-simplicial vertex of G. If G' is the graph obtained by making v simplicial and removing it, then G' is also N-T.*

Proof. Suppose by contradiction that G' fails to be an N-T graph, we will prove that G is not N-T. According to Theorem 3.6, there must be a chordless cycle C' in G' containing two consecutive probes. Let X' be the vertex set corresponding to C.

If $G_{X'}$ is also a chordless cycle (in G), then G fails to be N-T, by Theorem 3.6.

Otherwise, in cycle $G'_{X'}$ there exists a unique edge xy which was added while making v simplicial. (If several edges were added, C' would not be chordless). Let $X = X' \cup \{v\}$. Clearly, G_X is a cycle, call it C; suppose it fails to be chordless: v has a neighbor w on C, $w \neq x, y$; but in that case, edges xw and yw would have been added to G', which contradicts the fact that C' is chordless in G. Thus C is a chordless cycle with two consecutive probes, so by Theorem 3.6, G is not N-T. $\square$

Theorem 3.16. *Let $G = (P + N, E)$ be a graph. The following are equivalent:*

1. *G is an N-T graph.*
2. *G has a quasi-perfect elimination ordering.*
3. *A greedy elimination game on quasi-simplicial vertices succeeds.*

Proof. (2) ⇒ (1): Let $G = (P + N, E)$ be a graph with a qpeo α. Running the elimination game on (G, α) will add only edges between two non-probes, so it will produce an N-triangulation of G.

(1) ⇒ (2) Let $G = (P+N, E)$ be an N-T graph, let G' be an N-triangulation of G, and let $\beta = (v_1, \dots, v_n)$ be a peo of G'. We claim that β is a qpeo of G. By Property 2.3, $G' = G^+_\beta$. Since v_i is simplicial in $G'_{\{v_i,\dots,v_n\}}$, it is quasi-simplicial in $G_{\{v_i,\dots,v_n\}}$, because the elimination game only adds edges whose endpoints are non-probes at each step. Therefore, β is a qpeo of G.

(1) ⇒ (3): If the elimination game fails at some step, then the corresponding transitory graph has no quasi-simplicial vertex, so by the equivalence (1) $\iff$ (2), it fails to be N-T. Therefore, by Lemma 3.15, G is not N-T.

(3) ⇒ (2): Trivial. □

Complexity. A recognition algorithm can be given based on condition (3) of Theorem 3.16. This runs in $O(n^2m')$ time, where m' is the number of edges of the N-triangulation computed by the elimination game run on a qpeo, since a brute force approach will require $O(nm')$ time to find a quasi-simplicial vertex and process it. (A referee has pointed out that the complexity of recognizing N-T graphs in this way can be further reduced to $O(|P|m')$.)

In any case, this complexity is not as good as the $O(|P|m)$ time we found in section 3.1. However, there may be, as is the case for chordal graphs, a LEX M-type algorithm which could compute a qpeo in $O(|N|m)$ time—a question we leave open.

We now will use our results to extend Dirac's Theorem 2.1 to N-T graphs.

THEOREM 3.17. *Let $G = (P+N, E)$ be an N-T graph which is not a clique; then in G there are at least two non-adjacent quasi-simplicial vertices.*

Proof. By induction on the number of vertices. Clearly, any N-T graph on 4 vertices which is not a clique has two non-adjacent vertices which are quasi-simplicial.

Let us consider an N-T graph G with n vertices. By Theorem 3.16, G has a quasi-simplicial vertex x. If x sees all the other vertices in G, let G' be obtained by simply removing x from G. By the induction hypothesis, G' has two non-adjacent quasi-simplicial vertices, which are trivially also quasi-simplicial and non-adjacent in G.

Otherwise, let G' be obtained by saturating $\mathscr{N}_G(x)$ and removing x. According to Lemma 3.15, G' is an N-T graph. By the induction hypothesis, G' must have two non-adjacent quasi-simplicial vertices, at least one of which, call it z, is not in $\mathscr{N}_G(x)$, since in G', $\mathscr{N}_G(x)$ is a clique. We claim that z is quasi-simplicial in G. Suppose this is not the case: In G, z must see a non-edge $\{v, w\}$, with v a probe, which is not a non-edge of G', so edge vw must have been added to G' when making x simplicial. But since x is quasi-simplicial in G, there can be no such non-edge $\{v, w\}$ in $G_{\mathrm{N}\,(x)}$. Thus, in G, x and z are two non-adjacent quasi-simplicial vertices. □

4. Recognizing partitioned chordal probe graphs.

In this section, we apply our results from section 3 on N-triangulatable graphs to characterizing and recognizing partitioned chordal probe graphs.

THEOREM 4.1. *Let $G = (P+N, E)$, with N a stable set. The following conditions are equivalent:*

1. *G is chordal probe.*
2. *All probes of G are quasi-LB-simplicial.*
3. *G contains no chordless cycle with two adjacent probes.*

Proof. This follows directly from Theorem 3.6 and Remark 3.2 (4). □

Complexity. Provided we test that N is a stable set, the recognition algorithm for N-T graphs given in section 3.1 also recognizes chordal probe graphs, with the same $O(|P|m)$ complexity.

THEOREM 4.2. *Let $G = (P+N, E)$ be a graph, with N a stable set. The following three are equivalent:*

1. *G is a chordal probe graph.*
2. *G has a quasi-perfect elimination ordering.*
3. *A greedy elimination game on quasi-simplicial vertices succeeds.*

Proof. This follows immediately from Theorem 3.16. □

Remark 4.3. Theorem 4.2 defines an elimination process on neighborhoods which are split graphs, as the vertices in each neighborhood are partitioned into a clique of probes and a stable set of non-probes; moreover, they form a special kind of split graph, which can be qualified as "complete split graph," meaning that all possible edges between a probe and a non-probe belong to the graph. This extends the simplicial elimination process on chordal graphs, where the elimination is on complete neighborhoods.

Theorem 3.17 also trivially extends to chordal probe graphs:

COROLLARY 4.4. *Let $G = (P + N, E)$ be a chordal probe graph which is not a clique; then in G there are at least two non-adjacent quasi-simplicial vertices.*

Part II. The non-partitioned case. Having solved the partitioned case for recognizing chordal probe graphs in Part I, we will now go on to the non-partitioned case. Again, we do this by first solving a more general problem.

5. Decomposing an arbitrary graph into *C*-components.

As stated in the introduction, our approach to recognizing chordal probe graphs uses a lemma from [11] which remarks that in any valid partition of a chordal probe graph, probes and non-probes must alternate on every chordless cycle. In order to study this phenomenon, we first propose a partition of the vertices of a graph into components which group together cycles of the graph, thus introducing a new graph decomposition.

DEFINITION 5.1. *Let G be an arbitrary graph.*

1. *A C-*edge *is an edge which belongs to some chordless cycle.*
2. *A C-*path *is a path made out of C-edges.*
3. *A C-*component *is a set of vertices in which there is a C-path connecting each pair of vertices in the component.*
4. *An* external edge *is an edge which has its endpoints in two different C-components.*

Example 5.2. Figure 4 shows the partition into C-components of a graph.

PROPERTY 5.3. *Let G be an arbitrary graph; being connected by a C-path is an equivalence relation on the vertices of G; we will denote this relation by $\sim$.*

Proof. Trivially, if x and y are connected by a C-path, then y and x are also connected by the same path, thus ensuring symmetry. Transitivity: Let μ_1 be a C-path from x to y and μ_2 be a C-path from y to z; the concatenation of μ_1 with μ_2 is a C-path from x to z. □

PROPERTY 5.4. *Every chordless cycle is entirely contained in some C-component.*

Proof. Two vertices belonging to some chordless cycle are connected by a C-path, so by definition they must belong to some common C-component. □

PROPERTY 5.5. *Every antihole is entirely contained in some C-component.*

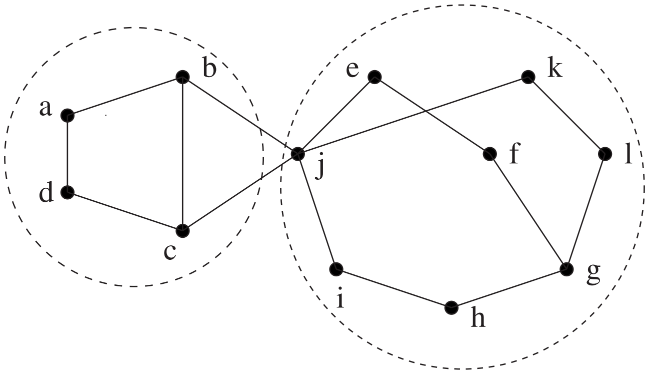


FIG. 4. *A graph and its partition into C-components.*

Proof. Every vertex of an antihole belongs to a C_4 along with every other vertex of the antihole. □

As a result of Properties 5.4 and 5.5, we have the following theorem.

THEOREM 5.6. *The decomposition into C-components is hole and antihole preserving.*

The partition induced by $\sim$ on the vertices of G very naturally defines a quotient graph, which we will denote by G^0.

DEFINITION 5.7. *Let $G = (V, E)$ be an arbitrary graph. Let us define the* quotient graph *$G^0 = (V', E')$ of G, where V' is the set of C-components of G and there is an edge between two C-components X_i and X_j if there is an edge in G with one endpoint in X_i and the other in X_j.*

THEOREM 5.8. *Let $G = (V, E)$ be an arbitrary graph. The quotient graph G^0 of G is chordal.*

Proof. Suppose graph G^0 is not chordal.

There must be a chordless cycle $C^0 = (X_1, \dots, X_k, X_1)$ in G^0. We will construct a corresponding chordless cycle in G. Let us consider three consecutive components X_i, X_{i+1}, and X_{i+2} of C^0, and let x be a vertex in X_{i+1}; x can see a vertex x' of another component of C^0 only if x' is either in X_i, or in X_{i+2}, else edge xx' corresponds to a chord of C^0.

In each component X_i of G^0, let us choose some vertex y_i which sees a vertex x_{i+1} of X_{i+1}. Thus our construction chooses in each component X_i two vertices, x_i and y_i; x_i is seen by y_{i-1}. Let P_i be a chordless path in X_i which connects x_i with y_i. Let us concatenate all of these paths: we obtain a cycle from which we can extract a chordless cycle C of G, which has vertices in different C-components, thus contradicting Property 5.4. □

Let us now discuss the case where a vertex does not belong to any chordless cycle. Recall that by Definition 2.13, a vertex is LB-simplicial if all its substars are cliques; by Property 2.12, a vertex is LB-simplicial if and only if it belongs to no chordless cycle. We will express this by the following property.

PROPERTY 5.9. *Let X be a C-component of a graph. The following propositions are equivalent:*

1. *X contains no chordless cycle.*
2. $|X| = 1$.
3. *X is an LB-simplicial vertex of the graph.*

We will call a C-component *trivial* when it contains no chordless cycle.

Complexity. Computing the C-components of a graph G can be done in $O(m^2)$ time using Definition 5.1. For each edge xy of the graph, one can determine whether it is part of a chordless cycle by removing the edge xy as well as the common neighbors of x and y; if in the resulting graph there is a path from x to y, then in the original graph xy belongs to a chordless cycle. This test requires $O(m)$ time for each edge, and thus all edges can be tested in $O(m^2)$ time. The C-components are then computed as being the connected components of the graph, obtained from G by removing all edges that do not belong to a chordless cycle.

6. Cycle-bicolorable graphs.

In this section, we present a new class of perfect graphs which generalizes chordal probe graphs in the case where no partition is given in advance. We exploit the property that in any valid partition, probes and non-probes must alternate on every chordless cycle.

In [11], the following lemma is shown for chordal probe graphs.

LEMMA 6.1 (see [11]). *If a graph $G = (P + N, E)$ is chordal probe with respect to the partition $\{P, N\}$ of its vertex set, then probes and non-probes alternate on every chordless cycle of G.*

We will use this property to introduce a new graph class.

DEFINITION 6.2. *We will say that a graph $G = (V, E)$ is* cycle-bicolorable *if and only if each vertex can be labeled with one of two colors in such a fashion that the colors alternate in every chordless cycle.*

Note that on a C-path in a cycle-bicolorable graph, the colors must alternate.

6.1. Recognition of cycle-bicolorable graphs.

The following proposition will allow us to characterize cycle-bicolorable graphs by considering each C-component separately.

PROPOSITION 6.3. *A graph is cycle-bicolorable if and only if each of its C-components is cycle-bicolorable.*

Proof. By Property 5.4, every chordless cycle of G is entirely contained in a unique C-component of G. Thus, coloring the chordless cycles inside each C-component is equivalent to coloring all chordless cycles. □

LEMMA 6.4. *Each cycle-bicolorable C-component has exactly two opposite bicolorings.*

Proof. Let us consider a C-component X_i, and let κ_1 be a bicoloring of X_i. By exchanging the colors of every vertex, another bicoloring κ_2 is obtained. Suppose there is a third possible coloring κ_3. Let x be a vertex whose color is different in κ_1 and κ_3. We claim that every other vertex in X_i has a different coloring in κ_1 and in κ_3. Suppose by contradiction that some vertex y of X_i has the same color in κ_1 as in κ_3. There is a C-path connecting x and y; since the colors in κ_3 must alternate on this path, the color of x uniquely determines the color of y, a contradiction. Thus, the color of every vertex of X_i is different in κ_1 as in κ_3, so κ_3 is the same as κ_2. □

Lemma 6.4 justifies the following definition.

DEFINITION 6.5. *Let X be a cycle-bicolorable C-component of a graph G. The bicoloring of G_X induce a unique partition of the vertices into $V_1 + V_2$, which we will call the* color bipartition *of X.*

To recognize cycle-bicolorable graphs, we determine the C-components as described in section 5, then check that each C-component is bicolorable. The correctness follows from Proposition 6.3. Figure 5 gives an easy algorithm to recognize cycle-bicolorable graphs.

```
Algorithm cycle-bicolorable recognition.
   input: A graph $G = (V, E)$, the set $\mathcal{X}$ of $C$-components of $G$.
  output: "Failure" if $G$ is not bicolorable, otherwise a black/white coloring of the
          vertices of $G$ such that the colors alternate on every chordless cycle.

//At the beginning, all vertices are uncolored and $Q$ is an empty queue;
while there remains some uncolored vertex do
    //$Q$ is empty.
    Choose a not yet colored $C$-component $\{X_i\}$ of $\mathcal{X}$;
    Choose a vertex $x$ of $\{X_i\}$, color it black and insert it into $Q$;
    while $Q$ is non-empty do
        Remove a vertex $y$ from $Q$;
        foreach neighbor $v$ of $y$ in $\{X_i\}$ do
            if $v$ has the same color as $y$ then
                return (failure);
            if $v$ is uncolored then
                Color $v$ with the color different from $y$'s, and insert $v$ into $Q$ ;
```

Fig. 5. *Algorithm CYCLE-BICOLORABLE RECOGNITION.*

Complexity. The complexity of recognizing cycle-bicolorable graphs is the same as that of computing the C-components of a graph, namely $O(m^2)$.

6.2. Some properties of cycle-bicolorable graphs.

Theorem 6.6. *The class of cycle-bicolorable graphs is perfect.*

In order to prove Theorem 6.6, we will need the following lemmas.

Lemma 6.7. *A cycle-bicolorable graph has no odd hole.*

Proof. Clearly, an odd chordless cycle cannot be labeled with two colors in a fashion that the colors alternate. □

Lemma 6.8. *A cycle-bicolorable graph has no antihole.*

Proof. Let $G = (V, E)$ be a cycle-bicolorable graph. By Lemma 6.7, G has no induced $\overline{C_5}$, since $\overline{C_5}$ is isomorphic to C_5. Suppose there exists $k \geq 6$, such that C_k is a chordless cycle of $\overline{G}$, with $C_k = (x_1, \ldots, x_k, x_1)$. Observe that $C' = (x_2, x_4, x_1, x_5, x_2)$ is a cycle of length 4 in G. In any bicoloring of V into black and white, x_1 and x_2 have the same color, w.l.o.g., black. Observe that x_1 sees all the vertices in $\overline{C_k}$, except for x_2 and x_k. Therefore, all the vertices in $\overline{C_k}$, except possibly x_2 and x_k, are white. But $C'' = (x_3, x_5, x_2, x_6, x_3)$ is also a chordless cycle of length 4 in G. In any bicoloring of V, either x_2 and x_3 are black or x_5 and x_6 are black, a contradiction. □

Theorem 6.6 follows directly from the Strong Perfect Graph Theorem and from Lemmas 6.7 and 6.8.

Remark 6.9. There are graphs with no odd antiholes and no odd holes which are not cycle-bicolorable, as is the case for an even antihole. Figure 6 shows a graph, for which we thank Frédéric Maffray, which has no antiholes and no odd holes, and which is a Meyniel graph and is perfectly orderable, but which is not cycle-bicolorable.

Proposition 6.10. *Let G be a cycle-bicolorable graph, where we arbitrarily call P and N the classes induced by a color-bipartition of each C-component of G. Then*

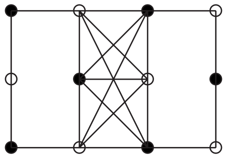


FIG. 6. *A graph with no antiholes and no odd holes, which is a Meyniel graph and is perfectly orderable, but which is not cycle-bicolorable.*

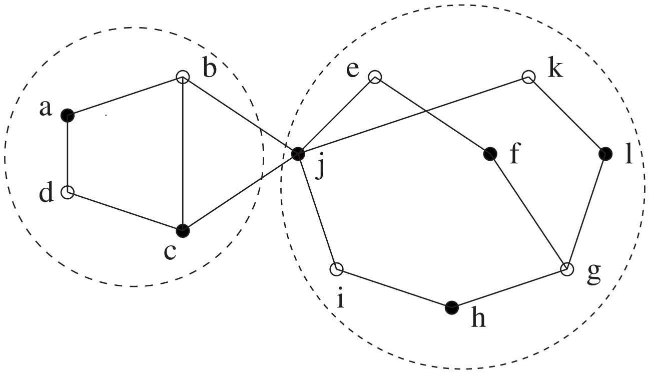


FIG. 7. *A cycle-bicolorable graph, its C-components, and a corresponding partition into white and black vertices.*

$G = (P + N, E)$ is an N-T graph.

Note that there are 2^t color-bipartitions where t is the number of C-components.

Proof of Proposition 6.10. By definition of a cycle-bicolorable graph, in the bicoloring of G, there can be no chordless cycle with two consecutive vertices which have the same color; let us arbitrarily call the color classes in each C-component of G probes and non-probes: there can be no chordless cycle with two consecutive probes, so by Theorem 3.6, graph G is N-T with respect to any partition induced by the bicolorings of its C-components. □

Note that the converse of Proposition 6.10 does not hold, as N-T graphs are not perfect and thus not always cycle-bicolorable, as is the case for the chordless cycle C_5.

7. Recognizing non-partitioned chordal probe graphs. From Lemma 6.1, we can easily deduce the following theorem.

THEOREM 7.1. *Chordal probe graphs are cycle-bicolorable graphs.*

The converse fails to hold: the complement of a P_6 is cycle-bicolorable but not chordal probe.

In section 6, we saw that a cycle-bicolorable graph can easily be bipartitioned, and we gave an $O(m^2)$ algorithm to do this. We will now apply our results to the recognition of chordal probe graphs.

LEMMA 7.2. *Let X_i be a C-component of a cycle-bicolorable graph. Then G_{X_i} is a chordal probe graph if and only if one of the colors of X_i forms a stable set.*

Example 7.3. In Figure 7, the white vertices form a stable set and can be labeled as non-probes; the graph is chordal probe.

Proof of Lemma 7.2. ⇒ Let G_{X_i} be a chordal probe graph, and let $P + N$ be a partition of X_i into probes and non-probes where N is a stable set. Since probes and non-probes alternate on every cycle, $P + N$ is the unique color bipartition of X_i, by Lemma 6.4. Thus, one of the colors, namely N, is the required stable set.

$\Leftarrow$ Let G_{X_i} be cycle-bicolorable such that one of the classes induced by the bicoloring, call it N, is a stable set. By Proposition 6.10, G_{X_i} is N-T with respect to this partition, so by definition G_{X_i} is chordal probe. □

In [11, 12], an algorithmic approach was presented to recognize chordal probe graphs in the case where the graph is weakly chordal. We observed in [2] (without proof) that their method (called Procedure *"Propagate Constraint Graph"*), can also be applied to arbitrary chordal probe graphs, using the following additional lemma. The recognition algorithm that we will present here is a further modification, and we provide a detailed proof.

LEMMA 7.4. *Let G be a chordal probe graph, let X_i and X_j be two bipartite C-components of the graph. Let x be a vertex of X_i which is an endpoint of at least two external edges connecting x to X_j. Then for any chordal probe partition $P + N$ of V, x is a probe.*

Proof. Let Y be the set of vertices of X_j which x sees. If G_Y has at least one edge e, then one of the endpoints of e is a probe and the other a non-probe, since X_j is bipartite; this forces x to be a probe. If G_Y has no edge, then let us choose a, b and a chordless path P in X_j such that P is shortest possible over all such pairs $\{a, b\}$; P together with edges ax and bx forms a chordless cycle, which is not fully contained in a C-component, a contradiction. □

Remark 7.5. In the case of N-triangulatable and cycle-bicolorable graphs, Proposition 6.10 showed that it was sufficient to combine any local assignment of $P + N$ to the cycle-bicoloring of each C-component to obtain an N-T graph. This is not the case for chordal probe graphs; we cannot simply apply Lemma 7.2 to each C-component and combine the results, since globally we must maintain N as a stable set. For this reason, we must insure that the external edges, which join one C-component to another, obey the constraint that their endpoints may not both be non-probes.

The considerations described in Remark 7.5 lead us to the algorithm NON-PARTITIONED CHORDAL PROBE GRAPH RECOGNITION presented later which decides whether an arbitrary graph G is chordal probe and if yes, computes a partition of the vertex set into probes and non-probes. Step 1 checks whether G is cycle-bicolorable, and if so, produces a bicoloring. Step 2 verifies that each C-component satisfies Lemma 7.2; if only one color is a stable set, then the labeling into probes and non-probes for that C-component is fixed by the LABEL-COMPONENT routine; if both colors are stable sets, then that component is a bipartite subgraph and no decision is made (yet) for its labeling. Step 3 applies the condition in Lemma 7.4; for every vertex x in a unlabeled (hence bipartite) C-component which sees two vertices in another unlabeled C-component, the labeling into probes and non-probes for the C-component containing x is fixed by the LABEL-COMPONENT routine. *Notice that LABEL-COMPONENT has the side effect of building a global queue Q of external edges which will have to be checked for consistency later in the algorithm.*

Step 4 uses the routine PROPAGATE to check the external edges uv for which u has been labeled a non-probe, to verify that v is either a probe or unlabeled; in the latter case, the labeling for the C-component containing v is fixed by the LABEL-COMPONENT routine. It terminates when the queue of all such edges is empty. Step 5 simply declares that the graph is now *recognized as being chordal probe*, although some C-components may still be unlabeled. Step 6 completes the labeling in a greedy manner. Figures 8 and 9 give the subroutines LABEL-COMPONENT and PROPAGATE which are used by the main algorithm.

```
ALGORITHM NON-PARTITIONED CHORDAL PROBE GRAPH RECOGNITION.
INPUT: A graph $G = (V, E)$.
OUTPUT: "NO" if $G$ is not chordal probe, otherwise a chordal probe labeling of
        $V$ with $P$ and $N$.
$Q$ is an empty queue;
STEP 1: CYCLE-BICOLORABLE RECOGNITION;
if "failure" is returned then
    return ("NO");
STEP 2: foreach $C$-component $X_i$ do
    if neither black nor white induces a stable set then
        return ("NO");
    if only one class of the color-bipartition induces a stable set then
        choose a vertex $x$ of this color and label it $N$;
        LABEL-COMPONENT ($x$, $N$);
// At this point, the only components left unlabeled are bipartite graphs;
STEP 3: foreach external edge $x_i x_j$, with $x_i$ in $C$-component $X_i$ and $x_j$ in
$C$-component $X_j$, $i \neq j$, where $X_i$ and $X_j$ are unlabeled and such that $x_i$ sees at
least two vertices in $X_j$ do
    LABEL-COMPONENT($x_i$, $P$);
STEP 4: PROPAGATE;
if "failure" is returned then
    return ("NO");
// At this point, any external edge $uv$ with $u$ labeled and $v$ unlabeled
// will have $u$ labeled as $P$;
STEP 5: $G$ is chordal probe;
//At this point, some of the $C$-components may remain unlabeled;
STEP 6: while there remain some unlabeled $C$-components do
    Arbitrarily choose an unlabeled vertex $x$;
    LABEL-COMPONENT ($x$, $P$);
    PROPAGATE;
```

Complexity. In the algorithm NON-PARTITIONED CHORDAL PROBE GRAPH RECOGNITION, the bottleneck is Step 1, which requires $O(m^2)$ time. All other steps have lower complexity. An N-triangulation can be obtained using the results for the partitioned case.

Correctness.

THEOREM 7.6. *The input graph G is chordal probe if and only if algorithm NON-PARTITIONED CHORDAL PROBE GRAPH RECOGNITION does not return "NO." Moreover, when G is a chordal probe graph, the algorithm produces a valid chordal probe partition $P + N$.*

Proof. ⇒ Suppose G is a chordal probe graph. We show that a "NO" answer gives a contradiction.

```
ALGORITHM LABEL-COMPONENT.
INPUT: A vertex $x$ and its label.
OUTPUT: Labels the vertices which are in the same $C$-component $X_i$ as $x$ and adds
        to global queue $Q$ all external edges with one endpoint in $X_i$ labeled $N$.

Label all the vertices of $X_i$ with $P$ or $N$ according to the label of $x$;
foreach external edge $uv$ with $u$ in $X_i$ and labeled $N$ do
    Add $(u, v)$ to $Q$;
```

Fig. 8. *Algorithm LABEL-COMPONENT.*

```
ALGORITHM PROPAGATE.
OUTPUT: Returns "failure" if there is a conflict, otherwise labels the vertices of
        some unlabeled $C$-component and may add some to-be-processed edges
        to the global queue $Q$.

while $Q$ is non-empty do
    $(u, v) \leftarrow$ dequeue($Q$); //$u$ is labeled $N$;
    if $v$ is labeled $N$ then
        return ("failure");
    if $v$ has no label then
        LABEL-COMPONENT($v$, $P$);
```

Fig. 9. *Algorithm PROPAGATE.*

If the algorithm returns "NO" in Step 1, then, by Theorem 7.1, G could not be a chordal probe graph. Therefore, Step 1 succeeds and produces a cycle-bicoloring of G. If the Algorithm returns "NO" in Step 2, then, by Lemma 7.2, G could not be a chordal probe graph. Therefore, Step 2 succeeds and assigns the probe/non-probe labeling for all non-bipartite C-components. Step 3 applies the condition in Lemma 7.4 and always succeeds.

Note that in Steps 2–3, a C-component is labeled with a probe/non-probe assignment only when the opposite assignment has been found to be contradictory.

If the Algorithm returns "NO" in Step 4, then the routine PROPAGATE returns "failure" for an external edge whose endpoints were both labeled non-probes, and by Remark 7.5, G could not be a chordal probe graph. Therefore, Step 4 succeeds.

At this point of the algorithm, the following properties hold.

Claim 1. *If X_i and X_j are unlabeled C-components, then there is at most one edge joining them; we call it the* exclusive *edge.*

Proof of Claim 1. If there were two such external edges uv and $u'v'$ with $u, u' \in X_i$ and $v, v' \in X_j$, by Lemma 7.4, having completed Steps 3–4 we have $u \neq u'$ and $v \neq v'$. From this it follows that the subgraph induced by uv and $u'v'$ and chordless paths in X_i and X_j connecting u with u' and v with v', respectively, contains a chordless cycle, contradicting Property 5.4.

Claim 2. *If $\{X_1, X_2, \ldots, X_k\}$ are unlabeled C-components forming a cycle C in the quotient graph G^0, i.e., the exclusive edges $u_i v_{i+1}$ exist joining X_i, X_{i+1} for all i (arithmetic mod k), then $u_i = v_i$ for all i.*

Proof of Claim 2 (by induction on k). If $k = 3$ and one of the equalities fails to hold, then combining shortest paths in X_1, X_2, X_3 connecting u_i with v_i, respectively, will yield a chordless cycle in G of length > 4, contradicting Property 5.4. If $k \geq 4$, then C has a chord, since the quotient graph G^0 is chordal (Property 5.8), thus splitting C into two smaller cycles C_1, C_2. So by induction, applying Claim 2 to C_1, C_2, we obtain all k equalities.

The remainder of the proof of Theorem 7.6 in this direction follows from the next claim.

Claim 3. *Step* 6 *never fails.*

Proof of Claim 3. Suppose Step 6 fails at an iteration where x was chosen to be labeled arbitrarily and where failure occurred for external edge uv, where vertex u is in component X_u and v is in X_v, both labeled N. The propagation defines a search tree in the subgraph of the quotient graph induced by the components labeled by that iteration of the propagation. Let X_j be the smallest common ancestor of X_u and X_v.

Consider the cycle C in G^0 formed by the exclusive edge uv and the two paths in the search tree from X_j to X_u and from X_j to X_v. Note, however, that the exclusive edges on these paths have the endpoint of the parent labeled N and the endpoint of the child labeled P, by the routine PROPAGATE. This contradicts Claim 2 and completes the proof of Claim 3.

$\Leftarrow$ If the algorithm succeeds, then the probe/non-probe partition $P+N$ is a cycle-bicoloring (Step 1), so by Proposition 6.10, G is an N-T graph with respect to $P+N$. Furthermore, N is a stable set since, if $u, v \in N$ and $uv \in E$, Step 2 implies uv is an external edge and PROPAGATE would cause a "failure." Therefore, by definition, G is a chordal probe graph, and we have produced a valid chordal probe partition $P + N$. □

8. Conclusion and open questions.

Though chordal probe graphs were originally defined as a generalization of interval probe graphs, they may have their own computational biology application as a special case of the chordal graph sandwich problem, which arises in reconstructing phylogenies, tree structures which model genetic mutations, when part of the information is missing [4]. In fact, the polynomiality of N-T graph recognition which we show in this paper also provides an interesting tractable subcase of the chordal graph sandwich problem [10].

Regarding the structure of chordal probe graphs and N-T graphs, it appears clearly from the results in this paper that they are similar to chordal graphs in many respects, with similar characterizations. The evident difference is that chordal graphs have no chordless cycles, but we have shown that such cycles can be structured into C-components, which enables us to handle them efficiently. We believe that C-components in a general graph may have many interesting properties, one example of which is the chordality of the quotient graph.

We have solved partitioned and non-partitioned chordal probe graph recognition. Our results also solve the problem of non-partitioned interval probe recognition in some subcases, for example, when G is asteroidal triple free or when the number of C-components is small. The general non-partitioned interval probe recognition problem has been solved in [5].

Acknowledgment. The authors thank the anonymous referee who pointed out an improvement in the complexity of computing the C-components of a graph.

REFERENCES

[1] A. Berry, J.-P. Bordat, P. Heggernes, G. Simonet, and Y. Villanger, *A wide-range algorithm for minimal triangulation from an arbitrary ordering*, J. Algorithms, 58 (2006), pp. 33–36.

[2] A. Berry, M.C. Golumbic, and M. Lipshteyn, *Two tricks to triangulate chordal probe graphs in polynomial time*, Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2004), New Orleans, LA, SIAM, Philadelphia, pp. 962–969.

[3] A. Berry, P. Heggernes, and G. Simonet, *The minimum degree heuristic and the minimal triangulation process*, in Graph-theoretic Concepts in Computer Science, Lecture Notes in Comput. Sci. 2880, Springer-Verlag, Berlin, pp. 58–70.

[4] H. L. Bodlaender, M. R. Fellows, and T. J. Warnow, *Two strikes against perfect phylogeny*, in Automata Languages and Programming, Lecture Notes in Comput. Sci. 623 (1992), Springer-Verlag, Berlin, pp. 273–283.

[5] G. J. Chang, A. J. J. Kloks, J. Liu, and S.-L. Peng, *The PIGs full Monty—A floor show of minimal separators*, in STACS 2005, Lecture Notes in Comput. Sci. 3404 (2005), pp. 521–532.

[6] M. Chudnovsky, N. Robertson, P. Seymour, and R. Thomas, *The strong perfect graph theorem*, Ann. of Math. (2), 164 (2006), pp. 51–229.

[7] G. A. Dirac, *On rigid circuit graphs*, Abh. Math. Sem. Univ. Hamburg, 25 (1961), pp. 71–76.

[8] D. R. Fulkerson and O. A. Gross, *Incidence matrices and interval graphs*, Pacific J. Math., 15 (1965), pp. 835–855.

[9] M. C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980. Second edition: *Annals of Discrete Mathematics* 57, Elsevier, Amsterdam, 2004.

[10] M. C. Golumbic, H. Kaplan, and R. Shamir, *Graph sandwich problems*, J. Algorithms, 19 (1995), pp. 449–473.

[11] M. C. Golumbic and M. Lipshteyn, *Chordal probe graphs (extended abstract)*, in Graph-theoretic concepts in Computer Science, Lecture Notes in Comput. Sci. 2880, Springer-Verlag, Berlin, pp. 249–260.

[12] M. C. Golumbic and M. Lipshteyn, *Chordal probe graphs*, Discrete Appl. Math., 143 (2004), pp. 221–237.

[13] M. C. Golumbic and A. N. Trenk, *Tolerance Graphs*, Cambridge University Press, Cambridge, UK, 2004.

[14] R. Hayward, *Weakly triangulated graphs*, J. Combin. Theory B, 39 (1985), pp. 200–208.

[15] J. L. Johnson and J. P. Spinrad, *A polynomial time recognition algorithm for probe interval graphs*, Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2001), Washington, D.C., SIAM, Philadelphia, pp. 477–486.

[16] T. Kloks, D. Kratsch, and J. Spinrad, *Treewidth and Pathwidth of Cocomparability Graphs of Bounded Dimension*, Res. Rep. 93-46, Eindhoven University of Technology, The Netherlands, 1993.

[17] C. G. Lekkerkerker and J. Ch. Boland, *Representation of a finite graph by a set of intervals on the real line*, Fund. Math., 51 (1962), pp. 45–64.

[18] R. M. McConnell and J. P. Spinrad, *Construction of probe interval models*, Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2002), San Francisco, CA, SIAM, Philadelphia, pp. 866–875.

[19] T. A. McKee and F. R. McMorris, *Topics in Intersection Graph Theory*, SIAM Monographs on Discrete Mathematics and Applications, Philadelphia, 1999.

[20] F. R. McMorris, C. Wang, and P. Zhang, *On probe interval graphs*, Discrete Appl. Math., 88 (1998), pp. 315–324.

[21] A. Parra and P. Scheffler, *How to use the minimal separators of a graph for its chordal triangulation*, in Automata, Languages and Programming, Lecture Notes in Comput. Sci., Springer-Verlag, Berlin, 944 (1995), pp. 123–134.

[22] L. Sheng, *Cycle free probe interval graphs*, Congr. Numer., 140 (1999), pp. 33–42.

[23] M. Steele, *The complexity of reconstructing trees from qualitative characters and subtrees*, J. Classification, 9 (1992), pp. 91–116.

[24] P. Zhang, *Probe Interval Graphs and Their Application to Physical Mapping of DNA*, manuscript, 1994.

SIAM J. DISCRETE MATH.
Vol. 21, No. 3, pp. 592–611

PREFIX REVERSALS ON BINARY AND TERNARY STRINGS[*]

COR HURKENS[†], LEO VAN IERSEL[†], JUDITH KEIJSPER[†], STEVEN KELK[‡], LEEN STOUGIE[†‡], AND JOHN TROMP[‡]

Abstract. Given a permutation π, the application of prefix reversal $f^{(i)}$ to π reverses the order of the first i elements of π. The problem of sorting by prefix reversals (also known as *pancake flipping*), made famous by Gates and Papadimitriou (*Discrete Math.,* 27 (1979), pp. 47–57), asks for the minimum number of prefix reversals required to sort the elements of a given permutation. In this paper we study a variant of this problem where the prefix reversals act not on permutations but on strings over a fixed size alphabet. We determine the minimum number of prefix reversals required to sort binary and ternary strings, with polynomial-time algorithms for these sorting problems as a result; demonstrate that computing the minimum prefix reversal distance between two binary strings is NP-hard; give an exact expression for the prefix reversal diameter of binary strings; and give bounds on the prefix reversal diameter of ternary strings. We also consider a weaker form of sorting called *grouping* (of identical symbols) and give polynomial-time algorithms for optimally grouping binary and ternary strings. A number of intriguing open problems are also discussed.

Key words. algorithms, genome comparison, pancake flipping, prefix reversals, sorting, strings

AMS subject classification. 68R05, 68R15, 68Q17, 92D20

DOI. 10.1137/060664252

1. Introduction.

For a permutation $\pi = \pi(0)\pi(1)\ldots\pi(n-1)$ the application of *prefix reversal* $f^{(i)}$, which we call *flip* for short, to π reverses the order of the first i elements: $f^{(i)}(\pi) = \pi(i-1)\ldots\pi(0)\pi(i)\ldots\pi(n-1)$. The problem of *sorting by prefix reversals* (MIN-SBPR), brought to popularity by Gates and Papadimitriou [7] and often referred to as the *pancake flipping problem*, is defined as follows: given a permutation π of $\{0, 1, \ldots, n-1\}$, determine its sorting distance, i.e., the smallest number of flips required to transform π into the identity permutation $01\ldots(n-1)$.[1]

MIN-SBPR has practical relevance in the area of efficient network design [9, 10], and arises in the context of computational biology when seeking to explain the genetic difference between two given species by the most parsimonious (i.e., shortest) sequence of gene rearrangements. The computational complexity of MIN-SBPR remains open. A recent 2-approximation algorithm [5] is currently the best-known approximation result.[2] Indeed, most studies to date have focused not on the computational complexity of MIN-SBPR but rather on determining the worst-case sorting distance $wc(n)$ over all length-n permutations, i.e., the "worst case scenario" for length-n permutations. From [7] and [9] we know that $(15/14)n \leq wc(n) \leq (5n+5)/3$.

A natural variant of MIN-SBPR is to consider the action of flips not on permutations but on strings over fixed size alphabets. The shift from permutations to strings alters the problem universe somewhat. With permutations, for example, the *distance*

[*] Received by the editors July 4, 2006; accepted for publication (in revised form) January 5, 2007; published electronically July 25, 2007. Part of this research has been funded by the Dutch BSIK/BRICKS project.
http://www.siam.org/journals/sidma/21-3/66425.html

[†] Technische Universiteit Eindhoven (TU/e), Den Dolech 2, 5612 AX Eindhoven, The Netherlands (wscor@win.tue.nl, l.j.j.v.iersel@tue.nl, j.c.m.keijsper@tue.nl).

[‡] Centrum voor Wiskunde en Informatica (CWI), Kruislaan 413, 1098 SJ Amsterdam, The Netherlands (S.M.Kelk@cwi.nl, Leen.Stougie@cwi.nl, John.Tromp@cwi.nl).

[1] We adopt the convention of numbering from 0 rather than from 1.

[2] Although not explicitly described as such, the algorithm provided ten years earlier in [3] is a 2-approximation algorithm for the *signed* version of the problem.

problem—i.e., given two permutations π_1 and π_2, determine the smallest number of flips required to transform π_1 into π_2—is equivalent to sorting, because the symbols can simply be relabeled to make either permutation equal to the identity permutation. For strings like 101, such a relabeling is not possible. Thus, the distance problem on string pairs appears to be strictly more general than the sorting problem on strings, naturally defined as putting all elements in nondescending order.

Indeed, papers by Christie and Irving [2] and Radcliffe, Scott, and Wilmer [11] explore the consequences of switching from permutations to strings; they both consider arbitrary (substring) reversals and *transpositions* (where two adjacent substrings are swapped). It has been noted that, viewed as a whole, such rearrangement operations on strings have bearing on the study of orthologous gene assignment [1], especially where the level of symbol repetition in the strings is low. There is also a somewhat surprising link with the relatively unexplored family of *string partitioning* problems [8]. To put our work in context, we briefly describe the most relevant (for this paper) results from [2] and [11].

The earlier paper [2] gives, in the case of both reversals and transpositions, polynomial-time algorithms for computing the minimum number of operations to sort a given binary string, as well as exact constructive diameter results on binary strings. Additionally, their proof that computing the reversal distance between strings is NP-hard supports the intuition that distance problems are harder than sorting problems on strings. They present upper and lower bounds for computing reversal and transposition distance on binary strings.

The more recent paper [11] gives refined and generalized reversal diameter results for non–fixed size alphabets. It also gives a polynomial-time algorithm for optimally sorting a ternary (3-letter alphabet) string with reversals. The authors refer to the prefix reversal counterparts of these (and other) results as interesting open problems. They further provide an alternative proof of Christie and Irving's NP-hardness result for reversals, and sketch a proof that computing the *transposition distance* between binary strings is NP-hard. As we later note, this proof can also be used to obtain a specific reducibility result for prefix reversals. They also have some first results on approximation (giving a PTAS—a *polynomial-time approximation scheme*—for computing the distance between *dense instances*) and on the distance between random strings, both of which apply to prefix reversals as well.

In this paper we supplement results of [2] and [11] by their counterparts on prefix reversals. In section 3 (*grouping*) we introduce a weaker form of sorting where identical symbols need only be grouped together, while the groups can be in any order. For grouping on binary and ternary strings we give a complete characterization of the minimum number of flips required to group a string, and provide polynomial-time algorithms for computing such an optimal sequence of flips. (The complexity of grouping over larger fixed size alphabets remains open, but as an intermediate result we describe how a PTAS can be constructed for each such problem.) Grouping aids in developing a deeper understanding of sorting, which is why we tackle it first. It was also mentioned as a problem of interest in its own right by Eriksson et al. [4]. Then, in section 4 (*sorting*), we give polynomial-time algorithms (again based on a complete characterization) for optimally sorting binary and ternary strings with flips. (The complexity of sorting also remains open for larger fixed size alphabets. As with grouping we thus provide, as an intermediate result, a PTAS for each such problem.) In section 5 we show that the flip diameter on binary strings is $n-1$, and on ternary strings (for $n > 3$) lies somewhere between $n-1$ and $(4/3)n$, with empirical support for the former. In section 6 we show that the flip distance problem on binary strings is NP-hard, and point out that a reduction in [11] also applies to prefix reversals, showing that the flip distance problem on *arbitrary* strings is polynomial-time reducible (in an approximation-preserving sense) to the binary problem. We conclude in section 7 with a discussion of some of the intriguing open problems that have emerged during this work. Indeed, our initial exploration has identified many basic (yet surprisingly difficult) combinatorial problems that deserve further analysis.

2. Preliminaries.

Let $[k]$ denote the first k nonnegative integers $\{0, 1, \ldots, k-1\}$. A k-ary string is a string over the alphabet $[k]$, while a string s is said to be *fully* k-ary, or to *have arity* k, if the set of symbols occurring in it is $[k]$.

We index the symbols in a string s of length n from 1 through n: $s = s_1 s_2 \ldots s_n$. Two strings are *compatible* if they have the same symbol frequencies (and hence the same length); e.g., 0012 and 1002 are compatible but 0012 and 0112 are not. For a given string s, let $I(s)$ be the string obtained by sorting the symbols of s in nondescending order, e.g., $I(1022011) = 0011122$. The prefix reversal (flip for short) $f^{(i)}(s)$ reverses the length i prefix of its argument, which should have length at least i. Alternatively, we denote application of $f^{(i)}(s)$ by underlining the length i prefix. Thus, $f^{(2)}(2012) = \underline{20}12 = 0212$ and $f^{(3)}(2012) = \underline{201}2 = 1022$. The *flip distance* $d(s, s')$ between two strings s and s' is defined as the smallest number of flips required to transform s into s' if they are compatible, and ∞ otherwise. Since a flip is its own inverse, flip distance is symmetric.

The *flip sorting distance* $d_{\mathrm{s}}(s) = d(s, I(s))$ of a string s is defined as the number of flips of an *optimal sorting sequence* needed to transform s into $I(s)$. An algorithm sorts s optimally if it computes an optimal sorting sequence for s.

In the next two sections we consider strings to be equivalent if one can be transformed into the other by repeatedly duplicating symbols and eliminating one of two adjacent identical symbols. As representatives of the equivalence classes we take the shortest string in each class. These are exactly the strings in which adjacent symbols always differ. We express all flip operations in terms of these *normalized* strings. For example, we write $f^{(3)}(2012) = \underline{201}2 = 102$. A flip that brings two identical symbols together, thereby shortening the string by 1, is called a 1-*flip*, while all others, which leave the string length invariant, are called 0-*flips*.

We follow the standard notation for regular expressions: superindex ^{i} on a substring denotes the number of repetitions of the substring, with * and $^+$ denoting 0-or-more and 1-or-more repetitions, respectively; ϵ denotes the empty string; brackets of the form $\{\}$ are used to denote that a symbol can be exactly one of the elements within the brackets; and the product sign $\prod$ denotes concatenation of an indexed series. For example, $\prod_{i=1}^{3}(10^i 2) = 102100210002$, and $\{1, 01\}^*\{\epsilon, 0\}$ denotes the set of binary strings with no 00 substring.

3. Grouping.

The task of sorting a string can be broken down into two subproblems: *grouping* identical symbols together and putting the groups of identical symbols in the right *order*. Notice that first grouping and then ordering may not be the most efficient way to sort strings. Although grouping appears to be slightly easier than the sorting problem, essentially the same questions remain open as in sorting. Grouping binary strings is trivial, and in section 3.1 we give the grouping distances of all ternary strings. As a result we give polynomial time algorithms for binary and ternary grouping. For larger alphabets the grouping problem remains open; as an intermediate result we describe in section 3.2 a PTAS for each such problem. While the problems of grouping and sorting are closely related for strings on small alphabets, the problems diverge when alphabet size approaches the string length, with permutations being the limit.

Recall that we consider only normalized strings, as representatives of equivalence classes. The *flip grouping distance* $d_{\mathrm{g}}(s)$ of a fully k-ary string s is defined as the minimum number of flips required to reduce the string to one of length k.

3.1. Grouping binary and ternary strings.

LEMMA 1. *$d_{\mathrm{g}}(s) \geq n-k$ for any fully k-ary string s of length n.*

Proof. The proof follows from the observations that, after grouping, fully k-ary string s has length k and that each flip can shorten s by at most 1. □

LEMMA 2. *$d_{\mathrm{g}}(s) \leq n-2$ for any fully k-ary string s of length n.*

Proof. Consider the following simple algorithm. If the leading symbol occurs elsewhere, then a 1-flip bringing them together exists, so perform this 1-flip. If not, then we use a 0-flip to put this symbol in front of a suffix in which we accumulate uniquely appearing symbols. Repeat until the string is grouped.

Clearly no more than $n-k$ 1-flips will be necessary. Also, no more than $k-2$ 0-flips will ever be necessary, because after $k-2$ 0-flips the prefix of the string will consist of only two types of symbol, and the algorithm will never perform a 0-move on such a string. Thus at most $(n-k)+(k-2) = n-2$ flips in total will be needed. □

As a corollary we obtain the grouping distance of binary strings.

THEOREM 1. *$d_{\mathrm{g}}(s) = n-2$ for any fully binary string s of length n.* □

We will now define a class of bad ternary strings and prove that these are the only ternary strings that need $n-2$ rather than $n-3$ flips to be grouped.

DEFINITION 1. *We define* bad *strings as all fully ternary strings of one of the following types, up to relabeling:*

I. *strings of length greater than* 3*, in which the leading symbol appears only once:* $0(12)^{\geq 2}$ *and* $02(12)^{+}$*;*

II. *strings having identical symbols at every other position, starting from the last:* $(\{0,1\}2)^{+}$ *and* $(2\{0,1\})^{+}2$*;*

III. *odd length strings whose leading symbol appears exactly once more, at an even position, and both occurrences are followed by the same symbol:* $0(21)^{+}02(12)^{*}$*;*

IV. *the following strings:* $X_1 = 210212$*,* $X_2 = 021012$*,* $X_3 = 0120212$*,* $X_4 = 1201212$*,* $X_5 = 02101212$*,* $X_6 = 20210212$*,* $X_7 = 020210212$*,* $X_8 = 120120212$*.*

All other fully ternary strings are good*. Strings of type* I, II, *and* III*, shortly* I-, II-, *and* III-*strings, respectively, are called generically bad, or* g-bad *for short.*

LEMMA 3. *$d_{\mathrm{g}}(s) = n-2$ if ternary string s of length n is bad.*

Proof. Because of Lemmas 1 and 2, it suffices to show that in each case a 0-flip is necessary: I-strings admit only 0-flips. A 1-flip on a II-string leads to a II-string and eventually to a I-string. Any III-string admits only one 1-flip leading to a II-string. For IV-strings, Table 1 shows that each possible 1-flip leads either to a shorter IV-string or to a I-,II-, or III-string. □

LEMMA 4. *$d_{\mathrm{g}}(s) = n-3$ if ternary string s of length n is good.*

Proof. The proof is by induction on n. The induction basis for $n = 3$ is trivial. We show the statement for strings of length $n+1$ by showing that if a bad string s' of length n can be obtained through a 1-flip from a good (*parent*) string s of length $n+1$, then s admits another 1-flip which leads to a good string. Note that a 1-flip $f^{(i)}(s) = s'$ brings symbols s_1 and s_{i+1} together; hence $s_1 = s_{i+1} \neq s_i = s'_1$, which shows that the symbol deleted from *parent* s differs from the leading symbol of *child*

TABLE 1
Type IV *strings and all their* 1*-flips.*

X_1	X_6
$\underline{210}212 = 01212$ is of type I	$\underline{20}210212 = 0210212$ is of type III
$\underline{21021}2 = 12012$ is of type III	$\underline{20210}212 = 0120212 = X_3$
X_2	$\underline{2021021}2 = 1201202$ is of type III
$\underline{021}012 = 12012$ is of type III	X_7
X_3	$\underline{02}0210212 = 20210212 = X_6$
$\underline{012}0212 = 210212 = X_1$	$\underline{020210}212 = 12020212$ is of type II
X_4	X_8
$\underline{120}1212 = 021212$ is of type I and II	$\underline{120}120212 = 02120212$ is of type II
$\underline{12012}12 = 210212 = X_1$	$\underline{1201202}12 = 20210212 = X_6$
X_5	
$\underline{021}01212 = 1201212 = X_4$	

TABLE 2
Type IV *strings, their parents, and for each good parent, a* 1*-flip to a good string.*

X_1	Parents	X_4	Parents	X_7	Parents
210212	$\underline{12102}12$	1201212	$\underline{21201}212$	020210212	$\underline{2020210}212$
	$0120212 = X_3$		$02101212 = X_5$		$\underline{2020210}212$
	$1201212 = X_4$		$\underline{2102121}2$		$\underline{12020102}12$
X_2	Parents		$\underline{21}210212$		$\underline{201}2020212$
021012	$\underline{202101}2$	X_5	Parents		$\underline{120}1202012$
	$\underline{12010}12$	02101212	$\underline{202101}212$		$\underline{20210}21212$
	$\underline{10}12012$		$\underline{12010}1212$	X_8	Parents
	$\underline{2010}212$		$\underline{10}1201212$	120120212	$\underline{21201}20212$
X_3	Parents		$\underline{2101}20212$		$\underline{021012}0212$
0120212	$\underline{101202}12$		$\underline{12}1012012$		$\underline{210}2120212$
	$\underline{21020}212$		$\underline{20}2010212$		$\underline{021}0210212$
	$20210212 = X_6$	X_6	Parents		$\underline{20210}21212$
	$\underline{1202}1012$	20210212	$020210212 = X_7$		$\underline{21}20210212$
	$\underline{20}212012$		$\underline{1202102}12$		
			$\underline{012}020212$		
			$120120212 = X_8$		

s'. We enumerate all possible bad child strings s' and distinguish cases based on the leading symbol of good parent s.

For IV-strings, Table 2 lists all parents with, for each good parent, a 1-flip to a good string. It remains to prove that for each g-bad string all parents are either bad or have a g-1-flip, defined as a 1-flip resulting in a string that is not g-bad (i.e., either good or of type IV).

Type I, odd: $0(12)^{\geq 2}$ has possible parents starting with:

1: $1(21)^i012(12)^j$ with $i + j > 0$.
If $i > 0$, there is a g-1-flip $\underline{12}1(21)^{i-1}012(12)^j = (21)^i012(12)^j$;
If $i = 0$ and $j > 0$, there is a g-1-flip $\underline{1012}(12)^j = 210(12)^j$;

2: $21(21)^i02(12)^j$ with $i + j > 0$.
If $i > 0$, there is a g-1-flip $\underline{21}(21)^i02(12)^j = 1(21)^i02(12)^j$;
If $i = 0$ and $j > 1$, there is a g-1-flip $\underline{21021}2(12)^{j-1} = 120(12)^j$;
If $i = 0$ and $j = 1$, the parent is $210212 = X_1$.

Type I, even: these strings are also of type II; see below.

Type II, odd: $(2\{0,1\})^+2$ has only parents of type II.

Type II, even: $02(\{0,1\}2)^*$ has possible parents starting with:

2: $2(\{0,1\}2)^*$ is of type II;

1: $12(\{0,1\}2)^*012(\{0,1\}2)^*$ with three cases for a possible third 1:

None: parent is $12(02)^*012(02)^*$, which is of type III;
Before 01**:** then there is a g-1-flip
$\underline{12(\{0,1\}2)^*}12(\{0,1\}2)^*012(\{0,1\}2)^* = 2(\{0,1\}2)^*12(\{0,1\}2)^*$
$012(\{0,1\}2)^*$;
After 01: then there is a g-1-flip
$\underline{12(\{0,1\}2)^*012(\{0,1\}2)^*}12(\{0,1\}2)^* = 2(\{0,1\}2)^*102(\{0,1\}2)^*$
$12(\{0,1\}2)^*$.

Type III: $0(21)^+02(12)^*$ has possible parents starting with:
1: $(12)^i01(21)^j02(12)^k$ with $i > 0$.
If $i > 1$, there is a g-1-flip
$\underline{12}(12)^{i-1}01(21)^j02(12)^k = 2(12)^{i-1}01(21)^j02(12)^k$;
If $i = 1, j > 0$, there is a g-1-flip
$\underline{12012}1(21)^{j-1}02(12)^k = 21021(21)^{j-1}02(12)^k$;
If $i = 1, j = 0, k > 0$, there is a g-1-flip $\underline{120102}(12)^k = 20102(12)^k$;
If $i = 1, j = k = 0$, then the parent is $120102 = X_2$ (relabeled);
1: $(12)^+0(12)^+0(12)^+$: there is a g-1-flip
$\underline{(12)^+0}(12)^+0(12)^+ = 0(21')^+20(12)^+$;
2: $2(12)^*0(21)^+02(12)^*$: there is a g-1-flip
$\underline{2(12)^*0(21)^+0}2(12)^* = 0(12)^+0(21)^*2$;
2: $(21)^i20(12)^j02(12)^k$ with $j > 0$.
If $i = 0, j = 1$, then the parent is $210212 = X_1$;
If $i + j > 1$, then $\underline{(21)^i2012}(12)^{j-1}02(12)^k = 102(12)^{i+j-1}02(12)^k$ is a g-1-flip. □

The following theorem results directly from Lemmas 3 and 4.

THEOREM 2. $d_g(s) = n - 2$ *if and only if fully ternary string* s *of length* n *is bad and* $d_g(s) = n - 3$ *otherwise. Moreover, there exists a polynomial-time algorithm for grouping ternary strings with a minimum number of flips.*

Proof. The first statement is direct from Lemmas 3 and 4. In case string s is bad, which by Definition 1 can be decided in polynomial time, the algorithm implicit in the proof of Lemma 2 shows how to group s optimally in polynomial time. Otherwise, we repeatedly find a 1-flip to a good string, as guaranteed by Lemma 4. The time complexity is $O(n^3)$, since grouping distance, number of choices for a 1-flip, and time to perform a flip and test whether its result is good are all $O(n)$. □

3.2. Grouping strings over larger alphabets. Lemmas 1 and 2 say that $n-k \leq d_g(s) \leq n-2$ for any fully k-ary string s. For any k there are fully k-ary strings that have flip grouping distance equal to $n-2$. For example, the length $n = 2(k-1)$ string $1020\dots(k-1)0$ requires that every 1-flip bring a 0 to the front first, and hence we need as many 0-flips as 1-flips, and $d_g(1020\dots(k-1)0) \geq 2(k-2) = 2k-4 = n-2$. Computer calculations suggest that for $k = 4$ and $k = 5$, for n large enough, the strings with grouping distance $n-2$ are precisely those having identical symbols at every other position, starting from the last (i.e., type II of Definition 1). Proving (or disproving) this statement remains open, as well as finding a polynomial-time algorithm for grouping k-ary strings for any fixed $k > 3$. We do, however, have the following intermediate result.

THEOREM 3. *For every fixed* k *there is a PTAS for grouping* k*-ary strings.*

Proof. We show that, for every fixed k and for every fixed $\epsilon > 0$ there is a polynomial-time algorithm that, given any k-ary string s of length n, computes a sequence of flips which groups s in at most $(1+\epsilon)d_g(s)$ flips. We assume $k \geq 4$ because for $k = 2$ and $k = 3$ the exact algorithms suffice. Let $N = (k-2)/\epsilon + k$. We distinguish two cases.

Case 1. If $n \geq N$, we use the simple "greedy" algorithm described in the proof of Lemma 2. This will group s in $d_{\mathrm{g}}^{G}(s)$ flips with $d_{\mathrm{g}}^{G} \leq n{-}2$ steps. This together with the lower bound of $n{-}k$ on $d_{\mathrm{g}}(s)$ from Lemma 1 gives $d_{\mathrm{g}}^{G}(s) \leq d_{\mathrm{g}}(s){+}(k{-}2) \leq (1{+}\epsilon)d_{\mathrm{g}}(s)$.

Case 2. If $n < N$, we compute $d_{\mathrm{g}}(s)$ by a brute force algorithm which simply chooses the best among all possible flip sequences of length $n-2$: there are n^{n-2} of these. This yields the optimal solution since $d_{\mathrm{g}}(s) \leq n-2$ (Lemma 2). The running time in this case is bounded by a constant. □

Clearly, there is a strong relationship between grouping and sorting. Understanding grouping may help us to understand sorting and lead to improved bounds (especially as the length of strings becomes large relative to their arity), because for a k-ary string s, we have $d_{\mathrm{g}}(s) \leq d_{\mathrm{s}}(s) \leq d_{\mathrm{g}}(s) + wc(k)$, with $wc(k)$ the flip diameter on permutations with k elements, as defined before.

Also $d_{\mathrm{g}}(s) = \min\{d_{\mathrm{s}}(t) : t \text{ a relabeling of } s\}$, which gives (for fixed k) a polynomial-time reduction from grouping to sorting. Thus every polynomial-time algorithm for sorting by prefix reversals directly gives a polynomial-time algorithm for the grouping problem (for fixed k).

4. Sorting.

In this section we present results on sorting similar to those on grouping in the previous section. Also flip sorting distance remains open for strings over alphabets of size larger than 3. As an intermediate result we thus provide at the end of this section a PTAS for each such problem.

Again a 1-*flip* brings identical symbols together and thus shortens the representative of the equivalence class under symbol duplication. But since symbol order matters for sorting, relabeled strings are no longer equivalent. As in grouping, sorting of binary strings is straightforward, as seen in the following.

THEOREM 4. *$d_{\mathrm{s}}(s) = n-2$ for every fully binary string s of length n with $s_n = 1$, and $d_{\mathrm{s}}(s) = n-1$ otherwise.*

Proof. Exactly $n-2$ 1-flips suffice and are necessary to arrive at length 2 string 01 or 10. If the last symbol is 0, an additional 0-flip is necessary, putting a 1 at the end. All these flips can be $f^{(2)}$. □

From Lemma 1 we know that $d_{\mathrm{g}}(s) \geq n-3$ and hence $d_{\mathrm{s}}(s) \geq n-3$ for every ternary string s of length n. In the upper bound on $d_{\mathrm{s}}(s)$ that we derive below we focus on strings s ending in a 2 ($s_n = 2$), since sorting distance is invariant under appending a 2 to a string. It turns out that, when sorting a ternary string ending in a 2, one needs at most one 0-flip, except for the string 0212.

LEMMA 5. *$d_{\mathrm{s}}(s) \leq n-2$ for every fully ternary string s of length n with $s_n = 2$, except* 0212.

Proof. It is easy to check that 0212 requires three flips to be sorted. By induction on n we prove the rest of the lemma. The basis case of $n = 3$ is trivial. For a string s of length $n > 3$ we distinguish three cases:

- $s_{n-1} = 0$: If $s = 20102$, it is sorted in three flips: $\underline{2010}2 \rightarrow \underline{01}02 \rightarrow \underline{10}2 \rightarrow 012$. Otherwise, by induction and relabeling $0 \leftrightarrow 2$, the string $s_1 \ldots s_{n-1}$ can be reduced to 210 in $n-3$ flips (to 20 or 10 by Theorem 4 if $s_1 \ldots s_{n-1}$ has only two symbols), and one more flip sorts s to 012.
- $s_{n-1} = 1$, $s_1 = 0$ and appears only once: Thus $s = 0(12)^{\geq 2}$ or $s = 02(12)^{\geq 2}$. Then s can be sorted with only one 0-flip: $\underline{0(12)^{+}1}2 \rightarrow \underline{1(2}1)^{+}02 \rightarrow \ldots \rightarrow \underline{210}2 \rightarrow 012$ or, respectively, $\underline{02}(12)^{\geq 2} \rightarrow \underline{20(12)^{+}1}2 \rightarrow \underline{(12)^{+}}102 \rightarrow \ldots \rightarrow \underline{210}2 \rightarrow 012$.

TABLE 3
Type IX *strings.*

$Y_1 = 210212$	$Y_{21} = 10212012$	$Y_{41} = 021202012$	$Y_{61} = 0210212012$
$Y_2 = 021012$	$Y_{22} = 02121012$	$Y_{42} = 021201012$	$Y_{62} = 1021202012$
$Y_3 = 212012$	$Y_{23} = 02120102$	$Y_{43} = 020210212$	$Y_{63} = 1021201012$
$Y_4 = 120102$	$Y_{24} = 10102102$	$Y_{44} = 101020212$	$Y_{64} = 1020210212$
$Y_5 = 201202$	$Y_{25} = 02010212$	$Y_{45} = 020212012$	$Y_{65} = 1010210202$
$Y_6 = 0210202$	$Y_{26} = 21202012$	$Y_{46} = 212010202$	$Y_{66} = 0202010212$
$Y_7 = 1021202$	$Y_{27} = 21201012$	$Y_{47} = 212012012$	$Y_{67} = 2120202012$
$Y_8 = 0212012$	$Y_{28} = 21201202$	$Y_{48} = 010210212$	$Y_{68} = 2120102012$
$Y_9 = 2120102$	$Y_{29} = 20210212$	$Y_{49} = 010210202$	$Y_{69} = 2021021212$
$Y_{10} = 0102102$	$Y_{30} = 01021202$	$Y_{50} = 010212012$	$Y_{70} = 2010212012$
$Y_{11} = 1212012$	$Y_{31} = 01020212$	$Y_{51} = 202010212$	$Y_{71} = 1201021202$
$Y_{12} = 2010212$	$Y_{32} = 20212012$	$Y_{52} = 121202012$	$Y_{72} = 1201202012$
$Y_{13} = 0120212$	$Y_{33} = 12120102$	$Y_{53} = 121201202$	$Y_{73} = 10202010212$
$Y_{14} = 1201012$	$Y_{34} = 12010212$	$Y_{54} = 201021202$	$Y_{74} = 02120102012$
$Y_{15} = 1201212$	$Y_{35} = 12010202$	$Y_{55} = 120212012$	$Y_{75} = 02021021212$
$Y_{16} = 2012012$	$Y_{36} = 20120102$	$Y_{56} = 012021212$	$Y_{76} = 21201202012$
$Y_{17} = 10210212$	$Y_{37} = 12012012$	$Y_{57} = 120102012$	$Y_{77} = 12120202012$
$Y_{18} = 21021212$	$Y_{38} = 021021202$	$Y_{58} = 201202012$	
$Y_{19} = 02102012$	$Y_{39} = 102120102$	$Y_{59} = 120120212$	
$Y_{20} = 02101212$	$Y_{40} = 102010212$	$Y_{60} = 201201012$	

- $s_{n-1} = 1$, s_1 not unique: If $s = 12012$, then three flips suffice: $\underline{12}012 \rightarrow \underline{2101}2 \rightarrow \underline{10}12 \rightarrow 012$. Otherwise, since the other two parents of 0212 can flip to 1202, there is a 1-flip to a string $\neq 0212$, to which we can apply the induction hypothesis. □

As in section 3, we characterize the strings ending in a 2 that need $n-2$ rather than $n-3$ flips to sort.

DEFINITION 2. *We define* bad *strings as all fully ternary strings ending in a* 2 *of the following types:*

I. $0(12)^{\geq 2}$,
II. $(\{0,1\}2)^+$ *and* $2(\{0,1\}2)^+$,
III. $(\{1,2\}0)^+2$ *and* $0(\{1,2\}0)^+$,
IV. $(\{1,2\}0)^+12$ *and* $(0\{1,2\})^+012$ *with at least two* 2*'s,*
V. $(01)^*0212$ *and* $(10)^+212$,
VI. $1(20)^+1(20)^*2$ *and* $0(21)^+0(21)^*2$,
VII. $1(02)^+1(02)^+$,
VIII. $1(02)^+12$,
IX. 77 *strings of length at most* 11, *shown in Table* 3.

All other fully ternary strings ending in a 2 *are* good *strings. Strings of type* I–VIII *(*I*-strings …*VIII*-strings, for short) are called generically bad, or* g-bad *for short.*

This definition makes 0212 a bad string as well. From Lemma 5 we know that 0212 is the only ternary string ending in a 2 with sorting distance $n-1$.

THEOREM 5. *String* 0212 *has sorting distance* 3*. Any other fully ternary string* s *of length* n *with* $s_n = 2$ *has prefix reversal sorting distance* $n-2$ *if it is bad and* $n-3$ *if it is good. A fully ternary string* s *ending in a* 0 *or* 1 *has the same sorting distance as* $s2$.

Proof. The proof follows directly from Lemmas 6 and 7 below. Note that every sorting sequence for s sorts $s2$ as well, while every sorting sequence for $s2$ can be modified to avoid flipping the whole string and thus works for s as well. □

LEMMA 6. $d_s(s) = n-2$ *for every bad ternary string* $s \neq 0212$ *of length* n.

Proof. Since $d_s(s) \geq n-3$ and any 1-flip decreases the length of the string by 1, Lemma 5 says it suffices to show that for each type in Definition 2 a 0-flip is necessary:

- For I-strings only 0-flips are possible.
- A 1-flip on a II- or III-string leads to a string of the same type, so that eventually no 1-flip is possible.
- A 1-flip on a IV-string leads either again to a IV-string or (when destroying the 12 suffix) to a III-string.
- A 1-flip on a V-string leads either again to a V-string or (when destroying the suffix with a $\underline{\ldots 02}12$ flip) to a IV-string. Flips $\underline{\ldots 0}212$ and $\underline{\ldots 021}2$ are not possible for lack of more 2's.
- For strings of VI-, VII- and VIII-strings only one 1-flip is possible, leading to II-, III- and IV-strings respectively.
- For IX-strings, Table 4 lists all possible 1-flips, ultimately leading to a string of type I–VIII. □

LEMMA 7. *$d_{\mathrm{s}}(s) = n - 3$ for every good ternary string s of length n.*

Proof. The proof is by induction on n and is similar to the proof of Lemma 4. The induction basis for $n = 3$ is again trivial. We prove that for each g-bad string of length n all parents (of length $n + 1$) either are bad or have a 1-flip to a string that is not g-bad (i.e., either good or of type IX). Remember that such a flip is called a *g-1-flip*. That for each IX-string all parents either are bad or have a 1-flip to a good string is proved by case checking in Table 4. Together this proves that every good string of length $n+1$ has a 1-flip to a good string of length n, and therefore the lemma is proved.

Type I: $0(12)^+$ has possible parents starting with:

1: $1(21)^i0(12)^j$ with $j > 0$.
If $i > 0$, there is a g-1-flip $\underline{12}1(21)^{i-1}0(12)^j = (21)^i0(12)^j$;
If $i = 0$, $j > 1$, there is a g-1-flip $\underline{1012}(12)^{j-1} = 210(12)^{j-1}$;
If $i = 0$, $j = 1$, there is a g-1-flip $\underline{10}12 = 012$;

2: $(21)^i02(12)^j$ with $i > 0$.
If $i > 1$, there is a g-1-flip $\underline{21}21(21)^{i-2}02(12)^j = 1(21)^{i-1}02(12)^j$;
If $i = 1$, $j > 0$, there is a g-1-flip $\underline{21021}2(12)^{j-1} = 120(12)^j$;
If $i = 1$, $j = 0$, there is a g-1-flip $\underline{210}2 = 012$.

Type II, even: $(\{0,1\}2)^+$ has possible parents starting with:

0: $0(2\{0,1\})^*2102(\{0,1\}2)^*$, with three cases for a possible third 0:
*None***:** the parent is of type VI;
Before 2102**:** there is a g-1-flip
$\underline{0(2\{0,1\})^*2}0(2\{0,1\})^*2102(\{0,1\}2)^* = 2(\{0,1\}2)^*0(2\{0,1\})^*$
$2102(\{0,1\}2)^*$;
After 2102**:** there is a g-1-flip
$\underline{0(2\{0,1\})^*2102(\{0,1\}2)^*}02(\{0,1\}2)^* = (2\{0,1\})^*2012(\{0,1\}2)^*$
$02(\{0,1\}2)^*$;

1: $1(2\{0,1\})^*2012(\{0,1\}2)^*$, with three cases for a possible third 1:
*None***:** the parent is of type VI;
Before 2012**:** there is a g-1-flip
$\underline{1(2\{0,1\})^*2}1(2\{0,1\})^*2012(\{0,1\}2)^* = 2(\{0,1\}2)^*1(2\{0,1\})^*$
$2012(\{0,1\}2)^*$;
After 2012**:** there is a g-1-flip
$\underline{1(2\{0,1\})^*2012(\{0,1\}2)^*}12(\{0,1\}2)^* = (2\{0,1\})^*2102(\{0,1\}2)^*$
$12(\{0,1\}2)^*$;

2: $2\{0,1\}(2\{0,1\})^*2(\{0,1\}2)^*$ is of type II.

Table 4

All strings of type IX *(first column). For each string all parents and all* 1*-flips are listed. Either each parent is bad, or a* 1*-flip to a good string is given. For each string of type* IX *is also shown that each* 1*-flip leads to a bad string. Here Pi denotes the parent you get by doubling the ith symbol and applying $p(i)$, and Ci denotes the string you get by applying the* 1*-flip $p(i-1)$. Note that if the ith symbol is* not *equal to the first symbol, there is a parent Pi, and if the ith symbol* is *equal to the first symbol, there is a* 1*-flip possible, leading to Ci.*

$Y_1 = 210212$	$P2 : \underline{12102}12$	$P3 = Y_{13}$	$C4$ is of type I	$P5 = Y_{15}$	$C6$ is of type VI
$Y_2 = 021012$	$P2 : \underline{202101}2$	$P3 = Y_{14}$	$C4$ is of type VI	$P5 : \underline{10}12012$	$P6 : \underline{2101}202$
$Y_3 = 212012$	$P2 = Y_{11}$	$C3$ is of type VI	$P4 = Y_8$	$P5 : \underline{10}21212$	$C6$ is of type V
$Y_4 = 120102$	$P2 = Y_9$	$P3 : \underline{02101}02$	$C4$ is of type VI	$P5 = Y_{10}$	$P6 = Y_{12}$
$Y_5 = 201202$	$P2 : \underline{0201}202$	$P3 = Y_7$	$C4$ is of type III	$P5 = Y_6$	$C6$ is of type VI
$Y_6 = 0210202$	$P2 : \underline{20210}202$ $P7 : \underline{20}201202$	$P3 = Y_{35}$	$C4$ is of type II	$P5 : \underline{2012020}2$	$C6 = Y_5$
$Y_7 = 1021202$	$P2 = Y_{30}$ $P7 = Y_{32}$	$P3 : \underline{20121}202$	$C4 = Y_5$	$P5 = Y_{28}$	$P6 = Y_{23}$
$Y_8 = 0212012$	$P2 = Y_{32}$ $P7 : \underline{210}21202$	$P3 = Y_{37}$	$P4 = Y_{26}$	$C5 = Y_3$	$P6 = Y_{21}$
$Y_9 = 2120102$	$P2 = Y_{33}$ $C7$ is of type V	$C3 = Y_4$	$P4 = Y_{23}$	$P5 : \underline{10}212102$	$P6 = Y_{30}$
$Y_{10} = 0102102$	$P2 = Y_{24}$ $P7 = Y_{36}$	$C3$ is of type VII	$P4 : \underline{2010210}2$	$P5 : \underline{120}10102$	$C6 = Y_4$
$Y_{11} = 1212012$	$P2 : \underline{2121201}2$ $P7 = Y_{18}$	$C3 = Y_3$	$P4 : \underline{2121201}2$	$P5 = Y_{22}$	$C6$ is of type I
$Y_{12} = 2010212$	$P2 = Y_{25}$ $C7 = Y_4$	$P3 = Y_{17}$	$P4 = Y_{31}$	$C5$ is of type V	$P6 = Y_{34}$
$Y_{13} = 0120212$	$P2 : \underline{101202}12$ $P7 : \underline{21}202102$	$P3 : \underline{21020}212$	$C4 = Y_1$	$P5 = Y_{29}$	$P6 : \underline{1202}1012$
$Y_{14} = 1201012$	$P2 = Y_{27}$ $P7 : \underline{21010}212$	$P3 : \underline{02101}012$	$C4 = Y_2$	$P5 : \underline{01}021012$	$C6$ is of type V
$Y_{15} = 1201212$	$P2 : \underline{21201}212$ $P7 : \underline{21}210212$	$P3 = Y_{20}$	$C4$ is of type I	$P5 = Y_{18}$	$C6 = Y_1$
$Y_{16} = 2012012$	$P2 : \underline{02012}012$ $C7$ is of type VII	$P3 = Y_{21}$	$C4$ is of type IV	$P5 = Y_{19}$	$P6 = Y_{17}$
$Y_{17} = 10210212$	$P2 = Y_{48}$ $C7 = Y_{16}$	$P3 : \underline{201210}212$ $P8 = Y_{47}$	$C4 = Y_{12}$	$P5 : \underline{012}010212$	$P6 : \underline{201}201212$
$Y_{18} = 21021212$	$P2 : \underline{12102}1212$ $P7 : \underline{12}1201212$	$P3 = Y_{56}$ $C8 = Y_{11}$	$C4$ is of type I	$P5 : \underline{1201212}12$	$C6 = Y_{15}$
$Y_{19} = 02102012$	$P2 : \underline{20210}2012$ $P7 : \underline{1020}12012$	$P3 = Y_{57}$ $P8 : \underline{210}201202$	$C4$ is of type VI	$P5 = Y_{58}$	$C6 = Y_{16}$
$Y_{20} = 02101212$	$P2 : \underline{202101}212$ $P7 : \underline{12}1012012$	$P3 : \underline{12010}1212$ $P8 : \underline{21}2101202$	$C4 = Y_{15}$	$P5 : \underline{10}1201212$	$P6 : \underline{2101}20212$
$Y_{21} = 10212012$	$P2 = Y_{50}$ $C7 = Y_8$	$P3 : \underline{20121}2012$ $P8 : \underline{210}212012$	$C4 = Y_{16}$	$P5 = Y_{47}$	$P6 = Y_{42}$
$Y_{22} = 02121012$	$P2 : \underline{2021}21012$ $P7 : \underline{10}1212012$	$P3 : \underline{12012}1012$ $P8 : \underline{2101}21202$	$P4 : \underline{21}2021012$	$P5 : \underline{1212010}12$	$C6 = Y_{11}$
$Y_{23} = 02120102$	$P2 : \underline{2021}20102$ $C7 = Y_7$	$P3 : \underline{120120}102$ $P8 = Y_{54}$	$P4 : \underline{21}2020102$	$C5 = Y_9$	$P6 = Y_{39}$
$Y_{24} = 10102102$	$P2 : \underline{0101021}02$ $P7 : \underline{012}010102$	$C3 = Y_{10}$ $P8 = Y_{60}$	$P4 : \underline{0101021}02$	$P5 : \underline{20101210}2$	$C6$ is of type III
$Y_{25} = 02010212$	$P2 = Y_{51}$ $P7 = Y_{57}$	$C3 = Y_{12}$ $P8 = Y_{46}$	$P4 = Y_{40}$	$C5$ is of type VIII	$P6 : \underline{20102021}2$
$Y_{26} = 21202012$	$P2 = Y_{52}$ $P7 : \underline{10202}1212$	$C3$ is of type VI $C8$ is of type VIII	$P4 = Y_{41}$	$C5 = Y_8$	$P6 = Y_{45}$
$Y_{27} = 21201012$	$P2 : \underline{1212010}12$ $P7 : \underline{10}1021212$	$C3 = Y_{14}$ $C8$ is of type V	$P4 = Y_{42}$	$P5 : \underline{102}121012$	$P6 = Y_{50}$
$Y_{28} = 21201202$	$P2 = Y_{53}$ $P7 = Y_{38}$	$C3$ is of type VI $C8$ is of type VI	$P4 : \underline{0212012}02$	$P5 : \underline{102}121202$	$C6 = Y_7$
$Y_{29} = 20210212$	$P2 = Y_{43}$ $P7 = Y_{59}$	$C3$ is of type VI $C8$ is of type VI	$P4 : \underline{1202102}12$	$P5 : \underline{012}020212$	$C6 = Y_{13}$
$Y_{30} = 01021202$	$P2 : \underline{10102}1202$ $C7 = Y_9$	$C3 = Y_7$ $P8 : \underline{2021}20102$	$P4 = Y_{54}$	$P5 : \underline{120}101202$	$P6 = Y_{46}$
$Y_{31} = 01020212$	$P2 = Y_{44}$ $P7 : \underline{12020}1012$	$C3$ is of type VIII $P8 : \underline{21}2020102$	$P4 : \underline{20102021}2$	$C5 = Y_{12}$	$P6 = Y_{51}$
$Y_{32} = 20212012$	$P2 = Y_{45}$ $P7 : \underline{102}120212$	$C3 = Y_8$ $C8 = Y_7$	$P4 = Y_{55}$	$C5$ is of type VI	$P6 = Y_{41}$

Type II, odd: $2(\{0,1\}2)^+$ has possible parents starting with:

0: $0(2\{0,1\})^*202(\{0,1\}2)^*$ is of type II;

1: $1(2\{0,1\})^*212(\{0,1\}2)^*$ is of type II.

Type III, even: $0(\{1,2\}0)^+2$ has possible parents starting with:

1: $1(0\{1,2\})^+010(\{1,2\}0)^*2$ is of type III;

2: $2(0\{1,2\})^+020(\{1,2\}0)^*2$ is of type III;

2: $2(0\{1,2\})^+02$ is of type III.

Type III, odd: $(\{1,2\}0)^+2$ has possible parents starting with:

0: $0\{1,2\}(0\{1,2\})^*0(\{1,2\}0)^*2$ is of type III;

1: $1(0\{1,2\})^*0210(\{1,2\}0)^*2$, there are three cases for a possible third 1:

*None***:** the parent is of type VII;

TABLE 4
Continued.

$Y_{33} = 12120102$	$P2 : \underline{21212010}2$ $P7 : \underline{01}0212102$	$C3 = Y_9$ $P8 : \underline{2010}21212$	$P4 : \underline{21212010}2$	$P5 : \underline{0212101}02$	$C6$ is of type VI
$Y_{34} = 12010212$	$P2 : \underline{212010}212$ $C7 = Y_{12}$	$P3 : \underline{02101}0212$ $P8 : \underline{212010}212$	$C4$ is of type VI	$P5 = Y_{48}$	$P6 : \underline{2010}21212$
$Y_{35} = 12010202$	$P2 = Y_{46}$ $P7 : \underline{02}0102102$	$P3 : \underline{02101}0202$ $P8 = Y_{51}$	$C4 = Y_6$	$P5 = Y_{49}$	$P6 = Y_{54}$
$Y_{36} = 20120102$	$P2 : \underline{02012}0102$ $P7 = Y_{49}$	$P3 = Y_{39}$ $C8 = Y_{10}$	$C4$ is of type III	$P5 : \underline{021}020102$	$P6 : \underline{102}102102$
$Y_{37} = 12012012$	$P2 = Y_{47}$ $C7$ is of type VI	$P3 : \underline{021012}012$ $P8 : \underline{210}210212$	$C4 = Y_8$	$P5 : \underline{210}212012$	$P6 : \underline{021}021012$
$Y_{38} = 021021202$	$P2 : \underline{20210}21202$ $P7 : \underline{21201}20202$	$P3 = Y_{71}$ $C8 = Y_{28}$	$C4$ is of type II $P9 : \underline{20}21201202$	$P5 : \underline{201}2021202$	$P6 : \underline{120}1201202$
$Y_{39} = 102120102$	$P2 : \underline{010212}0102$ $C7 = Y_{23}$	$P3 : \underline{20121}20102$ $P8 : \underline{010212}0102$	$C4 = Y_{36}$ $P9 = Y_{70}$	$P5 : \underline{21}20120102$	$P6 : \underline{0212}010102$
$Y_{40} = 102010212$	$P2 : \underline{0102}010212$ $P7 : \underline{2010}201212$	$P3 : \underline{201201021}2$ $C8$ is of type IV	$P4 : \underline{02}01010212$ $P9 = Y_{68}$	$C5 = Y_{25}$	$P6 : \underline{0102}010212$
$Y_{41} = 021202012$	$P2 : \underline{202120201}2$ $C7 = Y_{32}$	$P3 = Y_{72}$ $P8 : \underline{10202}12012$	$P4 = Y_{67}$ $P9 : \underline{210}2021202$	$C5 = Y_{26}$	$P6 : \underline{202120201}2$
$Y_{42} = 021201012$	$P2 : \underline{2021}201012$ $C7 = Y_{21}$	$P3 : \underline{120120}1012$ $P8 : \underline{10102}12012$	$P4 : \underline{21}20201012$ $P9 : \underline{21010}21202$	$C5 = Y_{27}$	$P6 = Y_{63}$
$Y_{43} = 020210212$	$P2 : \underline{2020210}212$ $P7 : \underline{20120}20212$	$C3 = Y_{29}$ $P8 = Y_{72}$	$P4 : \underline{2020210}212$ $P9 : \underline{21201}20202$	$P5 : \underline{12020102}12$	$C6$ is of type II
$Y_{44} = 101020212$	$P2 : \underline{010102}0212$ $P7 : \underline{20}20101212$	$C3 = Y_{31}$ $C8$ is of type IV	$P4 : \underline{010102}0212$ $P9 : \underline{21}20201012$	$P5 : \underline{2010120}212$	$P6 : \underline{02}01010212$
$Y_{45} = 020212012$	$P2 : \underline{202021201}2$ $C7 = Y_{26}$	$C3 = Y_{32}$ $P8 = Y_{62}$	$P4 : \underline{202021201}2$ $P9 : \underline{210}2120202$	$P5 : \underline{12020120}12$	$P6 = Y_{67}$
$Y_{46} = 212010202$	$P2 : \underline{12120}10202$ $C7 = Y_{30}$	$C3 = Y_{35}$ $P8 : \underline{0201}021202$	$P4 : \underline{021201}0202$ $C9 = Y_{25}$	$P5 : \underline{102}1210202$	$P6 : \underline{01}02120202$
$Y_{47} = 212012012$	$P2 : \underline{12120}12012$ $P7 = Y_{61}$	$C3 = Y_{37}$ $P8 : \underline{102}1021212$	$P4 : \underline{0212012}012$ $C9 = Y_{17}$	$P5 : \underline{102}1212012$	$C6 = Y_{21}$
$Y_{48} = 010210212$	$P2 : \underline{10102}10212$ $P7 : \underline{201201021}2$	$C3 = Y_{17}$ $P8 : \underline{120120}1012$	$P4 : \underline{2010210}212$ $P9 : \underline{21}20120102$	$P5 : \underline{120}1010212$	$C6 = Y_{34}$
$Y_{49} = 010210202$	$P2 = Y_{65}$ $P7 : \underline{201201020}2$	$C3$ is of type VII $C8 = Y_{36}$	$P4 : \underline{2010210}202$ $P9 : \underline{20}20120102$	$P5 : \underline{120}1010202$	$C6 = Y_{35}$
$Y_{50} = 010212012$	$P2 : \underline{10102}12012$ $C7 = Y_{27}$	$C3 = Y_{21}$ $P8 = Y_{63}$	$P4 = Y_{70}$ $P9 : \underline{210}2120102$	$P5 : \underline{120}1012012$	$P6 = Y_{68}$
$Y_{51} = 202010212$	$P2 = Y_{66}$ $C7 = Y_{31}$	$C3 = Y_{25}$ $P8 : \underline{120}1020212$	$P4 = Y_{66}$ $C9 = Y_{35}$	$P5 = Y_{64}$	$P6 : \underline{0102}020212$
$Y_{52} = 121202012$	$P2 : \underline{212120}2012$ $P7 : \underline{02}02121012$	$C3 = Y_{26}$ $C8$ is of type II	$P4 : \underline{212120}2012$ $P9 : \underline{21020}21212$	$P5 : \underline{0212102}012$	$P6 : \underline{20}21212012$
$Y_{53} = 121201202$	$P2 : \underline{2121201}202$ $P7 : \underline{210}2121202$	$C3 = Y_{28}$ $P8 : \underline{021}0212102$	$P4 : \underline{2121201}202$ $P9 = Y_{69}$	$P5 : \underline{02121012}02$	$C6$ is of type II
$Y_{54} = 201021202$	$P2 : \underline{0201}021202$ $C7 = Y_{35}$	$P3 : \underline{102102}1202$ $P8 : \underline{021201}0202$	$P4 : \underline{01}02021202$ $C9 = Y_{23}$	$C5 = Y_{30}$	$P6 = Y_{71}$
$Y_{55} = 120212012$	$P2 : \underline{2120}212012$ $P7 : \underline{0212}021012$	$P3 = Y_{61}$ $C8$ is of type II	$P4 : \underline{20}21212012$ $P9 : \underline{210}2120212$	$C5 = Y_{32}$	$P6 : \underline{2120}212012$
$Y_{56} = 012021212$	$P2 : \underline{101202}1212$ $P7 : \underline{21}20210212$	$P3 : \underline{21020}21212$ $P8 : \underline{12}12021012$	$C4 = Y_{18}$ $P9 : \underline{21}21202102$	$P5 = Y_{69}$	$P6 : \underline{1202}101212$
$Y_{57} = 120102012$	$P2 = Y_{68}$ $P7 : \underline{02}01021012$	$P3 : \underline{02101}02012$ $C8 = Y_{25}$	$C4 = Y_{19}$ $P9 : \underline{210}2010212$	$P5 : \underline{01}02102012$	$P6 = Y_{70}$
$Y_{58} = 201202012$	$P2 : \underline{02012}02012$ $P7 : \underline{02}02102012$	$P3 = Y_{62}$ $P8 = Y_{64}$	$C4$ is of type IV $C9$ is of type VII	$P5 : \underline{0210202}012$	$C6 = Y_{19}$
$Y_{59} = 120120212$	$P2 : \underline{21201}20212$ $P7 = Y_{69}$	$P3 : \underline{021012}0212$ $C8 = Y_{29}$	$C4$ is of type II $P9 : \underline{21}20210212$	$P5 : \underline{210}2120212$	$P6 : \underline{021}0210212$
$Y_{60} = 201201012$	$P2 : \underline{02012}01012$ $P7 : \underline{01}02102012$	$P3 = Y_{63}$ $P8 : \underline{10102}10212$	$C4$ is of type IV $C9 = Y_{24}$	$P5 : \underline{021}0201012$	$P6 : \underline{102}1021012$
$Y_{61} = 210212012$	$P2 : \underline{20210}212012$ $P7 = Y_{76}$	$P3 : \underline{120102120}12$ $C8 = Y_{47}$	$C4 = Y_{55}$ $P9 : \underline{102}12012012$	$P5 : \underline{201}20212012$ $P10 : \underline{210}21201202$	$P6 : \underline{120}12012012$

Before 0210**:** there is a g-1-flip
$\underline{1(0\{1,2\})^*01}(0\{1,2\})^*0210(\{1,2\}0)^*2 = 0(\{1,2\}0)^*1(0\{1,2\})^*$
$0210(\{1,2\}0)^*2$;
After 0210**:** there is a g-1-flip
$\underline{1(0\{1,2\})^*0210(\{1,2\}0)^*}10(\{1,2\}0)^*2 = (0\{1,2\})^*0120(\{1,2\}0)^*$
$10(\{1,2\}0)^*2$;

2: $\underline{2(0\{1,2\})^*0120(\{1,2\}0)^*}2 = (0\{1,2\})^*0210(\{1,2\}0)^*2$ is a g-1-flip unless this last string is 02102 (type VI), but then the parent is $201202 = Y_5$;

2: $2(0\{1,2\})^*012$ is of type IV.

Type IV, even: $(\{1,2\}0)^+12$ with a second 2 has possible parents starting with:

0: $0(1,20)^+12$, with a second 2, is of type IV;

1: $\underline{1(0\{1,2\})^*0210(\{1,2\}0)^*}12 = (0\{1,2\})^*0120(\{1,2\}0)^*12$ is a g-1-flip;

1: $\underline{1(0\{1,2\})^*0212}$, with three cases:
No third 2**:** the parent is of type V;

Table 4
Continued.

$Y_{62} = 1021202012$	$P2 : \underline{01021202}012$ $P7 : \underline{20}212012012$	$P3 : \underline{20121}202012$ $P8 : \underline{02}021201012$	$C4 = Y_{58}$ $C9 = Y_{45}$	$P5 = Y_{76}$ $P10 : \underline{210}20212012$	$P6 = Y_{74}$
$Y_{63} = 1021201012$	$P2 : \underline{010212}01012$ $C7 = Y_{42}$	$P3 : \underline{20121}201012$ $P8 : \underline{010212}01012$	$C4 = Y_{60}$ $C9 = Y_{50}$	$P5 : \underline{21}201201012$ $P10 : \underline{21010}212012$	$P6 : \underline{0212}0101012$
$Y_{64} = 1020210212$	$P2 : \underline{0102}0210212$ $P7 : \underline{012}02010212$	$P3 : \underline{20120}210212$ $P8 : \underline{201}20201212$	$P4 : \underline{02}010210212$ $C9 = Y_{58}$	$P5 : \underline{20}201210212$ $P10 = Y_{76}$	$C6 = Y_{51}$
$Y_{65} = 1010210202$	$P2 : \underline{0101021}0202$ $P7 : \underline{012}01010202$	$C3 = Y_{49}$ $P8 : \underline{201}20101202$	$P4 : \underline{0101021}0202$ $P9 : \underline{02}012010102$	$P5 : \underline{20101210}202$ $P10 : \underline{20}201201012$	$C6$ is of type III
$Y_{66} = 0202010212$	$P2 : \underline{20202010}212$ $C7$ is of type VIII	$C3 = Y_{51}$ $P8 : \underline{2010}2020212$	$P4 : \underline{20202010}212$ $P9 : \underline{120}10202012$	$C5 = Y_{51}$ $P10 : \underline{21}201020202$	$P6 = Y_{73}$
$Y_{67} = 2120202012$	$P2 = Y_{77}$ $C7 = Y_{45}$	$C3$ is of type VI $P8 : \underline{02}020212012$	$P4 : \underline{021202}02012$ $P9 : \underline{1020202}1212$	$C5 = Y_{41}$ $C10$ is of type VIII	$P6 : \underline{02}021202012$
$Y_{68} = 2120102012$	$P2 : \underline{12120}102012$ $C7 = Y_{50}$	$C3 = Y_{57}$ $P8 : \underline{0201}0212012$	$P4 = Y_{74}$ $P9 : \underline{1020}1021212$	$P5 : \underline{102}12102012$ $C10 = Y_{40}$	$P6 : \underline{01021202}012$
$Y_{69} = 2021021212$	$P2 = Y_{75}$ $P7 : \underline{120120212}12$	$C3$ is of type VI $C8 = Y_{59}$	$P4 : \underline{1202102}1212$ $P9 : \underline{12}120120212$	$P5 : \underline{012}02021212$ $C10 = Y_{53}$	$C6 = Y_{56}$
$Y_{70} = 2010212012$	$P2 : \underline{0201}0212012$ $C7 = Y_{57}$	$P3 : \underline{102102}12012$ $P8 = Y_{74}$	$P4 : \underline{01}020212012$ $P9 : \underline{102}12010212$	$C5 = Y_{50}$ $C10 = Y_{39}$	$P6 : \underline{120102120}12$
$Y_{71} = 1201021202$	$P2 : \underline{212010}21202$ $C7 = Y_{54}$	$P3 : \underline{02101}021202$ $P8 : \underline{212010}21202$	$C4 = Y_{38}$ $P9 : \underline{0212}0102102$	$P5 : \underline{01}021021202$ $P10 : \underline{20}212010212$	$P6 : \underline{2010}2121202$
$Y_{72} = 1201202012$	$P2 = Y_{76}$ $P7 : \underline{20210}212012$	$P3 : \underline{0210120}2012$ $P8 : \underline{02}021021012$	$C4 = Y_{41}$ $C9 = Y_{43}$	$P5 : \underline{210}21202012$ $P10 : \underline{210}20210212$	$P6 : \underline{021}02102012$
$Y_{73} = 10202010212$	$P2 : \underline{0102}02010212$ $C7 = Y_{66}$	$P3 : \underline{20120}2010212$ $P8 : \underline{0102}02010212$	$P4 : \underline{02}0102010212$ $P9 : \underline{2010}20201212$	$P5 : \underline{20}2012010212$ $C10$ is of type IV	$P6 : \underline{02}0201010212$ $P11 : \underline{21}2010202012$
$Y_{74} = 02120102012$	$P2 : \underline{2021}20102012$ $C7 = Y_{62}$	$P3 : \underline{120120}102012$ $P8 : \underline{2010}21202012$	$P4 : \underline{21}2020102012$ $C9 = Y_{70}$	$C5 = Y_{68}$ $P10 : \underline{1020}10212012$	$P6 : \underline{102}120102012$ $P11 : \underline{210}201021202$
$Y_{75} = 02021021212$	$P2 : \underline{2020210}21212$ $P7 : \underline{201}202021212$	$C3 = Y_{69}$ $P8 : \underline{120}120201212$	$P4 : \underline{2020210}21212$ $P9 : \underline{21}2012020212$	$P5 : \underline{12020102}1212$ $P10 : \underline{12120}1202012$	$C6$ is of type II $P11 : \underline{21}2120120202$
$Y_{76} = 21201202012$	$P2 : \underline{12120}1202012$ $P7 : \underline{021}021202012$	$C3 = Y_{72}$ $C8 = Y_{61}$	$P4 : \underline{0212012}02012$ $P9 : \underline{02}0210212012$	$P5 : \underline{102}121202012$ $P10 : \underline{10202}1021212$	$C6 = Y_{62}$ $C11 = Y_{64}$
$Y_{77} = 12120202012$	$P2 : \underline{212120}202012$ $P7 : \underline{02}0212102012$	$C3 = Y_{67}$ $P8 : \underline{20}2021212012$	$P4 : \underline{212120}202012$ $P9 : \underline{02}0202121012$	$P5 : \underline{0212102}02012$ $C10$ is of type II	$P6 : \underline{20}2121202012$ $P11 : \underline{210}202021212$

No third 1**:** the parent is of type VIII;

Otherwise: $\underline{1(0\{1,2\})^*0}1(0\{1,2\})^*0212 = 0(\{1,2\}0)^*1(0\{1,2\})^*0212$ (with a third 2) is a g-1-flip;

2: $2(0\{1,2\})^*0120(\{1,2\}0)^*12$, with four cases:

A fourth 2 *before* 0120**:** there is a g-1-flip $\underline{2(0\{1,2\})^*0}2(0\{1,2\})^*0120(\{1,2\}0)^*12 = 0(\{1,2\}0)^+120(\{1,2\}0)^*12$;

A fourth 2 *after* 0120**:** there is a g-1-flip $\underline{2(0\{1,2\})^*0120(\{1,2\}0)^*}20(\{1,2\}0)^*12 = (0\{1,2\})^*0210(\{1,2\}0)^+12$;

A third 1**:** $\underline{2(0\{1,2\})^*0120(\{1,2\}0)^*1}2 = 1(0\{1,2\})^*0210(\{1,2\}0)^*2$ is a g-1-flip;

*Otherwise***:** $2012012 = Y_{16}$;

2: $\underline{21(0\{1,2\})^*0}2(0\{1,2\})^*012 = 0(\{1,2\}0)^*12(0\{1,2\})^*012$ is a g-1-flip.

Type IV, odd: $0(\{1,2\}0)^+12$ with a second 2, has possible parents starting with:

1: $1(0\{1,2\})^+012$, with a second 2, is of type IV;

2: $2(0\{1,2\})^+012$ is of type IV;

2: $\underline{21(0\{1,2\})^*0}2(0\{1,2\})^*02 = 0(\{1,2\}0)^*12(0\{1,2\})^*02$ is a g-1-flip.

Type V, even: $0(10)^+212$ (0212 is also of type I), has possible parents starting with:

1: $(10)^+212$ is of type V;

1: $\underline{120}1(01)^*012 = 021(01)^*012$ is a g-1-flip;

2: $\underline{2(01)^+021}2 = 120(10)^+2$ is a g-1-flip;

2: $\underline{21}2(01)^+02 = 12(01)^+02$ is a g-1-flip.

Type V, odd: $(10)^+212$, has possible parents starting with:

0: $(01)^+0212$ is of type V;

2: $\underline{2(01)^+21}2 = 12(10)^+2$ is a g-1-flip;

2: $\underline{21}2(01)^i2 = 12(01)^i2$ is a g-1-flip unless $i = 1$, but then the parent is $212012 = Y_3$.

Type VI, $1(20)^+1(20)^*2$**:** has possible parents starting with:

0: $(02)^i10(20)^j1(20)^k2$ with $i > 0$.

If $i > 1$, $\underline{02}02(02)^{i-2}10(20)^j1(20)^k2 = 2(02)^{i-1}10(20)^j1(20)^k2$ is a g-1-flip;

If $i = 1$, $j > 0$, $\underline{02102}0(20)^{j-1}1(20)^k2 = 201(20)^j1(20)^k2$ is a g-1-flip;

If $i = 1$, $j = 0$, $k > 0$, $\underline{021012}0(20)^{k-1}2 = 2101(20)^k2$ is a g-1-flip;

If $i = 1$, $j = k = 0$, $021012 = Y_2$;

0: $\underline{(02)^+1}(02)^+1(02)^+ = 1(20)^+21(02)^+$ is a g-1-flip;

2: $\underline{2(02)^*1(20)^+1(20)^*}2 = (02)^*1(02)^+1(20)^*2$ is a g-1-flip unless this last string is 10212 (type V) or 0210212 (type VI), but then the parent is $212012 = Y_3$ or $21201202 = Y_{28}$ respectively;

2: $\underline{2(02)^*1(02)^+1(20)^+2} = (02)^+1(20)^+1(20)^*2$ is a g-1-flip;

2: $\underline{2(02)^*1}02(02)^*12 = 01(20)^*212$ is a g-1-flip unless there is no second 0, but then the parent is $210212 = Y_1$.

*Type VI***,** $0(21)^+0(21)^*2$**:** has possible parents starting with:

1: $(12)^i01(21)^j0(21)^k2$ with $i > 0$.

If $i > 1$, $\underline{12}12(12)^{i-2}01(21)^j0(21)^k2 = 2(12)^{i-1}01(21)^j0(21)^k2$ is a g-1-flip;

If $i = 1$, $j > 0$, $\underline{12012}1(21)^{j-1}0(21)^k2 = 210(21)^j0(21)^k2$ is a g-1-flip;

If $i = 1$, $j = 0$, $k > 0$, $\underline{120102}1(21)^{k-1}2 = 2010(21)^k2$ is a g-1-flip;

If $i = 1$, $j = k = 0$, $120102 = Y_4$;

1: $\underline{(12)^+0}(12)^+0(12)^+ = 0(21)^+20(12)^+$ is a g-1-flip;
2: $\underline{2(12)^*0(21)^+0(21)^*2} = (12)^*0(12)^+0(21)^*2$ is a g-1-flip unless this last string is 1201202 (type VI), but then the parent is $20210212 = Y_{29}$;
2: $\underline{2(12)^*0(12)^+0(21)^+}2 = (12)^+0(21)^+0(21)^*2$ is a g-1-flip;
2: $\underline{2(12)^*01}2(12)^*02 = 10(21)^*202$ is a g-1-flip unless this last string is 10202 (type III), but then the parent is $201202 = Y_5$.

Type VII: $1(02)^+1(02)^+$ has possible parents starting with:
0: $\underline{0(20)^*1(02)^+1}(02)^+ = 1(20)^+1(02)^+$ is a g-1-flip;
0: $\underline{0(20)^*12}0(20)^*1(02)^+ = 210(20)^*1(02)^+$ is a g-1-flip;
2: $\underline{(20)^+12(02)^*10}2(02)^* = 01(20)^*21(02)^+$ is a g-1-flip;
2: $\underline{(20)^+1}(20)^+12(02)^* = 1(02)^+012(02)^*$ is a g-1-flip.

Type VIII: $1(02)^+12$ has possible parents starting with:
0: $0(20)^i1(02)^j12$ with $j > 0$.
If $i > 0$, $\underline{02}0(20)^{i-1}1(02)^j12 = (20)^i1(02)^j12$ is a g-1-flip;
If $i = 0$, $j > 1$, $\underline{0102}(02)^{j-1}12 = 201(02)^{j-1}12$ is a g-1-flip;
If $i = 0$, $j = 1$, 010212 is of type V;
2: $\underline{(20)^+12(02)^*1}2 = 1(20)^*21(02)^+$ is a g-1-flip;
2: $21(20)^i12$ with $i > 0$.
If $i = 1$, $212012 = Y_3$;
If $i > 1$, $\underline{2120}(20)^{i-1}12 = 021(20)^{i-1}12$ is a g-1-flip. □

THEOREM 6. *There exists a polynomial-time algorithm for optimally sorting ternary strings.*

Proof. This follows rather easily from Theorem 5. □

Finally, in light of the fact that the complexity of the sorting problem on quaternary (and higher) strings remains open, the following serves as an intermediate result.

THEOREM 7. *For every fixed k there is a PTAS for sorting k-ary strings.*

Proof. The proof is very similar to the proof of Theorem 3. We assume that $k \geq 4$. Let $N = (3k-2)/\epsilon + k$. Let s, the string that we wish to sort, be of length n. We distinguish two cases. (In both cases it is useful to note that $d_{\mathrm{s}}(s) \leq 2n$ because we can always bring the greatest symbol not yet in its final position to the front and then to its correct position.)

Case 1. If $n \geq N$, we first group the string using the "greedy" algorithm from the proof of Lemma 2, which yields a permutation on k symbols. This permutation can then be easily sorted with at most $2k$ flips. Thus the total number of flips, denoted by $d_{\mathrm{s}}^G(s)$, is at most $(n-2) + 2k$. This, together with the grouping lower bound of Lemma 1 of $n-k$ on $d_{\mathrm{s}}(s)$, yields $d_{\mathrm{s}}^G(s) \leq d_{\mathrm{s}}(s) + (3k-2) \leq (1+\epsilon)d_{\mathrm{s}}(s)$.

Case 2. If $n < N$, we apply brute force by selecting the shortest sorting sequence from among all length-$2n$ sequences of flips; there are at most n^{2n} such sequences. Given that $d_{\mathrm{s}}(s) \leq 2n$, this is guaranteed to give an optimal solution. The running time in this case is bounded by a constant. □

5. Prefix reversal diameter. Let $S(n,k)$ be the set of fully k-ary strings of length n. We define $\delta(n,k)$ as the largest value of $d(s,t)$ ranging over all compatible $s, t \in S(n,k)$.

THEOREM 8. *For all $n \geq 2$, $\delta(n,2) = n-1$.*

Proof. To prove $\delta(n,2) \geq n-1$, consider compatible $s, t \in S(n,2)$ with $s = (10)^{n/2}$ in case n even and $s = 0(10)^{(n-1)/2}$ in case n odd and in both cases $t = I(s)$; i.e., t is the sorted version of s. By Theorem 4, $d(s,t) \geq n-1$.

The proof that $\delta(n,2) \leq n-1$ for all $n \geq 2$ is by induction on n. The lemma is trivially true for $n=2$. Consider two compatible binary strings of length n: $s = s_1 s_2 \dots s_n$ and $t = t_1 t_2 \dots t_n$. If $s_n = t_n$, then by induction $d(s,t) \leq n-2$. Thus, suppose (without loss of generality) $s_n = 0$ and $t_n = 1$. If $t_1 = 0$, then $f^{(n)}t$ and s both end with a 0, and using induction and symmetry $d(s,t) \leq 1 + d(f^{(n)}t, s) \leq n-2+1 = n-1$. An analogous argument holds if $s_1 = 1$.

There remains the case $s_1 = s_n = 0$ and $t_1 = t_n = 1$. First, suppose $t_{n-1} = 0$. Since s and t are compatible, there must exist an index i such that $s_i = 0$ and $s_{i+1} = 1$. Hence, $f^{(n)}(f^{(i+1)}(s))$ ends with 01 like t, and by induction $d(s,t) \leq 2 + d(f^{(n)}(f^{(i+1)}(s)), t) = 2 + n - 3$. Analogously, we resolve the case $s_{n-1} = 1$.

Finally, suppose $s = 0 \dots 00$ and $t = 1 \dots 11$. If s contains 11 as a substring, then flipping that 11 (in the same manner as above) to the end of s using two flips gives two strings that both end in 11. Alternatively, if s does not contain 11 as a substring, then s has at least two more 0's than 1's, which implies that t must contain 00 as a substring. In that case two prefix reversals on t suffice to create two strings that both end with 00. In both cases, the induction hypothesis gives the required bound. □

Note that, trivially, $d(s,t) \leq 2n$ for all compatible $s,t \in S(n,k)$, for all k, because two prefix reversals always suffice to increase the maximal common suffix between s and t by at least 1. The following tighter bound gives the best bound known on the diameter of ternary strings.

LEMMA 8. *For any two compatible $s,t \in S(n,k)$, for any k, let a be the most frequent symbol in s and α its multiplicity. Then $d(s,t) \leq 2(n-\alpha)$.*

Proof. We prove the lemma by induction on n. The lemma is trivially true for $n=2$. Consider $s,t \in S(n,k)$. If $s_n = t_n = a$, then $s_1 s_2 \dots s_{n-1}$ and $t_1 t_2 \dots t_{n-1}$ are compatible length-$(n-1)$ strings where the most frequent symbol occurs at least $\alpha - 1$ times. Thus, by induction $d(s,t) \leq 2((n-1)-(\alpha-1)) = 2(n-\alpha)$. In case $s_n = t_n \neq a$ induction even gives $d(s,t) \leq 2((n-1)-(\alpha)) = 2(n-\alpha)-2$. Thus, suppose $s_n \neq t_n$ implying without loss of generality that $t_n = b \neq a$. Suppose $s_i = b$; after two flips $s' = f^{(n)}(f^{(i)}(s))$ has b at the end; $s'_n = t_n$. Moreover, the length $n-1$ suffixes of s' and t still contain α a's. Hence by induction, $d(s,t) \leq 2 + d(s',t) \leq 2 + 2((n-1)-\alpha) = 2(n-\alpha)$. □

LEMMA 9. *For all $n > 3$, $n-1 \leq \delta(n,3) \leq (4/3)n$.*

Proof. Since in any ternary case $\alpha \geq \lceil n/3 \rceil$, Lemma 8 implies $\delta(n,3) \leq (4/3)n$. To prove $\delta(n,3) \geq n-1$ we distinguish between n odd and n even. For odd $n = 2h+1$, let s be $2(01)^h$, and for even $n = 2h$ let $s = 01(21)^{h-1}$. In both cases we let $t = I(s)$. We observe that, in the even and in the odd case, $s2$ is a bad I-string and a bad IV-string, respectively, in the sense of Definition 2. Thus, by Theorem 5 we have that $d(s,t) = d(s2,t2) = (n+1)-2 = n-1$. (Here $s2$ (respectively, $t2$) refers to the concatenation of s (respectively, t) with an extra 2 symbol.) □

Brute force enumeration has shown that, for $4 \leq n \leq 13$, $\delta(n,3) = n-1$. (Note that $\delta(3,3) = 3$ because $d(021,012) = 3$.) Proving or disproving the conjecture that $\delta(n,3) = n-1$ for $n > 3$ remains an intriguing open problem.[3]

6. Prefix reversal distance.

6. Prefix reversal distance. We show that computing flip distance is NP-hard on binary strings. We also point out, using a result from [11], that computing flip distance on *arbitrary* strings is polynomial-time reducible (in an approximation-preserving sense) to computing it on binary strings.

[3] Interestingly, initial experiments with brute force enumeration have also shown that, for $4 \leq n \leq 10$, $\delta(n,4) = n$, and for $5 \leq n \leq 9$, $\delta(n,5) = n$.

THEOREM 9. *The problem of computing the prefix reversal distance of binary strings is NP-hard.*

We prove NP-completeness of the corresponding decision problem:
Name: BINARY-PD (abbreviated 2PD).
Input: Two compatible strings $s, t \in S(n, 2)$, and a bound $B \in \mathbb{Z}^+$.
Question: Is $d(s, t) \leq B$?

2PD ∈ NP, since a certificate for a positive answer consists of at most B flips.[4] To show completeness we use a reduction from 3-PARTITION [6] (cf. [2] and [11]).
Name: 3-PARTITION (abbreviated 3P).
Input: A set $A = \{a_1, a_2, \ldots, a_{3k}\}$ and a number $N \in \mathbb{Z}^+$. Element a_i has size $r(a_i) \in \mathbb{Z}^+$ satisfying $N/4 < r(a_i) < N/2$, $i = 1, \ldots, 3k$, and $\sum_{i=1}^{3k} r(a_i) = kN$.
Question: Can A be partitioned into k disjoint triplet sets $A_1, A_2, \ldots, A_k$ such that $\sum_{a \in A_j} r(a) = N$, $j = 1, \ldots, k$?

Given instance $I = (A, N, r)$ of 3P, we create an instance of 2PD by setting $B = 6k$ and building two compatible binary strings s and t:

$$s = \Bigg(\prod_{1 \leq i \leq 3k} 0001^{r(a_i)} \Bigg) 000, \qquad t = 0^{3(3k+1)-k}(01^N)^k.$$

This construction is clearly polynomial in a unary encoding of the 3P instance; we use the *strong* NP-hardness of 3P [6]. We claim that $I = (A, N, r)$ is a positive instance of 3P ⇔ $d(s, t) \leq 6k$.

⇒) Let a_{ij} denote the jth element from triples A_i (in arbitrary order), $j = 1, 2, 3$, $i = 1, \ldots, k$, and let us abuse its name also to denote the corresponding 1-block of length $r(a_{ij})$ in s.

That s can be transformed to t in $6k$ flips follows directly from the correctness of the following claim for $h = k$.

CLAIM. *For $0 < h \leq k$, s can be transformed into a string $\psi_h = \alpha_h \omega_h$ in h phases, each consisting of six flips, where ψ_h has the following specific properties:*

(a) *The suffix (i.e., ω_h) is equal to $(01^N)^h$ and contains all $3h$ 1-blocks corresponding to the elements in $\cup_{j=1}^{h} A_j$.*

(b) *The prefix (i.e., α_h) contains the remaining $3(k - h)$ 1-blocks, each of them flanked by 0-blocks of length at least 3, except possibly a 0-block of length 2 at its right end. (Given that $\psi_h = \alpha_h \omega_h$, it follows that, in ψ_h,* all *these remaining 1-blocks are flanked by 0-blocks of length at least 3.)*

Proof. The proof is by induction. First we transform s into ψ_1 in six flips: flips 1 and 2 bring a_{11} to the back, flips 3 and 4 bring a_{12} to the back (just in front of a_{11}), and flips 5 and 6 bring a_{13} to the back (just in front of a_{12}). No 0-blocks are cut in this process, and only 1-blocks a_{11}, a_{12}, and a_{13} are affected (i.e., concatenated into a single length-N 1-block).

Now, suppose by induction that after $6(h - 1)$ flips we have created ψ_{h-1}. The next six flips (which form phase h) work exclusively on α_{h-1}. Flips 1 and 2 bring a_{h1} to the front and then to the back of α_{h-1}; flips 3 and 4 bring a_{h2} to the front and then to the back just in front of a_{h1}; flips 5 and 6 bring a_{h3} to the front and then to the back just in front of a_{h2}. These six flips (which do not cut any 0-blocks within α_{h-1})[5] thus transform α_{h-1} into a string with 01^N at the suffix, which, appended to ω_{h-1},

[4] Recall that, for all compatible strings $s, t \in S(n, 2)$, trivially $d(s, t) \leq 2n$.

[5] Observe that, in terms of its action on the *overall* string, flip 2 of phase h does cut a 0-block, cutting α_{h-1} from ω_{h-1}, creating the singleton 0-block in between two length-N 1-blocks.

gives a suffix equal to ω_h. The only question is whether the resulting overall string satisfies condition (b). The only obstacle to this is the possible length-2 0-block at the end of α_{h-1}. However, this block is not flipped in flip 1 of phase h; it is brought to the front in flip 2 and concatenated to another 0-block in flip 3, leaving the prefix string without a length-2 0-block. This completes the proof of the claim. □

$\Leftarrow$) Suppose that I is a negative instance of 3P. We show that $d(s,t) > 6k$. Notice that if I is not a positive instance, then in any sequence of flips taking s to t some flip must split a 1-block, i.e., $\underline{\ldots 1}1\ldots$. Below we add this to a list of tasks that any sequence of flips taking s to t must complete:

(0) split at least one 1-block;
(1) reduce the number of 1-blocks by $2k$;
(2) bring a 1 symbol to the end of the string (because t ends with a 1, but s does not);
(3) increase the number of singleton 0-blocks by $k-1$;
(4) reduce the number of *big* (i.e., of length at least 3) 0-blocks by $3k$.

To prove that at least $6k+1$ flips are needed to complete tasks (0)–(4), we show that flips which make progress towards completing one of the tasks cannot effectively be used to make progress on another task. From this (and other intermediate observations shown below) it will follow that at least $1+2k+1+(k-1)+3k = 6k+1$ flips will be needed.

It is immediately clear that task (2), requiring a flip of a whole string, cannot be combined with any of the other tasks in one flip. Notice that any task(0)-flip (which is of the form $\underline{1\ldots 1}1\ldots$ or of the form $\underline{0\ldots 1}1\ldots$) does not decrease the number of 1-blocks, while 0-blocks remain unaffected. So such flips do not contribute to tasks (1)–(4). Nor can any task(1)-flip (which is always of the form $\underline{1\ldots 0}1\ldots$) contribute to any of the other tasks from the list. It is also not too difficult to verify that it is not possible to reduce the number of big blocks by 2 or more in one flip. However, some types of task(3)-flip *can* at the same time also contribute to task (4), and some other types of task(3)-flip can increase the number of singleton 0-blocks by two, effectively contributing twice to task (3). Such flips we call (34)- and (33)-flips, respectively. We will show that all (34)- and (33)-flips necessarily have to be succeeded by at least one flip that does not, in an overall sense, help us with the completion of the tasks.

Any (33)-flip is of the type
(33.1) $\underline{1\ldots 0}0\ldots$ (where the 0s form a complete block).

Any (34)-flip is of the type
(34.1) $\underline{1\ldots 0}00\ldots$ (where the 0s form a complete block),
(34.2) $\underline{1\ldots 00}0\ldots$ (where the 0s form a complete block),
(34.3) $\underline{000\ldots 10}00\ldots$.
We emphasize here that 00 is not considered to be a big 0-block.

After a flip of type (33.1), (34.1), or (34.3) we have a single 0 at the front. In such a situation a task(1)- or task(2)-flip is not possible. We cannot perform a task(3)-flip because flips of the form $\underline{01\ldots}0\ldots$ will destroy the initial singleton 0, and flips of the form $\underline{01\ldots}1\ldots$ cannot create new singleton 0's. The only task(4)-flip possible is $\underline{01\ldots 00}0\ldots$ (where the second group of 0's forms a complete block), but this also reduces the number of singleton 0-blocks by 1, meaning that an extra task(3)-flip would then be needed. Termination is not an option (because t does not begin with 01). A task(0)-flip of the form $\underline{01\ldots 1}1\ldots$ is potentially possible, but, as noted, this increases the number of required task(1)-flips.

After a flip of type (34.2) we are left with 001 at the front. Again, a task(1)- or task(2)-flip is not possible in this situation, and neither is termination. A task(3)-flip is potentially possible, but this brings a single 0 to the front, which (by the earlier argument) cannot be followed by any useful flip. A task(4)-flip is not possible because, when the string begins with 001, a task(4)-flip must necessarily split a 00-adjacency in some big 0-block, but this simply creates a different big 0-block. □

For studying problems on arbitrary strings, let X and Y be two compatible, length-n strings, where we assume (without loss of generality) that each of the symbols from X and Y are drawn from the set $\{0, 1, \ldots, n-1\}$. We define $D(X, Y)$ as the smallest number of flips required to transform X to Y. The arity of the strings X and Y does not need to be fixed, and symbols may be repeated. Hence, sorting a permutation by flips (MIN-SBPR) and the flip distance problem over fixed arity strings are both special cases of computing D. Given that computing D is a generalization of computing distance d of binary strings, this immediately implies that it is NP-hard. However, an approximation-preserving reduction in the *other* direction is possible, meaning that inapproximability results for one of the problems will be automatically inherited by the other.

Theorem 10. *Given two compatible strings X and Y of length n with each symbol from X and Y drawn from $\{0, 1, \ldots, n-1\}$, it is possible to compute in time polynomial in n two binary strings x and y of length polynomial in n such that $D(X, Y) = d(x, y)$.*

As demonstrated shortly, the above result follows directly from work by Radcliffe, Scott, and Wilmer. A little background is necessary to understand the context. In Theorem 8 of [11] it is shown that sorting permutations by reversals is directly reducible to the reversal distance problem on binary strings. It is later argued (in Theorem 11 of [11]) that the same reduction technique can be used to reduce the transposition distance problem on a 4-ary alphabet to the transposition distance problem on a binary alphabet. The proof of Theorem 11 lacks detail, but personal communication with the authors [12] has since clarified that the result is correct. Furthermore, the reduction technique underpinning Theorems 8 and 11 from [11] can be directly applied to prove the present theorem. We show this by reproducing the reduction technique (complete with clarification) in the context of prefix reversals. We also use this opportunity to clarify the correctness of Theorem 11 from [11]. The following should thus be considered attributed to Radcliffe, Scott, and Wilmer.

Proof. The strings x and y are constructed as follows:

$$x = (10^{X_1+1}1)^{2n+1} \ldots (10^{X_n+1}1)^{2n+1},$$
$$y = (10^{Y_1+1}1)^{2n+1} \ldots (10^{Y_n+1}1)^{2n+1}.$$

In the above encoding, each symbol X_i is thus encoded as the fragment $(10^{X_i+1}1)^{2n+1}$, each fragment consisting of $2n+1$ subfragments. (This also holds for each symbol in Y.) Note that a fragment is reversal-invariant. To see that $d(x, y) \leq D(X, Y)$, observe that—by mapping to prefix reversals that cut at the boundaries between fragments—any sequence of m prefix reversals taking X to Y can be trivially mapped to m prefix reversals which take x to y.

The proof that $D(X, Y) \leq d(x, y)$ is more involved. Combining $d(x, y) \leq D(X, Y)$ with the trivial fact that $D(X, Y) \leq 2n$ yields $d(x, y) \leq 2n$. Now, consider any shortest sequence of prefix reversals taking x to y. This sequence of prefix reversals will cut the string x in at most $2n$ places. A subfragment within x is said to *survive* if and only if it is not cut by any of these prefix reversals. Now, construct a bipartite graph with vertex set $\{e_1, e_2, \ldots, e_n\} \cup \{f_1, f_2, \ldots, f_n\}$ and add an edge (e_i, f_j) if and only if some subfragment of the fragment corresponding to X_i survives and ends up in the fragment corresponding to Y_j. Observe that within any set of m fragments from x, strictly more than $(m-1)(2n+1)$ subfragments will survive, and hence at least m fragments from y will be required to absorb these surviving subfragments. Thus, by Hall's theorem, the graph has a perfect matching. For each edge (e_i, f_j) of the perfect matching, pick a subfragment from the fragment corresponding to X_i that survives and ends up in the fragment corresponding to Y_j. Considering the action of the flips only on these n subfragments, we see that there exists a sequence of $d(x, y)$ prefix reversals transforming the sequence of symbols in X into the sequence of symbols in Y, and thus $D(X, Y) \leq d(x, y)$. □

The correctness of Theorem 11 from [11] follows by using the same reduction but encoding each fragment as $3n$ subfragments rather than $2n+1$ subfragments. (The transposition distance between two compatible length-n strings is strictly less than n, and a transposition cuts a string in at most 3 places.) Indeed, it is easy to see that the reduction works for a whole family of string rearrangement operators, by ensuring that the number of subfragments per fragment is sufficiently large. For example, consider a rearrangement operator *op*, and let u be some upper bound on the number of places an *op*-operation can cut a string. Let v be any upper bound on the maximum value of $d_{op}(X, Y)$ ranging over all compatible length-n strings X, Y. Encoding each fragment with $uv+1$ subfragments is sufficient to generalize the above reduction.

7. Open problems.

In this study we have unearthed many rich (and surprisingly difficult) combinatorial questions which deserve further analysis. We discuss some of them here. The main unifying "umbrella" suggestion is that, to go beyond ad hoc (and case-based) proof techniques, it will be necessary to develop deeper, more structural insights into the action of flips on strings over fixed-size alphabets.

Grouping and sorting on higher arity alphabets. We have shown how to group and sort optimally binary and ternary strings, but characterizations and algorithms for quaternary (and higher) alphabets have so far eluded us. As observed in section 3.2, it *seems* that for $k = 4, 5$ and for sufficiently long strings, the strings with grouping distance $n-2$ settle into some kind of pattern, but this has not yet offered enough insights to allow the development either of a characterization or of an algorithm. Related problems include: for all fixed k, are there polynomial algorithms to optimally sort (optimally group) k-ary strings? Is grouping strictly easier than sorting, in a complexity sense? How does grouping function under other operators, e.g., reversals, transpositions? An upper bound on the grouping transposition distance has been presented in [4].

Diameter questions. Proving or disproving that $\delta(n, 3) = n-1$ for $n > 3$ remains the obvious open diameter question. Beyond that, diameter results for quaternary and higher arity alphabets are needed. How does the diameter $\delta(n, k)$ grow for increasing k? (At this point we conjecture that, for sufficiently long strings, the diameter of 3-ary, 4-ary, and 5-ary strings is $n-1$, n, and n, respectively.)

The suspicion also exists that, for all k and for all sufficiently long n, there exists a length-n fully k-ary string s such that $d(s, I(s)) = \delta(n, k)$. In other words, the set of all pairs of strings that are $\delta(n, k)$ flips apart includes some instances of the sorting problem. It should be noted, however, that, following empirical testing, it is apparent that there are also very many pairs of strings s, t with $s \neq I(t)$ and $t \neq I(s)$ that are $\delta(n, k)$ flips apart.

It also seems important to develop diameter results for subclasses of strings, perhaps (as in [11]) characterized by the frequency of their most frequent symbol. It may be that such refined diameter results for k-ary alphabets provide information that is important in determining $\delta(n, k+1)$.

Note finally that the diameter of strings over fixed size alphabets, i.e., $\delta(n, k)$, is always bounded from above by the diameter of permutations, $wc(n)$. This is because the distance problem on two length-n fixed size alphabet strings s, t can easily be rewritten as a sorting problem on a length-n permutation π such that a sequence of prefix reversals sorting the permutation also suffices to transform s into t. Indeed, because of this relabeling property, the flip distance between two fixed size alphabet strings can be viewed as being equal to the minimum permutation sorting distance, ranging over all such relabelings into a permutation π. Can this relationship between the fixed size alphabet and permutation world be further specified and exploited?

Signed strings. The problem of sorting signed permutations by flips (the *burnt* pancake flipping problem) is well known [3, 7, 9], but in this paper we have not yet attempted to analyze the action of flips on signed fixed size alphabet strings. Obviously, analogues of all the problems described in this paper exist for signed strings.

Complexity/approximation. In the presence of hardness results (e.g., Theorem 9) it is interesting to explore the complexity of restricted instances and to develop algorithms with guaranteed approximation bounds. For example, [11] gives a PTAS for dense instances. The development of approximation algorithms is also a useful intermediate strategy where the complexity of a problem remains elusive. In particular, this requires the development of improved lower bounds.

Acknowledgment. We thank Alex Scott for useful remarks made during the writing of this paper.

REFERENCES

[1] X. Chen, J. Zheng, Z. Fu, P. Nan, Y. Zhong, S. Lonardi, and T. Jiang, *Assignment of orthologous genes via genome rearrangement*, IEEE/ACM Trans. Comput. Biol. Bioinform., 2 (2005), pp. 302–315.

[2] D. A. Christie and R. W. Irving, *Sorting strings by reversals and by transpositions*, SIAM J. Discrete Math., 14 (2001), pp. 193–206.

[3] D. S. Cohen and M. Blum, *On the problem of sorting burnt pancakes*, Discrete Appl. Math., 61 (1995), pp. 105–120.

[4] H. Eriksson, K. Eriksson, J. Karlander, L. Svensson, and J. Wastlund, *Sorting a bridge hand*, Discrete Math., 241 (2001), pp. 289–300.

[5] J. Fischer and S. W. Ginzinger, *A 2-approximation algorithm for sorting by prefix reversals*, in Proceedings of ESA 2005: 13th Annual European Symposium, Lecture Notes in Comput. Sci. 3669, Springer, New York, 2005, pp. 415–425.

[6] M. R. Garey and D. S. Johnson, *Computers And Intractability: A Guide To The Theory Of NP-completeness*, W. H. Freeman, San Francisco, CA, 1979.

[7] W. H. Gates and C. H. Papadimitriou, *Bounds for sorting by prefix reversal*, Discrete Math., 27 (1979), pp. 47–57.

[8] A. Goldstein, P. Kolman, and J. Zheng, *Minimum common string partition problem: Hardness and approximations*, Electron. J. Combin., 12 (2005), #R50.

[9] M. H. Heydari and I. H. Sudborough, *On the diameter of the pancake network*, J. Algorithms, 25 (1997), pp. 67–94.

[10] L. Morales and I. H. Sudborough, *Comparing star and pancake networks*, in The Essence of Computation: Complexity, Analysis, Transformation, T. Mogensen, D. Schmidt, and I. H. Sudborough, eds., Lecture Notes in Comput. Sci. 2566, Springer, New York, 2002, pp. 18–36.

[11] A. J. Radcliffe, A. D. Scott, and E. L. Wilmer, *Reversals and transpositions over finite alphabets*, SIAM J. Discrete Math., 19 (2005), pp. 224–244.

[12] A. D. Scott, *personal communication*, Merton College, Oxford, 2006.

SIAM J. Discrete Math.
Vol. 21, No. 3, pp. 612–636

APPROXIMATION ALGORITHMS FOR NETWORK DESIGN WITH METRIC COSTS[*]

JOSEPH CHERIYAN[†] AND ADRIAN VETTA[‡]

Abstract. We study undirected networks with edge costs that satisfy the triangle inequality. Let n denote the number of nodes. We present an $O(1)$-approximation algorithm for a generalization of the metric-cost subset k-node-connectivity problem. Our approximation guarantee is proved via lower bounds that apply to the simple edge-connectivity version of the problem, where the requirements are for edge-disjoint paths rather than for openly node-disjoint paths. A corollary is that, for metric costs and for each $k = 1, 2, \ldots, n-1$, there exists a k-node connected graph whose cost is within a factor of 22 of the cost of any simple k-edge connected graph. Based on our $O(1)$-approximation algorithm, we present an $O(\log r_{\max})$-approximation algorithm for the metric-cost node-connectivity survivable network design problem, where $r_{\max}$ denotes the maximum requirement over all pairs of nodes. Our results contrast with the case of edge costs of 0 or 1, where Kortsarz, Krauthgamer, and Lee. [*SIAM J. Comput.*, 33 (2004), pp. 704–720] recently proved, assuming NP$\not\subseteq$ DTIME($n^{polylog(n)}$), a hardness-of-approximation lower bound of $2^{\log^{1-\epsilon} n}$ for the subset k-node-connectivity problem, where ϵ denotes a small positive number.

Key words. approximation algorithm, linear programming relaxation, network design, vertex-connectivity, edge-connectivity

AMS subject classification. Primary, 68W25, 05C40; Secondary, 68R10, 90C27

DOI. 10.1137/040621806

1. Introduction.

A basic problem in network design is to find a minimum-cost subnetwork H of a given network G such that H satisfies some prespecified connectivity requirements. Fundamental examples include the *minimum spanning tree* (MST) problem and the *traveling salesman problem* (TSP). By a *network* we mean an undirected graph together with nonnegative costs for the edges; we use V to denote the set of nodes, and n to denote $|V|$. Our focus is on networks where the edge costs are metric; that is, the input graph is a complete graph and the edge costs satisfy the triangle inequality. This special case is significant from both theoretical and practical viewpoints; metric costs arise in many applications of network design, and perhaps in most of the obvious ones, such as the design of telecommunication networks. Our goal is to design and analyze approximation algorithms for some key problems in network design. Moreover, we resolve a conjecture from the folklore on metric graphs, where by a *metric graph* we mean a complete graph K_n together with edge costs that satisfy the triangle inequality.

We attack the metric-cost *node-connectivity survivable network design problem* (NC-SNDP). In this problem, we are given a metric graph, as well as a connectivity requirement $r_{i,j}$ between every pair of nodes i and j. Let $r_{\max}$ denote $\max_{i,j \in V} r_{i,j}$. The goal is to find a minimum-cost subgraph H that satisfies these requirements;

[*] Received by the editors December 30, 2004; accepted for publication (in revised form) January 18, 2007; published electronically July 25, 2007. An extended abstract of this paper has appeared in the *Proceedings of the 35th ACM Symposium on the Theory of Computing*, 2005, pp. 167–175.
http://www.siam.org/journals/sidma/21-3/62180.html

[†] Department of Combinatorics and Optimization, University of Waterloo, Waterloo, ON, Canada, N2L 3G1 (jcheriyan@math.uwaterloo.ca). This author was supported in part by NSERC research grant 138432-04.

[‡] Department of Mathematics and Statistics and School of Computer Science, McGill University, Montreal, QC, Canada, H3A 2A7 (vetta@math.mcgill.ca). This author was supported in part by NSERC grant 288334-04 and FQRNT grant NC-98649.

that is, H should have $r_{i,j}$ openly node-disjoint paths between every pair of nodes i and j. There are two well-known special cases of NC-SNDP. The first is the *subset k-node-connectivity problem*, where we are given a set of terminal nodes $T \subseteq V$ and $r_{i,j} = k$ precisely if both i and j are in T; otherwise $r_{i,j} = 0$. The second is the classical *k-node connected spanning subgraph problem* (k-NCSS), where $r_{i,j} = k$ for every pair of nodes; this is the special case of the subset k-node-connectivity problem with $T = V$. We also study a new special case of NC-SNDP that we call the *subset $[k,\ 1.5k]$-node-connectivity* problem: given a set of terminal nodes $T \subseteq V$ and an (integer) requirement r_i for each node $i \in T$, where $1 \leq k \leq r_i \leq 1.5k$, the goal is to find a minimum-cost subgraph that has $\min(r_i, r_j)$ openly node-disjoint i, j-paths for every pair of nodes $i, j \in T$. (Thus the subset k-node-connectivity problem is the special case where $r_i = k\ \forall i \in T$.) See section 4 for more discussion.

Most network design problems stay NP-hard and APX-hard even assuming metric costs. This remains true even for small connectivity requirements; for example, Bern and Plassmann [3] showed that the Steiner tree problem (the classical special case of the subset k-node-connectivity problem with $k = 1$) is APX-hard even with edge costs of 1 and 2. Over the past decade, there has been significant research on approximation algorithms for network design, and there have been some notable successes in the design of networks that satisfy various types of "edge-connectivity" requirements; see, e.g., Goemans and Williamson [17] and Jain [18]. In particular, Jain [18] gives a 2-approximation algorithm for EC-SNDP, a problem similar to NC-SNDP except that the requirements are for edge-disjoint paths instead of openly node-disjoint paths. Nevertheless, from the perspective of approximation algorithms, the design of networks subject to "node-connectivity" requirements is a murky area. For example, Kortsarz, Krauthgamer, and Lee [21] recently proved a hardness-of-approximation lower bound of $2^{\log^{1-\epsilon} n}$ for the subset k-node connectivity problem in graphs with zero-one edge costs, provided that NP$\not\subseteq$ DTIME$(n^{polylog(n)})$, where ϵ denotes a small positive real number. (We give a detailed discussion on previous work in the area after stating our results.)

We present a 22-approximation algorithm for the metric-cost subset k-node-connectivity problem, and then we generalize this to get an $O(1)$-approximation algorithm for the metric-cost subset $[k,\ 1.5k]$-node-connectivity problem. These algorithms are deterministic and combinatorial; they do not use linear programming (LP) relaxations. Based on this, we present an $O(\log r_{\max})$-approximation algorithm for the metric-cost NC-SNDP. The algorithm for NC-SNDP is based on an LP relaxation. Also, it uses a 2-approximation algorithm of Goemans and Williamson [17] (see also Agrawal, Klein, and Ravi [1]) for the generalized Steiner tree problem. Moreover, we resolve the following folklore conjecture: In a metric graph and for each $k = 1, 2, \ldots, n-1$, the minimum cost of a k-node connected spanning subgraph is within a constant factor of the minimum cost of a simple k-edge connected spanning subgraph. Thus, for metric graphs, the requirements of k-node-connectivity and simple k-edge-connectivity are equivalent for the objective function, up to constant factors. A similar result holds for requirements of subset $[k,\ 1.5k]$-node-connectivity versus subset simple $[k,\ 1.5k]$-edge-connectivity.

We apply two lower bounds on the optimal value of the subset $[k, 1.5k]$-connectivity problem. We may assume (without loss of generality) that there exist at least two terminals with the maximum requirement. Hence, every solution subgraph has at least r_i edges incident to each terminal i, because there is another terminal j with $r_j \geq r_i$, so the solution subgraph must have r_i openly node-disjoint i, j-paths. Our first lower bound comes from the the minimum cost of a subgraph that has degree $\geq r_i$ for every terminal i. Our second lower bound comes from the cost of a MST of the subgraph induced by the terminals. For any node i, we use σ_i or $\sigma(i)$ to denote the sum of the costs of the r_i cheapest edges incident to i in the complete graph, and for any set of nodes S, we use $\sigma(S)$ to denote $\sum_{i \in S} \sigma_i$. We use the abbreviations MST and TSP as already defined. Let $mst(T)$ denote the cost of an MST of the subgraph induced by T. Our lower bounds are

(i) $\frac{1}{2}\,\sigma(T)$ and

(ii) $\frac{k}{2}\; mst(T)$.

Note that these lower bounds apply also to the simple edge-connectivity version of the subset $[k,\ 1.5k]$-connectivity problem, where the requirements are for $\min(r_i, r_j)$ edge-disjoint paths between every pair of nodes $i, j \in T$; note that multiedges are not allowed in the solution subgraph. See section 2 for more details. Throughout, we use OPT to denote the cost of an optimal solution. Next, we state our main results formally.

THEOREM 1. *There is a polynomial-time algorithm for computing a solution to the metric-cost subset k-node-connectivity problem of cost $\leq 9\sigma(T) + 4(\frac{k}{2})\; mst(T) \leq$ 22OPT.*

Consider k-NCSS, the special case of the subset k-node connectivity problem in which the terminal set T is V. Let k-ECSS be the problem of finding a minimum-cost *simple* k-edge connected spanning subgraph. Then our two lower bounds apply for both k-NCSS and k-ECSS. This gives the next result.

COROLLARY 2. *In a network with metric costs, there is a k-node connected spanning subgraph whose cost is at most 22 times the minimum cost of a simple k-edge connected spanning subgraph.*

Remarks. For metric graphs, it is well known that there exists a 2-node connected graph of cost $\leq$ the cost of any 2-edge connected graph on the same node set (see Appendix A), but this does not hold for $k \geq 3$ (see [4, Figure 1] and Appendix A for examples). Also, note that the $\frac{1}{2}\,\sigma(V)$ lower bound for k-ECSS does not apply for the version where multiedges are allowed. In more detail, if multiedges are allowed, then there exist k-edge connected graphs H such that any k-node connected graph on the same node set has cost $\geq \Theta(k)\, c(H)$. See Appendix A for more detail.

THEOREM 3. *There is a polynomial-time algorithm for computing a solution to the metric-cost subset $[k,\ 1.5k]$-node-connectivity problem of cost $\leq O(1) \cdot (\sigma(T) + \frac{k}{2}\; mst(T)) \leq O(1) \cdot$ OPT.*

Remark. A loose analysis gives a constant factor between 500 and 1000 in the above theorem. Possibly, an approximation guarantee of ≤ 100 can be obtained by some changes to the algorithm. We have not attempted to optimize the constants in the approximation guarantees.

THEOREM 4. *There is a polynomial-time algorithm for computing a solution to the metric-cost NC-SNDP of cost $\leq O(\ln r_{max}) \cdot$ OPT.*

Previous work. Over the past few decades, there has been significant research on approximation algorithms for network design. For early work in network design, see, for example, Dantzig, Ford, and Fulkerson [12]. A celebrated and still unsurpassed result was Christofides' $\frac{3}{2}$-approximation algorithm for the metric-cost TSP [8]. Partly motivated by Christofides' result, there followed a stream of research on related problems in the design of metric-cost networks. Most of this research focused on small connectivity requirements, such as 2-edge connectivity and 2-node connectivity; see Frederickson and Ja'Ja' [14], Monma and Shallcross [27], Monma, Munson, and Pulleyblank [26], and Bienstock, Brickell, and Monma [4]. For constant k, this last paper gives a constant-factor approximation algorithm for k-NCSS. Moreover, the proof also shows that for metric graphs and any constant k, there exists a k-node connected spanning subgraph of K_n whose cost is within a constant factor of the cost of any k-edge connected spanning subgraph; see [4, section 4]. They left open the question of extending these results to all k. This was followed by another burst of research, partly initiated by the work of Goemans and Bertsimas [15] who presented a logarithmic approximation algorithm for a general model called the edge-connectivity survivable network design problem (EC-SNDP), assuming metric costs. Soon after this, the research focus changed from metric costs to the more general setting of nonnegative costs. Agrawal, Klein, and Ravi [1] and Goemans and Williamson [17] built on the primal-dual method to obtain $O(1)$-approximation algorithms for some special cases of EC-SNDP with small (i.e., 0 and 1) connectivity requirements. Later, these methods were generalized to EC-SNDP, albeit with a logarithmic approximation guarantee, by Goemans et al. [16], based on work by Williamson et al. [31]. This line of research culminated with a 2-approximation algorithm for EC-SNDP by Jain [18]. Jain's approximation guarantee of 2 is tight in the sense that his algorithm and analysis are based on an LP relaxation, and this has an integrality ratio of 2. But there are no tight lower bounds from the hardness of approximation; i.e., an approximation guarantee less than 2 for EC-SNDP has not been ruled out.

Although there was considerable interest in extending these methods to the setting of node-connectivity, there was limited success even for rather special cases of NC-SNDP. We mention a few results and refer the interested reader to [6] for more references. For the case of nonnegative edge costs, the authors jointly with Vempala [7] and Kortsarz and Nutov [22] gave logarithmic (or worse) approximation guarantees for the k-NCSS problem. For metric costs, there is an $O(1)$-approximation algorithm due to Khuller and Raghavachari [20], and there are other related results in [5, 23]. Some explanation for this lack of good approximation algorithms for NC-SNDP comes from the recent hardness-of-approximation results of Kortsarz, Krauthgamer, and Lee [21]. Also, see the surveys by Frank [13], Khuller [19], and Stoer [28], and the book by Vazirani [30].

We briefly mention the relationship between our work and the stream of exciting recent results on PTAS (polynomial-time approximation schemes) for related problems. Beginning with the results of Arora [2] on the Euclidean TSP, many PTAS have been obtained for problems in "geometric network design" where the edge costs come from special metrics such as the Euclidean metric; see [9, 10, 11, 25] and the references in those papers. But, modulo P $\neq$ NP, such PTAS do not exist in the setting of interest to us, namely, (general) metric costs; this follows from APX-hardness results in [3, 21, 29].

The rest of the paper is structured as follows. In section 2, we discuss some preliminaries and give an overview of our method for the metric-cost subset k-node-connectivity problem. We present a constant-factor approximation algorithm for the problem in section 3. Section 4 gives a constant-factor approximation algorithm for a generalization. This leads to an $O(\log r_{\max})$-approximation algorithm for the metric-cost NC-SNDP in section 5.

2. Preliminaries and an overview of the algorithm for subset *k*-connectivity.

Apart from section 1, we omit the word "node" from terms such as "node-connectivity" when there is no danger of ambiguity.

Let the input graph be $G = (V, E)$. We denote the nodes by numbers $i =$

$1, 2, \ldots, n$, and for nodes i, j the edge between them is denoted ij. The cost of an edge $ij \in E$ is denoted c_{ij} or $c(i, j)$. Throughout, we assume $c_{ij} > 0 \; \forall ij \in E$. The costs are said to be *metric* if the triangle inequality holds: $c(v, w) \le c(v, u) + c(u, w) \; \forall u, v, w \in V$. Whenever we assume metric costs, we also assume that G is the complete graph. Let k be an integer such that $n > k \ge 1$ (k may be a function of n). For a graph H and a pair of nodes i, j, let $\kappa_H(i, j)$ denote the maximum number of openly node-disjoint i, j-paths in H. Recall that T denotes the set of terminal nodes. We use n' to denote $|T|$, and we assume $T = \{1, \ldots, n'\}$.

Let us formalize the lower bounds (i) and (ii) for the subset $[k,\ 1.5k]$-connectivity problem stated in section 1. For each terminal node i, let Γ_i denote the set of r_i nearest neighbors of i; by convention, $i \notin \Gamma_i$. (Thus $|\Gamma_i| = r_i$ and $\forall x \in \Gamma_i,\ y \notin \Gamma_i \cup \{i\},\ c_{iy} \ge c_{ix}$.) Then note that σ_i denotes $\sum_{x \in \Gamma_i} c_{ix}$. Also, for each terminal node i, let μ_i denote σ_i / r_i, namely, the average cost of an edge from i to one of its r_i nearest neighbors. Note that each terminal node i has at least r_i neighbors in an optimal subgraph; thus $\text{OPT} \ge \frac{1}{2}\sigma(T)$. This gives the first lower bound. Next, we claim that $\text{OPT} \ge \frac{k}{2} \mathit{mst}(T)$. In more detail, we have $\text{OPT} \ge \frac{1}{2}\text{ECOPT}(T, 2k) \ge \frac{k}{2}\mathit{mst}(T)$, where $\text{ECOPT}(T, \lambda)$ denotes the minimum cost of a λ-edge connected subgraph of $G[T]$ (allowing multiedges). To see this, start with a graph corresponding to OPT, and take two copies per edge to get an Eulerian multigraph H' that is $2k$-edge connected on T, then apply the Lovász–Mader splitting-off theorem [24, Example 6.51], [13] to eliminate all nodes of $V - T$ from H' to get a $2k$-edge connected multigraph on the node set T that has cost $\ge \text{ECOPT}(T, 2k)$. Then we apply the well-known fact that $\text{ECOPT}(T, \lambda) \ge \frac{\lambda}{2}\mathit{mst}(T)$. For metric costs, splitting off edges does not increase the cost. This gives the second lower bound: $\text{OPT} \ge \frac{k}{2}\mathit{mst}(T)$.

We first give an overview of our method for subset k-connectivity by describing a key special case where k is even, say $k = 2\ell$, and the sets $\{i\} \cup \Gamma_i$ of the terminals i are pairwise disjoint (that is, $(\{i\} \cup \Gamma_i) \cap (\{j\} \cup \Gamma_j) = \emptyset \; \forall i \ne j \in T$). Arbitrarily name the nodes in Γ_i as $i_1, i_2, \ldots, i_k \; \forall i \in T$. Construct a cheap cycle Q on the terminals using the well-known MST-doubling heuristic for the TSP. (Start with an MST of the subgraph induced by T, replace each edge by two copies, and shortcut the resulting connected Eulerian graph to get a cycle Q with $V(Q) = T$ and $c(Q) \le 2\mathit{mst}(T)$.) Let the sequence of terminals on Q be $1, 2, \ldots, n', 1$ (renumber the nodes if needed). For each $\tau = 1, \ldots, \ell$, construct a cycle Q_τ "parallel" to Q, where $Q_\tau = 1_\tau, 1_{\ell+\tau}, 2_\tau, 2_{\ell+\tau}, 3_\tau, \ldots, (n'-1)_{\ell+\tau}, n'_\tau, n'_{\ell+\tau}, 1_\tau$. (See Figure 1; informally, start with the cycle $1_\tau, 2_\tau, \ldots, n'_\tau, 1_\tau$, and then for each $i = 1, \ldots, n'$ insert the node $i_{\ell+\tau}$ between nodes i_τ and $(i+1)_\tau$; see the next paragraph for some discussion on our notation.) Let us refer to these cycles as *tracks*. It can be seen that a track Q_τ has cost $c(Q_\tau) \le c(Q) + \sum_{i=1}^{t} 2(c(i, i_\tau) + c(i, i_{\ell+\tau}))$ (see the second subroutine below), and the total cost of the tracks is $\sum_{\tau=1}^{\ell} c(Q_\tau) \le \ell \cdot c(Q) + 2\sigma(T)$. Finally, for each terminal $i \in T$, we add the k edges $ii_1, ii_2, \ldots, ii_k$. The resulting subgraph is our solution graph H; it has cost $c(H) \le 2\ell \cdot \mathit{mst}(T) + 3\sigma(T) \le 2\text{OPT} + 6\text{OPT} = 8\text{OPT}$. Note that each terminal has precisely two neighbors in each track. Thus H satisfies the connectivity requirements, because for every pair of terminals $i, j (i \ne j)$, each of the $k/2$ tracks contributes two openly disjoint i, j paths.

Our algorithms in sections 3 and 4 rely on the notion of tracks. As above, we start with a cycle Q on a subset of the terminals such that $c(Q) \le 2\mathit{mst}(T)$; Q is computed by the MST-doubling heuristic for the TSP. We index the terminals according to their ordering in Q as $1, 2, \ldots, n^*$, where $n^* = |V(Q)|$, and the terminals not in Q get indices higher than n^*. We have tracks $Q_1, Q_2, \ldots, Q_\ell$, where each is a cycle

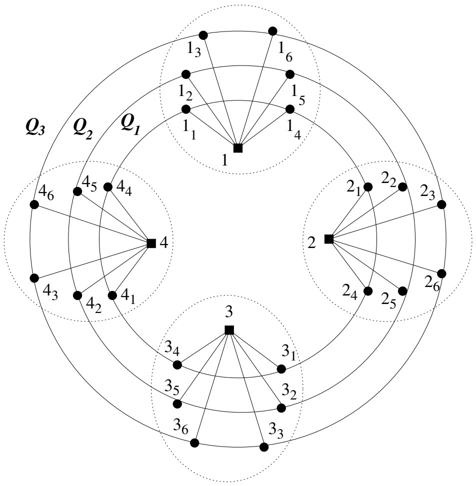


FIG. 1. *A key special case of the algorithm. Here, $k = 6$, $T = \{1, 2, 3, 4\}$, and the sets $\{i\} \cup \Gamma_i$ (indicated by dotted blobs) for $i \in T$ are pairwise disjoint. The tracks Q_1, Q_2, Q_3 are indicated by circles.*

"parallel" to Q. In more detail, track Q_τ (for $\tau = 1, \ldots, \ell$) has one or more nodes from Γ_i for each terminal $i = 1, 2, \ldots, n^*$; informally speaking, each terminal i in Q "places" one or two nodes from Γ_i into Q_τ, and it turns out that the name (or index) of the "first" of these two nodes is essential information (global information). We index the nodes in Γ_i as $i_1, i_2, \ldots, i_\ell, \ldots, i_k$ such that node i_τ (for $\tau = 1, \ldots, \ell$) is the "first" node from Γ_i placed in the track Q_τ; the index of the second node (if any) from Γ_i placed in Q_τ is implicit in the algorithm but is not relevant here. Thus the track Q_τ (for $\tau = 1, \ldots, \ell$) has the form $1_\tau, ., 2_\tau, ., 3_\tau, ., \ldots, n^*_\tau, ., 1_\tau$, where the "." denotes that there may or may not be a node i_q ($q > \ell$) between the nodes i_τ and $(i+1)_\tau$. (Although each terminal i in Q places ≤ 2 nodes from Γ_i into a track Q_τ, the track may have > 2 nodes from Γ_i since some other terminal j may place a node from $\Gamma_j \cap \Gamma_i$ as its second node in Q_τ; as stated already, we will explicitly reference only the "first" node placed into Q_τ by a terminal.)

The algorithm uses the following two subroutines. Note that the solution graph H is simple, so when we add edges to H we do so without creating multiedges.

- The first subroutine *copies a specified set of neighbors* of a terminal i to another terminal v (possibly, v is adjacent to i). More precisely, given a terminal i and a specified set of neighbors of i, call it N_i, and another terminal v, the subroutine adds an edge vx to H for each node $x \in N_i$ (without creating multiedges or loops in H). After this step, $\kappa_H(i, v) \geq |N_i|$. The cost of the new edges is $\leq |N_i|\, c(i, v) + \sum_{x \in N_i} c(i, x)$.
- The other subroutine starts with a cycle containing a terminal i and *inserts new node(s) into the cycle*. Given a cycle Q', a terminal i in Q', and a node $x \notin V(Q')$, we first add two copies of the edge ix to Q' to get a connected Eulerian graph. Then we shortcut this Eulerian graph (as in the MST-doubling heuristic for the TSP) to obtain a new cycle Q with node set $V(Q') \cup \{x\}$. The increase in cost is $\leq 2c(i, x)$.

It is important for our analysis to get good upper bounds on the costs of the tracks. Note that the tracks are pairwise node-disjoint; thus each terminal is in at most one track. But, for upper-bounding the track costs, we use the following accounting trick. Consider any track $Q_{\mathcal{T}}$. We assume that the track initially consists of all the terminals, thus $V(Q_{\mathcal{T}}) = T$, and using the MST-doubling heuristic we have $c(Q_{\mathcal{T}}) \leq 2mst(T)$. Subsequently, the algorithm may insert new nodes into the track—such insertions occur while we are processing some terminal—thus for inserting node x while processing terminal i the cost $c(Q_{\mathcal{T}})$ increases by $\leq 2c(i, x)$. Possibly, x may be another terminal—in that case, we implicitly remove x from $Q_{\mathcal{T}}$ and then insert x via the double-edge ix. At the end of the execution, we keep only those terminals that were explicitly inserted into $Q_{\mathcal{T}}$ and remove all the other terminals from $Q_{\mathcal{T}}$; clearly, this does not increase the cost $c(Q_{\mathcal{T}})$. Note that this "historical view" of $Q_{\mathcal{T}}$ is needed only for upper-bounding the cost. Other than this, it may be easier to view the tracks as being pairwise node-disjoint all through the execution, and this is the viewpoint we use in presenting the detailed algorithm.

3. The algorithm for subset *k*-connectivity. This section is devoted to an algorithm and proof for Theorem 1. The detailed algorithm follows. An analysis of the cost of the edges added to H (the solution graph) is given after the algorithm. A terminal may be in two states, *active* or *inactive*. Initially, all the terminals are active. Let ℓ denote $\lceil k/2 \rceil$. Initially, H is the graph consisting of all the terminal nodes and no edges; thus $H = (T, \emptyset)$.

(1) [DEACTIVATE TERMINALS AND CONSTRUCT DISJOINT BALLS FOR ACTIVE TERMINALS]
Renumber the terminals as $1, 2, \ldots, n'$ by increasing value of μ; thus $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_{n'}$.
Note: $\mu_h \leq \mu_j$ iff $\sigma_h \leq \sigma_j$.
Scan the terminals in the order $1, 2, \ldots, n'$, and skip the current terminal if it is inactive. Let $\alpha = 2$ and $\beta = 2$. For an active terminal i, let B_i be the subset of Γ_i consisting of the ℓ nearest neighbors of i (excluding i), and name the nodes in B_i as $i_1, i_2, \ldots, i_\ell$. Note that $|B_i| = \ell$, $i \notin B_i$, and $\forall x \in B_i, y \notin B_i \cup \{i\}$ we have $c_{ix} \leq c_{iy}$. A key fact is that $c_{ix} \leq \alpha\mu_i = 2\mu_i \ \forall x \in B_i$. (Otherwise, we have $\geq k/\alpha = k/2$ nodes x in Γ_i with $c(i, x) > \alpha\mu_i = \alpha\sigma_i/k$, so the sum of $c(i, x)$ over these nodes is $> \sigma_i$, and hence $\sum_{x\in\Gamma_i} c(i, x) > \sigma_i$, a contradiction.) Thus B_i may be viewed as a "ball" with center i and radius $\leq \alpha\mu_i$ that has the ℓ nearest neighbors of i (excluding i).
For each active terminal $v > i$, if $c_{iv} \leq (\alpha\mu_i + \beta\alpha\mu_v)$, then make v inactive and record i as the parent of v by assigning $p(v) = i$. (The aim is to ensure that the sets B_i of active terminals i are pairwise disjoint.) Also note that $\mu_{p(v)} \leq \mu_v$ for each inactive terminal v.

(2) [CONSTRUCT ℓ TRACKS ON THE DISJOINT BALLS]
After step (1), let T^* denote the set of active terminals, and let $n^* = |T^*|$. If $n^* < 3$, then apply step (2′) and stop. Otherwise, construct a cheap cycle Q on the active terminals by applying the MST-doubling heuristic for the TSP to the subgraph induced by T^*. Renumber the terminals such that $Q = 1, 2, \ldots, n^*, 1$; that is, the active terminals get the numbers in $\{1, \ldots, n^*\}$ according to their ordering in Q. Construct ℓ tracks $Q_1, Q_2, \ldots, Q_\ell$, where track $Q_{\mathcal{T}} = 1_{\mathcal{T}}, 2_{\mathcal{T}}, \ldots, n^*_{\mathcal{T}}, 1_{\mathcal{T}} (\mathcal{T} = 1, \ldots, \ell)$. Add all the tracks (but not the cycle Q) to H. The cost of the tracks constructed in this step is analyzed in Proposition 6 below.

(2′) [SPECIAL HANDLING FOR ONE OR TWO ACTIVE TERMINALS]
Skip this step if $n^* \geq 3$. Suppose $n^* = 1$. Let the active terminal be i. Add all the edges $iv, v \in \Gamma_i$, and then for each inactive terminal j, copy the set

Γ_i of neighbors of i to j. The resulting graph H satisfies the connectivity requirements.

Suppose $n^* = 2$. Let the active terminals be h, i, with $\sigma_h \leq \sigma_i$. Add all the edges $hq, q \in \Gamma_h$, and $iv, v \in \Gamma_i$. Then add a matching M of maximum size between the nodes in $\Gamma_i - (\Gamma_h \cup \{h\})$ and in $\Gamma_h - (\Gamma_i \cup \{i\})$; now, each matching edge qv (say, $q \in \Gamma_h - \{i\}$ and $v \in \Gamma_i - \{h\}$) gives an h, i path, namely, h, q, v, i. Finally, for each inactive terminal j, copy the set $\Gamma_{p(j)}$ of neighbors of $p(j)$ to j. The resulting graph H satisfies the connectivity requirements.

(3) [AUGMENT DISJOINT BALLS AND ASSIGN TOKEN ARCS]

In summary, this step scans the active terminals i and augments each "ball" B_i to get an "augmented ball" B'_i (that ideally has $|B'_i| = r_i = k$) such that these augmented balls are pairwise disjoint. The obvious construction for B'_i is to start with B_i and then add the nodes from $\Gamma_i - B_i$, but then the augmented balls may intersect. We "deintersect" two intersecting sets B'_h and B'_i, while preserving the balls B_h and B_i, by assigning so-called *token arcs* to the active terminals such that for each active terminal i, $|B'_i|$ plus the number of token arcs assigned to i is equal to $r_i = k$; for each node $q \in \Gamma_i$, we either keep q in B'_i or not, and if we do not keep q, then we assign to i a token arc of cost $\leq 3c_{iq}$. Later, in step (4), we handle the token arcs, by examining each token arc (i, j) and adding an edge iv to H such that the cost of the new edge is no more than the cost of the token arc; the end nodes of the new edge are partly specified by the end nodes of the token arc. The details follow.

Scan the active terminals $i = 1, 2, \ldots, n^*$ in order of increasing μ_i values; we resolve any "ties" by using the index; i.e., if $h < j$ and $\mu_h = \mu_j$, then h precedes j in our ordering. Start the scan of $i \in T^*$ by defining $B'_i := \Gamma_i$. If B'_i is disjoint from B'_h for all active terminals h that precede i in the ordering, then continue with the next active terminal; otherwise, for each active terminal h that precedes i in the ordering and has $B'_h \cap B'_i \neq \emptyset$, examine the nodes in $B'_h \cap B'_i$ in any order.

Note that h, i are active terminals, and $\mu_h \leq \mu_i$; hence, $c_{hi} \geq \alpha\mu_h + \beta\alpha\mu_i$. Let q be any node in $B'_h \cap B'_i$. We have two cases: either $c_{iq} \geq c_{hq}$ or not. Suppose $c_{iq} \geq c_{hq}$; then we have

$$c_{iq} \geq \frac{1}{2}(c_{iq} + c_{hq}) \geq \frac{1}{2}c_{hi} \geq \frac{1}{2}(\alpha\mu_h + \beta\alpha\mu_i) > \alpha\mu_i \geq \alpha\mu_h,$$

where the last two inequalities hold because $\beta = 2$ and $\mu_i \geq \mu_h > 0$. Note that $q \notin B_i$, since B_i has radius $\leq \alpha\mu_i$. We remove q from B'_i and give to i a token arc (i, h) with cost $3c_{iq}$. (Later, this token arc will be replaced by an edge ix, where $x \in B_h$; note that the cost of ix is $\leq c_{iq}+c_{hq}+c_{hx} \leq 2c_{iq}+\alpha\mu_h \leq 3c_{iq}$.) See Figure 2 for an illustration.

In the other case, we have $c_{hq} \geq c_{iq}$, and moreover, $c_{hq} \geq \frac{1}{2}(\alpha\mu_h + \beta\alpha\mu_i) > \alpha\mu_i \geq \alpha\mu_h$. Note that $q \notin B_h$, since B_h has radius $\leq \alpha\mu_h$. We remove q from B'_h and give to h a token arc (h, i) with cost $3c_{hq}$. (Later, this token arc will be replaced by an edge hx, where $x \in B_i$; note that the cost of hx is $\leq c_{hq} + c_{iq} + c_{ix} \leq 2c_{hq} + \alpha\mu_i \leq 3c_{hq}$.)

After step (3), note that the sets B'_i of the active terminals i are pairwise disjoint; for each active terminal i, the number of token arcs given to i plus $|B'_i|$ is $r_i = k$, and the cost of each token arc (i, j) given to i depends on the

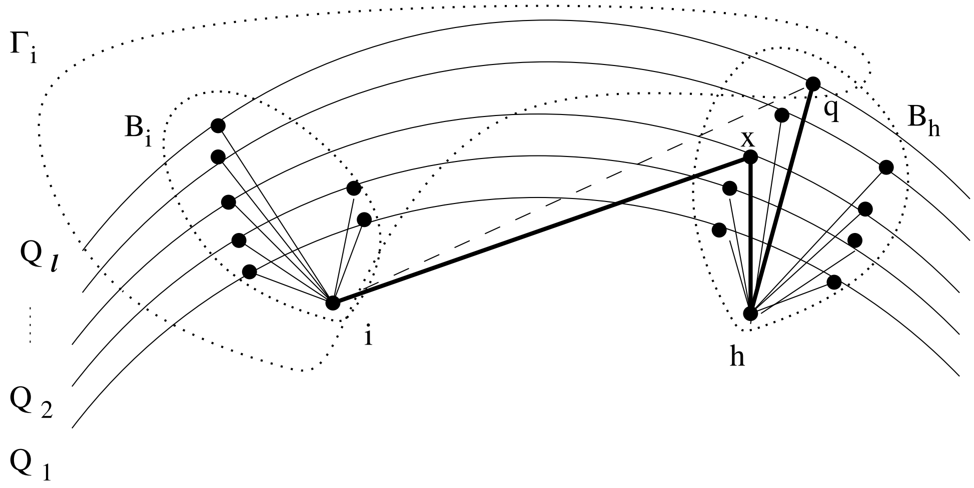


FIG. 2. *An illustration of step* (3) *in section* 3*: the "dashed edge" iq is replaced by a token arc ih that is later (in step* (4)*) replaced by an edge ix, $x \in B_h$.*

cost of the associated edge iq, where $q \in \Gamma_i$, and is $3c_{iq}$.

(4) [ATTACH ACTIVE TERMINALS TO TRACKS]
In summary, this step scans each active terminal i and adds edges from i to the tracks such that each track $Q_\tau, \tau = 1, \ldots, \lfloor k/2 \rfloor$, gets two neighbors of i, and the last track Q_ℓ gets ≥ 1 neighbor of i.
First add edges from i to each of $i_1, i_2, \ldots, i_\ell$; also, mark the nodes $i_1, i_2, \ldots, i_\ell$ as used.
Then for each $\tau = 1, 2, \ldots, \lfloor k/2 \rfloor$ do the following. If an unused token arc (i, h) is available, then choose it, mark it as used, and add the edge ih_τ; note that h_τ is in B_h and is the "first neighbor" of h in track Q_τ. By the discussion in step (3), $c(i, h_\tau)$ is $\leq$ the cost of the token arc (i, h). If no unused token arcs are available, then choose an unused node $q \in B'_i$, mark it as used, insert q into track Q_τ, and add the edge iq. (Note that the number of token arcs given to i plus $|B'_i|$ is k; hence, this step will find $\lfloor k/2 \rfloor$ token arcs or unused nodes, excluding the nodes $i_1, i_2, \ldots, i_\ell$.)
For each active terminal i, let N_i denote the set of neighbors of i in the tracks, just after step (4) is applied to i.

(5) [ATTACH INACTIVE TERMINALS TO TRACKS]
Finally, "attach" the inactive terminals to the tracks. Note that an inactive terminal may already be in one of the tracks. For each inactive terminal j, copy the set of neighbors $N_{p(j)}$ of the parent $p(j)$ to j.

PROPOSITION 5. *The cost of the graph constructed in step* $(2')$ *is* ≤ 16OPT.

Proof. Suppose $n^* = 1$, and let i be the (unique) active terminal. Then $c(H) \leq \sigma_i + \sum_{j \in T-T^*}(kc_{ij} + \sigma_i) \leq \sigma_i + \sum_{j \in T-T^*}(k(\alpha\mu_i + \beta\alpha\mu_j) + \sigma_i) \leq \sigma_i + \sum_{j \in T-T^*}(\alpha(1+\beta)\sigma_j + \sigma_j) \leq 7\sigma(T) \leq 14$OPT (we have $\alpha = 2, \beta = 2$, and we used $\sigma_i \leq \sigma_j$ for an inactive terminal j).

Suppose $n^* = 2$, and let i, h be the two active terminals. Then recall that M denotes a matching of maximum size between the nodes in $\Gamma_i - (\Gamma_h \cup \{h\})$ and in $\Gamma_h - (\Gamma_i \cup \{i\})$; note that an edge $qv \in M$ (say, $q \in \Gamma_h, v \in \Gamma_i$) has cost $\leq c_{hq} + c_{hi} + c_{iv}$; hence, $c(M) \leq \sigma_h + \sigma_i + k \cdot mst(T)$. The other edges in H contribute a cost of

$\leq \sigma_h + \sigma_i + \sum_{j \in T-T^*}(\alpha(1+\beta)\sigma_j + \sigma_j)$ (as in the analysis for $n^* = 1$), and hence, $c(H) \leq 7\sigma(T) + k \cdot mst(T) \leq 16$OPT. □

PROPOSITION 6. (i) *The total cost of the edges added by step* (4) *and incident to an active terminal* i *is* $\leq 3\sigma_i$. (ii) *At the end of step* (4), *the total cost of the* ℓ *tracks is* $\leq 2\ell \cdot mst(T) + 2\sigma(T^*)$.

Proof. For an active terminal i and a node $x \in \Gamma_i$, we either (a) add the edge ix to H or (b) find that x has been removed from B_i' (in step (3)), instead a token arc of cost $\leq 3c_{ix}$ being given to i, and we add an edge incident to i of cost $\leq 3c_{ix}$ (in step (4), when we handle the token arc). Thus the total cost of the added edges incident to i is $\leq 3\sigma_i$.

The total cost of the ℓ tracks (that were constructed in step (2) and modified in step (4)) is $\leq 2\ell \cdot mst(T) + 2\sigma(T^*)$. To see this, first consider the term $2\ell \cdot mst(T)$. Recall (from section 2) the accounting trick we use for upper-bounding the cost of a track; due to this, we take the upper bound on the cost of Q (the cheap cycle on T^* in step (2)) to be $2mst(T)$ rather than $2mst(T^*)$. Summed over ℓ tracks, this gives $2\ell \cdot mst(T)$. For the second term, note that $i \in T^*$ contributes $\leq \sum_{q \in B_i'} 2c(i,q)$, and this is $\leq 2\sigma_i$ (since $B_i' \subseteq \Gamma_i$). □

PROPOSITION 7. *The total cost of the edges added by step* (5) *and incident to the inactive terminals is* $\leq 9\sigma(T - T^*)$.

Proof. Suppose that the cost of the added edges incident to an active terminal i is $\leq \gamma\sigma_i$. (From Proposition 6, we have $\gamma = 3$.) Then the cost of the edges added for an inactive terminal j with parent i is $\leq k \cdot c_{ij} + \gamma\sigma_i \leq k(\alpha\mu_i + \beta\alpha\mu_j) + \gamma\sigma_i \leq (\alpha(\beta+1)+\gamma)\sigma_j$, using the fact that $\sigma_{p(j)} \leq \sigma_j$. Thus the total cost of the edges added in this step is $\leq 9\sigma(T - T^*)$, using $\alpha = 2, \beta = 2, \gamma = 3$. □

Proof of Theorem 1. If step (2′) is executed, then Proposition 5 shows that the total cost of H is ≤ 16OPT. Otherwise, by the above propositions, the total cost of H is $\leq 2\ell \cdot mst(T) + 2\sigma(T^*) + \gamma\sigma(T^*) + 9\sigma(T-T^*) \leq (k+1)mst(T) + 5\sigma(T^*) + 9\sigma(T-T^*) \leq (k+1)mst(T) + 9\sigma(T) \leq (2+\frac{2}{k})$OPT $+ 18$OPT ≤ 22OPT.

We claim that the graph H has the required connectivity property, namely, $\kappa_H(i,j) \geq k \ \forall i \neq j \in T$. To see this, consider any pair of terminals i, j, and consider any one track Q_τ. Suppose that either i is in Q_τ, or i is not in Q_τ but has two neighbors in Q_τ. Suppose that the same statement holds for j (that is, j is in Q_τ, or j is not in Q_τ but has two neighbors in Q_τ). Then, Q_τ (together with the edges from i and j to Q_τ) contributes two openly disjoint i,j-paths. Similarly, Q_τ contributes one i,j-path if both i and j either are in Q_τ or have a neighbor in Q_τ. By construction, each active terminal has two neighbors in each of the tracks Q_τ for $\tau = 1, \ldots, \lfloor k/2 \rfloor$, and has a neighbor in Q_ℓ; similarly, each inactive terminal is either in Q_τ or has two neighbors in Q_τ for $\tau = 1, \ldots, \lfloor k/2 \rfloor$, and is in Q_ℓ or has a neighbor in Q_ℓ. Then, for any two terminals i and j, H has k openly disjoint i,j-paths, since each of the tracks Q_τ for $\tau = 1, \ldots, \lfloor k/2 \rfloor$ contributes two openly disjoint i,j-paths, Q_ℓ contributes an i,j-path, and these k paths together are openly disjoint. □

4. The algorithm for subset $[k, 1.5k]$-connectivity. In this section, we extend the methods of the previous section to obtain an $O(1)$-approximation algorithm for the the subset $[k, 1.5k]$-connectivity problem. It seems likely that these methods will give similar results for the subset $[k, \rho k]$-connectivity problem, for any constant ρ, $1 \leq \rho < 2$, but they do not extend to $\rho = 2$ for the following reason: As in section 3, we choose some terminals to be active, and we construct pairwise-disjoint sets B_i of radius $O(1)\mu_i$ for the active terminals i, where B_i has at least a fraction ϕ of the nodes in Γ_i ($\phi = \frac{1}{2}$ in section 3); our method assumes $\phi \geq \frac{\rho}{2}$, i.e., that the size of every set B_i is at least half the maximum requirement. Then, for $\rho = 2$ and an active terminal i with $r_i = k$ we need $|B_i| \geq k = |\Gamma_i|$, and this is not possible for sets of radius $O(1)\mu_i$. Our main application is to the NC-SNDP, and for this any constant $\rho > 1$ suffices; we choose $\rho = 1.5$ for convenience.

The main difficulty in extending the methods of section 3 comes from the fact that an active terminal i may have an inactive terminal v with $r_v > r_i$ as a child. Then we cannot satisfy the connectivity requirement of v by copying the neighbors of i to v. Roughly speaking, We handle this as follows: We pick a child v^* of i with the maximum requirement and copy all the neighbors of i to v^*; then, if needed, we add new neighbors for v^* in the tracks by examining the nodes $x \in \Gamma_{v^*}$. If $x \in B'_h$ for some active terminal h, then we proceed similarly to step (3) of section 3 (though there are new complications); otherwise, either we insert x as a new node into a track or we "transform" to the case of $x \in B'_h$. For any other inactive child v of i, we attempt to copy the "first" r_v neighbors of v^* to v. This is an informal (and inaccurate) overview; the details are given below. The main contribution of this section is an algorithm and proof for the following restricted case of Theorem 3.

THEOREM 8. *Let k be an integer multiple of* 4, *thus* $k = 0 \pmod 4$. *There is polynomial-time algorithm for computing a solution to the metric-cost subset* $[k,\ 1.5k]$-*connectivity problem of cost* $\leq O(1) \cdot \text{OPT}$.

Remark. A loose analysis gives a constant factor between 500 and 600 in the above theorem.

Theorem 3 follows by combining this theorem with Theorem 1. To see this, suppose that $k \neq 0 \pmod 4$ (otherwise, we are done). Let $\hat{k} \geq k$ denote the next integer multiple of 4; clearly, $\hat{k} - k \leq 3$. Then for each $\rho = k, k+1, \ldots, \hat{k} - 1$, we apply the algorithm in Theorem 1 to the following instance $\Pi(\rho)$ of the subset ρ-connectivity problem to obtain a solution subgraph $H(\rho)$: we take the requirement of a terminal i in $\Pi(\rho)$ to be $r'_i = 0$ if $r_i < \rho$, and we take $r'_i = \rho$ if $r_i \geq \rho$; the rest of the instance stays the same. Finally, we apply the algorithm of this section to the instance of subset $[\hat{k},\ 1.5\hat{k}]$-connectivity, where we take the requirement of a terminal i to be $r'_i = 0$ if $r_i < \hat{k}$, and we take $r'_i = r_i$ if $r_i \geq \hat{k}$; the rest of the instance stays the same. Let H' be the solution subgraph. Then, for the original instance (of subset $[k,\ 1.5k]$-connectivity), we output the solution subgraph $H^* = H(k) \cup H(k+1) \cup \cdots \cup H(\hat{k}-1) \cup H'$, whose cost is at most $O(1)\text{OPT}$. To see that H^* satisfies the connectivity requirements, note that for every pair of terminals i, j, one of the subgraphs forming H^* (namely, one of $H(k), H(k+1), \ldots, H(\hat{k}-1), H'$) has $\min(r_i, r_j)$ openly disjoint i,j-paths.

Assume that k is an integer multiple of 4. Let ℓ denote $3k/4$. For any terminal i and any edge ix of the complete graph, let $\widetilde{c}_{ix} = \widetilde{c}(i, x)$ denote the normalized edge cost $\max(c_{ix}, \mu_i)$.

The detailed algorithm follows. A terminal may be in two states, *active* or *inactive.* Initially, all the terminals are active, and H is the graph consisting of all the terminal nodes and no edges; thus $H = (T, \emptyset)$. See Appendix B for a summary of the notation.

(1) [DEACTIVATE TERMINALS AND CONSTRUCT DISJOINT BALLS FOR ACTIVE TERMINALS]

Renumber the terminals as $1, 2, \ldots, n'$ by increasing value of μ; thus $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_{n'}$.

Note: $\mu_h \leq \mu_j$ does not imply $\sigma_h \leq \sigma_j$ since the requirements may differ, but we do have $\sigma_h \leq 1.5\sigma_j$ since $k \leq r_h, r_j \leq 1.5k$.

Scan the terminals in the order $1, 2, \ldots, n'$, and skip the current terminal if it is inactive. Let $\alpha = 4$ and $\beta = 2$. For an active terminal i, let B_i be the subset of Γ_i consisting of the ℓ nearest neighbors of i (excluding i), and name the nodes in B_i as $i_1, i_2, \ldots, i_\ell$. Note that $|B_i| = \ell$, $i \notin B_i$, and $\forall x \in B_i, y \notin B_i \cup \{i\}$ we have $c_{ix} \leq c_{iy}$. A key fact is that $c_{ix} \leq \alpha\mu_i\ \forall x \in B_i$. (Otherwise, we have $\geq r_i - \ell \geq r_i - \frac{3}{4}r_i = r_i/4 \geq k/\alpha$ nodes x in Γ_i with $c(i,x) > \alpha\mu_i = \alpha\sigma_i/k$, so these nodes contribute $> \sigma_i$ to $\sum_{x\in\Gamma_i} c(i,x)$.) Thus B_i may be viewed as a "ball" with center i and radius $\leq \alpha\mu_i$ that has the ℓ nearest neighbors of i (excluding i).

For each active terminal $v > i$, if $c_{iv} \leq (\alpha\mu_i + \beta\alpha\mu_v)$, then make v inactive and record i as the parent of v by assigning $p(v) = i$. (The aim is to ensure that the sets B_i of active terminals i are pairwise disjoint.) Also note that $\mu_{p(v)} \leq \mu_v$ for each inactive terminal v.

(2) [CONSTRUCT ℓ TRACKS ON THE DISJOINT BALLS]

After step (1), let T^* denote the set of active terminals, and let $n^* = |T^*|$. If $n^* < 3$, then apply step $(5')$ and stop. Otherwise, construct a cheap cycle Q on the active terminals by applying the MST-doubling heuristic for the TSP to the subgraph induced by T^*. Renumber the terminals such that $Q = 1, 2, \ldots, n^*, 1$; that is, the active terminals get the numbers in $\{1, \ldots, n^*\}$ according to their ordering in Q. Construct ℓ tracks $Q_1, Q_2, \ldots, Q_\ell$, where track $Q_\mathcal{T} = 1_\mathcal{T}, 2_\mathcal{T}, \ldots, n^*_\mathcal{T}, 1_\mathcal{T}\ (\tau = 1, \ldots, \ell)$. Moreover, we have a special track $Q_0 = Q$; this track is used to satisfy the requirements of inactive terminals, but not the requirements between active terminals. Add all the tracks to H. Note that an inactive terminal may be in one of the tracks $Q_1, Q_2, \ldots, Q_\ell$, although none of the active terminals are in those tracks. (Although the tracks are similar to each other, our construction distinguishes between the tracks and relies on the ordering of the tracks given by the track indices $0, 1, 2, 3, \ldots$.)

(3) [AUGMENT DISJOINT BALLS AND ASSIGN TOKEN ARCS]

This step is the same as step (3) in the algorithm for subset k-connectivity in section 3, except that some parameters are different: Here, we have $\alpha = 4$, $\beta = 2$, $\ell = \frac{3k}{4}$.

After step (3), note that the sets B'_i of the active terminals i are pairwise disjoint; each such set has size $\geq (3k/\alpha) = (3k/4)$ (since $B'_i \supseteq B_i$ and $|B_i| = 3k/\alpha$); moreover, for each active terminal i, the number of token arcs given to i plus $|B'_i|$ is equal to r_i, and the cost of each token arc (i,j) given to i depends on the cost of the associated edge iq, where $q \in \Gamma_i$, and is $3c_{iq}$.

(4) [ATTACH ACTIVE TERMINALS TO TRACKS]

In summary, this step scans each active terminal i and adds edges from i to the tracks such that i has a neighbor $i_\mathcal{T}$ in each of the tracks $Q_\mathcal{T}, \tau = 1, \ldots, \ell$, and moreover, i has a second neighbor in each of the $r_i - \ell$ tracks $Q_\mathcal{T}, \tau = 1, \ldots, r_i - \ell$. We call $i_\mathcal{T}$ the *inner neighbor* of i in $Q_\mathcal{T}$, and if i has another neighbor x in $Q_\mathcal{T}$, then we call x the *outer neighbor* of i in $Q_\mathcal{T}$.

Note that the sets B'_j of the active terminals j are pairwise disjoint. In step (4), every node added to a track is in B'_j for some active terminal j (this can be seen from the description below). We call a node x *free* if x is in none of the tracks of the current graph H. While processing a terminal v we may find a free node $x \in \Gamma_v$, and we may insert x as the outer neighbor of v in a track. Throughout the execution, x stays in the same track and stays as the outer neighbor of v, but other terminals too may add x as their outer neighbor on that track.

We examine the active terminals in any order. Let i be the current active terminal. First, we add edges from i to each of $i_1, i_2, \dots, i_\ell$; also, we mark the nodes $i_1, i_2, \dots, i_\ell$ as used (with respect to i). We start with the variable $\tau = 1$; this variable denotes the number of the track where the next outer neighbor of i is placed.

If an unused token arc (i, h) is available, then we choose it, mark it as used, add the edge ih_τ, and increase τ by one; note that h_τ is in B_h and is the inner neighbor of h in track Q_τ; also, note that $c(i, h_\tau)$ is $\leq$ the cost of the token arc (i, h). We repeat this step until there are no unused token arcs.

We choose an unused node $x \in B'_i - B_i$ with minimum $c(i, x)$, and mark it as used w.r.t. i (note that x is a free node). Then we insert x into track Q_τ and add the edge ix, *provided* there exists *no* suitable "target terminal" $h \neq i$ (the details are given below; note that the target terminal is defined with respect to the cost c_{ix}). If a suitable h exists, then we discard x and add the edge ih_τ; that is, we take the inner neighbor of h in Q_τ to be the outer neighbor of i in Q_τ. (The reason for using an edge ih_τ rather than ix is that x is a free node now, but later we may find that x is essential for attaching some inactive terminal v to the tracks, and at that step, we will be forced to "replace" the edge ix by some other edge iy; to avoid such "replacements" we look ahead, and we use the edge ix only if there are no "future conflicts" for x.)

The details are as follows. We check whether there exists an active terminal $h \neq i$ such that

$$\text{hop-rule} \qquad c(i,h) \leq \left(2 + \frac{\alpha(\beta+1)}{2}\right) c(i,x) \leq 8c(i,x) \qquad \text{and} \qquad \mu_h \leq c(i,x).$$

If such an h exists, then we add the edge ih_τ. Note that $c(i, h_\tau) \leq c(i, h) + \alpha\mu_h \leq (8 + \alpha)c(i, x) \leq 12c(i, x)$. If no such h exists, then (as mentioned before) we insert x into track Q_τ and add the edge ix. (This causes an increase of $2c_{ix}$ in the cost of the tracks and an increase of c_{ix} in the cost of the edges from i to the tracks; we use these facts in the proof of Proposition 9 below.) In either case, we increase τ by one.

Note that the number of token arcs given to i plus $|B'_i|$ is r_i; hence, this step will find r_i token arcs or unused nodes (including the nodes $i_1, \dots, i_\ell$).

After all active terminals have been examined by step (4), it can be seen that H satisfies the connectivity requirements of all active terminals.

For each active terminal i, let N_i denote the (ordered) set of neighbors of i in the graph $H - V(Q_0)$ at the end of step (4). (Thus, N_i is the set of neighbors attaching i to the other tracks—excluding the track Q_0 containing i.) Note that $|N_i| = r_i \ \forall i \in T^*$. Moreover, we order the nodes in each set N_i such that the inner neighbors of i come first in the order $i_1, i_2, \dots, i_\ell$, followed by the outer neighbors in the order of their track numbers (the outer neighbor in Q_1, followed by the outer neighbor in Q_2, ...).

Remark. The ordered sets N_i for $i \in T^*$ are used in step (5), and there it is critical that the total cost of the edges from i to the nodes in N_i is $\leq \gamma\sigma_i$ for a constant γ; in particular, none of these edge costs should be "charged" to the *mst* lower bound.

(5) [ATTACH INACTIVE TERMINALS TO TRACKS]
Finally, "attach" each inactive terminal to the tracks.

By a *sibling* of an inactive terminal v we mean either the parent $p(v)$ or another child of $p(v)$. In summary, we first copy to v the neighbors of a sibling, and then, if needed, we add additional neighbors via Γ_v—note that v's requirement r_v may be much greater than that of any of its siblings, and hence copying the neighbors of a sibling may not suffice. We also use the special track Q_0 to satisfy the requirements of inactive terminals. To see the need for Q_0 consider a child v of an active terminal i with $r_v = r_i + 1$ and $\Gamma_v = \{i\} \cup \Gamma_i$; we handle the requirement of v by adding the edge vi, thus attaching v to Q_0, and then copying the neighbors of i to v.

Focus on an active terminal i and its children, and let v^* have the maximum requirement among these terminals; assume that $v^* \neq i$ (the other case is easy). Step (5) attaches v^* to the tracks via a neighbor in each of the $\ell+1$ tracks $Q_0, Q_1, \ldots, Q_\ell$ and two neighbors in each of the $r_{v^*} - (\ell+1)$ tracks $Q_1, Q_2, \ldots, Q_{r_{v^*}-(\ell+1)}$. These neighbors of v^* constitute the ordered set N_{v^*}; we use our standard ordering; i.e., the neighbor $i = p(v^*)$ in Q_0 comes first, followed by the inner neighbors in the tracks $Q_1, \ldots, Q_\ell$, followed by the outer neighbors, and further, the neighbors are ordered by their track number. Similarly, we have an ordered set of neighbors N_v for each inactive terminal v, where N_v is the (ordered) set of nodes x such that step (5)—while processing v—adds an edge from v to x. (Possibly, v occurs in a track, but then neither of the two neighbors of v in the track occurs in N_v unless step (5)—while processing v—adds the edge from v to that node.) A key property of our construction is that for each sibling v of v^*, N_v is a prefix of N_{v^*}; in particular, for each $\tau \in \{1, 2, \ldots, r_v - (\ell+1)\}$, the outer neighbor of v in track Q_τ is the same as the outer neighbor of v^* in that track. (Thus, for siblings $v_1, v_2, \ldots$, our construction makes the sets $N_{v_1}, N_{v_2}, \ldots$ "consistent" even though the sets $\Gamma_{v_1}, \Gamma_{v_2}, \ldots$ may have arbitrary intersections.)

We examine the active terminals i in order of increasing μ_i values, and we examine the children v (of i) in order of increasing μ_v values. By a *prior sibling* of v we mean either the parent $p(v)$ or another child of $p(v)$ that precedes v in this ordering. For each child v of i, define the source terminal of v, denoted $\hat{p}(v)$, to be a prior sibling with the maximum requirement; furthermore, define the ordered set N_v^0 to be $\{p(v)\} \cup N_{p(v)}$ if $\hat{p}(v) = p(v)$ (i.e., the source terminal is the parent), and let $N_v^0 = N_{\hat{p}(v)}$ otherwise.

If the requirement of v is $\leq |N_v^0|$, then we "copy" the first r_v nodes of N_v^0 to v; i.e., for each of the first r_v nodes in the ordered set N_v^0, we add an edge from v to that node. Step (5) for v is finished after this.

If the requirement of v is $> |N_v^0|$, then we "copy" all the nodes of N_v^0 to v; i.e., for each of the nodes in N_v^0, we add an edge from v to that node. We mark all these new neighbors of v as used w.r.t. v.

Let $\tau \in \{1, \ldots, \ell\}$ be the next available track for v; i.e., v has two neighbors in each of the tracks $Q_1, \ldots, Q_{\tau-1}$ but has only one neighbor in each of the tracks $Q_\tau, \ldots, Q_\ell$.

We repeat the following until step (5) has added a total of r_v neighbors of v. We pick an unused node $x \in \Gamma_v$ with minimum c_{vx}, and mark x as used w.r.t. v. First, suppose that x is free. If the following version of the hop-rule does *not* apply to c_{vx} (i.e., there exists no h satisfying the rule), then we insert x into track Q_τ and add the edge vx. Also, we increase τ by one. This causes an increase of $2c_{vx}$ in the cost of the tracks, and an increase of c_{vx} in the cost of the edges from v to the tracks; we use these facts in the proof of Proposition 10 below.

To apply the modified hop-rule, we check whether there exists an active terminal $h \neq i = p(v)$ such that

$$c(v,h) \leq (2 + (\beta+1)\alpha)\widetilde{c}(v,x) \leq 14\widetilde{c}(v,x) \qquad \text{and} \qquad \mu_h \leq \widetilde{c}(v,x);$$

recall that $\widetilde{c}(j,y)$ denotes $\max(c_{jy},\ \mu_j)$ for any terminal j and any $y \in V$. If such an h exists, then we add the edge $vh_{\mathcal{T}}$ and we have $c(v,h_{\mathcal{T}}) \leq c(v,h) + \alpha\mu_h \leq (2+(\beta+2)\alpha)\widetilde{c}(v,x) \leq 18\widetilde{c}(v,x)$. Also, we increase τ by 1.

Now, suppose that x is not free. Then one of the following mutually exclusive cases applies:

(a) $x \in \{h\} \cup N_h$ for some active terminal h.

(b) x is in one of the tracks, and (a) does not apply.

Consider case (a). Note that $h \neq p(v) = i$, because we added edges from v to all nodes in $N_v^0 \supseteq N_i \cup \{i\}$ (and marked all those nodes as used w.r.t. v) before picking x; hence, $x \notin N_i \cup \{i\}$. First, suppose that $2\widetilde{c}(v,x) \geq \widetilde{c}(h,x)$. (*Remark*. Parts of the following analysis remain valid if we replace 2 by $\frac{9}{7}$, but the approximation guarantee does not seem to improve substantially, so we use 2 for convenience.) Then we discard x as a neighbor of v, and we add the edge (v,h_τ) to H; i.e., the inner neighbor of h in Q_τ is made the outer neighbor of v in Q_τ. Also, we increase τ by one. The new edge has cost $c(v,h_\tau) \leq c(v,x)+c(h,x)+\alpha\mu_h \leq (1+2)\widetilde{c}(v,x)+\alpha\mu_h \leq (1+2+2\alpha)\widetilde{c}(v,x) \leq 11\widetilde{c}(v,x)$.

Now, suppose that $2\widetilde{c}(v,x) < \widetilde{c}(h,x)$. Then we have a contradiction. To see this, consider two mutually exclusive cases: (I) $c_{hx} < \mu_h$, or (II) $c_{hx} \geq \mu_h$. In case (I), we have $c(h,i) \leq c(h,x) + c(v,x) + c(v,i) < \mu_h + c(v,x) + (\beta\alpha\mu_v + \alpha\mu_i) \leq \mu_h + (1+\beta\alpha)\widetilde{c}(v,x) + \alpha\mu_i = \mu_h + 9\widetilde{c}(v,x) + \alpha\mu_i$ (by $\alpha = 4, \beta = 2$) $< \alpha\mu_i + (1+5)\mu_h$ (by $9\widetilde{c}(v,x) < 4.5\widetilde{c}(h,x) = 4.5\mu_h$) $< \alpha\mu_i + \beta\alpha\mu_h$, and moreover, $\mu_i \leq \mu_v \leq \widetilde{c}(v,x) < \mu_h$. To verify the contradiction, recall from step (1) that for active terminals i,h with $\mu_i \leq \mu_h$, we have $c(h,i) \geq \alpha\mu_i + \beta\alpha\mu_h$. See Figure 3 for an illustration. In case (II), we have a contradiction by the hop-rule of step (4), because $c(h,i) \leq c(h,x) + c(v,x) + c(v,i) < \frac{3}{2}c(h,x) + (\beta+1)\alpha\mu_v < \frac{(3+(\beta+1)\alpha)}{2}c(h,x)$ (by $\widetilde{c}(v,x) < \frac{1}{2}\widetilde{c}(h,x) = \frac{1}{2}c_{hx}$) $< 8c(h,x)$ and $\mu_i \leq \mu_v \leq \widetilde{c}(v,x) < c(h,x)$. Thus the hop-rule of step (4) applies to h and hx, so the active terminal h cannot use the edge hx. Hence, we cannot have $2\widetilde{c}(v,x) < \widetilde{c}(h,x)$.

Now, consider case (b). Let w be the first inactive terminal whose processing by step (5) results in the addition of the edge wx (i.e., x changes from a free node to a nonfree node when step (5) processes w). Let $p(w) = h$. Note that $h \neq i$ (i.e., $p(w) \neq p(v)$), otherwise, w is a prior sibling of v (since the edge wx was added during the processing of w by step (5)), and for every prior sibling u of v, each node in N_u is already used w.r.t. v (since we added edges from v to all nodes in $N_v^0 \supseteq N_{\hat{p}(v)}$). First, suppose that $\widetilde{c}(v,x) \geq \widetilde{c}(w,x)$. Then we discard x as a neighbor of v, and we add the edge (v,h_τ) to H. Also, we increase τ by one. The new edge has cost $c(v,h_\tau) \leq c(v,x)+c(w,x)+c(h,w)+\alpha\mu_h \leq 2\widetilde{c}(v,x)+\alpha(\beta+2)\mu_w \leq 2\widetilde{c}(v,x)+\alpha(\beta+2)\widetilde{c}(v,x) \leq (2+\alpha(\beta+2))\widetilde{c}(v,x) \leq 18\widetilde{c}(v,x)$. Now, suppose that $\widetilde{c}(v,x) < \widetilde{c}(w,x)$. Then we have a contradiction by the modified hop-rule of step (5), because $c(w,i) \leq c(w,x)+c(v,x)+c(v,i) < 2\widetilde{c}(w,x)+(\beta+1)\alpha\mu_v <$

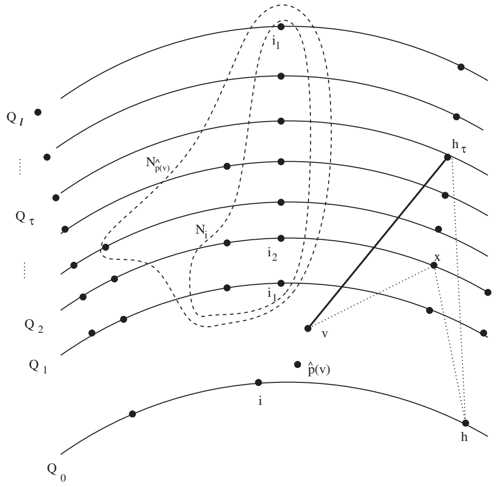


FIG. 3. *An illustration of step* (5) *case* (a) *in section* 4, *showing* v, $p(v) = i$, *the prior sibling* $\hat{p}(v)$ *of* v, N_i, *and* $N_{\hat{p}(v)}$; *the "dashed edge"* vx *is replaced by the edge* vh_τ *since the active terminal* h *has* $x \in B_h$.

$(2 + (\beta + 1)\alpha)\widetilde{c}(w, x) \le 14\widetilde{c}(w, x)$ and $\mu_i \le \mu_v \le \widetilde{c}(v, x) < \widetilde{c}(w, x)$. Thus the modified hop-rule of step (5) applies to w and wx, so the inactive terminal w cannot use the edge wx. Hence, we cannot have $\widetilde{c}(v, x) < \widetilde{c}(w, x)$.

This completes the description of step (5).

(5′) [SPECIAL HANDLING FOR ONE OR TWO ACTIVE TERMINALS]

Suppose that $n^* = 1$ and $T^* = \{i\}$. Then we ignore the tracks altogether, but we compute the ordered set N_i via step (4) applied to i, and the ordered set N_v for each inactive terminal v by applying step (5) to v. We add the edges from each terminal v (where $v = i$ or v is a child of i) to all the nodes in N_v.

Now, suppose that $n^* = 2$ and $T^* = \{h, i\}$. We proceed as in steps (2)–(5), except that we temporarily allow tracks that consist of exactly two nodes and two copies of the edge between them. In particular, the special track Q_0 consists of nodes h, i and two copies of the edge hi. At the end, for each track consisting of exactly two nodes, we keep only one copy of the edge between them; thus the solution graph H is simple.

PROPOSITION 9. (i) *The total cost of the edges added by step* (4) *and incident to an active terminal* i *is* $\le 12\sigma_i$. (ii) *At the end of step* (4), *the total cost of the* $\ell + 1$ *tracks is* $\le (2\ell + 2)mst(T) + 2\sigma(T^*)$.

Proof. Recall the proof of Proposition 6. Similarly to that, for an active terminal i and a node $x \in \Gamma_i$, we either add the edge ix to H, or we add an edge incident to i of cost $\le 3c_{ix}$ (via a token arc), or we add an edge incident to i of cost $\le 12c_{ix}$ (via the hop-rule of step (4)). Thus the total cost of the added edges incident to i is $\le 12\sigma_i$. This proves part (i).

For part (ii), observe that the total cost of the $\ell + 1$ tracks (that were constructed in step (2) and modified in step (4)) is $\le (2\ell + 2)mst(T) + 2\sigma(T^*)$; for the second term, note that the contribution of $i \in T^*$ is $\le 2\sigma_i$. (In steps (4) and (5), while processing a terminal v, we may insert a node $x \in \Gamma_v$ into a track; this increases the cost of the track by $\le 2c_{vx}$.) □

PROPOSITION 10. (i) *The total cost of the edges added by step* (5) *or* (5′) *and* *incident to an inactive terminal v is $\leq 294\sigma_v$.* (ii) *The total increase in the cost of the tracks in step* (5) *or* ($5'$) *is at most $2\sigma(T - T^*)$.*

Proof. We claim that the cost of the added edges for an inactive terminal v with parent i is $\leq 294\sigma_v$. Let γ be a constant such that the cost of the added edges incident to an active terminal i is $\leq \gamma\sigma_i$. (From step (4) and Proposition 9, we have $\gamma = 12$.) First, note that if $r_v \leq r_{p(v)}$, then the cost of the added edges incident to v is given by the cost of copying r_v neighbors from the parent $p(v)$, and this cost is $\leq \gamma\sigma_{p(v)} + r_v \cdot c(v, p(v)) \leq 1.5\gamma\sigma_v + r_v \cdot (\beta + 1)\alpha\mu_v \leq (1.5\gamma + (\beta + 1)\alpha)\sigma_v \leq 30\sigma_v$. Now, assume that $r_v \geq r_{p(v)}$, and hence $\sigma_v \geq \sigma_{p(v)}$.

First, consider the cost incurred in copying the neighbors of the source terminal $\hat{p}(v)$. This cost consists of two components, (i) the cost of copying $r_{p(v)} \leq r_v$ neighbors from the parent $p(v)$, and (ii) the cost of copying the remaining (at most $r_{\hat{p}(v)} - r_{p(v)} \leq k/2$) nodes from $N_{\hat{p}(v)}$. The component (i) is $\leq \gamma\sigma_{p(v)} + r_v \cdot c(v, p(v)) \leq \gamma\sigma_v + r_v \cdot (\beta + 1)\alpha\mu_v \leq (\gamma + (\beta + 1)\alpha)\sigma_v \leq 24\sigma_v$.

Now, consider component (ii). We claim that component (ii) is $\leq 246\sigma_v$. Consider any node $y \in N_{\hat{p}(v)} - N_{p(v)}$. Let w be the first (earliest processed) sibling of v that has an edge wy (i.e., step (5) added the edge wy while processing w, and no prior sibling u of w has $y \in N_u$); possibly, $w \neq \hat{p}(v)$. Call w the *sponsor* of y. By examining the details of step (5), it can be seen that for each node $y \in N_w - N_{\hat{p}(w)}$, there exists a distinct node $y' \in \Gamma_w$ such that $c(w, y) \leq 18\widetilde{c}(w, y')$. Thus for each node $y \in N_{\hat{p}(v)} - N_{p(v)}$, the sponsor w of y has a distinct node $y' \in \Gamma_w$ such that $c(w, y) \leq 18\widetilde{c}(w, y')$. Moreover, there is a distinct node $x' \in \Gamma_v$ such that $c(w, y') \leq 24\mu_v + c(v, x')$. To see this, recall that $\mu_w \leq \mu_v$, and first note that $c(v, w) \leq c(v, p(v)) + c(p(v), w) \leq 2(\beta + 1)\alpha\mu_v \leq 24\mu_v$; next, focus on the nodes x_j in Γ_v ordered by increasing cost of the edge vx_j, say $x_1, x_2, \ldots, x_{r_v}$; suppose that y' is the sth closest neighbor of w. Then note that $c(w, y') \leq 24\mu_v + c(v, x_s)$ because each of the nodes x_j in Γ_v has $c(w, x_j) \leq c(v, w) + c(v, x_j)$; hence, for each of the s nodes x_j, $j = 1, \ldots, s$, we have $c(w, x_j) \leq 24\mu_v + c(v, x_j)$. Moreover, $c(v, y) \leq c(v, w) + c(w, y) \leq 24\mu_v + c(w, y)$. Hence, for each node $y \in N_{\hat{p}(v)} - N_{p(v)}$, there is distinct node $x' \in \Gamma_v$ such that $c(v, y) \leq 24\mu_v + 18(24\mu_v + c(v, x')) \leq 456\mu_v + 18c(v, x')$ (since $\mu_w \leq \mu_v$ and $c(w, y) \leq 18\widetilde{c}(w, y') \leq 18(24\mu_v + c(v, x'))$, where w and y' are as above). Then, summing over all nodes $y \in N_{\hat{p}(v)} - N_{p(v)}$, we see that the total cost of these edges vy is $\leq (|N_{\hat{p}(v)} - N_{p(v)}|)(456\mu_v) + 18\sigma_v \leq (k/2)(456\mu_v) + 18\sigma_v \leq 228\sigma_v + 18\sigma_v \leq 246\sigma_v$.

Finally, consider the total cost of the edges from v to the nodes in $N_v - N_{\hat{p}(v)}$. As mentioned above, for each node $y \in N_v - N_{\hat{p}(v)}$, there exists a distinct node $y' \in \Gamma_v$ such that $c(v, y) \leq 18\widetilde{c}(v, y')$. Also, $|N_v - N_{\hat{p}(v)}| \leq r_v - k$, and for each node $y' \in \Gamma_v$ we have $\widetilde{c}(v, y') \leq \mu_v + c(v, y')$. Hence, $\sum\{c(v, y) \ : \ y \in N_v - N_{\hat{p}(v)}\} \leq (r_v - k) \cdot 18\mu_v \ + \ 18\sigma_v \leq 36\sigma_v - 18k\mu_v \leq 36\sigma_v - 18(\frac{2r_v}{3})\mu_v = 24\sigma_v$.

Summing the three contributions (from components (i), (ii), and the previous paragraph), we see that the total cost of the edges added (by step (5) or ($5'$)) incident to an inactive terminal v is $\leq (24 + 246 + 24)\sigma_v \leq 294\sigma_v$.

The total increase in the cost of the tracks in step (5) or ($5'$) is at most $2\sigma(T - T^*)$, because during the processing of an inactive terminal v, the step may insert each node $x \in \Gamma_v$ into the tracks at an incremental cost of $2c(v, x)$. This completes the proof of the proposition. □

Remarks. The constant factor in the above proposition is not optimal. We did not optimize the analysis, to avoid further complications.

Proof of Theorem 8. We claim that the cost of the solution subgraph H is $c(H) \leq 600\text{OPT} = O(1)\text{OPT}$. By Propositions 9 and 10 and using $k \geq 4$, we have $c(H) \leq$

$(2\sigma(T^*) + 2\sigma(T - T^*) + (2\ell + 2)mst(T)) + (12\sigma(T^*) + 294\sigma(T - T^*)) \leq 296\sigma(T) + (4)(\frac{k}{2})mst(T) \leq 600\textsc{opt}$.

We claim that the solution subgraph H satisfies the connectivity requirements. Consider any pair of inactive terminals s, t. (The proof is similar but simpler for a pair of active terminals, or for one active and one inactive terminal.) First assume that there are at least three active terminals (that is, $|T^*| \geq 3$). Without loss of generality let $r_s = \min(r_s, r_t)$. We claim that H has r_s openly disjoint s,t-paths. Recall that each inactive terminal v has inner neighbors on all $\ell + 1$ tracks, and has outer neighbors on the first $r_v - (\ell + 1)$ tracks among $Q_1, \ldots, Q_\ell$. (An active terminal v has at least $r_v - \ell$ tracks that have outer neighbors.) It follows that we have have $\ell + 1 + r_s - (\ell + 1) = r_s$ openly disjoint s,t-paths using these tracks. (One of these s,t-paths consists of a path of the special track $Q_0 = Q$ and the edges $sp(s)$ and $tp(t)$.)

Clearly, the connectivity requirements hold for the case of $|T^*| = 1$. Now, suppose that $|T^*| = 2$. The above arguments still apply unless both s and t have inner and outer neighbors on a track that consists of exactly two nodes, call them x and y. In this case, our track consists of a single edge xy (since we discarded the second copy of xy in step (5′)). Still, this track gives two openly disjoint s,t-paths, namely, s, x, t and s, y, t. Thus it can be seen that the connectivity requirements hold. This completes the proof of Theorem 8. □

5. The algorithm for NC-SNDP. This section presents a proof of Theorem 4, based on (the algorithms in) Theorems 1 and 3. For the sake of motivation, let us obtain an $O(\ln r_{max})$-approximation algorithm for a restricted version of NC-SNDP, where every terminal has a requirement r_i and every pair of terminals i, j has the requirement $r_{i,j} = \min(r_i, r_j)$. The method is similar to the method for proving Theorem 3 from Theorems 1 and 8.

Let $\textsc{opt}$ denote the optimal value of the instance (of restricted NC-SNDP). First, for each $\rho = 1, 2, \ldots, 7$, we apply the algorithm in Theorem 1 to the following instance $\Pi(\rho)$ of the subset ρ-connectivity problem to obtain a solution subgraph $H(\rho)$: We take the requirement of a terminal i in $\Pi(\rho)$ to be $r'_i = 0$ if $r_i < \rho$, and we take $r'_i = \rho$ if $r_i \geq \rho$; the rest of the instance stays the same. By Theorem 1, the cost of $H(\rho)$ is $O(1) \cdot \textsc{opt}$. After this, we repeatedly apply the algorithm in Theorem 8 to solve an instance (specified below) of subset $[\rho,\ 1.5\rho]$-connectivity, where ρ is an integer multiple of 4 ($\rho = 8, 12, 16, 24, \ldots$; details later), to obtain a solution subgraph $H'(\rho)$. The instances of subset $[\rho,\ 1.5\rho]$-connectivity are as follows: We take the requirement of a terminal i to be $r'_i = 0$ if $r_i < \rho$, we take $r'_i = r_i$ if $\rho \leq r_i \leq 1.5\rho$, and we take $r'_i = 1.5\rho$ if $r_i > 1.5\rho$. By Theorem 8, the cost of $H'(\rho)$ is $O(1) \cdot \textsc{opt}$. We start with $\rho = 8$, and we iterate until $r_{max} \leq 1.5\rho$; after each iteration, we update ρ to the largest integer multiple of 4 that is ≤ 1.5 times the previous ρ. Clearly, the number of iterations is $O(\ln r_{max})$. Finally, we output the solution subgraph H^* for the instance (of restricted NC-SNDP); H^* is the union of all the solution subgraphs $H(\rho)$, $\rho = 1, \ldots, 7$, and $H'(\rho)$, $\rho = 8, 12, \ldots$. Thus H^* is the union of $O(\ln r_{max})$ subgraphs such that each of these subgraphs has cost $O(1) \cdot \textsc{opt}$, and so H^* has cost $O(\ln r_{max}) \cdot \textsc{opt}$. To see that H^* satisfies the connectivity requirements, note that for every pair of terminals i, j, one of the subgraphs forming H^* has $\min(r_i, r_j)$ openly disjoint i,j-paths, namely, the subgraph $H(\min(r_i, r_j))$ if $\min(r_i, r_j) \leq 7$, and otherwise, any subgraph $H'(\rho)$, where ρ satisfies $\rho \leq \min(r_i, r_j) \leq 1.5\rho$.

Our algorithm for metric-cost NC-SNDP is similar to the algorithm described above for the restricted version of NC-SNDP. Let Π^* be an instance of NC-SNDP, and let OPT denote its optimal value. We use k_f to denote an integer multiple of 4 such that $r_{max} \leq 1.5k_f$. We repeatedly apply the algorithm of Theorem 1 (for subset k-connectivity) for $k = 1, \ldots, 7$, and derived instances $\Pi(1), \ldots, \Pi(7)$ to obtain solution subgraphs $H(1), \ldots, H(7)$. Then we repeatedly apply the algorithm of Theorem 8 (for subset $[k,\ 1.5k]$-connectivity) for $k = 8, 12, 16, 24, \ldots, k_f$ and derived instances $\Pi'(8), \Pi'(12), \ldots, \Pi'(k_f)$ to obtain solution subgraphs $H'(8), \ldots, H'(k_f)$. We start these iterations with $k = 8$, and we iterate until $k = k_f$; after each iteration, we update k to the largest integer multiple of 4 that is ≤ 1.5 times the previous k. The construction of the derived instances $\Pi(\rho)$ and $\Pi'(k)$ is described below.

Finally, we output the solution subgraph H^* for Π^*; H^* is the union of all the solution subgraphs $H(k)$, $k = 1, \ldots, 7$, and $H'(k)$, $k = 8, 12, \ldots, k_f$; we call these solution subgraphs the *constituent subgraphs* of H^*. Below, we prove that the cost of each of the constituent subgraphs is at most $O(1) \cdot \text{OPT}$. Clearly, the number of iterations is $O(\ln r_{max})$. Thus H^* is the union of $O(\ln r_{max})$ subgraphs such that each of these subgraphs has cost $O(1) \cdot \text{OPT}$, and so H^* has cost $O(\ln r_{max}) \cdot \text{OPT}$. Below, we prove that H^* satisfies the connectivity requirements, because for every pair of terminals i, j one of the constituent subgraphs of H^* has $\geq r_{i,j}$ openly disjoint i, j-paths.

We define the derived instances via a well-studied problem in network design, namely, the *generalized Steiner tree* problem, which is as follows: We are given a graph $G = (V, E)$, edge costs c, and $\hat{q}$ sets of terminal nodes $\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_{\hat{q}}$; the goal is to compute an (approximately) minimum-cost forest F of G such that each terminal set $\hat{D}_m, m = 1, \ldots, \hat{q}$, is contained in a (connected) component of F. Goemans and Williamson [17], based on earlier work by Agrawal, Klein, and Ravi [1], gave 2-approximation algorithms for this problem based on the primal-dual method.

Here is the construction for one of the derived instances $\Pi'(k)$; recall that this is an instance of the subset $[k, 1.5k]$-connectivity problem, where k is a fixed parameter. We start from Π^* and construct a requirements graph R with node set T and edge set $E(R)$ as follows. For each terminal pair i, j with $k \leq r_{i,j} \leq 1.5k$ (i.e., the requirement for the pair is within the valid range for our derived instance), we add the edge ij to R. Denote the node sets of the (connected) components of R by $\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_{\hat{q}}$. Next, we define an instance $\Pi(gst)$ of the *generalized Steiner tree* problem on the graph G with edge costs c (here, G, c are as in Π^*), and with terminal sets $\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_{\hat{q}}$. We solve this auxiliary problem $\Pi(gst)$ by applying the primal-dual algorithm of Goemans and Williamson [17]. Let $F \subseteq E(G)$ be the forest computed by the Goemans–Williamson algorithm, and let $F_1, F_2, \ldots, F_q$ denote the partition of F into connected components. Let the set of terminals in the component of F_m be denoted by D_m, $m = 1, \ldots, q$; thus each set D_m is the union of one or more of the terminal sets $\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_{\hat{q}}$. For each $m = 1, \ldots, q$, we define an instance $\Pi'_m(k)$ of the subset $[k,\ 1.5k]$-connectivity problem as follows: The graph G and the edge costs c are as in Π^*, the set of terminal nodes is D_m, and the requirement r'_i of a terminal $i \in D_m$ is defined to be $\max(r_{i,j} : \{i, j\} \in E(R))$; clearly, $k \leq r'_i \leq 1.5k\ \forall i \in D_m$. We take the derived instance $\Pi'(k)$ to be the disjoint union of these instances $\Pi'_m(k)$, $m = 1, \ldots, q$; i.e., we assume that each instance $\Pi'_m(k)$ has its own copy of G and c. To solve $\Pi'(k)$, we take each $m = 1, \ldots, q$ and apply the algorithm in Theorem 8 separately to $\Pi'_m(k)$ to obtain a solution subgraph, call it $H'_m(k)$. (These instances $\Pi'_m(k)$ are pairwise disjoint, and we solve them separately, one by one.) Then we take the union of the subgraphs $H'_1(k), \ldots, H'_m(k)$ and call it $H'(k)$; this is the solution subgraph of $\Pi'(k)$. The cost of the subgraphs $H'_m(k)$, $m = 1, \ldots, q$, is analyzed below.

Our reasons for using the auxiliary problem $\Pi(gst)$ for defining the instance $\Pi'(k)$

may be seen from the following example. Suppose that k is large (say, $k = \sqrt{n}$) and the edges in $E(R)$ form a matching, say $\{\{s_1, t_1\}, \{s_2, t_2\}, \dots, \{s_{\hat{q}}, t_{\hat{q}}\}\}$, where $\hat{q} = \Theta(n)$. Moreover, suppose that G has a cut $\delta(S)$ such that each edge in this cut is expensive, some of the edges in $E(R)$ have both ends in S, and the remaining edges in $E(R)$ have both ends in $V - S$. Say that the optimal solution consists of two disjoint subgraphs, one contained in the subgraph induced by S, and the other contained in the subgraph induced by $V - S$. Then we cannot take $\Pi'(k)$ to be a single instance with terminal set $\{s_1, \dots, s_{\hat{q}}, t_1, \dots, t_{\hat{q}}\}$, because then every solution subgraph will have $\geq k$ edges from the expensive cut $\delta(S)$. Also, we cannot take $\Pi'(k)$ to consist of $\hat{q}$ separate subinstances with one subinstance for each connected component of $R = (T, E(R))$, because the optimal values of these subinstances may sum to $\hat{q} \cdot \text{OPT}$, and the solution subgraph computed by our algorithm may have cost as high as this (assuming that the algorithm returns the union of the solution subgraphs of these $\hat{q}$ subinstances). We get around this difficulty by using the Goemans–Williamson algorithm to merge the connected components of $R = (T, E(R))$ into appropriate "clusters," and then we construct a separate subinstance for each of these "clusters" (these are the subinstances that we called $\Pi'_1(k), \dots, \Pi'_q(k)$). The key point is that (i) these subinstances have pairwise-disjoint terminal sets $D_1, \dots, D_q$, and hence the sum of the $\sigma()$ lower bounds (used in Theorem 8), namely, $\sum_{m=1}^{q} \sigma(D_m)$, is $\leq$ the $\sigma()$ lower bound of Π^*; and (ii) the following proof (which is based on the 2-approximation guarantee of Goemans and Williamson) shows that the sum of the *mst*() lower bounds for these subinstances, namely, $\sum_{m=1}^{q} \mathit{mst}(D_m)$, is $\leq O(1)$ times the *mst*() lower bound of Π^*. Also, for each subinstance, the solution subgraph has cost within an $O(1)$ factor of the sum of its $\sigma()$ and *mst*() lower bounds. Hence, the union of the solution subgraphs of these subinstances has cost within an $O(1)$ factor of the optimal value of Π^*.

The construction of the instances $\Pi(\rho)$, $\rho = 1, \dots, 7$, is similar to that of the instances $\Pi'(k)$. We start with $R = (T, E(R))$, where $E(R)$ consists of terminal pairs $\{i, j\}$ with $r_{i,j} = \rho$. Then we obtain a family of pairwise disjoint subinstances $\Pi_1(\rho), \Pi_2(\rho), \dots$ and these subinstances together form $\Pi(\rho)$.

Proof of Theorem 4. Recall that Π^* denotes the instance of NC-SNDP, OPT denotes the optimal value of Π^*, and H^* denotes the solution subgraph of Π^* found by our algorithm. The goal is to analyze the cost of the constituent subgraphs of H^* and show that each has cost $\leq O(1) \cdot \text{OPT}$, and then to show that H^* satisfies the connectivity requirements. The proof is based on the following LP relaxation P^* of Π^*, which interprets each requirement $r_{i,j}$ as a requirement for $r_{i,j}$ edge-disjoint i, j paths. Thus the optimal value of P^* gives a lower bound on OPT. The linear program has a variable x_e, $0 \leq x_e \leq 1$, for each edge $e \in E$; the intention is that each feasible solution H of Π^* gives a zero-one vector $x \in \Re^E$ that satisfies two conditions: $x_e = 1$ iff $e \in H$, and x satisfies the constraints of the LP relaxation (though feasible zero-one solutions of the linear program may not give feasible solutions of Π^*).

$$P^*: \qquad z^* = \min \sum_{e \in E} c_e x_e$$

$$\text{subject to}$$

$$x(\delta(S)) \geq \max\{r_{i,j} \ : \ i \in S, j \notin S\} \quad \forall S \subseteq V,$$

$$x_e \geq 0 \quad \forall e \in E.$$

Focus on one of the derived instances $\Pi'(k)$ and its associated generalized Steiner tree instance $\Pi(gst)$. We use the notation from the construction of $\Pi'(k)$ given above.

Goemans and Williamson [17] proved that the cost of the forest computed by their algorithm is ≤ 2 times the optimal value $z(gst)$ of the following LP relaxation $P(gst)$ of $\Pi(gst)$. The linear program has a variable x_e, $0 \leq x_e \leq 1$, for each edge $e \in E$; the intention is that each feasible solution F of $\Pi(gst)$ corresponds to a zero-one vector $x \in \Re^E$ that satisfies two conditions: $x_e = 1$ iff $e \in F$, and x satisfies the constraints of the linear program.

$$\begin{aligned} P(gst): \qquad\qquad z(gst) &= \min \sum_{e \in E} c_e x_e \\ \text{subject to} \qquad\qquad & \\ x(\delta(S)) &\geq 1 \quad \forall S \subseteq V : \exists m = 1, \ldots, \hat{q} : \emptyset \neq S \cap \hat{D}_m \neq \hat{D}_m, \\ x_e &\geq 0 \quad \forall e \in E. \end{aligned}$$

A key observation is that $k \cdot z(gst) \leq \text{OPT}$. To see this, note that multiplying the right-hand side of any constraint of the linear program $P(gst)$ by k gives a constraint that is valid for the LP P^*. (This follows because whenever we have a constraint $x(\delta(S)) \geq 1$ in the LP $P(gst)$, then the node set S separates two terminals v, w such that the requirements graph R has an v, w-path consisting of terminal-pairs $\{i, j\}$ such that $r_{i,j} \geq k$; since the v, w-path of R "crosses" S, one of the terminal-pairs $\{i, j\}$ in the v, w-path "crosses" S; therefore, $\max\{r_{i,j} \,:\, i \in S, j \notin S\} \geq k$, and hence the constraint "$x(\delta(S)) \geq k$" is a valid constraint for the LP P^*.) Consequently, for every feasible solution x^* of the LP P^*, we see that $\frac{1}{k}x^*$ is a feasible solution of the LP $P(gst)$. Moreover, if x^* is an optimal solution of the LP P^*, then we have $z(gst) \leq \frac{1}{k}c(x^*) = \frac{1}{k}z^* \leq \frac{1}{k}\text{OPT}$, or equivalently, $k \cdot z(gst) \leq \text{OPT}$.

Focus on the cost of the solution subgraph $H'(k) = H'_1(k) \cup H'_2(k) \cup \cdots \cup H'_q(k)$, and note that for each $m = 1, \ldots, q$ the cost of $H'_m(k)$ is $O(k) \cdot mst(D_m) + O(1) \cdot \sigma(D_m)$ (by Theorem 8), where D_m denotes the terminal set of $H'_m(k)$. Then the cost of $H'(k)$ is

$$\begin{aligned} O(k) \cdot \sum_{m=1}^{q} mst(D_m) \; &+ O(1) \cdot \sum_{m=1}^{q} \sigma(D_m) \\ &\leq O(k) \cdot \sum_{m=1}^{q} c(F_m) + O(1) \cdot \sigma(T) \\ &\qquad (\text{since } mst(D_m) \leq 2c(F_m) \ \forall m = 1, \ldots, q) \\ &\leq O(k) \cdot c(F) + O(1) \cdot \sigma(T) \\ &\leq O(1) \cdot \text{OPT} + O(1) \cdot \sigma(T) \\ &\qquad (\text{since } c(F) \leq 2z(gst) \text{ and } z(gst) \leq \text{OPT}/k) \\ &\leq O(1) \cdot \text{OPT}. \end{aligned}$$

A similar analysis for the solution subgraphs $H(1), \ldots, H(7)$ shows that each has cost $\leq O(1) \cdot \text{OPT}$.

Thus our claim for the cost of the solution subgraph H^* follows: $c(H^*) = O(\ln r_{max}) \cdot \text{OPT}$.

Finally, let us verify that H^* satisfies the connectivity requirements. Consider any pair of terminals i, j and their requirement $r_{i,j}$. Assume that $r_{i,j} \geq 8$ (otherwise, we are done by a similar but simpler analysis). Focus on an iteration of the algorithm that fixes the parameter k such that $k \leq r_{i,j} \leq 1.5k$. In that iteration, the requirements graph R has the edge $\{i, j\}$, and hence both i and j must be contained in one of the terminal sets $D_1, \ldots, D_q$, say, D_1. Now, consider the subinstance $\Pi'_1(k)$ and its solution subgraph $H'_1(k)$, and note that $H'_1(k)$ must have $\geq r_{i,j}$ openly disjoint i, j-paths because both r'_i and r'_j are $\geq r_{i,j}$ (here, r'_i and r'_j denote the requirements of i and j in $\Pi'_1(k)$). Thus, H^* has $\geq r_{i,j}$ openly disjoint i, j-paths.

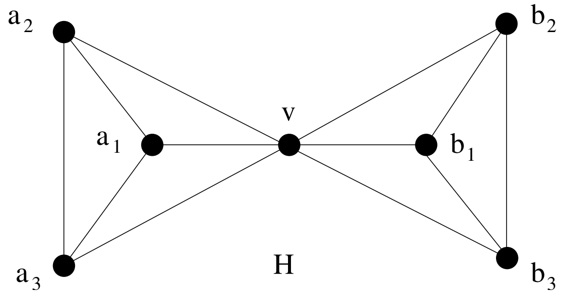


FIG. 4. *A metric-cost* 3*-edge connected graph that is strictly cheaper than any* 3*-node connected (spanning) graph. The edges in* H *have cost* 1*, and the edges in* $E(K_n) - E(H)$ *have cost* 2*.*

This completes the proof of Theorem 4. □

Appendix A. Examples illustrating claims in section 1. This appendix has details pertaining to Corollary 2 and the remarks following it (in section 1). In particular, we include a proof of the claim on 2-connected graphs with metric costs, and give examples to show that this claim does not apply to k-connected graphs for $k \geq 3$. Also, we give examples showing that for metric costs a k-connected graph may be a factor of $\Theta(k)$ times more expensive than a k-edge-connected *multi*-graph. The next result is well known, but we include a proof for the reader's convenience.

PROPOSITION 11. *In a metric graph, a minimum-cost* 2*-edge connected spanning subgraph has the same cost as a minimum-cost* 2*-node connected spanning subgraph.*

Proof. Take a counterexample such that the minimum-cost 2-edge connected spanning subgraph H contains as few cut nodes as possible. Clearly H contains at least one cut node v. Let W_1 and W_2 be connected components in $H - \{v\}$. Clearly, v lies on a cycle C_1 in $W_1 \cup \{v\}$ and a cycle C_2 in $W_2 \cup \{v\}$. Let w_1 and w_2 be neighbors of v on C_1 and C_2, respectively. Now, split off the edge-pair vw_1, vw_2; that is, add the edge w_1w_2 and remove the edges vw_1 and vw_2. This creates a cycle C on the node set $V(C_1) \cup V(C_2)$. Thus the resulting graph stays 2-edge connected. Note that the number of components in $H - \{v\}$ decreases by one. We repeat this step until $H - \{v\}$ is connected. By the triangle inequality, the cost of the subgraph does not increase. This contradicts our original choice of H. □

For $k \geq 3$, however, there exist k-edge connected spanning subgraphs of K_n that have lower cost than that of a minimum-cost k-node connected spanning subgraph. To see this let H be the union of two $k+1$ cliques that share exactly one node v. Let the nodes of these cliques be labeled $a_1, a_2, \ldots, a_k, v$ and $b_1, b_2, \ldots, b_k, v$, respectively. Next consider the complete graph K_n on $2k+1$ nodes whose edges costs are given by the shortest-path distances induced by H. That is, every edge in H has cost 1, and every edge in $E(K_n) - E(H)$ has cost exactly 2. Since H itself is k-edge connected we see that K_n contains a k-edge connected spanning subgraph of cost $2\binom{k+1}{2} = k^2 + k$. Now, any k-node connected spanning subgraph of K_n contains at least $\frac{1}{2}(2k+1)k = k^2 + \frac{1}{2}k$ edges. Moreover, there must be at least $k-1$ edges of cost 2 between nodes in $a_1, a_2, \ldots, a_k$ and nodes in $b_1, b_2, \ldots, b_k$; otherwise we obtain a node-cut containing less than k nodes. So any k-node connected spanning subgraph of K_n has cost at least $k^2 + \frac{1}{2}k + (k-1)$. This is strictly greater than the cost of the k-edge connected graph H if $k \geq 3$. The case of $k = 3$ is shown in Figure 4.

Clearly, if the edge costs do not satisfy the triangle inequality, then the minimum cost of a k-node connected spanning subgraph of K_n cannot be bounded in terms of the cost of a k-edge connected spanning subgraph. To see this take any k-edge

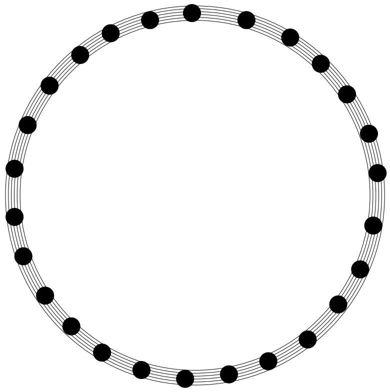


FIG. 5. *A metric-cost k-edge connected multigraph that is a factor of $\Theta(k)$ cheaper than any k-node connected spanning subgraph. The edge costs are given by the shortest-path distances in the cycle.*

connected graph H that is not also k-node connected (e.g., see Figure 4 for $k = 3$). Let every edge in H have cost 1 and every edge in $E(K_n) - E(H)$ have cost L. Since any k-node connected spanning subgraph of K_n has cost $\geq L$, the claim follows by the choice of L.

Corollary 2 and the other results do not extend to multigraphs. To see this, let k be an even number, $n - 1 \geq k \geq 2$, and let H be obtained from a cycle on n nodes by taking $\frac{1}{2}k$ copies of each edge. See Figure 5. If each edge in H has cost 1, then a minimum-cost k-edge connected multigraph has cost $\frac{1}{2}nk$. Let the cost of the other edges of K_n be given by the shortest-path distances in H. Each node has at least k different neighbors in a k-node connected spanning subgraph, and so the cost of the edges incident to any node is $\geq 2\sum_{i=1}^{\frac{k}{2}} i = k(\frac{k}{2} + 1)$. Hence, the minimum cost of a k-node connected spanning subgraph is $\geq \frac{1}{4}nk^2$. This is a factor of $\Theta(k)$ times the cost of the k-edge connected graph H.

Appendix B. Table of notation and symbols for section 4.

Node set	$V \quad (\lvert V\rvert = n)$
Set of terminal nodes	$T \quad (\lvert T\rvert = n')$
Set of active terminal nodes	$T^* \quad (\lvert T^*\rvert = n^*)$
Terminal nodes (usually active)	h, i, j
Inactive terminal nodes	u, v, w
Arbitrary nodes (terminals/nonterminals)	x, y
Requirement of terminal i	r_i
Requirement of terminal pair i, j	$r_{i,j}$
Connectivity parameter	$k \quad (k = 0 \pmod 4)$ in section 4)
Edge incident to nodes x, y	xy
Cost of edge xy	c_{xy} or $c(x, y)$
Set of r_i nearest neighbors of i	Γ_i
Total cost of edges from i to nodes in Γ_i	σ_i
Average cost of an edge from i to nodes in Γ_i	μ_i
Normalized cost of edge ix	$\widetilde{c}(i, x) := \max(c_{ix},\ \mu_i)$ (or $\widetilde{c}_{ix}$)
Parameters of algorithm in section 4	$\alpha, \beta, \gamma \quad (\alpha = 4, \beta = 2)$
Number of tracks	$\ell \quad (\ell = 3k/4$ in section 4)
Set of ℓ nearest neighbors of i (excluding i)	B_i
Tracks	$Q_0, Q_1, Q_2, \ldots, Q_\ell$
Index of current track	τ
Inner neighbors of active terminal i	$i_1, i_2, \ldots, i_\ell$
Parent of inactive terminal v	$p(v)$
Ordered set of nodes attaching terminal i to tracks	N_i
Cost of MST of subgraph induced by node set X	mst(X)

Acknowledgments. We thank Bill Cunningham, Michel Goemans, Balaji Raghavachari, R. Ravi, and Santosh Vempala for useful discussions over the years. We also thank the referees for their comments.

REFERENCES

[1] A. Agrawal, P. Klein, and R. Ravi, *When trees collide: An approximation algorithm for the generalized Steiner problem on networks*, SIAM J. Comput., 24 (1995), pp. 440–456.

[2] S. Arora, *Polynomial-time approximation schemes for Euclidean TSP and other geometric problems*, J. ACM, 45 (1998), pp. 753–782.

[3] M. Bern and P. Plassmann, *The Steiner problem with edge lengths* 1 *and* 2, Inform. Process. Lett., 32 (1989), pp. 171–176.

[4] D. Bienstock, E. F. Brickell, and C. L. Monma, *On the structure of minimum-weight k-connected spanning networks*, SIAM J. Discrete Math., 3 (1990), pp. 320–329.

[5] J. Cheriyan, T. Jordán, and Z. Nutov, *On rooted node-connectivity problems*, Algorithmica, 30 (2001), pp. 353–375.

[6] J. Cheriyan, S. Vempala, and A. Vetta, *Approximation algorithms for minimum-cost k-vertex connected subgraphs*, in Proceedings of the 34th ACM Symposium on Theory of Computing, Montréal, QC, ACM, New York, 2002, pp. 306–312.

[7] J. Cheriyan, S. Vempala, and A. Vetta, *An approximation algorithm for the minimum-cost k-vertex connected subgraph*, SIAM J. Comput., 32 (2003), pp. 1050–1055.

[8] N. Christofides, *Worst Case Analysis of a New Heuristic for the Traveling Salesman Problem*, Report 388, Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh, PA, 1976.

[9] A. Czumaj and A. Lingas, *A polynomial time approximation scheme for Euclidean minimum cost k-connectivity*, in Proceedings of the 25th International Colloquium on Automata, Languages and Programming, Lecture Notes in Comput. Sci. 1443, Springer, New York, 1998, pp. 682–694.

[10] A. Czumaj, A. Lingas, and H. Zhao, *Polynomial-time approximation schemes for the Euclidean survivable network design problem*, Proceedings of the 29th International Colloquium on Automata, Languages and Programming, Lecture Notes in Comput. Sci. 2380, Springer, New York, 2002, pp. 973–984.

[11] A. Czumaj, M. Grigni, P. Sissokho, and H. Zhao, *Approximation schemes for minimum 2-edge-connected and biconnected subgraphs in planar graphs*, in Proceedings of the 15th ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 2004, pp. 489–498.

[12] G. B. Dantzig, L. R. Ford, and D. R. Fulkerson, *Solution of a large-scale traveling-salesman problem*, Oper. Res., 2 (1954), pp. 393–410.

[13] A. Frank, *Connectivity augmentation problems in network design*, in Mathematical Programming: State of the Art 1994, J. R. Birge and K. G. Murty, eds., The University of Michigan, Ann Arbor, MI, 1994, pp. 34–63.

[14] G. L. Frederickson and J. Ja'Ja', *On the relationship between the biconnectivity augmentation and traveling salesman problems*, Theoret. Comput. Sci., 19 (1982), pp. 189–201.

[15] M. Goemans and D. J. Bertsimas, *Survivable networks, linear programming relaxations and the parsimonious property*, Math. Programming, 60 (1993), pp. 145–166.

[16] M. X. Goemans, A. V. Goldberg, S. Plotkin, D. B. Shmoys, É. Tardos, and D. P. Williamson, *Improved approximation algorithms for network design problems*, in Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, Arlington, VA, SIAM, Philadelphia, 1994, pp. 223–232.

[17] M. X. Goemans and D. P. Williamson, *A general approximation technique for constrained forest problems*, SIAM J. Comput., 24 (1995), pp. 296–317.

[18] K. Jain, *A factor* 2 *approximation algorithm for the generalized Steiner network problem*, Combinatorica, 21 (2001), pp. 39–60.

[19] S. Khuller, *Approximation algorithms for finding highly connected subgraphs*, in Approximation Algorithms for NP-Hard Problems, D. S. Hochbaum, ed., PWS Publishing, Boston, 1996, pp. 236–265.

[20] S. Khuller and B. Raghavachari, *Improved approximation algorithms for uniform connectivity problems*, J. Algorithms, 21 (1996), pp. 434–450.

[21] G. Kortsarz, R. Krauthgamer, and J. R. Lee, *Hardness of approximation for vertex-connectivity network design problems*, SIAM J. Comput., 33 (2004), pp. 704–720.

[22] G. Kortsarz and Z. Nutov, *Approximating k-node connected subgraphs via critical graphs*, SIAM J. Comput., 35 (2005), pp. 247–257.

[23] G. Kortsarz and Z. Nutov, *Approximating node connectivity problems via set covers*, Algorithmica, 37 (2003), pp. 75–92.

[24] L. Lovász, *Combinatorial Problems and Exercises*, North–Holland, Amsterdam, and Akadémiai Kiadó, Budapest, 1979.

[25] J. S. B. Mitchell, *Guillotine subdivisions approximate polygonal subdivisions: A simple polynomial-time approximation scheme for geometric TSP, k-MST, and related problems*, SIAM J. Comput., 28 (1999), pp. 1298–1309.

[26] C. L. Monma, B. S. Munson, and W. R. Pulleyblank, *Minimum-weight two-connected spanning networks*, Math. Programming, 46 (1990), pp. 153–171.

[27] C. L. Monma and D. F. Shallcross, *Methods for designing communication networks with certain two-connectivity survivability constraints*, Oper. Res., 37 (1989), pp. 531–541.

[28] M. Stoer, *Design of Survivable Networks*, Lecture Notes in Math. 1531, Springer-Verlag, Berlin, 1992.

[29] L. Trevisan, *When Hamming meets Euclid: The approximability of geometric TSP and Steiner tree*, SIAM J. Comput., 30 (2000), pp. 475–485.

[30] V. V. Vazirani, *Approximation Algorithms*, Springer-Verlag, Berlin, 2001.

[31] D. Williamson, M. Goemans, M. Mihail, and V. Vazirani, *A primal-dual approximation algorithm for generalized Steiner network problems*, Combinatorica, 15 (1995), pp. 435–454.

SIAM J. DISCRETE MATH.
Vol. 21, No. 3, pp. 637–646

LOCATING SERVERS FOR RELIABILITY AND AFFINE EMBEDDINGS[*]

KENNETH A. BERMAN[†]

Abstract. Consider the problem of locating servers in a network for the purpose of storing data, performing an application, etc., so that at least one server will be available to clients even if up to k component failures occur throughout the network. Letting $G = (V, E)$ be the graph with vertex set V and edge set E representing the topology of the network, and letting $L \subseteq V$ be a set of potential locations for the servers, a fundamental problem is to determine a minimum-size set $S \subseteq L$ such that the network remains connected to S even if up to k component failures occur throughout the network. We say that such a set S is *k-fault-tolerant*. In this paper we present an algebraic characterization of k-fault-tolerant sets in terms of affine embeddings of G in k-dimensional Euclidean space. Employing this characterization, we present a polynomial-time Monte Carlo algorithm for computing a minimum-size k-fault-tolerant subset S of L. In fact, we solve the following more general problem for directed networks: given a digraph $G = (V, E)$ (an undirected graph is equivalent to a symmetric digraph) and a subset $L \subseteq V$, we find a k-fault-tolerant subset S of L having minimum cost, where a unary integer cost $c(v)$ is associated with locating a server at vertex $v \in V$.

Key words. graph algorithms, server location, fault-tolerance, convex embeddings, affine embeddings, randomized algorithms, parallel algorithms

AMS subject classifications. 05C85, 05C50, 05C10, 68M10

DOI. 10.1137/060670158

1. Introduction.

One method used to increase reliability of access to information, data base objects, applications, and so forth, in a network such as the Internet is to replicate them on multiple servers at different locations in the network. A fundamental problem is to locate servers in the network so that every client is guaranteed to have access to at least one such server even if up to k component failures occur in the network. In this paper we consider the problem of finding a minimum-size set with this property.

Let $G = (V, E)$ be the digraph with vertex set V and edge set E representing the topology of the network. Note that an undirected network is represented by a symmetric digraph; i.e., each undirected edge $\{u, v\}$ is replaced by two directed edges uv and vu. Let $L \subseteq V$ be a set of potential server locations. We say that a set of vertices $S \subseteq L$ is *k-fault-tolerant* if the network remains connected to at least one vertex $s \in S$; i.e., there is a directed path from each vertex $v \in V - S$ to s, even after the deletion of any k vertices and their incident edges (where we allow the deletion of vertices from S). If the network is undirected, i.e., if G is a symmetric digraph, then connectivity is two-way; i.e., there exists a directed from v to s if and only if there exists a directed path from s to v. Applications of k-fault-tolerant sets in nonsymmetric digraphs occur in networks such as wireless sensor networks, where data from sensors needs to reach at least one sink or egress vertex in the network. A k-fault-tolerant set has the property that data from every sensor that has not failed

[*] Received by the editors September 18, 2006; accepted for publication (in revised form) February 6, 2007; published electronically July 25, 2007. This research was supported in part by NSF grant CCR-9877139 and by an Ohio Board of Regents' Research Investment grant.
http://www.siam.org/journals/sidma/21-3/67015.html

[†] Department of Computer Science, Mail Location 30, 813 Rhodes Hall, University of Cincinnati, Cincinnati, OH 45221 (ken.berman@uc.edu).

can reach, via a multihop path, at least one sink vertex, even if up to k sensors fail throughout the network.

The following proposition is easily proved using Menger's theorem (see [4]).

PROPOSITION 1. *Let $G = (V, E)$ be a digraph and S a subset of vertices (of size at least $k + 1$). Then the following statements are equivalent:*

(i) *S is a k-fault-tolerant set.*

(ii) *There exist $k+1$ directed paths from every vertex $v \in V - S$ to $k+1$ (distinct) vertices in S such that any two paths have only the vertex v in common.*

(iii) *For any set $T \subseteq V$ of size $k+1$ there exists a set $\mathcal{P}$ of $k+1$ pairwise vertex-disjoint directed paths from T to S, i.e., having initial vertex in T and terminal vertex in S.*

In statement (iii) of Proposition 1 we do not make the assumption that S and T are disjoint. However, we do allow paths in $\mathcal{P}$ to be trivial, i.e., consist of a single vertex.

In this paper we consider the problem of finding a minimum-size k-fault-tolerant subset of L. In fact, we consider the following somewhat more general problem, where we associate an integer cost $c(v)$ with locating a server at vertex $v \in V$.

k-FAULT-TOLERANT SERVER LOCATION PROBLEM. *Let $G = (V, E)$ be a digraph with vertex set V and edge set E. Given a positive integer cost $c(v)$ associated with locating a server at each vertex $v \in V$, a set $L \subseteq V$ of potential locations for the servers, and a specified positive integer k, find a k-fault-tolerant subset S of L such that the cost of S, given by $c(S) = \sum_{s \in S} c(s)$, is minimized (or determine that no such set S exists).*

Note that the k-fault-tolerant set problem is no less general if it is stated with no restriction on the server placement, i.e., with $L = V$. To ensure that the server placement is restricted to a given set L, we can simply add a sufficiently large cost to each vertex not in L. The latter observation is true even for unary costs. Also, note that when $L = V$ a minimum-cost k-fault-tolerant set always exists, since V is k-fault-tolerant.

Given two sets of vertices L and M, Bar-Ilan and Peleg [1] introduced the concept of a *k-tolerant L-set for M* in an undirected graph, i.e., a set $S \subseteq L$ such that there exist k vertex-disjoint paths from every vertex v in $M - S$ to (k distinct vertices in) S. They showed that this problem is NP-hard and presented an approximation algorithm with ratio $k(\log n + 1)$ (also see [2]). Note that a k-fault-tolerant subset of L is equivalent to a $(k + 1)$-tolerant L-set for V. As far as the author knows, no deterministic polynomial-time algorithm has been given for finding a minimum-size $(k+1)$-tolerant L-set for V, or equivalently a minimum-size k-fault-tolerant subset of L, even for undirected graphs.

In this paper we present a polynomial-time Monte-Carlo algorithm for solving the k-fault-tolerant server location problem when the costs are all unary integers. Further, we show that a parallelized version of the algorithm is in RNC^2 . The problem of finding a minimum-size k-fault-tolerant set corresponds to the special case where all vertices have unit cost. Our algorithm is based on the following characterization of k-fault-tolerant sets. This characterization involves an "affine" embedding of the digraph G in k-dimensional Euclidean space.

The *out-neighborhood N_v of a vertex* $v \in V$ is the set of all the vertices $x \in V$ such that $vx \in E$. Let d_v denote the *out-degree* of v, i.e., the cardinality of N_v. An *embedding* of G in Euclidean k-dimensional space $\mathbf{R}^k$ is a mapping f from V to $\mathbf{R}^k$; i.e, f maps each $v \in V$ onto the point $(f_1(v), \ldots, f_k(v))$. Such an embedding is said to be in *general position* if no $k+1$ points are on a hyperplane. Throughout this paper we will assume that the embedding f is in general position. For $U \subseteq V$, let $f(U) = \{f(u) | u \in U\}$. The *affine hull* of a set of points $P = \{P_1, \ldots, P_p\}$ in $\mathbf{R}^k$ is the set of all points $\sum_{i=1}^{p} \lambda_i P_i$, where $\lambda_1, \ldots, \lambda_p$ are real numbers such that $\sum_{i=1}^{p} \lambda_i = 1$.

Given a set $S \subseteq V$, we say that f is an *affine* S-embedding of G if every vertex u not belonging to S lies in the affine hull of its neighbors, i.e., for each $u \in V - S$, there exists a weighting α of E over the real numbers such that $\sum_{v \in N_u} \alpha(uv) = 1$ and $f(u) = \sum_{v \in N_u} \alpha(uv) f(v)$. We will refer to such a weighting α of the edges as a *k-dimensional affine weighting*. Note that if f is in general position and the out-degree of each vertex $u \in V - S$ is at least $k+1$, then the affine hull of $f(N_u)$ is the entire space $\mathbf{R}^k$, so that f is necessarily an affine S-embedding.

A *branching rooted at S* is a acyclic subdigraph B having out-degree one at every vertex not in S and out-degree zero at every vertex in S. Equivalently, letting $S = \{s_0, s_1, \ldots, s_{p-1}\}$, B is a spanning forest consisting of p trees, where the ith tree is directed into root vertex $s_i, i = 0, \ldots, p-1$. Let $\mathcal{B}_S$ denote the set of all branchings B rooted at S. The weight $\alpha(B)$ of a branching with respect to an edge weighting α is the product of the weights on the edges of B. We define the α-weighted *branching number $\beta_S(\alpha)$ for S* to be the sum of the weights over the set $\mathcal{B}_S$ of all branchings B rooted at S, i.e.,

$$(1) \qquad \beta_S(\alpha) = \sum_{B \in \mathcal{B}_S} \alpha(B).$$

In the next section we show that S is a k-fault-tolerant set if and only if $\beta_S(\alpha)$ is nonzero for some k-dimensional affine weighting α.

Given a set of vertices $S = \{s_0, \ldots, s_k\}$, a *convex S-embedding* is an embedding f such that, for each $u \in V - S$, $f(u)$ is in the convex hull of $f(N_u)$, $f(s_0)$ is the zero vector, and $f(s_i)$ is the vector with a "1" in position i and zeros in all other positions (see [10, 7]). Clearly, a convex S-embedding f in general position is also an affine S-embedding in general position. Further, it has an associated affine weighting α such that $\alpha(e) \geq 0$ for every edge e with tail in $V - S$. Clearly all branchings rooted at S have nonnegative weight with respect to α. Further, there exists a branching B having strictly positive weight, i.e., all of whose edges have strictly positive weight. To show that such a branching exists, remove all edges having zero weight. Clearly, f is also a convex embedding of the resulting digraph G'. Further, a branching in G' rooted at S determines a branching of G rooted at S having strictly positive weight. To show that G' contains such a branching, it is sufficient to show that there exists a path P in G' from any vertex $v \in V - S$ to some vertex $s \in S$. Letting $P = u_1 \ldots u_j$, where $u_1 = v$ and $u_j = s$, we inductively choose u_{i+1} to be a vertex in the out-neighborhood of u_i that is on the boundary of the convex hull and has the largest first component of all the points on the boundary. Since the embedding is in general position it follows that $f_1(u_i) < f_1(u_{i+1})$. Thus, since the graph is finite, we must eventually choose a vertex in S. We have just shown that G contains a branching rooted at S having strictly positive weight with respect to α. Hence, $\beta_S(\alpha) > 0$. By the theorem we prove in the next section, this implies that S is a k-fault-tolerant set of G, a result given by Linial, Lovász, and Widgerson in [10] for undirected graphs and generalized by Cheriyan and Rief in [7] to directed graphs.

In the third section we employ the characterization result of section 2 to obtain a parallel Monte Carlo algorithm for solving the k-tolerant server problem (with $B = V$) for unary integer weights.

2. Characterization theorem.

THEOREM 1. *Let G be a digraph and S be a subset of vertices of G. Then, S is a k-fault-tolerant set if and only if there exists an affine S-embedding of G in general position in $\mathbf{R}^k$ with an associated affine weighting α of the edges such that the α-weighted branching number for S is nonzero, i.e., $\beta_S(\alpha) \neq 0$.*

We prove the theorem with the aid of a lemma. Before stating the lemma, we establish some definitions. The *volume* of a set of $k+1$ points $P = \{P_0, P_1, \ldots, P_k\}$, denoted by $vol(P)$, is the determinant of the $k \times k$ matrix whose ith row is $P_i - P_0, i = 1, \ldots, k$. Note that, up to a change in sign, $vol(P)$ is invariant under permutations of the elements of P and equals the volume of the parallelepiped containing the edges $P_0P_i, i = 1, \ldots, k$.

We extend the definition of the α-weighted branching number β_S for S to the α-weighted branching number $\beta_{S,T}$ for pairs of vertex sets S and T, where S and T have the same cardinality, as follows. Let $S = \{s_0, \ldots, s_{p-1}\}$ and $T = \{t_0, \ldots, t_{p-1}\}$ be two subsets of V of size p (not necessarily disjoint), where $s_0 < s_1 < \cdots < s_{p-1}$ and $t_0 < t_1 < \cdots < t_{p-1}$. A branching B with root set S has *coroot set* T if for each vertex $t_i \in T$ the unique path from t_i to S in B does not intersect the path from any other vertex of T to S, i.e., each of the p trees of B contains exactly one vertex of T. Equivalently, for some permutation π of $\{0, \ldots, p-1\}$, there exists a path in B from t_i to $s_{\pi(i)}, i = 0, \ldots, p-1$. We define the *sign* of B, denoted $sign(B)$, to be the sign of the permutation π, and define $\beta_{S,T}(\alpha)$ to be the signed sum of the weight of a branching over the set $\mathcal{B}_{S,T}$ of all branchings having root set S and coroot set T, i.e.,

$$\beta_{S,T}(\alpha) = \sum_{B \in \mathcal{B}_{S,T}} sign(B)\alpha(B). \tag{2}$$

Note that $\beta_S(\alpha) = \beta_{S,S}(\alpha)$.

LEMMA 1. *Let f be an affine embedding of a digraph G in general position in $\mathbf{R}^k$, and let α be an associated affine weighting of the edge set E. Let S and T be any vertex sets of size $k+1$, and let $\beta_S(\alpha)$ and $\beta_{S,T}(\alpha)$ be defined by* (1) *and* (2), *respectively. If $\beta_S(\alpha)$ is nonzero, then*

$$\frac{\beta_{S,T}(\alpha)}{\beta_S(\alpha)} = \frac{\pm vol(f(T))}{vol(f(S))}. \tag{3}$$

Proof. We utilize the following result known as the all-minors matrix-tree theorem (see [6]). Given a matrix M whose rows and columns are indexed by V and given $X, Y \subseteq V$, let $M[X : Y]$ denote the submatrix with rows indexed by X and columns indexed by Y. Given $u, v \in V$, let $M(u, v)$ denote the entry in row u and column v; i.e., $M(u, v) = M[\{u\} : \{v\}]$. Given a weighting w of the edges with real numbers, the *(generalized) Kirchhoff* matrix is the $n \times n$ matrix K_w whose rows and columns are both indexed by V such that

$$K_w(u, v) = \begin{cases} -w(u, v) & \text{if } u \neq v, \\ \\ \sum_{x \in V} w(u, x) & \text{otherwise,} \end{cases} \tag{4}$$

where $w(u, v) = w(uv)$ if $uv \in E$, and $w(u, v) = 0$ otherwise.

All-minors matrix-tree theorem. Given two vertex sets S and T of the same cardinality,

$$\beta_{S,T}(w) = \pm \det K_w[V - S : V - T]. \tag{5}$$

A formula for the actual sign on the right-hand side of the above equality is given in [6], but we omit it here since it does not affect our arguments. As a special case of the all-minors matrix-tree theorem we have that

$$\beta_S(w) = \det K_w[V - S : V - S]. \tag{6}$$

Now let $K = K_\alpha$, and let K_S be the matrix obtained from K by replacing each diagonal entry indexed by S with a 1 and every other entry in a row indexed by S with a 0; i.e., for all $s \in S, u \in V - S, v \in V$, $K_S(s,s) = 1, K_S(s,u) = 0$, and $K_S(u,v) = K(u,v)$. Then we have that

$$\det K_S = \det K[V - S, V - S] = \beta_S(\alpha). \tag{7}$$

Since by hypothesis $\beta_S(\alpha) \neq 0$, it follows that K_S is invertible. Let h be the embedding of G in $\mathbf{R}^{k+1}$ defined by

$$h(v) = \left(f_1(v), \ldots, f_k(v), 1 - \sum_{j=1}^{k} f_j(v) \right),$$

$v \in V$. Let H be the $n \times (k+1)$ matrix with rows indexed by V and columns indexed by $\{1, \ldots, k+1\}$ such that the row vector indexed by $v \in V$ is $h(v)$. Then it follows directly from the definitions of K and H that

$$K[V - S : V]H = 0. \tag{8}$$

Let H_S and H_T denote the submatrices of H consisting of all the rows indexed by vertices in S and T, respectively. It is easily verified that

$$vol(f(S)) = \pm\det(H_S), \qquad vol(f(T)) = \pm\det(H_T). \tag{9}$$

Let $Z = (z_{ij})$ denote the $n \times (k+1)$ matrix whose rows corresponding to S determine the identity matrix and whose remaining $n-k-1$ rows are all zeros. Then it follows immediately from the definitions of the matrices involved and from (8) that

$$K_S H = Z H_S. \tag{10}$$

Letting J denote the adjoint of K_S, i.e., $J = adj K_S$, and using the fact that $K_S^{-1} = adj K_S / \det K_S$, we have

$$H = K_S^{-1} Z H_S = J Z H_S / \det K_S = J[V : S] H_S / \det K_S. \tag{11}$$

Thus, by deleting all rows of on both sides of (11) not belonging to T, we have

$$H_T = J[T : S] H_S / \det K_S. \tag{12}$$

Taking the determinant of both sides yields

$$\det H_T = \det J[T : S] \det H_S / (\det K_S)^{k+1}. \tag{13}$$

But by Jacobi's theorem (see [12]), we have

$$\det J[T : S] = \det K_S[V - S : V - T](\det K_S)^k. \tag{14}$$

Substituting (14) into (13), we obtain

$$\det H_T = \det K_S[V-S:V-T]\det H_S/\det K_S. \tag{15}$$

Observing that $K_S[V-S:V-T] = K[V-S:V-T]$ and applying (5) and (7), we have that

$$\det H_T = \pm\beta_{S,T}(\alpha)\det H_S/\beta_S(\alpha). \tag{16}$$

Equation (3) of Lemma 1 is obtained by substituting (9) into (16). □

We are now ready to prove Theorem 1. Consider any set of vertices $S = \{s_0, \ldots, s_{p-1}\}$ such that $\beta_S(\alpha) \neq 0$. First suppose $p = k+1$. Now let T be any set of $k+1$ vertices (possibly having intersection with S). Since f is an embedding in general position, $vol(S)$ and $vol(T)$ are both nonzero. Therefore, by Lemma 1, $\beta_{S,T}(\alpha)$ is nonzero. However, this implies that there exists at least one branching with root set S and coroot set T. Thus, there exists a set of $k+1$ vertex-disjoint paths with initial vertex set T and terminal vertex set S (we allow some of the paths to consist of a single vertex). Since T was chosen to be an arbitrary set of $k+1$ vertices it follows from Proposition 1 that S is a k-fault-tolerant set.

Now suppose that S is a set of cardinality $p > k+1$. Construct the graph $G' = (V', E')$ from G by removing all edges having both tail and head in S, adding $k+1$ new vertices $s'_0, \ldots, s'_k$, and for every pair of vertices s_i and $s'_j, i \in \{0, \ldots, p-1\}, j \in \{0, \ldots, k\}$, adding an edge directed from s_i to s'_j. Let f' be an embedding of G' in $\mathbf{R}^k$ such that $f'(v) = f(v), v \in V$, and $f'(s'_0), \ldots, f'(s'_k)$ do not lie on a hyperplane, and let α' be an associated affine weighting of G' such that $\alpha'(e) = \alpha(e), e \in E$. Since each vertex of S has out-degree $k+1$ in G', such a weighting α' exists. It follows easily from the definition of the branching function that

$$\beta_{S'}(\alpha') = \left(\prod_{i=0}^{p}\sum_{j=0}^{k}\alpha'(s_i s'_j)\right)\beta_S(\alpha) = \left(\prod_{i=0}^{p} 1\right)\beta_S(\alpha) = \beta_S(\alpha).$$

By hypothesis, $\beta_S(\alpha) \neq 0$. Thus, $\beta_{S'}(\alpha') \neq 0$. Since $|S'| = k+1$, which is the case $p = k+1$, it follows from the above argument that S' is k-fault-tolerant. This implies that S is k-fault-tolerant.

Conversely, suppose S is a k-fault-tolerant set. Then S' is a k-fault-tolerant set in G'. Thus, by a result in [7], there exists a convex S-embedding f' of G' in k-dimensional space in general position; i.e., there exists a weighting α' of the edges of G' such that, for each vertex $u \in V'$, $f'(u) = \sum_{uv \in E'} \alpha'(uv) f'(v)$, where $\alpha'(uv) > 0$ for each $uv \in E'$. Let α be the affine weighting of G such that $\alpha(uv) = \alpha'(uv)$, for every edge $uv \in E$, not directed out of a vertex in S. Since S is k-fault-tolerant there exists at least one branching B rooted at S. Further, since $\alpha(uv)$ is nonnegative for each edge uv not directed out of S, it follows that $\alpha(B) > 0$ for every branching rooted at S. Hence, $\beta_S(\alpha) > 0$. □

3. Monte Carlo algorithm. For convenience let $V = \{1, \ldots, n\}$, and let $i_1, \ldots, i_{d_i}$ denote the vertices in the out-neighborhood N_i of $i \in V$. Associate the indeterminant f_{ij} with the jth component of $f(i)$, i.e., $f_j(i) = f_{ij}, i \in V, j = 1, \ldots, k$, and associate the variable a_{ij} with the affine weight on each edge $ii_j \in E$, i.e., $\alpha(ii_j) = a_{ij}, i \in V, j = 1, \ldots, d_i$. Note that if the out-degree d_i of vertex i is less than $k+1$, then i must necessarily be contained in every k-fault-tolerant set S. Also note that setting $\alpha(ij) = a_{ij} = 0$ will not affect the value of $\beta_S(\alpha)$ for S a set containing i, and will result in $\beta_S(\alpha) = 0$ for every set S that does not contain i. Now assume that $d_i \geq k+1$. It follows from the definition of α that

$$\sum_{j=1}^{k+1} a_{ij} f_{i_j l} = f_{il} - \sum_{j=k+2}^{d_i} a_{ij} f_{i_j l}, \quad l = 1, \ldots, k,$$

$$\sum_{j=1}^{k+1} a_{ij} = 1 - \sum_{j=k+2}^{d_i} a_{ij}.$$

We express these equations using matrices. Letting M_i be the $(k+1) \times (k+1)$ matrix whose ljth entry is $f_{i_j l}, j = 1, \ldots, k+1, l = 1, \ldots, k$, and whose last row consists entirely of 1's; $\mathbf{a}$ be the column vector whose jth entry is $a_{ij}, j = 1, \ldots, k+1$; and $\mathbf{b}$ be the column vector whose lth entry is $f_{il} - \sum_{j=k+2}^{d_i} a_{ij} f_{i_j l}, l = 1, \ldots, k$, and whose last entry is $1 - \sum_{j=k+2}^{d_i} a_{ij}$, we have

$$M_i \mathbf{a} = \mathbf{b}. \tag{17}$$

Letting $M_i^{(j)}$ be the matrix obtained from M_i by replacing the jth column of M_i with $\mathbf{b}$, $j = 1, \ldots, k+1$, and employing Cramer's rule, we obtain

$$a_{ij} = \det M_i^{(j)} / \det M_i, \quad j = 1, \ldots, k+1. \tag{18}$$

Note that $\det M_i$ is a polynomial of degree k in the indeterminates $f_{ii_1}, \ldots, f_{ii_k}$, and $\det M_i^{(j)}$ is a polynomial of degree $k+1$ in the indeterminates $a_{ii_{k+2}}, \ldots, a_{ii_{d_i}}$, $f_{ii_1}, \ldots, f_{ii_k}$.

Let A denote the set of indeterminates $a_{ij}, i = 0, \ldots, n-1, j = k+2, \ldots, d_i$, and let F denote the set of indeterminates $f_{ij}, i = 0, \ldots, n-1, j = 1, \ldots, k$. Based on the above discussion, we define the *indeterminate affine weighting* $\alpha_{F,A}$ of E as follows: for each edge $v_i v_{i_j} \in E$,

$$\alpha_{F,A}(v_i v_{i_j}) = \begin{cases} 0 & \text{if } d_i \leq k, \\ a_{ij} & \text{if } j \in \{k+2, \ldots, d_i\}, \\ \det M_i^{(j)} / \det M_i & \text{otherwise.} \end{cases} \tag{19}$$

THEOREM 2. *Let G be a digraph with vertex set V, and let $\alpha = \alpha_{F,A}$ be the indeterminate affine weighting. Then S is a k-fault-tolerant set if and only if $\beta_S(\alpha)$ is not identically equal to zero.*

Proof. $\beta_S(\alpha_{F,A})$ is not identically equal to zero if and only if $\beta_S(\alpha_{F_0,A_0})$ is nonzero for some choice of F_0 and A_0. However, we have just shown that there exists such a weighting if and only if S is k-fault-tolerant. □

We define the *scaled indeterminate affine weighting* $\sigma_{F,A}$ of E as follows: for each $v_i v_j \in E$,

$$\sigma_{F,A}(v_i v_{i_j}) = \begin{cases} 0 & \text{if } d_i \leq k, \\ a_{ij} \det M_i & \text{if } j \in \{k+2, \ldots, d_i\}, \\ \det M_i^{(j)} & \text{otherwise.} \end{cases} \tag{20}$$

It is easily verified that

$$\beta_S(\sigma_{F,A}) = \left(\prod_{v_i \in V-S} detM_i \right) \beta_S(\alpha_{F,A}).$$

With this observation the following result is an immediate corollary of Theorem 2.

COROLLARY 1. *Let G be a digraph with vertex set V, and let $\sigma = \sigma_{F,A}$ be the scaled indeterminate affine weighting. Then, S is a k-fault-tolerant set if and only if $\beta_S(\sigma)$ is not identically equal to zero.*

Now take a random embedding f of G; i.e., f_{ij} is assigned a random value $f_{ij}^{(0)}$ chosen independently and uniformly from $[1,\dots,p], i = 0,\dots,n-1, j = 1,\dots,k$. Further, assign a_{ij} a random value $a_{ij}^{(0)}$ chosen independently and uniformly from $[1,\dots,p], v_iv_j \in E, j = k+2,\dots,d_i$. Letting $F_0 = (f_{ij}^{(0)})$ and $A_0 = (a_{ij}^{(0)})$, we will refer to the resultant edge weighting $\sigma = \sigma_{F_0,A_0}$ as a *random scaled k-dimensional affine weighting.*

THEOREM 3. *Let σ be a random scaled k-dimensional affine weighting as defined above. If $S \in \mathcal{S}$ is not a k-fault-tolerant set, then $\beta_{\mathcal{S}}(\sigma)$ equals 0. Otherwise, $\beta_S(\sigma)$ is nonzero with probability at least $1 - n(k+1)/p$.*

Proof. If S is not a k-fault-tolerant set, then it follows immediately from Corollary 1 that $\beta_S(c) = 0$. Now suppose that S is a k-fault-tolerant set. Then, by Corollary 1, $\beta_S(\sigma_{F,A})$ is not identically equal to zero. Note that $\beta_S(\sigma_{F,A})$ is a polynomial of degree at most $(n-|S|)(k+1) \leq n(k+1)$ in indeterminates $f_{ij}, i = 1,\dots,n, j = 1,\dots,k$, and $a_{ij}, i = 1,\dots,n, j = k+2,\dots,d_i$, where $n = |V|$. It follows from the Swartz–Zippel theorem (see [11]) that the probability that $\beta_S(\sigma_{F_0,A_0})$ is zero is less than $n(k+1)/p$. □

We now design a Monte Carlo algorithm for finding a minimum-cost k-fault-tolerant set S. Let X be the $n \times n$ diagonal matrix whose ith diagonal entry is $x_i, i = 1,\dots,n$. Then, applying (6), we have

$$\det(X + K_\sigma) = \sum_{S \subseteq V} \det K_\sigma[V-S : V-S] \prod_{v_i \in S} x_i = \sum_{S \subseteq V} \beta_S(\sigma) \prod_{v_i \in S} x_i. \tag{21}$$

Given a cost weighting c of the vertex set, i.e., vertex v_i is assigned the positive integer weight $c_i = c(v_i)$, let X_c be the diagonal matrix obtained from X by setting $x_i = x^{c_i}, i = 0,\dots,n-1$. Then,

$$\det(X_c + K_\sigma) = \sum_{S \subseteq V} \beta_S(\sigma) x^{c(S)}. \tag{22}$$

The determinant $\det(X_c + K_\sigma)$ can be computed in time $O(nM(n)|c|log|c|)$, where $M(n)$ denotes the time needed to multiply two $n \times n$ matrices, i.e., $M(n) \in O(n^{2.376})$ (see [8]) and $|c| = \sum_{i=0}^{n-1} c_i$. In parallel on an EREW PRAM (see [4, 9]), $\det(X_c + K_\sigma)$ can be computed in time $O(\log|c| \log^2 n)$ using $O(|c|nM(n))$ processors. For references on parallel computation of determinants over rings such as the ring of polynomials, see [3, 5].

By (22), the minimum power μ of x having nonzero coefficient in the polynomial $\det(X_c + K_\sigma)$ is the minimum cost of a set $S \subseteq V$ such that $\beta_S(\sigma)$ is nonzero. However, by Theorem 3 such a set S is k-fault-tolerant with probability at least $1 - n(k+2)/p$. Thus, to find, with high probability, a k-fault-tolerant set of minimum cost we need to find a set S corresponding to a minimum power μ of x in $\det(X_c + K_\sigma)$. To do this, we use a technique introduced by Mulmuley, Vazirani, and Vazirani in [13] (also see [14]), which is based on applying the following result, known as the isolating lemma.

LEMMA 2 (isolating lemma). *Let $\mathcal{S}$ be a family of subsets of a set $V = \{v_0, \ldots, v_{n-1}\}$. Choose integers $r_0, r_1, \ldots, r_{n-1}$ randomly and independently from $[1, \ldots, q]$, where $q > n$. Then, with high probability, i.e., probability at least $1 - n/q$, there is a unique minimum-weight set $S \in \mathcal{S}$.*

Using the isolating lemma, we can ensure that there exists with high probability a unique minimum-cost k-fault-tolerant set S by replacing the cost c_i of placing a server at vertex v_i with the cost c'_i, given by $c'_i = rc_i + r_i, i = 1, \ldots, n$, where $r = 1 + r_0 + \cdots + r_{n-1}$. Note that a k-fault-tolerant set S has minimum cost with respect to c' if and only if it has minimum cost with respect to c.

Based on the above discussion, we have the following polynomial-time Monte Carlo algorithm for the k-fault-tolerant server location problem for unary integer weights. In the algorithm we assume that the set L of potential server locations equals the entire vertex set V. As pointed out earlier, the k-fault-tolerant set problem is no less general if it is stated with the restriction that $L = V$, since we can ensure that the server placement is restricted to a given set L by simply adding a sufficiently large cost to each vertex not in L. The probability of correctness of the algorithm is at least $(1 - (k+1)n/p)(1 - n/q)$; e.g., for $p \geq 3(k+1)n$ and $q \geq 4n$, the probability of correctness is at least 1/2.

MONTE CARLO ALGORITHM FOR THE k-TOLERANT SERVER LOCATION PROBLEM.

Input: a digraph $G = (V, E), n = |V|$,
a unary cost weighting c of V,
a positive integer k.

Output: a minimum-cost k-fault-tolerant set S.
The probability of correctness is at least $(1 - (k+1)n/p)(1 - n/q)$.

Step 1. Compute a random scaled k-dimensional affine weighting $\sigma = \sigma_{F_0, A_0}$, where the values in $F_0 = (f_{ij}^{(0)})$ and $A_0 = (a_{ij}^{(0)})$ are chosen independently and uniformly from $[1, \ldots, p]$.

Step 2. Compute the random vertex weighing r, where r_i is chosen randomly and independently from $[1, \ldots, q], i = 1, \ldots, n$, and the vertex weighting c' given by $c'_i = rc_i + r_i, i = 1, \ldots, n$, where $r = 1 + r_0 + \cdots + r_{n-1}$.

Step 3. Compute $\det(X_{c'} + K_\sigma)$ and let μ be the smallest power of x having nonzero coefficient.

Step 4. For each j, compute the weighting $c^{(j)}$ obtained from c' by decreasing the weight of vertex v_j by 1; i.e., $c^{(j)}(v_i) = c'_i$ for $i \neq j$, and $c^{(j)}(v_j) = c'_j - 1$. Then, compute $\det(X_{c^{(j)}} + K_\sigma)$ and let μ_j be the smallest power of x having nonzero coefficient. OUTPUT the set S consisting of all vertices v_j such that $\mu_j < \mu$.

It follows from the previous discussion that the complexity of the algorithm is $O(n^2 M(n)|c'| \log |c'|) = O(n^3 M(n) q|c| \log(nq|c|))$. Further, a parallel implementation of the algorithm on the EREW PRAM using $O(|c'|n^2 M(n)) = O(|c|n^3 q)$ processors has parallel complexity $O(|c'| \log^2 n) = O(\log |c| \log^2 n)$.

Acknowledgments. I would like to thank Fred Annexstein, Chad Yoshikawa, and the referees for many helpful comments.

REFERENCES

[1] J. BAR-ILAN AND D. PELEG, *Approximation Algorithms for Selecting Network Centers*, Lecture Notes in Comput. Sci. 519, Springer, New York, 1991, pp. 343–354.

[2] J. BAR-ILAN, G. KORTSARZ, AND D. PELEG, *How to allocate network centers*, J. Algorithms, 15 (1993), pp. 385–412.

[3] S. J. Berkowitz, *On computing the determinant in small parallel time using a small number of processors*, Inform. Process. Lett., 12 (1984), pp. 147–150.

[4] K. A. Berman and J. L. Paul, *Algorithms: Sequential, Parallel and Distributed*, Thomson Course Technology, Boston, 2005.

[5] A. Borodin, S. A. Cook, and N. Pippinger, *Parallel computation for well-endowed rings and space bounded probabilistic machines*, Inform. and Control, 58 (1983), pp. 113–136.

[6] S. Chaiken, *A combinatorial proof of the all minors matrix tree theorem*, SIAM J. Alg. Disc. Methods, 3 (1982), pp. 319–329 (now SIMAX).

[7] J. Cheriyan and J. H. Reif, *Directed* s-t *numberings, rubber band, and testing digraph k-vertex connectivity*, Combinatorica, 14 (1994), pp. 435–451.

[8] D. Coppersmith and S. Winograd, *Matrix multiplication via arithmetic progressions*, J. Symbolic Comput., 9 (1990), pp. 23–52.

[9] R. M. Karp and V. Ramachandran, *A survey of parallel algorithms for shared-memory machines*, in Handbook of Theoretical Computer Science, MIT Press/Elsevier, New York, 1990, pp. 869–941.

[10] N. Linial, L. Lovász, and A. Wigderson, *Rubber bands, convex embeddings and graph connectivity*, Combinatorica, 8 (1988), pp. 91–102.

[11] M. R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, Cambridge, UK, 1995.

[12] T. Muir, *A Treatise on the Theory of Determinants*, Dover, New York, 1960.

[13] K. Mulmuley, U. V. Vazirani, and V. V. Vazirani, *Matching is as easy as matrix inversion*, Combinatorica, 7 (1987), pp. 105–114.

[14] H. Narayanan, H. Saran, and V. V. Vazirani, *Randomized parallel algorithms for matroid union and intersection, with applications to arboresences and edge-disjoint spanning trees*, SIAM J. Comput., 23 (1994), pp. 387–397.

SIAM J. DISCRETE MATH.
Vol. 21, No. 3, pp. 647–661

DISCRETE LINES AND WANDERING PATHS*

A. VINCE [†]

Abstract. The problem of finding an approximation to a geometric line by a discrete line using pixels is ubiquitous in computer graphics applications. We show that this discrete line problem in $\mathbb{R}^{n+1}$, for grids of any shape, is equivalent to a geometry problem in $\mathbb{R}^n$ concerning the minimization of the distance that a certain type of closed polygonal path wanders from the origin. This geometry problem is solved completely in dimension 1 (corresponding to 2-dimensional grids), and two simple and efficient algorithms provide near optimum solutions in higher dimensions.

AMS subject classifications. 68R05, 52C07, 65D18, 68U05

Key words. discrete line, polygonal path

DOI. 10.1137/050642009

1. Introduction.

This paper concerns a geometry problem in n-dimensional Euclidean space motivated by the drawing of a discrete line with pixels. The generation of such line segment raster images is ubiquitous in computer graphics applications, the first such algorithm due to Bresenham [1, 5]. In 2001, Bresenham wrote:

> I was working in the computation lab at IBM's San Jose development lab. A Calcomp plotter had been attached to an IBM 1401 via the 1407 typewriter console. [The algorithm] was in production use by summer 1962, possibly a month or so earlier. Programs in those days were freely exchanged among corporations so Calcomp (Jim Newland and Calvin Hefte) had copies. When I returned to Stanford in Fall 1962, I put a copy in the Stanford comp center library. A description of the line drawing routine was accepted for presentation at the 1963 ACM national convention in Denver, Colorado. It was a year in which no proceedings were published, only the agenda of speakers and topics in an issue of *Communications of the ACM.* A person from the IBM Systems Journal asked me after I made my presentation if they could publish the paper. I happily agreed, and they printed it in 1965.

The Bresenham algorithm was designed for rectangular grids in the plane. More recent applications in visualization of 3-dimensional medical image data and in global image processing have led to an interest in nonrectangular grids, for example the hexagonal grid, and in higher-dimensional grids. That is the motivation for this paper. There has also been an interest in issues not directly addressed in this paper, for example efficient implementation of algorithms [2, 3], discrete approximation of curves [6], and alternate approaches to constructing discrete lines [7].

The discrete line problem. Given two points $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^{n+1}$, the discrete line problem is to find a discrete line, in terms of cells (pixels), that is in some sense the best approximation to the Euclidean line $\overline{\mathbf{ab}}$. To precisely formulate the problem, let the points of a lattice L represent the "centers" of the cells in our $(n+1)$-dimensional

*Received by the editors October 5, 2005; accepted for publication (in revised form) February 5, 2007; published electronically August 1, 2007. This work was partially supported by PYXIS Innovation, Kingston, ON, Canada.

http://www.siam.org/journals/sidma/21-3/64200.html

[†] Department of Mathematics, University of Florida, P.O. Box 118105, Gainesville, FL 32611-8105 (vince@math.ufl.edu).

grid. By a *lattice* in $\mathbb{R}^{n+1}$ we mean the set of all integer linear combinations of $n+1$ linear independent vectors. The cells are the Voronoi cells of the lattice, the *Voronoi cell* at lattice point $\mathbf{x}$ being the set of points at least as close to $\mathbf{x}$ as to any other lattice point in L. Each Voronoi cell is a polytope P, and the grid is obtained by translation of P by the lattice L.

Two lattice points will be considered *neighbors* if their respective Voronoi cells share a common facet. Given lattice points $\mathbf{a}$ and $\mathbf{b}$, define a *discrete line* joining $\mathbf{a}$ and $\mathbf{b}$ as a sequence $\mathbf{a} = \mathbf{u}_1, \dots, \mathbf{u}_N = \mathbf{b}$ of lattice points (cells) such that $\mathbf{u}_i$ and $\mathbf{u}_{i+1}$ are neighbors for $i = 1, \dots, N-1$. This is a reasonable definition, especially in situations in which the cells can be viewed at variable resolutions—multiscale. This is the point of view taken in [4]. For a discrete line to be a "good approximation" to the geometric line $\overline{\mathbf{ab}}$, the discrete line should be as "short" as possible and as "close" as possible to the geometric line. More precisely, we impose the following requirements:

A. the length N should be minimum, and

B. of all such discrete lines $\mathbf{a} = \mathbf{u}_1, \dots, \mathbf{u}_N = \mathbf{b}$ of minimum length, the points $\mathbf{u}_k$ should be chosen so as to minimize

$$\max_{1 \le k \le N} d(\mathbf{u}_k, \overline{\mathbf{ab}}),$$

where $d(\mathbf{u}_k, \overline{\mathbf{ab}})$ is the orthogonal distance from $\mathbf{u}_k$ to $\overline{\mathbf{ab}}$.

The above optimization problem will be referred to as the *discrete line problem.*

In section 2 a geometry problem is posed concerning minimization of the distance that a certain type of closed polygonal path wanders from the origin. This geometry problem in $\mathbb{R}^n$ is shown to be equivalent to the discrete line problem in $\mathbb{R}^{n+1}$. The remainder of the paper concerns this "wandering path problem." After definitions and preliminary results in section 3, sections 4 and 6 contain two simple and efficient algorithms whose output is close to an optimum solution of the wandering path problem. An optimum solution is found for the dimension 1 case in section 5, which implies a complete solution to the discrete line problem for any grid in dimension 2. Theorems 10, 11, and 12 provide upper bounds on the output of Algorithms 1, 1.1, and 2, respectively.

2. The wandering path problem. A geometry problem in $\mathbb{R}^n$ will be posed which is equivalent to the discrete line problem in $\mathbb{R}^{n+1}$. Consider any set V of vectors in $\mathbb{R}^n$. A V-*multiset* is a finite ordered multiset $W = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_N)$ of elements from V. If we set

$$\mathbf{u}_k := \mathbf{w}_1 + \mathbf{w}_2 + \cdots + \mathbf{w}_k,$$

$1 \le k \le N$, then joining points $\mathbf{0} = \mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_N$ successively by line segments results in a polygonal path $P = (\mathbf{0} = \mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_N)$ in $\mathbb{R}^n$ called a V-*path* of *length* N. If $\mathbf{u}_N = \mathbf{0}$, then the V-path is called a *closed* V-path. Define

$$w(P) := \max_{1 \le k \le N} |\mathbf{u}_k| .$$

Then $w(P)$ is the furthest that path P wanders from the origin.

The case of interest for our application is where V is a set of exactly $n+1$ vectors in $\mathbb{R}^n$ satisfying the following two properties:

1. each subset of n vectors in V is linearly independent, and
2. there exists a closed V-path.

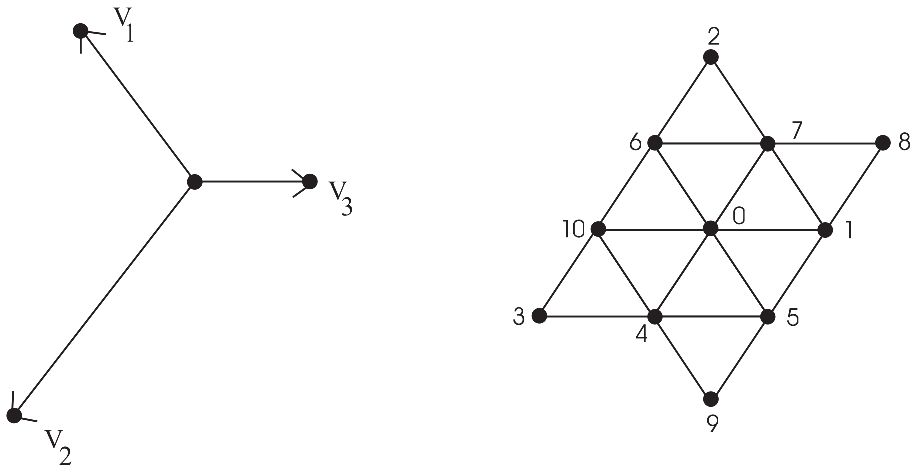


FIG. 1. *An optimum wandering path.*

Note that condition 2 is equivalent to the existence of a set $\{m_{\mathbf{v}} \mid \mathbf{v} \in V\}$ of positive integers such that

$$\sum_{\mathbf{v}\in V} m_{\mathbf{v}}\mathbf{v} = \mathbf{0}.$$

A set V satisfying properties 1 and 2 will be called a *basic set.* This paper concerns minimizing $w(P)$ over all closed V-paths P. In other words, we seek a closed V-path that stays as close as possible to the origin. Define

$$w(V) := \min\{w(P) \mid P \text{ is a closed } V\text{-path}\}.$$

Call $w(V)$ the *optimum wandering distance* for V. A closed path that realizes this distance will be called an *optimum wandering path.* The problem of finding the optimum wandering distance and optimum wandering path will be referred to as the *wandering path problem.*

A 1-dimensional example. Let $V = \{-4, 6\}$. For the closed V-path $P = (0,\ -4,\ +2,\ -2,\ 4,\ 0)$, we have $w(P) = 4$. In fact this is an optimum wandering path for V, so $w(V) = 4$. A complete solution to the wandering path problem for the 1-dimensional case appears in section 5. As will be shown below, this implies a complete solution to the discrete line problem in two dimensions.

A 2-dimensional example. Let $V = \{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2\}$, where

$$\begin{aligned}\mathbf{v}_0 &= (1,\ 0),\\ \mathbf{v}_1 &= (-1,\ \sqrt{3}),\\ \mathbf{v}_2 &= (-\tfrac{3}{2},\ -\tfrac{3}{2}\sqrt{3}).\end{aligned}$$

The optimum wandering path $(\mathbf{0} = \mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{11} = \mathbf{0})$ is shown in Figure 1, where the labels indicate the indices. The vectors of V are successively added in the order $(\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_0, \mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_0, \mathbf{v}_0, \mathbf{v}_2, \mathbf{v}_1, \mathbf{v}_0)$. The optimum wandering distance is $w(V) = \sqrt{3}$.

Relation between the discrete line and the optimum wandering path problems. Recall that the discrete line problem is to find a sequence $\mathbf{a} = \mathbf{u}'_1, \ldots, \mathbf{u}'_N$

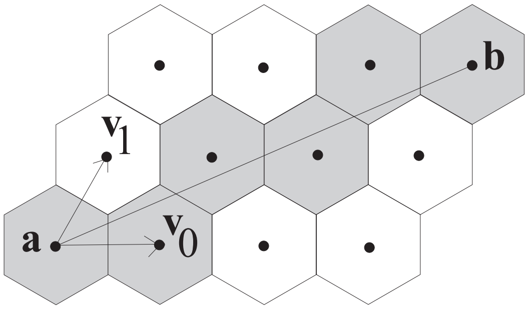


Fig. 2. *A discrete line that approximates* $\overline{\mathbf{ab}}$.

$= \mathbf{b}$ of points (cells) in an $(n+1)$-dimensional lattice L such that $\mathbf{u}'_i$ and $\mathbf{u}'_{i+1}$ are neighbors for $i = 1, \dots, N-1$ and that satisfies conditions (A) and (B) from the Introduction. Because the cells are the Voronoi cells of L, there exists a set Y of vectors that generates L such that $Y = -Y := \{-\mathbf{y} \mid \mathbf{y} \in Y\}$ and such that two lattice points $\mathbf{x}_1, \mathbf{x}_2$ are neighbors if and only if $\mathbf{x}_2 = \mathbf{x}_1 + \mathbf{y}$ for some $\mathbf{y} \in Y$. It also follows that there is a subset V' of Y consisting of $n+1$ *direction vectors* such that $\mathbf{b} - \mathbf{a}$ lies in the $(n+1)$-dimensional polyhedral cone C spanned by V'. In Figure 2 the lattice is the hexagonal lattice, the (Voronoi) cells regular hexagons. The figure shows a line $\overline{\mathbf{ab}}$ in $\mathbb{R}^2$ and its set $V' = \{\mathbf{v}_0, \mathbf{v}_1\}$ of direction vectors.

We will assume that $\mathbf{b} - \mathbf{a}$ lies in no cone spanned by a proper subset of V'; otherwise the problem reduces to the same problem in a lower dimension. So

$$\mathbf{b} - \mathbf{a} = \sum_{\mathbf{v}' \in V'} m_{\mathbf{v}'} \, \mathbf{v}',$$

where the $m_{\mathbf{v}'}$ are uniquely determined positive integers. Let $W' = (\mathbf{w}'_1, \mathbf{w}'_2, \dots, \mathbf{w}'_N)$ be any ordered multiset of elements from V' such that the vector $\mathbf{v}'$ appears exactly $m_{\mathbf{v}'}$ times in W'. Let

$$\mathbf{u}'_k = \mathbf{a} + \mathbf{w}'_1 + \mathbf{w}'_2 + \cdots + \mathbf{w}'_k.$$

Then the lattice points $\mathbf{a}$, $\mathbf{u}'_1$, $\mathbf{u}'_2$, $\dots, \mathbf{u}'_N = \mathbf{b}$ form a discrete line joining $\mathbf{a}$ and $\mathbf{b}$. Moreover, the number $N = \sum_{\mathbf{v}' \in V'} m_{\mathbf{v}'}$ is the length of a shortest discrete line joining $\mathbf{a}$ and $\mathbf{b}$.

To solve the discrete line problem it remains to satisfy condition (B). To find the orthogonal distance $d(\mathbf{u}'_k, \overline{\mathbf{ab}})$ from each of the points $\mathbf{u}'_k$ to the line $\overline{\mathbf{ab}}$, let H be the n-dimensional hyperplane orthogonal to vector $\mathbf{b} - \mathbf{a}$, and let $\texttt{proj}_H$ denote the orthogonal projection onto H. Further let $V = \{\texttt{proj}_H(\mathbf{v}') \mid \mathbf{v}' \in V'\}$. By the assumption that $\mathbf{b} - \mathbf{a}$ lies in no cone spanned by a proper subset of V', every set of n vectors from V is linearly independent. Moreover, defining $m_{\mathbf{v}} := m_{\mathbf{v}'}$ if $\mathbf{v} = \texttt{proj}_H(\mathbf{v}')$, note that

$$\sum_{\mathbf{v} \in V} m_{\mathbf{v}} = N \qquad \text{and} \qquad \sum_{\mathbf{v} \in V} m_{\mathbf{v}} \mathbf{v} = \mathbf{0}, \tag{1}$$

the latter because $\sum_{\mathbf{v}' \in V'} m_{\mathbf{v}'} \mathbf{v}' = \mathbf{b} - \mathbf{a}$ and H is orthogonal to $\mathbf{b} - \mathbf{a}$. Hence V is a basic set. Let $\mathbf{w}_i = \texttt{proj}_H(\mathbf{w}'_i)$ for each i and let

$$\mathbf{u}_k := \mathbf{w}_1 + \mathbf{w}_2 + \cdots + \mathbf{w}_k.$$

Note that $P := (\mathbf{0} = \mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N = \mathbf{0})$ is a closed V-path and that $d(\mathbf{u}'_k, \overline{\mathbf{ab}}) = \mathbf{u}_k$. Hence the problem of satisfying condition (B) is exactly the problem of minimizing

$$\max_{1 \le k \le N} |\mathbf{u}_k|$$

over all closed V-paths $P = (\mathbf{0} = \mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N = \mathbf{0})$ of length N. In other words, solving the discrete line problem in $\mathbb{R}^{n+1}$ reduces to solving the wandering path problem in $\mathbb{R}^n$. The requirement that the closed V-path P have the same length N as the discrete line is not a serious restriction, as explained in Remark 3 of section 4.

In the example of Figure 2 the basic set is $V = \{3c, -2c\}$, where $c = \sqrt{57}/38$. The solution to the wandering path problem for basic set V is $P = (0, -2c, c, -c, 2c, 0)$, obtained by successive additions $(-2c, +3c, -2c, 3c, -2c)$. The corresponding discrete line is

$$\begin{array}{l}
\mathbf{a} \\
\mathbf{a} + \mathbf{v}_0 \\
\mathbf{a} + \mathbf{v}_0 + \mathbf{v}_1 \\
\mathbf{a} + \mathbf{v}_0 + \mathbf{v}_1 + \mathbf{v}_0 \\
\mathbf{a} + \mathbf{v}_0 + \mathbf{v}_1 + \mathbf{v}_0 + \mathbf{v}_1 \\
\mathbf{a} + \mathbf{v}_0 + \mathbf{v}_1 + \mathbf{v}_0 + \mathbf{v}_1 + \mathbf{v}_0 = \mathbf{b},
\end{array}$$

indicated by the shaded hexagons in Figure 2.

3. The lattices L and Λ, multiplicity, and modulus. For a real number α, vector $\mathbf{y}$, and set X of vectors we use the notation $\alpha X = \{\alpha \mathbf{x} \mid \mathbf{x} \in X\}$ and $\mathbf{y} + X = \{\mathbf{y} + \mathbf{x} \mid \mathbf{x} \in X\}$. In the first lemma, the V-multiset $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)$ is used as an alternate way to denote the V-path $(\mathbf{0} = \mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N = \mathbf{0})$, where $\mathbf{u}_k = \mathbf{w}_1 + \mathbf{w}_2 + \cdots + \mathbf{w}_k$. The proof of Lemma 1 is clear.

LEMMA 1. *If V is a basic set and α is a positive real number, then $w(\alpha V) = \alpha\, w(V)$. Moreover, if V-multiset $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m)$ is an optimum wandering path for V, then $(\alpha\mathbf{w}_1, \alpha\mathbf{w}_2, \ldots, \alpha\mathbf{w}_m)$ is an optimum wandering path for αV.*

LEMMA 2. *Let V be a basic set such that $\sum_{\mathbf{v}\in V} m_{\mathbf{v}}\, \mathbf{v} = \mathbf{0}$, with the $m_{\mathbf{v}}$ relatively prime integers. Then $\sum_{\mathbf{v}\in V} m'_{\mathbf{v}}\, \mathbf{v} = \mathbf{0}$ if and only if there exists an integer c such that $m'_{\mathbf{v}} = c\, m_{\mathbf{v}}$ for all $\mathbf{v} \in V$.*

Proof. Letting $V = \{\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_n\}$, we have

$$\sum_{i=1}^{n} \frac{m_i}{m_0}\, \mathbf{v}_i = -\mathbf{v}_0 = \sum_{i=1}^{n} \frac{m'_i}{m'_0}\, \mathbf{v}_i,$$

which implies by the linear independence of $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ that $m_i/m_0 = m'_i/m'_0$ or $m'_i = (m'_0/m_0)m_i$ for all i. Since the m_i have no nontrivial common divisor, $c := m'_0/m_0$ is an integer. □

If V is a basic set, then, by definition, there exist positive integers $m_{\mathbf{v}}$ such that $\sum_{\mathbf{v}\in V} m_{\mathbf{v}}\mathbf{v} = \mathbf{0}$. In light of Lemma 2, call $m_{\mathbf{v}} := m_{\mathbf{v}}(V)$ the *multiplicity* of $\mathbf{v}$ in V if the $m_{\mathbf{v}}$ are relatively prime. Further denote

$$m := m(V) = \sum_{\mathbf{v}\in V} m_{\mathbf{v}},$$

where $m_{\mathbf{v}}$ is the multiplicity of $\mathbf{v}$ in V. Call $m := m(V)$ the *modulus* of V.

COROLLARY 3. *The length of any closed wandering V-path is divisible by the modulus $m(V)$.*

Proof. If P is a closed V-path of length N and $m_{\mathbf{v}}(P)$ is the number of occurrences of $\mathbf{v}$ in P, then $\sum_{\mathbf{v}\in V} m_{\mathbf{v}}(P)\mathbf{v} = \mathbf{0}$. By Lemma 2 there is a constant c such that $m_{\mathbf{v}}(P) = c\, m_{\mathbf{v}}(V)$. Therefore $N = \sum_{\mathbf{v}\in V} m_{\mathbf{v}}(P) = c \sum_{\mathbf{v}\in V} m_{\mathbf{v}}(V) = c\, m(V)$. $\square$

CONJECTURE 4. *The length of an optimum wandering path for a basic set V is equal to $m(V)$.*

Comments relevant to Conjecture 4 appear in Remark 3.

LEMMA 5. *If V is a basic set in $\mathbb{R}^n$, then the set of all integer linear combinations of elements from V is an n-dimensional lattice.*

Proof. Let L be the set of all integer linear combinations of elements from $V = \{\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_n\}$, and let $\{m_0, m_1, \dots, m_n\}$ be a set of integers such that $\sum_{i=0}^n m_i\, \mathbf{v}_i = \mathbf{0}$. Then by eliminating $\mathbf{v}_0$,

$$\sum_{i=0}^n a_i\, \mathbf{v}_i = \sum_{i=1}^n \frac{m_0 a_i - m_i a_0}{m_0}\, \mathbf{v}_i$$

for any integers a_i. Therefore $\mathbf{x} \in L$ if and only if $\mathbf{x} = \sum_{i=1}^n \beta_i \mathbf{v}_i$, where $\beta_i = b_i/m_0$ and $(b_1, \dots, b_n)$ is a multiple of $(m_1, m_2, \dots, m_n)$ modulo m_0. It readily follows that L is a sublattice of the lattice generated by the vectors $\frac{1}{m_0}\, \mathbf{v}_i$, $1 \le i \le n$. $\square$

The following is a converse of Lemma 5.

LEMMA 6. *Let V be any set of $n+1$ points of an n-dimensional lattice, every n of which are linearly independent. Then V is a basic set if and only if V is not contained in any closed half-space determined by a hyperplane through the origin.*

Proof. Clearly, if V is contained in some half-space, then $\sum_{i=0}^n m_i\, \mathbf{v}_i = \mathbf{0}$ is impossible for positive integers m_i. Conversely, assume that V is not a basic set. Since V is a dependent set of lattice points, $\sum_{i=0}^n a_i\, \mathbf{v}_i = \mathbf{0}$ for some integers a_i. It is not possible that all the a_i are positive, since there exists no closed V-path. Thus $\mathbf{v}_0 = \sum_{i=1}^n b_i\, \mathbf{v}_i$, where at least one of the b_i, say b_n without loss of generality, is positive. Let $\mathbf{x}$ be a vector orthogonal to $\mathbf{v}_1, \dots, \mathbf{v}_{n-1}$. Then $\langle \mathbf{x}, \mathbf{v}_i \rangle = 0$ for $i = 1, 2, \dots, n-1$ and $\langle \mathbf{x}, \mathbf{v}_0 \rangle = \sum_{i=1}^n b_i \langle \mathbf{x}, \mathbf{v}_i \rangle = b_n \langle \mathbf{x}, \mathbf{v}_n \rangle$. This shows that V is contained in the closed positive half-space determined by $\mathbf{x}$. $\square$

In light of Lemma 5, let $L := L(V)$ denote the lattice of all integer linear combinations of elements from V:

$$L(V) := \left\{ \sum_{\mathbf{v}\in V} a_{\mathbf{v}} \mathbf{v} \mid a_{\mathbf{v}} \in \mathbb{Z} \right\}.$$

The sublattice $\Lambda := \Lambda(V) \subset L(V)$ defined by

$$\Lambda(V) := \left\{ \sum_{\mathbf{v}\in V} a_{\mathbf{v}} \mathbf{v} \mid a_{\mathbf{v}} \in \mathbb{Z},\ \sum_{\mathbf{v}\in V} a_{\mathbf{v}} = 0 \right\}$$

also plays an important role in the wandering path problem. For any $\mathbf{v}_0 \in V$ the lattice $\Lambda(V)$ is generated by the set $\{\mathbf{v} - \mathbf{v}_0, \mid \mathbf{v} \in V \setminus \{\mathbf{v}_0\}\}$ of vectors. For $\mathbf{x}, \mathbf{y} \in L$ define

$$\mathbf{x} \equiv \mathbf{y} \pmod{\Lambda} \quad \text{if} \quad \mathbf{x} - \mathbf{y} \in \Lambda.$$

LEMMA 7. *If V is a basic set with modulus m, then*

$$\sum_{\mathbf{v}\in V} a_{\mathbf{v}} \mathbf{v} \equiv \mathbf{0}\ (mod\ \Lambda) \quad if\ and\ only\ if \quad \sum_{\mathbf{v}\in V} a_{\mathbf{v}} \equiv 0\ (mod\ m).$$

Proof. By definition, $\sum_{\mathbf{v}\in V} a_{\mathbf{v}}\mathbf{v} \equiv \mathbf{0}$ if and only if $\sum_{\mathbf{v}\in V} a_{\mathbf{v}}\mathbf{v} \in \Lambda$ if and only if there exist integers $b_{\mathbf{v}}$ such that $\sum_{\mathbf{v}\in V} a_{\mathbf{v}}\mathbf{v} = \sum_{\mathbf{v}\in V} b_{\mathbf{v}}\mathbf{v}$, where $\sum_{\mathbf{v}\in V} b_{\mathbf{v}} = 0$. This occurs if and only if $\sum_{\mathbf{v}\in V}(a_{\mathbf{v}}\mathbf{v} - b_{\mathbf{v}}\mathbf{v}) = \mathbf{0}$, which, by Lemma 2, occurs if and only if there is an integer c such that $a_{\mathbf{v}} - b_{\mathbf{v}} = c\,m_{\mathbf{v}}$ for all $\mathbf{v} \in V$. In one direction this implies that $\sum_{\mathbf{v}\in V} a_{\mathbf{v}} = \sum_{\mathbf{v}\in V} b_{\mathbf{v}} + c\sum_{\mathbf{v}\in V} m_{\mathbf{v}} = cm$. So $\sum_{\mathbf{v}\in V} a_{\mathbf{v}} \equiv 0 \pmod m$. In the other direction, if $\sum_{\mathbf{v}\in V} a_{\mathbf{v}} \equiv 0 \pmod m$, then take $b_{\mathbf{v}} = a_{\mathbf{v}} - c\,m_{\mathbf{v}}$, in which case $\sum_{\mathbf{v}\in V} b_{\mathbf{v}} = \sum_{\mathbf{v}\in V} a_{\mathbf{v}} - c\sum_{\mathbf{v}\in V} m_{\mathbf{v}} = 0$. □

COROLLARY 8. *If V is a basic set with modulus $m(V)$, then the order of the quotient group L/Λ is $m(V)$.*

4. First algorithm. The first result in this section is a lower bound on the optimum wandering distance. The first of two algorithms for the wandering path problem is then presented. The input is a basic set V, the output a closed V-path that is near optimum.

PROPOSITION 9. *If V is a basic set, then*

$$w(V) \geq \frac{1}{2}\max_{\mathbf{v}\in V}|\mathbf{v}|\,.$$

Proof. Let $\mathbf{v} \in V$ be an element that realizes the maximum in $\max_{\mathbf{v}\in V}|\mathbf{v}|$, and let $\mathbf{u}$ and $\mathbf{u}'$ be two consecutive points in an optimum wandering path such that $\mathbf{v} = \mathbf{u} - \mathbf{u}'$. By the triangle inequality $2\,w(V) \geq |\mathbf{u}| + |\mathbf{u}'| \geq |\mathbf{v}|$. □

Although this lower bound is somewhat trivial, it is, in a sense, best possible. For the following family in $\mathbb{R}^2$, for example, the bound is achieved: $V = \{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2\}$,

$$\begin{aligned} \mathbf{v}_0 &= (1, \beta), \\ \mathbf{v}_1 &= (1, -\beta), \\ \mathbf{v}_2 &= (-4k, 0), \end{aligned}$$

where $k \geq 2$ is an integer and $0 < \beta \leq \sqrt{3}$. In this case the optimum wandering path is

$$(\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_0, \mathbf{v}_1, \ldots \mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_0, \mathbf{v}_1, \ldots \mathbf{v}_0, \mathbf{v}_1),$$

where $\mathbf{v}_0, \mathbf{v}_1$ is repeated k time on each side of $\mathbf{v}_2$. Then

$$w(V) = w(P) = 2k = \frac{1}{2}|\mathbf{v}_2| = \frac{1}{2}\max_{\mathbf{v}\in V}|\mathbf{v}|.$$

In Algorithm 1 below, the notation $C'_{\mathbf{v}}$ stands for the polyhedral cone spanned by the vectors in $V \setminus \{\mathbf{v}\}$:

$$C'_{\mathbf{v}} := \left\{ \sum_{\mathbf{u}\in V\setminus\{\mathbf{v}\}} \alpha_{\mathbf{u}}\,\mathbf{u} \;\middle|\; \alpha_{\mathbf{u}} \geq 0 \right\}.$$

Let

$$\mathbf{x}_0 = \frac{1}{2}\sum_{\mathbf{v}\in V}\mathbf{v} \qquad \text{and} \qquad C_{\mathbf{v}} = -\mathbf{x}_0 + C'_{\mathbf{v}}.$$

Then $C_{\mathbf{v}}$ is a copy of $C'_{\mathbf{v}}$ translated by $-\mathbf{x}_0$. It follows from Lemma 6 that $\bigcup_{\mathbf{v}\in V} C'_{\mathbf{v}} = \mathbb{R}^n$, and therefore

$$\bigcup_{\mathbf{v}\in V} C_{\mathbf{v}} = \mathbb{R}^n. \tag{2}$$

Algorithm 1.

```
Input: A basic set $V$ in $\mathbb{R}^n$.
Output: A closed $V$-path $P$ in $\mathbb{R}^n$.
initialize: $i = 0$, $\mathbf{u}_0 = \mathbf{0}$
until $i = m(V)$ do
    find a $\mathbf{v} \in V$ such that $\mathbf{u}_i \in C_{\mathbf{v}}$
    $\mathbf{u}_{i+1} \leftarrow \mathbf{u}_i + \mathbf{v}$
    $i \leftarrow i + 1$
end
return: path $P = (\mathbf{0}, \mathbf{u}_1, \dots, \mathbf{u}_{m(V)} = \mathbf{0})$.
```

Theorem 10. *If V is a basic set with modulus $m(V)$, then Algorithm 1 finds a closed V-path P of length $m(V)$ with*

$$w(P) \le \frac{1}{2} \max \left| \sum_{\mathbf{v} \in V} \pm \mathbf{v} \right|,$$

where the maximum is taken over all choices of signs ±.

Proof. Note that (2) insures that the main step in the algorithm (find a $\mathbf{v} \in V$ such that $\mathbf{u}_i \in C_{\mathbf{v}}$) is always possible.

Let D' be the zonotope generated by V; in other words,

$$D' = \left\{ \sum_{\mathbf{v} \in V} \alpha_{\mathbf{v}} \mathbf{v} \mid 0 \le \alpha_{\mathbf{v}} \le 1 \right\},$$

and let $D = -\mathbf{x}_0 + D'$ be the translate of D' by $-\mathbf{x}_0$. We first show by induction that, at each iteration of Algorithm 1, the point $\mathbf{u}_i \in D$. The point $\mathbf{0} \in D$; in fact, it is the barycenter of D because $\mathbf{x}_0$ is the barycenter of D'. If $\mathbf{u}_i \in D \cap C_{\mathbf{v}}$, then $\mathbf{x}_0 + \mathbf{u}_i \in D' \cap C'_{\mathbf{v}}$, which, by the definition of D', implies that $\mathbf{x}_0 + \mathbf{u}_i + \mathbf{v} \in D'$. Therefore $\mathbf{u}_{i+1} = \mathbf{u}_i + \mathbf{v} \in D$.

Let G be the group consisting of translations of $\mathbb{R}^n$ by the vectors in $\Lambda(V)$. The quotient space $\mathbb{R}^n/G$ is often referred to as a *fundamental domain.* Such a fundamental domain contains exactly one representative from each coset of L/Λ. Note that D', and therefore D, is such a fundamental domain.

The points in the path constructed by Algorithm 1 are of the form $\mathbf{u}_i = \sum_{\mathbf{v} \in V} a_{i,\mathbf{v}} \mathbf{v}$, where the $a_{i,\mathbf{v}}$ are positive integers such that $\sum_{\mathbf{v} \in V} a_{i,\mathbf{v}} = i$. Letting m be the modulus, $\sum_{\mathbf{v} \in V} a_{m,\mathbf{v}} \equiv 0 \pmod{m}$, which by Lemma 7 implies that $\mathbf{u}_m \in \Lambda$. By the facts proved above, $\mathbf{0} \in D$ and $\mathbf{u}_m \in D$. But D can contain only one representative from each coset of L/Λ. Therefore $\mathbf{u}_m = \mathbf{0}$. Thus Algorithm 1 returns to $\mathbf{0}$ in $m(V)$ steps; it cannot, by Lemma 2, return sooner.

Also $w(P)$ is bounded above by

$$\max_{\mathbf{x} \in D} |\mathbf{x}| = \max_{\mathbf{x} \in D'} |-\mathbf{x}_0 + \mathbf{x}| = \left| -\frac{1}{2} \sum_{\mathbf{v} \in V} \mathbf{v} + \max_{Y \subsetneq V} \sum_{\mathbf{y} \in Y} \mathbf{y} \right| = \frac{1}{2} \max \left| \sum_{\mathbf{v} \in V} \pm \mathbf{v} \right|. \qquad \square$$

Remark 1. In dimension n, the difference between the upper bound in Theorem 10 and the lower bound in Proposition 9 is

$$\frac{1}{2} \max \left| \sum_{\mathbf{v} \in V} \pm \mathbf{v} \right| - \frac{1}{2} \max_{\mathbf{v} \in V} |\mathbf{v}| \le \frac{n}{2} \max_{\mathbf{v} \in V} |\mathbf{v}|.$$

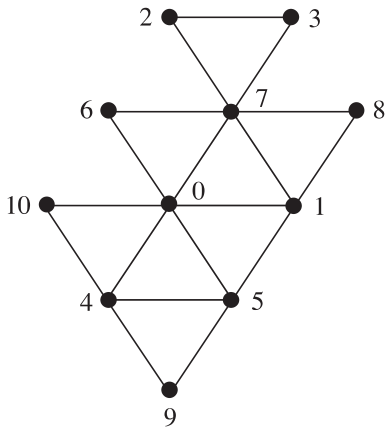


FIG. 3. *A closed V-path using Algorithm* 1.

Remark 2. In section 5 it will be shown that Algorithm 1 always finds an optimum wandering path in the 1-dimensional case. This is not necessarily true in higher dimensions, although finding examples is not completely trivial. In dimension 2 consider the basic set consisting of the following three vectors:

$$\begin{aligned}\mathbf{v}_0 &= (4,0),\\ \mathbf{v}_1 &= (0,3),\\ \mathbf{v}_2 &= (-7,-8).\end{aligned}$$

Algorithm 1 produces a closed wandering path P of length $m = 65$ with $w(P) = 5\sqrt{2} \approx 7.07$, where the point of P furthest from the origin is $(5,5)$. However, the optimum wandering path Q (of the same length 65) has $w(Q) = \sqrt{41} \approx 6.40$, where the point furthest from the origin is $(5,4)$. Another example is the 2-dimensional example

$$\begin{aligned}\mathbf{v}_0 &= (1,\ 0),\\ \mathbf{v}_1 &= (-1,\ \sqrt{3}),\\ \mathbf{v}_2 &= (-\tfrac{3}{2},\ -\tfrac{3}{2}\sqrt{3}),\end{aligned}$$

mentioned in section 2. The closed V-path $P = (\mathbf{0} = \mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{11} = \mathbf{0})$ found by Algorithm 1 is shown in Figure 3, where the labels indicate the indices. The vectors of V are successively added in the order $(\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_0, \mathbf{v}_2, \mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_0, \mathbf{v}_0, \mathbf{v}_2, \mathbf{v}_1, \mathbf{v}_0)$. The optimum wandering path for V is shown in Figure 1. Note that $w(P) = 2 > \sqrt{3} = w(V)$.

Remark 3. Algorithm 1 for the wandering path problem implies a corresponding algorithm for the discrete line problem. To obtain a solution to the discrete line problem from the wandering path problem, what is required is a closed V-path P of the same length N as the discrete line. By (1), the length N of the discrete line is also the length of some closed wandering V-path. Corollary 3 then implies that N is a multiple of $m(V)$. According to Theorem 10, the output of Algorithm 1 is a closed V-path P_0 of length $m(V)$. Therefore it suffices to take for P the path P_0 concatenated with itself $N/m(V)$ times, in which case $w(P) = w(P_0)$.

A discussion of the upper bound on $w(V)$ given in Theorem 10 is postponed until Algorithm 2 is introduced in section 6. In that section the bounds for the two algorithms are compared. The output of Algorithm 1 is a closed V-path P_0 of length $m(V)$. If Conjecture 4 is true, then the same can be said of the optimum wandering path.

5. The wandering path problem in dimension 1. In one dimension a basic set is just a pair $\{\alpha,\, \beta\}$ of real numbers for which there exist integers m_α and m_β such that $m_\alpha \alpha + m_\beta \beta = 0$. If $\alpha,\, \beta$ is such a basic set, then it is easy to find a relatively prime pair of integers $a,\, b$ and a real number λ such that $\alpha = \lambda\, a$ and $\beta = \lambda\, b$. Therefore, by Lemma 1, there is no loss of generality in assuming, in the 1-dimensional case, that the basic set is $V = \{a,\, b\}$, where a and b are a relatively prime pair of integers.

THEOREM 11. *Let $V = \{a, b\}$ be a basic set in $\mathbb{R}$ with a and b relatively prime. Then $w(V) = \lfloor (|a| + |b|)/2 \rfloor$. Moreover, Algorithm* 1.1 *below finds an optimum wandering path of length $|a| + |b|$.*

ALGORITHM 1.1.

```
Input: A basic set $V = \{a, b\}$
      (without loss of generality, $a, b$ are relatively prime and $a > 0,\ b < 0$).
Output: An optimum wandering path for V.
initialize: $i = 0,\ u_0 = 0$
until $i = |a| + |b|$ do
   if $u_i \geq -(a+b)/2$ then
      $u_{i+1} \leftarrow u_i + b$
   else
      $u_{i+1} \leftarrow u_i + a$
   $i \leftarrow i + 1$
end
return: Path $P = (0, u_1, u_1, \ldots, u_{|a|+|b|} = 0)$.
```

Proof. Algorithm 1.1 is exactly the 1-dimensional case of Algorithm 1 in the previous section. In this case the modulus $m(V) = |a| + |b|$. Theorem 10 then implies that Algorithm 1.1 finds a closed V-path P of length $|a|+|b|$ with $w(P) \leq \frac{1}{2}(|a|+|b|)$. Therefore

$$w(V) \leq \left\lfloor \frac{|a| + |b|}{2} \right\rfloor .$$

However, in the closed interval between $-\lfloor \frac{|a|+|b|}{2} \rfloor$ and $+\lfloor \frac{|a|+|b|}{2} \rfloor$ there are at most $|a|+|b|+1$ integers. By Corollary 3, any closed V-path P has at least $m(V) = |a|+|b|$ distinct points. Hence one of these points must be either $-\lfloor \frac{|a|+|b|}{2} \rfloor$, $+\lfloor \frac{|a|+|b|}{2} \rfloor$, or a point even further from the origin . Thus

$$w(V) \geq \left\lfloor \frac{|a| + |b|}{2} \right\rfloor . \qquad \square$$

6. Second algorithm. Our second algorithm for the wandering path problem is a "greedy" algorithm, choosing at each step the vector that brings the path closest to the origin.

ALGORITHM 2.

```
Input: A basic set $V$ in $\mathbb{R}^n$.
Output: A closed $V$-path $P$ in $\mathbb{R}^n$.
initialize: $i = 0,\ \mathbf{u}_0 = \mathbf{0}$
until $\mathbf{u}_i = \mathbf{u}_{i_0}$ for some $i_0 < i$ do
```

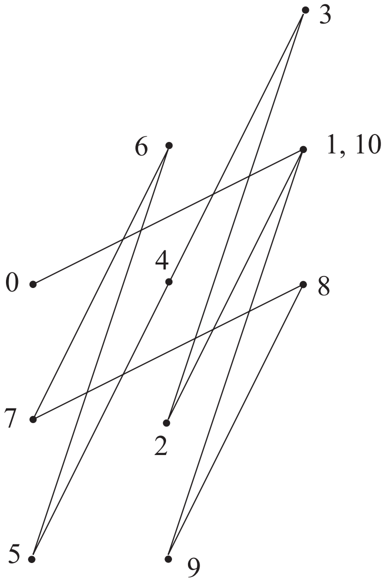


FIG. 4. *Closed V-path found using Algorithm* 2.

```
    choose $\mathbf{v} \in V$ such that $|\mathbf{u}_i + \mathbf{v}| \leq |\mathbf{u}_i + \mathbf{v}'|$ for all $\mathbf{v}' \in V$
    $\mathbf{u}_{i+1} \leftarrow \mathbf{u}_i + \mathbf{v}$
    $i \leftarrow i + 1$
end
return: path $P = (0, \mathbf{u}_{i_0+1} - \mathbf{u}_{i_0}, \mathbf{u}_{i_0+2} - \mathbf{u}_{i_0}, \ldots, \mathbf{0})$.
```

Remark 4. The translation by $-\mathbf{u}_{i_0}$ in Algorithm 2 is sometimes necessary. An example in dimension 2 is $V = \{(1,3), (2,1), (-1,-2)\}$. The V-path $(\mathbf{0} = \mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{10} = \mathbf{u}_1)$ found by Algorithm 2 is shown in Figure 4. The labels indicate the order in which the points of the path are found by Algorithm 2. In this case $i_0 = 1$. The output from Algorithm 2 is the closed V-path that starts and ends at $\mathbf{u}_1$ (labeled 1 in the figure). This V-path is translated so that $\mathbf{u}_1$ sits at the origin.

If V is a basic set, then by Lemma 6 the convex hull of V is an n-simplex $\Delta := \Delta(V)$ containing the origin. For $\mathbf{v} \in V$ let $\Delta_{\mathbf{v}}$ denote the n-simplex with vertex set $V \cup \{\mathbf{0}\} \setminus \{\mathbf{v}\}$. For any n-simplex Δ, let $R(\Delta)$ denote its circumradius.

THEOREM 12. *If V is a basic set, then Algorithm* 2 *finds a closed V-path that is contained in a ball of radius*

$$r(V) \leq \max\{R(\Delta), R(\Delta_{\mathbf{v}}) \,|\, \mathbf{v} \in V\}.$$

Before proving Theorem 12 several more remarks are in order.

Remark 5. Unlike Theorem 10, Theorem 12 does not state the length of the constructed path. We conjecture that the length is the modulus $m(V)$.

Remark 6. It is possible to find an explicit formula for the upper bound in Theorem 12. Let $V = \{\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_n\}$. The circumcenter $(x_1, \ldots, x_n)$ of $\Delta_k := \Delta_{\mathbf{v}_k}$ can be found by solving a system of linear equations as follows. Because each vertex of Δ_k is equidistant from the circumcenter of Δ_k,

$$\sum_j x_j^2 = \sum_j (x_j - v_{ij})^2$$

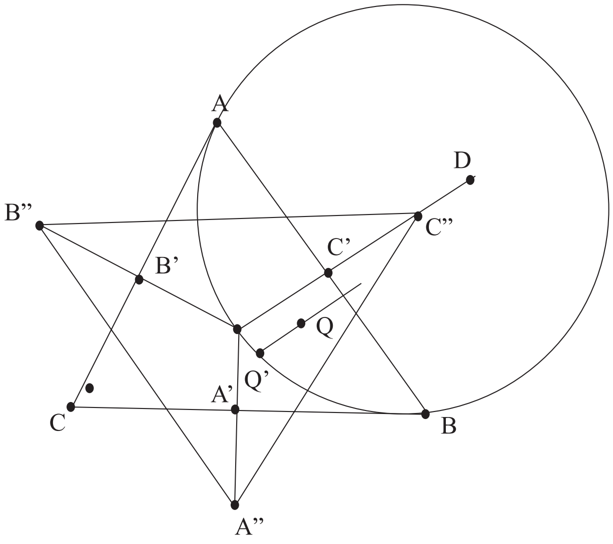


FIG. 5. *Proof of Theorem* 13.

for $i = 0, 1, 2, \dots, k-1, k+1, \dots, n$, where $\mathbf{v}_i = (v_{i1}, \dots, v_{in})$. This simplifies to the $n \times n$ linear system

$$\sum_j v_{ij}\, x_j \;=\; \frac{1}{2} \sum_j v_{ij}^2 .$$

Cramer's rule provides a determinantal formula for the circumradius $R(\Delta_k)$. Let $M_k = (v_{ij})$, and let M_{kjt} denote the result of replacing the jth column of M_k by the coordinatewise square of the tth column of M_k. Then

$$\begin{aligned} x_j \;&=\; \left(\textstyle\sum_{t=1}^n \det M_{kjt}\right)/(2 \det M_k),\\ R(\Delta_k) \;&=\; \left(\textstyle\sum_{j=1}^n x_j^2\right)^{\frac{1}{2}}. \end{aligned}$$

A similar formula is easily obtained for $R(\Delta)$ by translating one vertex to the origin.

Remark 7. The following theorem allows the upper bound $\max\{R(\Delta), R(\Delta_{\mathbf{v}}) \,|\, \mathbf{v} \in V\}$ in Theorem 12 to be simplified to $\max\{R(\Delta_{\mathbf{v}}) \,|\, \mathbf{v} \in V\}$ in the 2-dimensional case. We conjecture that the statement is true for a simplex Δ of arbitrary dimension $n \geq 2$, but the proof given below for a triangle does not seem to extend to higher dimensions.

THEOREM 13. *Let V denote the set of vertices of a 2-simplex Δ. If $\mathbf{v} \in V$ and $\mathbf{x}$ is any point in Δ, let $\Delta_{\mathbf{v}}^{\mathbf{x}}$ denote the n-simplex with vertex set $V \cup \{\mathbf{x}\} \setminus \{\mathbf{v}\}$. Then*

$$R(\Delta) \leq \max_{\mathbf{v} \in V} R(\Delta_{\mathbf{v}}^{\mathbf{x}}).$$

Proof. Denote the vertices of triangle Δ by A, B, C and its circumradius by R. We will prove the theorem in the case that Δ is an acute triangle, leaving the easy case of Δ obtuse to the reader. Let A', B', C' be the midpoints of sides BC, CA, AB, respectively. (See Figure 5.) Then triangle $\Delta' = \triangle(A'B'C')$ is similar to triangle Δ with ratio $1/2$, and with the same circumcenter O as triangle Δ. Since Δ is acute, O lies within triangle Δ. Let A'' be a point on the line OA' such that the distance

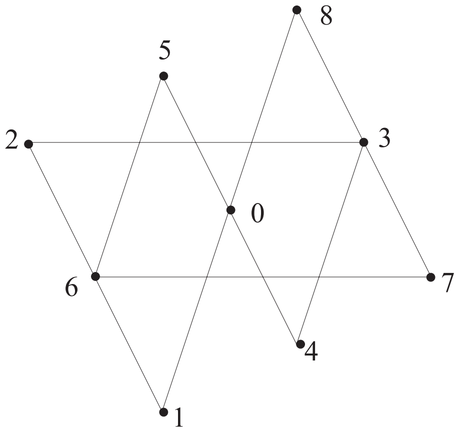


FIG. 6. *An optimum wandering path.*

OA'' is twice the distance OA' and such that A' lies between O and A'', similarly for B'' and C''. Note that $d(C'', A) = d(O, A) = R$ and $d(C'', B) = d(O, B) = R$, where d denotes Euclidean distance. Triangles Δ' and $\Delta'' = \triangle(A'', B''C'')$ are similar with ratio 2 implying that triangles Δ and Δ'' are congruent. Let O'' be the circumcenter of Δ''. Since $\Delta'' \cong \Delta$ we have $d(O'', C'') = R$. (Although not necessary for this proof, it is not hard to show that O'' is, in fact, the orthocenter of Δ, the intersection of the three altitudes.) Let S be the circumscribed circle of $\triangle(AO''B)$, which we have proved has center C''.

Let Q be an arbitrary point inside Δ. The three line segments AO'', BO'', CO'' subdivide Δ into three (some perhaps degenerate) triangles. Without loss of generality, assume that Q lies in triangle $AO''B$. Let R_Q be the circumradius of $\triangle(AQB)$, and let Q' be the point of intersection (inside Δ) of S with the line $\mathcal{L}$ through Q perpendicular to AB. The intersection of the perpendicular bisectors of the three line segments $AB, AQ', Q'B$ is C'', the center of the circle S. The circumcenter D of AQB lies on the ray OC'' because D is the intersection of the perpendicular bisectors of the three sides of $\triangle(AQB)$, and OC'' is the perpendicular bisector of side AB. On line $\mathcal{L}$, the point Q is closer to line AB than is Q'. It easily follows that $d(O, D) \geq d(O, C'')$, which in turn implies that $R_Q = d(D, A) \geq d(C'', A) = R$. This completes the proof. ▯

Remark 8. There is the question of which of the two algorithms, Algorithm 1 or 2, gives the better result. Often they both find the optimum wandering path. Consider the following examples in $\mathbb{R}^2$, where the basic set is $V = \{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2\}$:

$$\begin{aligned} \mathbf{v}_0 &= (5,\, 0), & \qquad \mathbf{v}_0 &= (5, 0), \\ \mathbf{v}_1 &= (-2,\, 4), & \mathbf{v}_1 &= (-2, 4), \\ \mathbf{v}_2 &= (-1,\, -3), & \mathbf{v}_2 &= (-2, -3). \end{aligned}$$

For the first example both algorithms find the same optimum wandering path,

$$P = (0,0),\ (-1,-3),\ (-3,1),\ (2,1),\ (1,-2),\ (-1,2),\ (-2,-1),\ (3,-1),\ (1,3),\ (0,0),$$

of length 9 with $w(V) = \sqrt{10} \approx 3.16$. This optimum wandering path, is shown in Figure 6.

For the second example both algorithms find the same optimum wandering path of length 49 with $w(V) = \sqrt{17} \approx 4.12$. For this example the upper bound provided for Algorithm 1 in Theorem 10 is $\sqrt{82}/2 \approx 4.53$, while the upper bound provided for Algorithm 2 in Theorem 12 is $\sqrt{389}/4 \approx 4.93$. In all the examples we have tried, the upper bound from Theorem 10 is less than or equal to the upper bound from Theorem 12.

It is not always the case that both algorithms find an optimum wandering path. It was noted in Remark 2 of section 4 that in neither of the following two examples does Algorithm 1 find an optimum wandering path. However, Algorithm 2 does find an optimum wandering path in both cases:

$$\begin{aligned} \mathbf{v}_0 &= (1,\, 0), & \qquad \mathbf{v}_0 &= (4, 0),\\ \mathbf{v}_1 &= (-1,\, \sqrt{3}), & \mathbf{v}_1 &= (0, 3),\\ \mathbf{v}_2 &= (-\tfrac{3}{2},\, -\tfrac{3}{2}\sqrt{3}), & \mathbf{v}_2 &= (-7, -8). \end{aligned}$$

For the first example, the optimum wandering path found using Algorithm 2 is shown in Figure 1, while the nonoptimum path found by Algorithm 1 is shown in Figure 3. For the second example, Algorithm 2 finds an optimum wandering path P_2 of length 65 with $w(P_2) = w(V) = \sqrt{41} \approx 6.40$, less than $w(P_1) = 5\sqrt{2} \approx 7.07$ for the closed V-path P_1 of the same length found by Algorithm 1.

On the other hand, there are examples for which Algorithm 1 finds an optimum wandering path while Algorithm 2 does not. For example, let $V = \{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2\}$, where

$$\begin{aligned} \mathbf{v}_0 &= (3,\, 1),\\ \mathbf{v}_1 &= (4,\, 1),\\ \mathbf{v}_2 &= (-14,\, -4). \end{aligned}$$

Algorithm 2 finds the closed V-path

$$P_2 = (0,0), (3,1), (6,2), (-8,-2), (-4,-1), (0,0),$$

with $w(P_2) = 2\sqrt{17} \approx 8.25$, while Algorithm 1 finds an optimum wandering path

$$P_1 = (0,0), (4,1), (7,2), (-7,2), (-3,-1), (0,0),$$

with $w(P_1) = w(V) = \sqrt{53} \approx 7.28$.

Proof of Theorem 12. The strategy is to find a ball B of radius $\max\{R(\Delta),\, R(\Delta_{\mathbf{v}}) \mid \mathbf{v} \in V\}$ such that, for all $\mathbf{x} \in B$, there is a $\mathbf{v} \in V$ such that $\mathbf{x} + \mathbf{v} \in B$. For distinct $\mathbf{u}, \mathbf{v} \in V$ let $H_{\mathbf{uv}}$ denote the hyperplane that is the perpendicular bisector of the segment joining $-\mathbf{u}$ and $-\mathbf{v}$. The intersection of all the $H_{\mathbf{uv}}$ is the circumcenter O of the simplex $-\Delta$, which is the convex hull of $-V$. Let $H^+_{\mathbf{uv}}$ be the closed half-space that contains $-\mathbf{v}$ determined by $H_{\mathbf{uv}}$, and let

$$P_{\mathbf{v}} = \bigcap_{\mathbf{u} \in V \setminus \{\mathbf{v}\}} H^+_{\mathbf{uv}}.$$

Note that

$$P_{\mathbf{v}} = \{\mathbf{x} : |\mathbf{x} + \mathbf{v}| \le |\mathbf{x} + \mathbf{u}| \text{ for all } \mathbf{u} \in V \setminus \{\mathbf{v}\}\}.$$

The $P_{\mathbf{v}}$, $\mathbf{v} \in V$, partition $\mathbb{R}^n$ into $n+1$ polyhedral cones with vertex at O.

For $\mathbf{v} \in V$, let $H_{\mathbf{v}}$ denote the hyperplane that is the perpendicular bisector of the segment joining $-\mathbf{v}$ and $\mathbf{0}$. Then

$$O_{\mathbf{v}} = \bigcap_{\mathbf{u} \in V \setminus \{\mathbf{v}\}} H_{\mathbf{u}}$$

is the circumcenter of $-\Delta_{\mathbf{v}}$. Therefore $O_{\mathbf{v}}$ also lies on each hyperplane $H_{\mathbf{u}\mathbf{u}'}$ with $\mathbf{u}, \mathbf{u}' \in V \setminus \{\mathbf{v}\}$. Let $H_{\mathbf{v}}^{+}$ denote the closed half-space that contains O determined by the hyperplane $H_{\mathbf{v}}$, and let $H_{\mathbf{v}}^{-}$ denote the complementary closed half-space. Further partition each $P_{\mathbf{v}}$ as follows. Let

$$P_{\mathbf{v}}^{+} = P_{\mathbf{v}} \cap H_{\mathbf{v}}^{+}, \qquad\qquad P_{\mathbf{v}}^{-} = P_{\mathbf{v}} \cap H_{\mathbf{v}}^{-}.$$

Note that $P_{\mathbf{v}}^{+}$ is a simplex with vertex set $\{O\} \cup \bigcup_{\mathbf{u} \in V \setminus \{\mathbf{v}\}} O_{\mathbf{u}}$. Any point $\mathbf{x}$ in $P_{\mathbf{v}}^{-}$ satisfies the properties

a. $|\mathbf{x} + \mathbf{v} \leq |\mathbf{x} + \mathbf{u}|$ for all $\mathbf{u} \in V \setminus \{\mathbf{v}\}$, and

b. $|\mathbf{x} + \mathbf{v}| \leq |\mathbf{x}|$.

Moreover, any point $\mathbf{x}$ in $P_{\mathbf{v}}^{+}$ satisfies the properties

c. $|\mathbf{x} + \mathbf{v}| \leq |\mathbf{x} + \mathbf{u}|$ for all $\mathbf{u} \in V \setminus \{\mathbf{v}\}$, and

d. $|\mathbf{x} + \mathbf{v}| \leq \max_{\mathbf{u} \in V \setminus \{\mathbf{v}\}} \{|O + \mathbf{v}|, |O_{\mathbf{u}} + \mathbf{v}|\}$,

inequality (d) following from the fact that $\mathbf{x}$ lies within the simplex $P_{\mathbf{v}}^{+}$ with vertex set $\{O\} \cup \bigcup_{\mathbf{u} \in V \setminus \{\mathbf{v}\}} O_{\mathbf{u}}$. But for $\mathbf{u} \in V \setminus \{\mathbf{v}\}$ we have $|O_{\mathbf{u}} + \mathbf{v}| = |O_{\mathbf{u}}|$, because $O_{\mathbf{u}}$ lies on the hyperplane $H_{\mathbf{v}}$, which is the perpendicular bisector of the line segment joining $\mathbf{0}$ and $-\mathbf{v}$. Also $|O_{\mathbf{u}}|$ is the circumradius of $-\Delta_{\mathbf{u}}$, and $|O + \mathbf{v}|$ is the circumradius of $-\Delta$. From property (d) it follows that for any $\mathbf{x} \in P_{\mathbf{v}}^{+}$ we have

e. $|\mathbf{x} + \mathbf{v}| \leq \max\{R(-\Delta), R(-\Delta_{\mathbf{v}}) \,|\, \mathbf{v} \in V\} = \max\{R(\Delta), R(\Delta_{\mathbf{v}}) \,|\, \mathbf{v} \in V\}$.

Properties (a) and (c) are the greedy property, and (b) and (e) insure that at each iteration i of Algorithm 2,

$$|\mathbf{u}_i| \leq \max\{R(\Delta), R(\Delta_{\mathbf{v}}) \,|\, \mathbf{v} \in V\}.$$

Because $\{\mathbf{0} = \mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2, \ldots\}$ is a bounded set and is a set of lattice points in L (by Lemma 5), it follows that there is a point of the constructed V-path that appears at least twice in this sequence. If $\mathbf{u}_{i_0}$ is the first such point, then the closed V-path $P = (\mathbf{u}_{i_0}, \mathbf{u}_{i_0+1}, \mathbf{u}_{i_0+2}, \ldots, \mathbf{u}_{i_0})$ is contained in a ball or radius $\max_{\mathbf{v} \in V} R(\Delta_{\mathbf{v}})$. □

REFERENCES

[1] J. E. Bresenham, *Algorithm for computer control of a digital plottter*, IBM Systems J., 4 (1965), pp. 25–30.

[2] G. W. Gill, *N-step and incremental straight-line algorithms*, IEEE Comput. Graph. Appl., 14 (1994), pp. 66–72.

[3] P. Graham and S. Iyengar, *Double- and triple-step incremental linear interpolation*, IEEE Comput. Graph. Appl., 14 (1994), pp. 49–53.

[4] L. Ibáñez, C. Hamitouche, and C. Roux, *A vectorial algorithm for tracing discrete straight lines in N-dimensional generalized grids*, IEEE Trans. Visualization and Comput. Graphics, 7 (2001), pp. 97–108.

[5] K. S. Reid-Green, *Three early algorithms*, IEEE Ann. Hist. Comput., 24 (2002) pp. 10–13.

[6] J.-P. Reveills and D. Richard, *Back and forth between continuous and discrete for the working computer scientist*, Ann. Math. Artificial Intelligence, 16 (1996), pp. 89–152.

[7] C. Yao and J. G. Rokne, *An integral linear interpolation approach to the design of incremental line algorithms*, J. Comput. Appl. Math., 102 (1999), pp. 3–19.

SIAM J. DISCRETE MATH.
Vol. 21, No. 3, pp. 662–675

THE MAXIMUM INDUCED BIPARTITE SUBGRAPH PROBLEM WITH EDGE WEIGHTS[*]

DENIS CORNAZ[†] AND A. RIDHA MAHJOUB[‡]

Abstract. Given a graph $G = (V, E)$ with nonnegative weights on the edges, the maximum induced bipartite subgraph problem (MIBSP) is to find a maximum weight bipartite subgraph $(W, E[W])$ of G. Here $E[W]$ is the edge set induced by W. An edge subset $F \subseteq E$ is called independent if there is an induced bipartite subgraph of G whose edge set contains F. Otherwise, it is called dependent. In this paper we characterize the minimal dependent sets, that is, the dependent sets that are not contained in any other dependent set. Using this, we give an integer linear programming formulation for MIBSP in the natural variable space, based on an associated class of valid inequalities called dependent set inequalities. Moreover, we show that the minimum dependent set problem with nonnegative weights can be reduced to the minimum circuit problem in a directed graph, and can then be solved in polynomial time. This yields a polynomial-time separation algorithm for the dependent set inequalities as well as a polynomial-time cutting plane algorithm for solving the linear relaxation of the problem. We also discuss some polyhedral consequences.

Key words. induced bipartite subgraph, edge weight, minimal dependent set, separation algorithm, polytope

AMS subject classifications. 05C85, 90C27, 90C57

DOI. 10.1137/060650015

1. Introduction.

A graph is called *bipartite* if its node set can be partitioned into two nonempty sets V_1 and V_2 such that no two nodes in V_i are linked by an edge, for $i = 1, 2$. Let $G = (V, E)$ be a graph. A subgraph (W, F) of G is said to be *induced* if F is the set of edges having both endnodes in W. Given w, a function that associates with each edge $e \in E$ a nonnegative weight $w(e)$, the *maximum induced bipartite subgraph problem* (MIBSP) is to find an induced bipartite subgraph with maximum weight.

An edge subset $F \subseteq E$ is called *independent* if there is an induced bipartite subgraph of G with edge set $B \subseteq E$ such that $F \subseteq B$. Otherwise, it is called *dependent*. In this paper we characterize the minimal dependent sets of a graph $G = (V, E)$. Using this, we give a 0-1 linear programming formulation for MIBSP in the natural variable space, based on an associated class of valid inequalities called dependent set inequalities. We also show that the minimum dependent set problem with nonnegative weights can be reduced to the minimum circuit problem in a directed graph and can then be solved in polynomial time. This yields a polynomial-time separation algorithm for the dependent set inequalities as well as a polynomial-time cutting plane algorithm for solving the linear relaxation of the problem.

To the best of our knowledge, MIBSP has not been considered before in the literature. However, the maximum bipartite subgraph problem has been extensively investigated. Here, given a graph $G = (V, E)$ and weights on the edges of G, the

[*] Received by the editors January 15, 2006; accepted for publication (in revised form) March 26, 2007; published electronically August 29, 2007.
http://www.siam.org/journals/sidma/21-3/65001.html

[†] Équipe Combinatoire, UFR 921, Université Pierre et Marie Curie, 4 place Jussieu, 75252 Paris Cedex 05, France. Current address: Laboratoire LIMOS, CNRS UMR 6158, Université Blaise Pascal, Clermont II, 63177 Aubière Cedex, France (cornaz@isima.fr).

[‡] Laboratoire LIMOS, CNRS UMR 6158, Université Blaise Pascal, Clermont II, 63177 Aubière Cedex, France (Ridha.Mahjoub@math.univ-bpclermont.fr).

problem is to find a bipartite subgraph (not necessarily induced) with maximum weight. In [3] Barahona, Grötschel, and Mahjoub describe several classes of facet defining inequalities of the associated bipartite subgraph polytope. They also present some methods with which new facet defining inequalities of that polytope can be constructed from known ones.

A graph is said to be weakly *bipartite* if the bipartite subgraph polytope coincides with the polytope given by the trivial inequalities and the so-called odd cycle inequalities. Grötschel and Pulleyblank [18] showed that the bipartite subgraph problem can be solved in polynomial time in that class of graphs. Barahona showed that planar graphs [1] and graphs G that contain two nodes which cover all the odd-cycles of G [2] belong to that class of graphs. In [10] Fonlupt, Mahjoub, and Uhry generalize these results by showing that the graphs noncontractible to K_5 are weakly bipartite. Recently Guenin [19] gave a characterization for that class of graphs.

The closely related MIBSP with node weights has also been studied. Here we suppose that the weights are associated with the nodes of the graph, and the problem is to determine an induced bipartite subgraph with maximum weight. This problem has applications to the via-minimization problem which arises in the design of integrated circuits and printed circuit boards [6], [11]. In [4] Barahona and Mahjoub study the polytope $\mathrm{BP}(G)$ associated with this problem. They exhibit some basic classes of facet defining inequalities for $\mathrm{BP}(G)$ and describe several lifting methods. In [5] they study a composition technique for $\mathrm{BP}(G)$ in the graphs which are decomposable by one- and two-node cutsets. Fouilhoux and Mahjoub [12] (see also Fouilhoux [11]) study the polytope $\mathrm{BP}(G)$. They describe new classes of facet defining inequalities and discuss separation procedures. Using this, they develop a branch-and-cut algorithm for the problem and present some computational results. In [13] Fouilhoux and Mahjoub consider the via-mimization problem and show that this can be reduced to the MIBSP with appropriate node weights. Further applications of the MIBSP with node weights to the via-minimization problem and DNA sequencing are also discussed in [11].

A related work has been done by Cornaz and Fonlupt [7] on the maximum biclique problem. (A biclique is the edge set of a complete bipartite (not necessarily induced) subgraph). Although the MIBSP and the maximum biclique problem are different, this paper gives rise to some structural relations between the minimal dependent sets associated to both problems.

The paper is organized as follows. In the following section we give some notation, definitions, and preliminary results. In section 3 we study the dependent sets and give a characterization for these sets. In section 4 we show that the minimum dependent set problem with nonnegative weights can be reduced to the minimum odd circuit problem and can then be solved in polynomial time. In section 5 we discuss some polyhedral consequences and give some concluding remarks.

2. Definitions, notation, and preliminary results.

2.1. Definitions and notation. Throughout the paper we consider only simple graphs and digraphs. We will denote a graph by $G = (V, E)$, where V is the *node set* and E is the *edge set*. An edge with endnodes u and v will be denoted by uv. For $W \subseteq V$, we let $E[W]$ denote the set of edges having both nodes in W. The graph $G[W] = (W, E[W])$ is the subgraph of G induced by W. If $F \subseteq E$, we let $V(F)$ denote the set of nodes incident to edges of F, and $G(F) = (V(F), F)$. Note that $G(F) = G[V(F)]$ holds if and only if $G(F)$ is an induced subgraph of G.

We denote a directed graph (or digraph) by $D = (V, A)$, where V is the node set and A the arc set of D. An arc with initial node u and terminal node v will be denoted by uv. (Note that $uv \neq vu$ for digraphs.)

A path in G (resp., D) is an alternate sequence of nodes and edges (resp., arcs) $P = v_1, e_1, v_2, \ldots, v_k, e_k, v_{k+1}$ such that $k \geq 1$, all the nodes v_i are distinct, and $e_i = v_i v_{i+1} \in E$ (resp., $e_i = v_i v_{i+1} \in A$) for $i = 1, 2, \ldots, k$. The nodes v_1 and v_{k+1} are the *extremities* of P, and we will say that P links v_1 and v_{k+1} (resp., v_1 to v_{k+1}). The integer k is called the *length* of P, and P is said to be *even* (*odd*) if k is even (odd). If $v_1 = v_{k+1}$, P is called a *cycle* (resp., *circuit*). An edge linking two nonconsecutive nodes of a path (cycle) P is called a *chord* of P. A chordless cycle is also called a *hole*. Given a path P, we let $E(P)$ (resp., $A(P)$) and $V(P)$ denote the sets of edges (resp., arcs) and nodes of P, respectively.

Given a vector $x \in \mathbb{R}^E$ and $T \subseteq E$, we let $x(T)$ denote $\sum_{e \in T} x(e)$. Bipartite graphs have the following property.

REMARK 2.1. *A graph is bipartite if and only if it does not contain an odd cycle.*

2.2. Signed digraphs. A *signed* digraph consists of a digraph $D = (V, A)$ and a subset $\Sigma \subseteq A$ of arcs called *signed arcs*. The arcs in $A \setminus \Sigma$ are said to be *unsigned*. Given a signed digraph $D = (V, A)$, a circuit is said to be *odd* if it contains an odd number of signed arcs. Note that if $\omega \in \mathbb{R}^A$ is a weight vector, then finding a minimum weight odd circuit in D reduces to finding a minimum weight odd circuit in an unsigned digraph. In fact, for this, it suffices to replace every unsigned arc $uv \in A \setminus \Sigma$ by a path u, uw, w, wv, v, where w is a new node, and associate to the new arcs uw, wv the weight $\frac{\omega(uv)}{2}$. Moreover, finding a minimum weight odd circuit in a digraph reduces to a shortest path problem [17] (see also [16]). As the weights are nonnegative, it can then be solved in polynomial time, using, for instance, Dijkstra's algorithm [9].

2.3. Independent sets. Given a graph $G = (V, E)$, we let $\mathcal{B}(G)$ denote the set of the edge sets of the induced bipartite subgraphs of G, i.e.,

$$\mathcal{B}(G) = \{B \subseteq E : \ G(B) = G[V(B)] \text{ and } G(B) \text{ is bipartite}\}.$$

Hence the MIBSP is equivalent to

$$\text{maximize } \{\omega(B) : B \in \mathcal{B}(G)\}.$$

Given a graph $G = (V, E)$, a node subset $W \subseteq V$ is called a *stable set* if $E[W] = \emptyset$. The *stable set problem* in G consists in finding a stable set of maximum cardinality. Note that the stable set problem can be reduced to the MIBSP. In fact, consider the graph $\bar{G} = (\bar{V}, \bar{E})$ obtained from G by adding a universal node (a node adjacent to all the other nodes of G) and associate with the edges of $\bar{E}$ the weight $\omega(e) = 1$ if $e \in \bar{E} \setminus E$ and $\omega(e) = 0$ if not. It is easy to see that an optimum solution of the MIBSP in $\bar{G}$ with respect to weight vector ω corresponds to a maximum cardinality stable set in G. This implies that the MIBSP is NP-hard. The maximum cardinality MIBSP is to find a set in $\mathcal{B}(G)$ with maximum cardinality. In what follows we shall show that the maximum cardinality MIBSP is also NP-hard, which implies that MIBSP is strongly NP-hard.

PROPOSITION 2.2. *The maximum cardinality MIBSP is NP-hard.*

Proof. We show that the stable set problem in a graph $G = (V, E)$ reduces to the maximum cardinality MIBSP. As the former problem is NP-hard [14], the latter is also NP-hard.

Let $\tilde{G} = (\tilde{V}, \tilde{E})$ be the graph obtained from $G = (V, E)$ by considering a copy $G' = (V', E')$ of G and adding all the possible edges between V and V', that is,

$\tilde{V} = V \cup V'$ and $\tilde{E} = E \cup E' \cup \{vv' : v \in V,\, v' \in V'\}$. Note that for every stable set $S \subseteq V$ of G and its copy $S' \subseteq V'$ of G', $\tilde{G}[S \cup S']$ is a complete bipartite graph. Thus for every maximum cardinality solution $B \in \mathcal{B}(\tilde{G})$ and every stable set $S \subseteq V$ of G, we have that $|B| \geq |S|^2$. In particular $|B| \geq |S^*|^2$, where S^* is a maximum stable set of G.

Now let $B \in \mathcal{B}(\tilde{G})$ be of maximum cardinality and $\tilde{S} \subseteq \tilde{V}$ a maximum cardinality stable set of $\tilde{G}$ such that every node $v \in \tilde{S}$ is incident to an edge in B. Obviously, either $\tilde{S} \subseteq V$ or $\tilde{S} \subseteq V'$. Thus $|\tilde{S}| \leq |S^*|$. As $|B| \leq |\tilde{S}|^2$, it follows that $|B| \leq |S^*|^2$. Consequently, $|B| = |\tilde{S}|^2 = |S^*|^2$. □

Denote by $\mathcal{I}(G)$ the set of the independent sets of G, i.e.,

$$\mathcal{I}(G) = \{I \subseteq E : \exists\, B \in \mathcal{B}(G),\ I \subseteq B\}.$$

Obviously, $\mathcal{B}(G) \subseteq \mathcal{I}(G)$. However, in general we have that $\mathcal{B}(G)$ is a strict subset of $\mathcal{I}(G)$. For instance, consider the graph G consisting of a path of three edges (e_1, e_2, e_3). Clearly, the edge subset $I = \{e_1, e_3\}$ is independent. However, $G(I) \neq G[V(I)]$, and hence $I \notin \mathcal{B}(G)$. Also note that, since the weights are nonnegative,

$$\max\{\omega(\mathcal{I}) \ : \ \mathcal{I} \in \mathcal{I}(G)\} = \max\{\omega(B) \ : \ B \in \mathcal{B}(G)\}.$$

Therefore the MIBSP is equivalent to finding a maximum weight independent set in G. Moreover, we have the following which is a direct consequence of Remark 2.1.

LEMMA 2.3. *Given an edge set $I \subseteq E$, $I \in \mathcal{I}(G)$ if and only if $G[V(I)]$ contains no odd cycle.*

In what follows we will denote by $\mathcal{C}(G)$ the set of the minimal dependent sets of G, i.e.,

$$\mathcal{C}(G) = \{C \subseteq E : C \notin \mathcal{I}(G) \text{ and } C' \in \mathcal{I}(G)\ \forall\, C' \subset C\}.$$

We have the following.

LEMMA 2.4. *Given an edge set $C \subseteq E$, $C \in \mathcal{C}(G)$ if and only if*

(i) *there exists at least one odd cycle in $G[V(C)]$, and*

(ii) *for every odd cycle Q of $G[V(C)]$ and every edge $f \in C$, there exists a node $v_f \in V(Q)$ such that f is the unique edge of C incident to v_f.*

Proof.

Necessity.

(i) This follows from Lemma 2.3.

(ii) Let Q be an odd cycle of $G[V(C)]$ and f an edge of $C \in \mathcal{C}(G)$. If the statement does not hold, it is not hard to see that $V(Q) \subseteq V(C \setminus \{f\})$. But this implies that $C \setminus \{f\}$ is dependent, contradicting the minimality of C.

Sufficiency. By Lemmas 2.4(i) and 2.3, we have that $C \notin \mathcal{I}(G)$. Now suppose that C is not minimal. Then there exists an edge $f = uv \in C$ such that $C' = C \setminus \{f\} \notin \mathcal{I}(G)$. By Lemma 2.3, this implies that $G[V(C')]$ contains an odd cycle, say Q, and hence $V(Q) \subseteq V(C')$. Moreover, by Lemma 2.4(ii), it follows that one of the nodes of f, say v, belongs to $V(Q)$ and is not incident to any edge in C'. But this implies that $v \in V(Q) \setminus V(C')$, a contradiction. □

Figure 1 shows a subgraph which is induced by the node set of a minimal dependent set. The dependent set is presented by bold lines. We can remark that the subgraph contains an odd cycle, and that for every edge f of the dependent set, there is a node of the cycle such that f is the only edge incident to it.

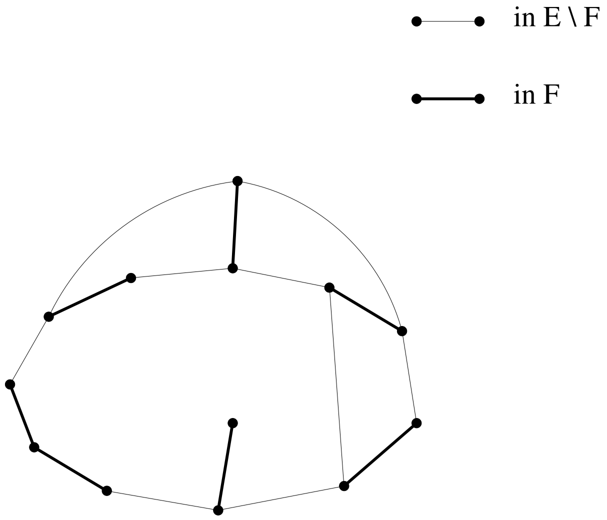


FIG. 1. *A subgraph induced by a minimal dependent set.*

3. Minimal dependent sets. The characterization of the minimal dependent sets, given by Lemma 2.4, is not strong enough to obtain certain polyhedral results for the MIBSP which we will present in the next sections. In this section we give a stronger characterization of the minimal dependent sets. This will be given in Theorem 3.2. To this end, we first introduce some definitions.

Let $F \subseteq E$ and Q be an odd cycle of G. A node $v \in V(Q)$ is said to be *unsaturated* with respect to F and Q if v is not incident to any edge of $F \cap E(Q)$; otherwise v is said to be *saturated*.

DEFINITION 3.1. *Given an edge set $F \subseteq E$, we say that F induces an obstruction with respect to an odd cycle Q if Q is an odd cycle of $G[V(F)]$ and if conditions* (1) *and* (2) *below are satisfied.*

(1) *Every edge $f \in F \backslash E(Q)$ is of the form $f = vw$ where $v \in V(Q)$, $w \in V \backslash V(Q)$, and there is no edge in $F \setminus \{f\}$ adjacent to f.*

(2) *Every edge in $F \cap E(Q)$ is adjacent to at most one edge of F.*

Figure 2 shows an obstruction induced by an edge set F with respect to the odd cycle on seven edges. We can remark here that F does not correspond to a minimal dependent set. The edges e, f induce a minimal dependent set.

Let $F \subseteq E$ be an edge set. And suppose that F induces an obstruction with respect to an odd cycle

$$Q = v_1, e_1, v_2, e_2, \ldots, v_k, e_k, v_{k+1}$$

of $G[V(F)]$, where $v_{k+1} = v_1$. Let $e \in E[V(F)] \setminus (F \cup E(Q))$. Edge e is called a *diagonal* (with respect to F and Q) if there is $i \in \{1, \ldots, k\}$ such that $e = v_i v_{i+3}$, the edges e_i, e_{i+2} are in F, and the edges $e_{i-1}, e_{i+1}, e_{i+3}$ are in $E \setminus F$ (the indices are taken modulo k). And edge e is called a *forward* (resp., *backward*) *wing* (with respect to F and Q) if there is $i \in \{1, \ldots, k\}$ and a node $w \in V \setminus V(Q)$ such that $e = wv_i$ with $e_i, wv_{i+2} \in F$ (resp., $wv_{i-2}, e_{i-1} \in F$), and $e_{i-1}, e_{i+1}, e_{i+2} \in E \setminus F$ (resp., $e_{i-3}, e_{i-2}, e_i \in E \setminus F$). An edge is called a *wing* if it is either a forward or a backward wing.

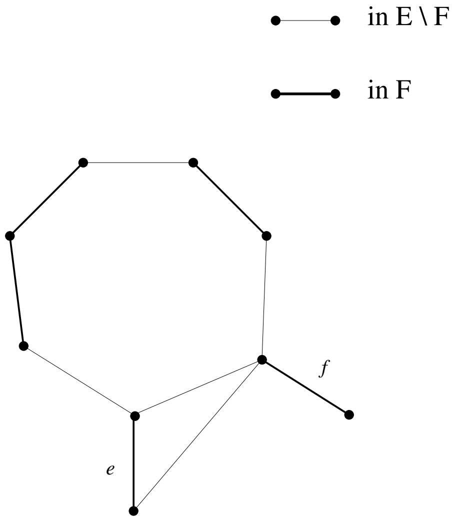


Fig. 2. *An obstruction induced by F.*

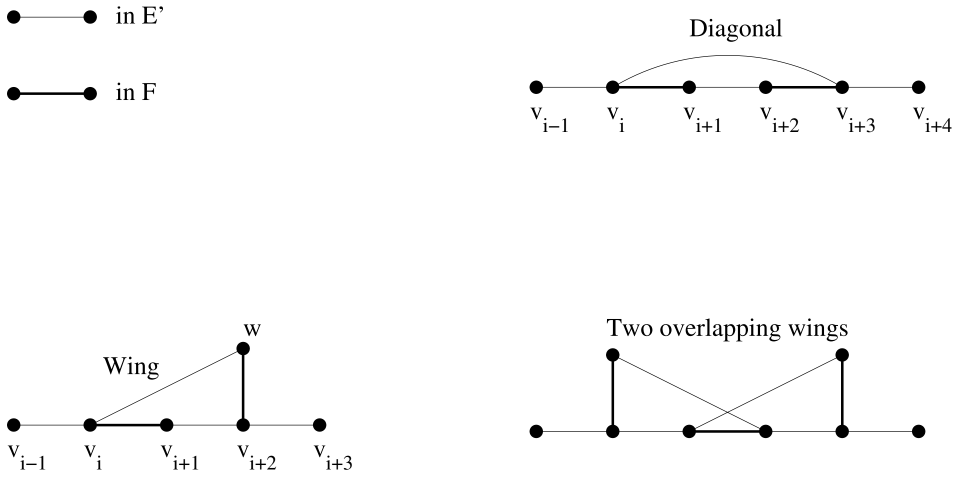


Fig. 3. *Wings and diagonals.*

We say that two wings wv_i and $w'v_j$ *overlap* if $v_iv_j \in F$. Note that if two wings overlap, then necessarily one is forward and the other is backward (see Figure 3 for an illustration).

The following theorem gives a characterization of the set of minimal dependent sets of G.

Theorem 3.2. *Let $G = (V, E)$ be a graph and $F \subseteq E$ an edge subset of E. Then F is a minimal dependent set if and only if F induces an obstruction with respect to an odd cycle Q such that*

(i) *every edge of $E[V(F)] \setminus (F \cup E(Q))$ is either a diagonal or a wing, and*

(ii) *no wings overlap.*

Proof.

Necessity. Suppose $F \in \mathcal{C}(G)$. Let $W = V(F)$. By Lemma 2.3, $G[W]$ contains an odd cycle. Let

$$Q = v_1, e_1, v_2, e_2, \ldots, v_k, e_k, v_{k+1},$$

with $v_{k+1} = v_1$, be an odd cycle of $G[W]$ such that $|F \cap E(Q)|$ is maximum. Let $F_1 = F \cap E(Q)$, $F_2 = F \setminus F_1$, and $E' = E[W] \setminus (F \cup E(Q))$. Note that by Lemma 2.4(ii), no edge of F_1 is adjacent to an edge of F_2.

For the rest of the proof, we will need to consider paths of $G[W]$ with extremities in $V(Q)$ and internal nodes in $W \setminus V(Q)$. If P is a path of $G[W]$ linking two nodes v_i and v_j of $V(Q)$ such that $E(P) \cap E(Q) = \emptyset$ and none of the internal nodes of P belongs to $V(Q)$, we let $P_1 = v_j, e_{j+1}, \ldots, e_{i-1}, v_i$ and $P_2 = v_i, e_i, \ldots, e_{j-1}, v_j$ (where the indices are modulo k) denote the edge-disjoint paths of Q between v_i and v_j. Note that $Q = P_1 \cup P_2$. We also denote by $Q_1 = P \cup P_1$ and $Q_2 = P \cup P_2$ the cycles obtained by adding P to Q. Note that Q_1 and Q_2 are of opposite parity. We will suppose, without loss of generality, that Q_1 is odd and Q_2 is even.

By Lemma 2.4(ii), for every edge $f \in F$, there exists a node $v_f \in V(Q)$ such that f is the only edge of F incident to v_f. By a similar argument, there is also a node $v_f^1 \in V(Q_1)$ such that f is the only edge of F incident to v_f^1 (v_f^1 and v_f may be the same). We have the following claims.

CLAIM 1. *Let P be a path of $G[W]$ linking two nodes v_i and v_j of $V(Q)$ whose internal nodes belong to $W \setminus V(Q)$. Let $f_2, f_2' \in F_2$ and $e \in E'$. Then the following cases cannot occur:*

(a) *$P = v_i, f_2, v_j$,*

(b) *$P = v_i, f_2, w, f_2', v_j$ with $w \in W \setminus V(Q)$, or*

(c) *$P = v_i, f_2, w, e, w', f_2', v_j$ with $w, w' \in W \setminus V(Q)$.*

Proof. Assume that P is of type (a), (b) or (c). Notice that if $f \in F \cap E(P_2)$, then $v_f^1 = v_i$ or v_j. Also note that v_i and v_j are both incident to an edge of $F \setminus E(P_2)$. Then it follows that $F \cap E(P_2) = \emptyset$. Since $|F \cap E(P)| \geq 1$, $|F \cap E(Q)| < |F \cap E(Q_1)|$. But this contradicts the maximality of $|F \cap E(Q)|$. □

CLAIM 2. *F induces an obstruction with respect to Q.*

Proof. Let $f = wv_f \in F \setminus E(Q)$. (Recall that v_f is the node of $V(Q)$ such that f is the only edge of F incident to it.) If $w \in V(Q)$, then w, f, v_f is a path of $G[W]$. But this is impossible by Claim 1(a). So suppose that $w \notin V(Q)$. If there is an edge $f' = wv_{f'} \in F$, then $P = v_f, f, w, f', v_{f'}$ is a path of $G[W]$, contradicting Claim 1(b). In consequence, f is adjacent to no edge in F, and thus condition (1) of Definition 3.1 is satisfied.

Now let $f = uv_f$ be an edge of $F \cap E(Q)$. Since F satisfies condition (1) of Definition 3.1, f is adjacent to no edge in F_2. Moreover, we have that v_f is incident to no edge in F. Hence f is adjacent to at most one edge of $F \cap E(Q)$. Therefore condition (2) of Definition 3.1 is satisfied. □

CLAIM 3. *Every edge of E' is incident to a node of Q.*

Proof. Suppose that for an edge $e = ww'$ of E', we have $\{w, w'\} \cap V(Q) = \emptyset$. Since $w, w' \in W \setminus V(Q)$, there exist two edges $f = wv_f$ and $f' = w'v_{f'}$ of F. Note that $v_f \neq v_{f'}$. Hence $v_f, f, w, e, w', f', v_{f'}$ is a path of $G[W]$, which contradicts Claim 1(c). □

CLAIM 4. *Let $e \in E'$, $f \in F$, and $v_i, v_j \in V(Q)$.*

(1) *If $P = v_i, e, w, f, v_j$ is a path of $G[W]$, then e is a forward wing.*

(2) *If $P = v_i, f, w, e, v_j$ is a path of $G[W]$, then e is a backward wing.*

Proof. We prove (1), the proof of (2) is similar. As Q_2 is even, the path P_2 must contain an even number of edges, and hence $|E(P_2)| \geq 2$. Moreover, as by Claim 2 F

induces an obstruction with respect to Q, f is the only edge of F incident to v_j (w). In consequence, v_j is unsaturated. Moreover, v_j is the unique unsaturated node of P_2. Indeed, if v was a further unsaturated node of P_2, then there must exist an edge, say f', of F_2 incident to v. As by Claim 2 F induces an obstruction with respect to Q, $v^1_{f'} \in V \setminus V(Q)$. (Recall that $v^1_{f'}$ is the node of $V(Q_1)$ such that f' is the only edge of F incident to it.) Thus $v^1_{f'} = w$. But this is impossible since f is (the only edge of F) incident to w. In consequence, e_i is the unique edge of F in P_2 and thus $|E(P_2)| = 2$. Therefore e_{i-1}, e_{j-1}, e_j are not in F, and hence e is a forward wing. □

Claim 5. *Let $e \in E'$. If $P = v_i, e, v_j$ is a path of $G[W]$ with $v_i, v_j \in V(Q)$, then e is a diagonal.*

Proof. As $|P|$ is odd and Q_2 is even, P_2 must be odd, and therefore $|E(P_2)| \geq 3$. Also P_2 contains no unsaturated nodes. In fact, if P_2 contains an unsaturated node v, then there must exist $f \in F$ incident to v such that $v^1_f \in V(P_1)$. But this contradicts Claim 1(a). In consequence, two consecutive edges of $E(P_2)$ cannot both be in E'. From Lemma 2.4(ii), it then follows that e_i and e_{j-1} are the only edges of F in $E(P_2)$ and that $|E(P_2)| = 3$. We also have that e_{i-1} and e_j are not in F, $j = i + 3$, and $e_{i+1} \notin F$. Thus e is a diagonal. □

Claim 6. *No wings overlap.*

Proof. Suppose that there are two wings $e = wv_{i+1}$ and $e' = w'v_i$ that overlap. Note that $w, w' \in W \setminus V(Q)$, $w \neq w'$, and $e_i = v_iv_{i+1} \in F$. The path P' with edge set $E(P') = \{wv_{i-1}, e, e_i, e', w'v_{i+2}\}$ has three edges in F. And the path P'' of Q with edge set $E(P'') = \{e_{i-1}, e_i, e_{i+1}\}$ has only one edge in F. Note that both P' and P'' are odd. Hence the cycle $\tilde{Q}$ obtained from Q by replacing P'' by P' is odd. As $V(\tilde{Q}) \subseteq W$ and $|E(\tilde{Q}) \cap F| > |E(Q) \cap F|$, we have a contradiction. □

By Claim 2, F induces an obstruction with respect to Q. If $e \in E'$, then by Claim 3, e belongs to a path P of the form either v_i, e, w, f, v_j or v_i, f, w, e, v_j or v_i, e, v_j with $v_i, v_j \in V(Q)$, $f \in F$, and $w \in W \setminus V(Q)$. It then follows by Claims 4 and 5 that e is either a wing or a diagonal. Moreover, by Claim 6, no wings overlap.

Sufficiency. Suppose that F induces an obstruction with respect to an odd cycle Q satisfying (i) and (ii). By Lemma 2.3, F is a dependent set. We will show that $F' = F \setminus \{f\}$ is an independent set for every $f \in F$. Let $W' = V(F')$ and let us set as before $Q = v_1, e_1, v_2, e_2, \ldots, v_k, e_k, v_{k+1}$ with $v_1 = v_{k+1}$. First remark that if f is an edge of $E(Q)$, as F induces an obstruction with respect to Q, at most one edge of F is adjacent to f. Thus Q cannot be a subgraph of $G[W']$. If f links a node not in $V(Q)$ to an unsaturated node, say v, of $V(Q)$, then v is not a node of $G[W']$ and again Q is not a subgraph of $G[W']$.

Now assume, by contradiction, that $F \setminus \{f\}$ is not in $\mathcal{I}(G)$. By Lemma 2.3, $G[W']$ contains an odd cycle, say D. Suppose that D contains an edge $e = v_iv_{i+3}$ which is a diagonal with respect to F and Q. Thus $e_i, e_{i+2} \in F$ and $e_{i-1}, e_{i+1}, e_{i+3} \notin F$. Also, since F is an obstruction, e_i and e_{i+2} are the only edges of F incident to v_i and v_{i+3}, respectively. In consequence, f can be neither e_i nor e_{i+2}. Now if we replace in D e by the path $v_i, e_i, v_{i+1}, e_{i+1}, v_{i+2}, e_{i+2}, v_{i+3}$, we get a new cycle in $G[W']$ which does not contain e and which is still odd. We can reiterate this procedure until we get an odd cycle in $G[W']$, still denoted by D, without diagonals with respect to F and Q.

Suppose that $V(D)$ contains a node $w \notin V(Q)$. As $w \in V(F')$, there is an edge, say g', belonging to $F_2 \cap E(D)$ incident to w. By condition (1) of Definition 3.1, this edge is the only edge of F incident to w. Consequently, there must exist an edge $g \in E(W) \setminus ((F \cup E(Q)) \cap E(D))$ incident to w. By our hypothesis, g is then a wing with respect to Q. Suppose, without loss of generality, that $g = wv_i$ is a forward wing. Hence $g' = wv_{i+2} \in E(D)$. Also note that e_i is the only edge of F incident to v_i, which implies that $f \neq v_i v_{i+1}$. If we replace in D the path v_i, g, w, g', v_{i+2} by $v_i, e_i, v_{i+1}, e_{i+1}, v_{i+2}$, we get an odd cycle in $G[W']$. Moreover, this new cycle does not contain the wing g. So, if we reiterate this procedure, we get an odd cycle D in $G[W']$ which contains neither a diagonal nor a wing with respect to F and Q and whose nodes are all in $V(Q)$. But this implies that D contains only edges of Q, which contradicts the fact that Q is not a subgraph of $G[W']$. □

4. Finding a minimum dependent set. In this section we consider the problem of finding a minimum dependent set in a graph with nonnegative weights. Using the characterization of the minimal dependent sets given in section 3, we will show that this problem reduces to the minimum odd circuit problem in a signed directed graph, and can then be solved in polynomial time. Some polyhedral and algorithmic consequences of this result will be discussed in the next section.

Let $G = (V, E)$ be a graph and $(c(e),\ e \in E)$ a nonnegative weight vector associated with the edges of E. In what follows we are going to construct from G a signed digraph $D = (U, A)$. For convenience we will use the following notation.

For every node u of G, $c(u)$ will denote the minimum weight of an edge incident to u, and e_u a minimum weight edge incident to u. That is, $c(e_u) = c(u)$. Given a minimal dependent set $F \in \mathcal{C}(G)$ of G, we let

$$Q = v_1, e_1, v_2, e_2, \dots, v_k, e_k, v_{k+1}$$

denote the odd cycle of the obstruction induced by F. (Such a cycle exists by Theorem 3.2.) And we suppose that the sequence $e_1, \dots, e_k$ follows the clockwise order. We say that a node $v_i \in V(Q)$ is *left-saturated* (resp., *right-saturated*) if e_{i-1} (resp., e_i) is in F; otherwise v_i is said to be *left-unsaturated* (resp., *right-unsaturated*). Note that, since by Definition 3.1 $E(Q) \setminus F \neq \emptyset$, Q has at least one left-unsaturated node and one right-unsaturated node. If v_i is unsaturated (i.e., left- and right-unsaturated), we denote by f_i the unique edge in F incident to v_i. A node v is said to be a *left-node* (resp., *right-node*) if v is either left-saturated or left-unsaturated (resp., right-saturated or right-unsaturated).

Now we define the signed digraph $D = (U, A)$ (see Figure 4 for an illustration).

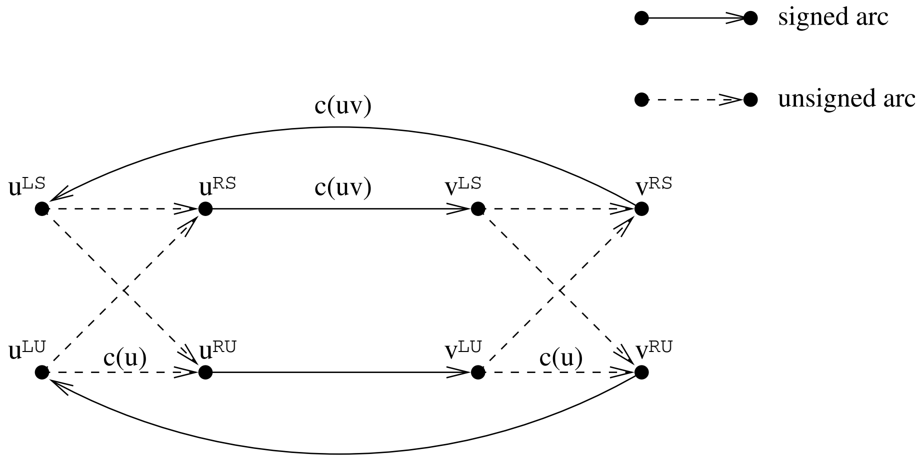


FIG. 4. *The subdigraph of D corresponding to an edge uv of G.*

- For every node $u \in V$, consider four nodes $u^{\mathtt{LS}}, u^{\mathtt{RS}}, u^{\mathtt{LU}}, u^{\mathtt{RU}}$ which correspond to the four possible states of u: left-saturated, right-saturated, left-unsaturated, and right-unsaturated. And consider four unsigned arcs: $u^{\mathtt{LS}}u^{\mathtt{RS}}$, $u^{\mathtt{LS}}u^{\mathtt{RU}}, u^{\mathtt{LU}}u^{\mathtt{RS}}$ with weight 0 and $u^{\mathtt{LU}}u^{\mathtt{RU}}$ with weight $d(u^{\mathtt{LU}}u^{\mathtt{RU}}) = c(u)$.
- For every edge $uv \in E$ consider four signed arcs: $u^{\mathtt{RU}}v^{\mathtt{LU}}, v^{\mathtt{RU}}u^{\mathtt{LU}}$ with weight $d(u^{\mathtt{RU}}v^{\mathtt{LU}}) = d(v^{\mathtt{RU}}u^{\mathtt{LU}}) = 0$ and $u^{\mathtt{RS}}v^{\mathtt{LS}}, v^{\mathtt{RS}}u^{\mathtt{LS}}$ with weight $d(u^{\mathtt{RS}}v^{\mathtt{LS}}) = d(v^{\mathtt{RS}}u^{\mathtt{LS}}) = c(uv)$.

Observe that the tail of an unsigned arc is always a left-node and the head is always a right-node. Also note that any sequence of arcs in D alternates between signed and unsigned arcs. In consequence, any circuit of D has an even number of arcs.

We let Σ denote the set of signed arcs and $U = A \setminus \Sigma$ the set of unsigned arcs. So a circuit Q^D of D is odd if $|A(Q^D) \cap \Sigma|$ is odd.

Let Q^D be a minimum weight odd circuit of D with respect to the weight vector d. Let F be a minimum weight dependent set of G with respect to the weight vector c. In what follows we are going to show that $d(Q^D) = c(F)$. As a consequence, we can compute minimum dependent sets by calculating minimum weight odd circuits in an auxiliary graph.

First we show that $c(F) \leq d(Q^D)$. Let F' be the set of edges uv of G such that either

(i) $u^{\mathtt{RS}}v^{\mathtt{LS}}$ is an arc of Q^D,
(ii) $v^{\mathtt{RS}}u^{\mathtt{LS}}$ is an arc of Q^D, or
(iii) the edge uv is the minimum weight edge e_u incident to u and $u^{\mathtt{LU}}u^{\mathtt{RU}}$ is an arc of Q^D.

The way we defined the cost vector d yields $c(F') \leq d(Q^D)$. Let $w_1, \ldots, w_q$ be the sequence of nodes of Q^D (where the indices are modulo q). Then the sequence $u_1, \ldots, u_{q'}$ of nodes of G, obtained by taking the node u_i if w_i is either $u_i^{\mathtt{LS}}$, $u_i^{\mathtt{RS}}$, $u_i^{\mathtt{LU}}$, or $u_i^{\mathtt{RU}}$ (note that $q' \leq q$), induces a subgraph H of G whose edges correspond to the signed arcs of Q^D. Since Q^D is odd, H contains an odd cycle Q'. Since Q' is a cycle of the graph $G[V(F')]$, by Lemma 2.3, F' is dependent. As F is chosen minimum, $c(F) \leq c(F')$ and therefore

$$c(F) \leq d(Q^D).$$

Now we show that $c(F) \geq d(Q^D)$. Since c is nonnegative we can assume that F is minimal. By Theorem 3.2, F induces an obstruction with respect to an odd cycle Q.

Let $P_1 = v_i, e_i, \ldots, v_{j-1}, e_{j-1}, v_j$ be a path of Q such that all the edges of P_1 are in F and $e_{i-1}, e_j \notin F$. The node v_i is left-unsaturated and right-saturated, the nodes $v_{i+1}, \ldots, v_{j-1}$ are left- and right-saturated, and v_j is left-saturated and right-unsaturated. In the digraph D, the path P_1 corresponds to a path P_1^D with node set $V(P_1^D) = \{v_i^{\mathtt{LU}}, v_i^{\mathtt{RS}}, v_{i+1}^{\mathtt{LS}}, v_{i+1}^{\mathtt{RS}}, \ldots, v_{j-1}^{\mathtt{LS}}, v_{j-1}^{\mathtt{RS}}, v_j^{\mathtt{LS}}, v_j^{\mathtt{RU}}\}$. The arc set of P_1^D is $A(P_1^D) = \{a_i, \sigma_i, \ldots, a_{j-1}, \sigma_{j-1}, a_j\}$, where $a_i, \ldots, a_j \in U$ are unsigned arcs with weight 0 and σ_l is a signed arc with cost $d(\sigma_l) = c(e_l)$ for $l = i, \ldots, j-1$. Thus

$$\begin{aligned} d(P_1^D) &= d(a_i) + d(\sigma_i) + \cdots + d(a_{j-1}) + d(\sigma_{j-1}) + d(a_j) \\ &= c(e_i) + \cdots + c(e_{j-1}) \\ &= c(P_1). \end{aligned}$$

Let $P_2 = v_i, e_i, \ldots, v_{j-1}, e_{j-1}, v_j$ be a path of Q such that no edge of P_2 is in F. The node v_i is right-unsaturated, the nodes $v_{i+1}, \ldots, v_{j-1}$ are left- and right-unsaturated, and v_j is left-unsaturated. In D, there is a path P_2^D with node set

$V(P_2^D) = \{v_i^{\mathtt{RU}}, v_{i+1}^{\mathtt{LU}}, v_{i+1}^{\mathtt{RU}}, \dots, v_{j-1}^{\mathtt{LU}}, v_{j-1}^{\mathtt{RU}}, v_j^{\mathtt{LU}}\}$. The arc set of P_2^D is $A(P_2^D) = \{\sigma_i, a_{i+1}, \sigma_{i+1}, \dots, a_{j-1}, \sigma_{j-1}\}$, where $a_l \in U$ is an unsigned arc with cost $c(v_l)$ for $l = i+1, \dots, j-1$ and $\sigma_i, \dots, \sigma_{j-1}$ are signed arcs with cost 0.

Thus

$$\begin{aligned} d(P_2^D) &= d(\sigma_i) + d(a_{i+1}) + d(\sigma_{i+1}) + \cdots + d(a_{j-1}) + d(\sigma_{j-1}) \\ &= c(v_{i+1}) + \cdots + c(v_{j-1}) \\ &\leq \sum_{l=i+1,\dots,j-1} c(f_l). \end{aligned}$$

(Recall that f_l is the unique edge of F incident to v_l.) Observe now that Q decomposes into paths of types P_1 and P_2. The associated paths of D of types P_1^D and P_2^D form a circuit R^D of D whose weight $d(R^D)$ is less than or equal to $c(F)$. Since $|\Sigma \cap A(R^D)| = |E(Q)|$, R^D is an odd circuit of D. Then $d(Q^D) \leq d(R^D)$, and therefore

$$c(F) \geq d(Q^D).$$

So we can state the following theorem.

THEOREM 4.1. *The minimum dependent set problem with nonnegative weights can be solved in polynomial time.*

5. Polyhedral consequences and concluding remarks. Given a graph $G = (V, E)$, let IBSP(G) be the convex hull of the incidence vectors of the edge sets of induced bipartite subgraphs of G.

Let $\mathcal{P}(G)$ be the polyhedron given by

$$(1) \qquad 0 \leq x(e) \leq 1 \qquad \forall\, e \in E,$$

$$(2) \qquad x(C) \leq |C| - 1 \qquad \forall\, C \in \mathcal{C}(G).$$

Obviously, inequalities (1) and (2) are valid for IBSP(G). Constraints (1) are called the *trivial inequalities*. Constraints (2) will be called the *dependent set inequalities*.

Moreover, we have that MIBSP is equivalent to the integer program

$$\max\ \{wx\ :\ x \in \mathcal{P}(G),\ x \text{ integer}\}.$$

The *separation problem* for a class of inequalities is to decide whether a given vector $\bar{x} \in \mathbb{Q}^E$ satisfies the inequalities and, if not, to find an inequality that is violated by $\bar{x}$.

Given a vector $\tilde{x} \in \mathbb{R}_+^E$, let $\bar{x} \in \mathbb{R}_+^E$ such that $\bar{x}(e) = 1 - \tilde{x}(e)$ for all $e \in E$. Clearly, there is an inequality of type (2) violated by $\tilde{x}$ if and only if the minimum weight of a dependent set with respect to $\bar{x}$ is less than 1. It thus follows by Theorem 4.1 that the separation problem associated with inequalities (2) is solvable in polynomial time. From [15] we then have the following corollary.

COROLLARY 5.1. *The problem*

$$\max\ \{wx\ :\ x \in \mathcal{P}(G)\}$$

can be solved in polynomial time.

A natural question that may be posed is to characterize the graphs for which the polytope $\mathcal{P}(G)$ is integral. As it will turn out, these graphs are precisely the bipartite graphs.

PROPOSITION 5.2. *$\mathcal{P}(G)$ is integral if and only if G is bipartite.*

Proof. If $G = (V, E)$ is bipartite, then $\mathcal{P}(G)$ is given by the trivial inequalities, and hence any extreme point of $\mathcal{P}(G)$ is integer.

Now suppose that $G = (V, E)$ is not bipartite, and let $Q = v_1, e_1, v_2, e_2, \dots, v_{2k+1}, e_{2k+1}, v_{2k+2}$, where $v_{2k+2} = v_1$, be an odd cycle of G. We can assume that Q is a hole. Consider the solution $\bar{x} \in \mathbb{R}^E$ given by

$$\bar{x}(e) = \begin{cases} \frac{k}{k+1} & \text{if } e \in E(Q), \\ 0 & \text{if not.} \end{cases}$$

Let $C_i = \{e_i, e_{i+2}, \dots, e_{i+2k}\}$ for $i = 1, \dots, 2k+1$ (the indices are modulo $2k+1$). Note that $|C_i| = k+1$. By Theorem 3.2, the C_i's are in $\mathcal{C}(G)$. We also have that $\bar{x}$ satisfies the system

$$\begin{aligned} x(C_i) &= |C_i| - 1 \qquad && \text{for } i = 1, \dots, 2k+1, \\ x(e) &= 0 && \forall e \in E \setminus E(Q). \end{aligned}$$

Furthermore it is not hard to see that $\bar{x}$ is the unique solution of that system. Hence $\bar{x}$ is an extreme point of $\mathcal{P}(G)$. □

In contrast with the classical bipartite subgraph problem, the linear relaxation of the MIBSP does not seem to be quite strong. As it has been shown by Guenin [19], for the former problem, the trivial and the cycle inequalities suffice to describe the bipartite subgraph polytope in a large class of graphs (the weakly bipartite graphs) which contains, for instance, planar graphs and bipartite graphs. However, for the MIBSP, as shown by Proposition 5.2, the graphs for which the trivial and the dependent set inequalities completely describe IBSP(G) are reduced to the bipartite graphs. We can also notice that any inequality that is valid for the bipartite subgraph polytope is also valid for the IBSP(G). These inequalities are not, however, so strong for the MIBSP. To see this, consider, for instance, a clique (W, T) on p nodes in a graph G. The inequality $x(T) \le \lfloor \frac{p}{2} \rfloor \lceil \frac{p}{2} \rceil$ is valid for the bipartite subgraph polytope on G and is facet defining [3], whereas, any solution for the MIBSP can take at most one edge from T.

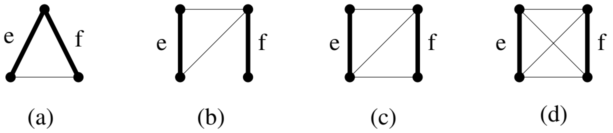


FIG. 5. *The minimal dependent sets of size* 2.

These negative observations motivated us to investigating new valid inequalities for IBSP(G). By Theorem 3.2, $\{e, f\} \in \mathcal{C}(G)$ if and only if the subgraph induced by the endnodes of e, f is one of the four graphs of Figure 5. Let us consider from G an auxiliary graph $A(G)$ whose nodes correspond to the edges of G and such that two nodes e, f are linked by an edge if and only if $\{e, f\} \in \mathcal{C}(G)$. Remark that any independent set of G is a stable set of $A(G)$. (Note that the converse is not true.) Hence the so-called clique and odd cycle inequalities of the stable set polytope of $A(G)$ (see [20]) given by

(3) $\qquad x(K) \le 1 \qquad$ for every clique K of $A(G)$,

(4) $\qquad x(C) \le \dfrac{|C| - 1}{2} \qquad$ for every odd cycle C of $A(G)$

are valid for IBSP(G). Notice that the edge set of a clique of G corresponds to a clique of $A(G)$. (Note that the converse is not true.) As it will turn out, the separation problem for inequalities (3) is NP-hard. The separation problem for inequalities (4) can be solved in polynomial time (see, for instance, [17]).

Proposition 5.3. *The separation problem for inequalities* (3) *is NP-hard.*

Proof. We use a reduction from the maximum clique problem. Let $G = (V, E)$ be a graph. Add a node u and the edges uv for each $v \in V$ to obtain the graph $\tilde{G} = (\tilde{V}, \tilde{E})$. Let $\tilde{x} \in \mathbb{R}^{\tilde{E}}$ given by

$$\tilde{x}(e) = \begin{cases} 1/k & \text{if } e \in \tilde{E} \setminus E, \\ 0 & \text{if } e \in E. \end{cases}$$

Clearly, there is a clique K in $A(\tilde{G})$ with $\tilde{x}(K) > 1$ if and only if there is a clique of size $k+1$ in G. □

Let $G = (V, E)$ be a (nonbipartite) graph and let Q and C_i be as defined in the proof of Proposition 5.2, $i = 1, \ldots, 2k+1$. Observe that e belongs to $k+1$ different C_i's for each $e \in E(Q)$. As the C_i's are dependent sets in G, the following inequalities are valid for IBSP(G):

$$x(C_i) \le k \qquad \text{for } i = 1, \ldots, 2k+1.$$

By summing these inequalities, we obtain the inequality

$$(k+1)x(E(Q)) \le k(2k+1).$$

Therefore the inequalities

$$(5) \qquad x(E(Q)) \le |E(Q)| - 2, \qquad \text{for every odd cycle } Q \text{ of } G,$$

are valid for IBSP(G). Inequalities (5) also arise naturally since any independent set of G uses at most $|V(Q)| - 1$ nodes (and thus at most $|E(Q)| - 2$ edges) of Q. Note that inequalities (5) are different from the inequalities induced by the odd cycles of $A(G)$. (If, for instance, $G = (V, E)$ is an odd cycle with edge set, say $E = \{e_1, \ldots, e_5\}$, then $A(G)$ has no edge, while G produces the inequality $x(E) \le 3$ of type (5) which is facet defining.) Inequalities (5) will be called *cycle inequalities*.

Proposition 5.4. *The separation problem for inequalities* (5) *can be solved in polynomial time.*

Proof. Let $\bar{x}$ be a vector associated with the edges of G. We may suppose that $\bar{x}$ satisfies the trivial inequalities. Let $y \in \mathbb{R}^E$ such that $y(e) = 1 - \bar{x}(e)$ for all $e \in E$. An inequality (5) is violated by $\bar{x}$ if and only if $y(E(Q)) < 2$. Thus the separation problem for inequalities (5) reduces to finding a minimum odd cycle in G with respect to the weight vector y. As $y(e) \ge 0$ for all $e \in E$, this can be done in polynomial time as shown in [18]. □

It would be interesting to determine when the dependent, cycle, and clique inequalities are facet defining for the polytope IBSP(G).

The approach presented in this paper can be adapted to handle the maximum induced forest and the maximum induced acyclic subgraph problems (see [8]).

Acknowledgments. We would like to thank the anonymous referees for their constructive comments. We are grateful to Hervé Kerivin for many comments that have improved the presentation.

REFERENCES

[1] F. Barahona, *On the Complexity of Max Cut*, Rapport de recherche 186, IMAG, Université Joseph Fourier, Grenoble, France, 1980.

[2] F. Barahona, *On some weakly bipartite graphs*, Oper. Res. Lett., 2 (1983), pp. 107–111.

[3] F. Barahona, M. Grötschel, and A. R. Mahjoub, *Facets of the bipartite subgraph polytope*, Math. Oper. Res., 10 (1985), pp. 340–358.

[4] F. Barahona and A. R. Mahjoub, *Facets of the balanced (acyclic) induced subgraph polytope*, Math. Programming, 45 (1989), pp. 21–33.

[5] F. Barahona and A. R. Mahjoub, *Compositions of graphs and polyhedra* I: *Balanced induced subgraphs and acyclic subgraphs*, SIAM J. Discrete Math., 7 (1994), pp. 344–358.

[6] H.-A. Choi, K. Nakajima, and C. S. Rim, *Graph bipartization and via minimization*, SIAM J. Discrete Math., 2 (1989), pp. 38–47.

[7] D. Cornaz and J. Fonlupt, *Chromatic characterization of biclique covers*, Discrete Math., 306 (2006), pp. 495–507.

[8] D. Cornaz, H. Kerivin, and A. R. Mahjoub, *The Maximum Induced Forest and Acyclic Subgraph Problems*, LIMOS, Université Blaise Pascal, Clermont-Ferrand, France, in preparation.

[9] E. W. Dijkstra, *A note on two problems in connection with graphs*, Numer. Math., 1 (1959), pp. 269–271.

[10] J. Fonlupt, A. R. Mahjoub, and J. P. Uhry, *Composition in the bipartite subgraph polytope*, Discrete Math., 105 (1992), pp. 71–91.

[11] P. Fouilhoux, *Graphs k-partis et conception de circuits VLSI*, Ph.D. dissertation, D. U. 1555, Université Blaise Pascal, Clermont-Ferrand, France.

[12] P. Fouilhoux and A. R. Mahjoub, *Polyhedral results for the bipartite induced subgraph problem*, Discrete Appl. Math., 154 (2006), pp. 2128–2149.

[13] P. Fouilhoux and A. R. Mahjoub, *Bipartization of Graphs, VLSI Design and DNA Sequencing*, LIMOS, Université Blaise Pascal, Clermont-Ferrand, France, preprint, 2005.

[14] M. R. M. Garey and D. S. Johnson, *Computers and Intractibility. A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, 1979.

[15] M. Grötschel, L. Lovasz, and A. Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 169–197.

[16] M. Grötschel, L. Lovasz, and A. Schrijver, *Polynomial algorithms for perfect graphs*, Ann. Discrete Math., 21 (1984), pp. 325–356.

[17] M. Grötschel, L. Lovasz, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag, Berlin, 1988.

[18] M. Grötschel and W. R. Pulleyblank, *Weakly bipartite graphs and the max-cut problem*, Oper. Res. Lett., (1981), pp. 23–27.

[19] B. Guenin, *A characterization of weakly bipartite graphs*, J. Combin. Theory Ser. B, (2001), pp. 112–168.

[20] M. Padberg, *On the facial structure of set packing polyhedra*, Math. Programming, 5 (1973), pp. 199–215.

SIAM J. DISCRETE MATH.
Vol. 21, No. 3, pp. 676–687

ON THE CHROMATIC NUMBER OF GEOMETRIC HYPERGRAPHS[*]

SHAKHAR SMORODINSKY[†]

Abstract. A finite family $\mathcal{R}$ of simple Jordan regions in the plane defines a hypergraph $H = H(\mathcal{R})$ where the vertex set of H is $\mathcal{R}$ and the hyperedges are all subsets $S \subset \mathcal{R}$ for which there is a point p such that $S = \{r \in \mathcal{R} | p \in r\}$. The chromatic number of $H(\mathcal{R})$ is the minimum number of colors needed to color the members of $\mathcal{R}$ such that no hyperedge is monochromatic. In this paper we initiate the study of the chromatic number of such hypergraphs and obtain the following results: (i) Any hypergraph that is induced by a family of n simple Jordan regions such that the maximum union complexity of any k of them (for $1 \leq k \leq m$) is bounded by $\mathcal{U}(m)$ and $\frac{\mathcal{U}(m)}{m}$ is a nondecreasing function is $O(\frac{\mathcal{U}(n)}{n})$-colorable. Thus, for example, we prove that any finite family of pseudo-discs can be colored with a constant number of colors. (ii) Any hypergraph induced by a finite family of planar discs is four colorable. This bound is tight. In fact, we prove that this statement is equivalent to the four-color theorem. (iii) Any hypergraph induced by n axis-parallel rectangles is $O(\log n)$-colorable. This bound is asymptotically tight. Our proofs are constructive. Namely, we provide deterministic polynomial-time algorithms for coloring such hypergraphs with only "few" colors (that is, the number of colors used by these algorithms is upper bounded by the same bounds we obtain on the chromatic number of the given hypergraphs). As an application of (i) and (ii) we obtain simple constructive proofs for the following: (iv) Any set of n Jordan regions with near linear union complexity admits a conflict-free (CF) coloring with polylogarithmic number of colors. (v) Any set of n axis-parallel rectangles admits a CF-coloring with $O(\log^2(n))$ colors.

Key words. hypergraphs, conflict-free, coloring, wireless

AMS subject classification. combinatorics

DOI. 10.1137/050642368

1. Introduction.

A *hypergraph* is a pair $(V, \mathcal{E})$, where V is a finite set and $\mathcal{E} \subset 2^V$. The set V is called the *ground set* or the *vertex set* and the elements of $\mathcal{E}$ are called *hyperedges*. A k-coloring of a hypergraph $H = (V, \mathcal{E})$, for some positive integer k, is a function $\chi : V \to \{1, 2, \ldots, k\}$ such that no $S \in \mathcal{E}$ with $|S| \geq 2$ is monochromatic. Let $\chi(H)$ denote the minimum integer k for which H has a k-coloring. $\chi(H)$ is called the *chromatic number* of H.

Let $\mathcal{R}$ be a set of regions in the plane. For a point $p \in \cup_{r\in\mathcal{R}} r$, put $r(p) = \{r \in \mathcal{R} \mid p \in r\}$. Let $H(\mathcal{R})$ denote the hypergraph $(\mathcal{R}, \{r(p) \mid p \in \cup_{r\in\mathcal{R}}\})$. We say that $H(\mathcal{R})$ is the hypergraph *induced* by $\mathcal{R}$.

DEFINITION 1.1. *Let $\mathcal{R}$ be a family of n simple Jordan regions in the plane. The* union complexity *of $\mathcal{R}$ is the number of vertices (i.e., intersection of boundaries of pairs of regions in $\mathcal{R}$) that lie on the boundary $\partial \bigcup_{r\in\mathcal{R}} r$.*

In this work we initiate the study of the chromatic number of hypergraphs that are induced by simple geometric regions such as discs, pseudo-discs, axis-parallel rectangles, etc. Our main result (section 5) is a theorem correlating the chromatic number of the underlying hypergraphs with the union complexity of the regions inducing those hypergraphs. Specifically, we prove the following theorem.

[*] Received by the editors October 10, 2005; accepted for publication (in revised form) March 13, 2007; published electronically September 5, 2007. This work was supported by the NSF Mathematical Sciences Postdoctoral Fellowship award 0402492.

http://www.siam.org/journals/sidma/21-3/64236.html

[†] Courant Institute for Mathematical Sciences, New York University, 251 Mercer St, New York, NY 10012.

THEOREM 1.2. *Let $\mathcal{R}$ be a set of n regions and let $\mathcal{U} : \mathbb{N} \to \mathbb{N}$ be a function such that $U(m)$ is the maximum complexity of any k regions in $\mathcal{R}$ over all $k \leq m$, for $1 \leq m \leq n$. We assume that $\frac{\mathcal{U}(m)}{m}$ is a nondecreasing function. Then, $\chi(H(\mathcal{R})) = O(\frac{\mathcal{U}(n)}{n})$. Furthermore, such a coloring can be computed in polynomial time.*

In section 3 we study the chromatic number of a hypergraph that is induced by discs and prove the following theorem.

THEOREM 1.3. *Let $\mathcal{D}$ be a finite family of discs in the plane. Then the hypergraph $H(\mathcal{D})$ is four colorable.*

As one can easily see, this bound is tight by taking four pairwise touching discs. In such a case, every pair of discs needs to be colored with distinct colors. This bound is somewhat surprising in the following sense. In the restricted case where we are given a finite family $\mathcal{R}$ of discs such that every two are either touching (i.e., the boundaries of the two discs share a common point but the interiors of the discs are disjoint) or disjoint, it is easy to see that bounding the chromatic number of $H(\mathcal{R})$ by four is equivalent to bounding the chromatic number of the "kissing" graph induced by the discs (i.e., the vertex set of the graph is $\mathcal{R}$ and the edges are the touching pairs) by four. However, this graph is planar. On the other hand, a classical theorem due to Koebe [12] asserts that every planar graph can be realized as a kissing discs graph. In section 3 we show that the four-color theorem is equivalent to coloring a hypergraph induced by a finite family of discs (not necessarily interior disjoint but also discs that might have arbitrary overlaps) with at most four colors. As mentioned above, one direction of the proof easily follows from Koebe's theorem [12].

In section 4 we study the chromatic number of a hypergraph induced by n axis-parallel rectangles and prove the following theorem.

THEOREM 1.4. *Let $\mathcal{R}$ be a family of n axis-parallel rectangles. Then $\chi(H(\mathcal{R})) = O(\log n)$.*

This bound is asymptotically tight as demonstrated recently by a lower bound construction of Pach and Tardos [14].

To the best of our knowledge, these questions were not addressed previously. Beyond their purely theoretical interest, we apply our results to obtain a simple framework for tackling the problems of conflict-free colorings.

DEFINITION 1.5 (see [7, 16]). *A coloring of regions is* conflict-free *(CF) if for any covered point in the plane, there exists a region that covers it with a unique color (i.e., no other region covering that point has the same color).*

In section 6 we show how to apply our results on proper colorings of regions to obtain simple deterministic, polynomial-time algorithms for CF-colorings of those regions.

CF-coloring problems were recently introduced in [7, 16] in the context of frequency assignment in cellular networks. In addition to this practical motivation, this new coloring model has drawn much attention of researchers through its own theoretical interest and such colorings have been the focus of several recent papers [1, 4, 6, 8, 9, 10, 13].

Even et al. [7] have shown that any family of n discs in the plane admits a CF-coloring with only $O(\log n)$ colors and that this bound is tight in the worst case. Furthermore, such a coloring can be computed in deterministic polynomial time.[1] The results of Even et al. [7] were further extended by Har-Peled and Smorodinsky

[1] In [7] it is shown that finding the minimum number of colors needed to CF-color a given collection of discs is NP-hard even when all discs are congruent, and an $O(\log n)$ approximation-ratio algorithm is provided.

[9] by combining more involved probabilistic and geometric ideas. The main result of [9] is a randomized algorithm which CF-colors any set of n pseudo-discs with $O(\log n)$ colors with high probability. As an application of our main result, we obtain a simple deterministic, polynomial-time algorithm for CF-colorings regions. The performance (i.e., the number of colors used by our algorithm) depends on the union complexity of the underlying regions. For example, we obtain a simple determnistic, polynomial-time algorithm that CF-colors any set of n pseudo-discs with only $O(\log n)$ colors.

2. Preliminaries. We start with some basic definitions and lemmas.

Definition 2.1. *The* Delaunay graph $G(H)$ *of a hypergraph* $H = (V, \mathcal{E})$ *is a simple graph* $G = (V, E)$*, where the edge set* E *is defined as* $E = \{(x, y) \mid \{x, y\} \in \mathcal{E}\}$ *(i.e.,* G *is the graph on the vertex set* V *whose edges consist of all hyperedges in* H *of cardinality two).*

Lemma 2.2. *For every hypergraph* H *we have*

$$\chi(G(H)) \leq \chi(H).$$

Proof. Simply because any proper coloring of the vertices of H is also a proper coloring for $G(H)$. □

Definition 2.3. *We say that a hypergraph* $H = (V, \mathcal{E})$ *has rank* i *for* $i \geq 2$ *if for any hyperedge* $S \in \mathcal{E}$ *with* $|S| > i$ *there exists a hyperedge* $S' \in \mathcal{E}$ *such that* $S' \subset S$ *and* $|S'| = i$.

Lemma 2.4. *Let* $H = (V, \mathcal{E})$ *be a hypergraph of rank two. Then* $\chi(H) = \chi(G(H))$.

Proof. By Lemma 2.2 we have $\chi(G(H)) \leq \chi(H)$. It remains to prove that $\chi(H) \leq \chi(G(H))$. Let χ be a proper coloring of $G(H)$ with $k = \chi(G(H))$ colors. This coloring is also a proper coloring of H. Indeed, let $e \in \mathcal{E}$ be a hyperedge with cardinality > 1. Then, by assumption, there exists an edge $e' \subset e$ in $G(H)$ and this edge is nonmonochromatic. Then, obviously, e is nonmonochromatic. This completes the proof of the lemma. □

Definition 2.5. *A simple graph* $G = (V, E)$ *is called* k*-degenerate for some positive integer* k*, if every (vertex-induced) subgraph of* G *has a vertex of degree at most* k.

Lemma 2.6. *Let* $G = (V, E)$ *be a* k*-degenerate graph. Then* $\chi(G) \leq k + 1$.

Proof. Proceed by induction on $n = |V|$. Let $v \in V$ be a vertex of degree at most k. By the induction hypothesis, the graph $G \setminus v$ (obtained by removing v and all of its incident edges from G) is $k + 1$ colorable. Since v has at most k neighbors there is always a free color that can be assigned to v which is distinct from the colors of its neighbors. □

3. Hypergraphs induced by discs. In this section we show that any hypergraph induced by a finite family of discs in the plane is four colorable.

Let H^+ denote the set of all positive halfspaces in $\mathbb{R}^3$ (i.e., those halfspaces consisting of all points that lie above some fixed plane). For a given set $P \subset \mathbb{R}^3$, put $H^+(P) = \{h \cap P | h \in H^+\}$.

A transformation to points and half-spaces. In what follows, we show that the problem of coloring n arbitrary discs in the plane reduces (but is not equivalent) to that of coloring a set of points P in $\mathbb{R}^3$ with respect to the set of ranges $H^+(P)$ (i.e., coloring the hypergraph $H = (P, H^+(P))$).

We transform a point $p = (a, b)$ in $\mathbb{R}^2$ to the plane p^* in $\mathbb{R}^3$, with the parametrization $z = -2ax - 2by + a^2 + b^2$ and transform a disc S in $\mathbb{R}^2$, with center (x, y) and radius $r \geq 0$, to the point S^* in $\mathbb{R}^3$, with coordinates $(x, y, r^2 - x^2 - y^2)$.

It is easily seen that in this transformation a point $p \in \mathbb{R}^2$ lies inside (respectively, on the boundary of, outside) a disc S, if and only if the point $S^* \in \mathbb{R}^3$ lies above (respectively, on, below) the plane p^*. Indeed, assume that point $p = (a, b)$ lies inside (respectively, on the boundary of, outside) the disc S with center (x, y) and radius r. This can be formulated by the inequality: $(a - x)^2 + (b - y)^2 < r^2$ or $-2ax - 2by + a^2 + b^2 < r^2 - x^2 - y^2$ (respectively, an equality $=$, or inequality with $>$), which is equivalent to that of the point $(x, y, r^2 - x^2 - y^2) = S^*$ lies above (respectively on, or below) the plane $z = -2ax - 2by + a^2 + b^2$ (which is the dual p^* of p), as asserted.

Given a collection $\mathcal{S} = \{S_1, \ldots, S_n\}$ of n distinct discs in the plane, one can use the above transformation to obtain a collection $\mathcal{S}^* = \{S_1^*, \ldots, S_n^*\}$ of n points in $\mathbb{R}^3$, such that any valid coloring of $\mathcal{S}^*$ with respect to $H^+(\mathcal{S}^*)$ with k colors induces a coloring of the discs of $\mathcal{S}$ with the same set of k colors.

Remark. We note that the two coloring problems are not equivalent. Indeed, the set of all planes in $\mathbb{R}^3$ that are images (under the above transformation) of points in the plane are such that they are all tangent to the paraboloid $z = -x^2 - y^2$. Since we color the points in $\mathbb{R}^3$ so that *any* positive halfspace is not monochromatic (not only positive halfspaces bounded by planes which are tangent to the paraboloid), we actually result in a coloring of a hypergraph with potentially more hyperedges than the original hypergraph.

LEMMA 3.1. *Let $P \subset \mathbb{R}^3$ be a finite set. Let H be the hypergraph induced by $H^+(P)$ (that is, $H = (P, H^+(P))$). Then $\chi(H) \leq 4$.*

Proof. Recall that $G(H)$ is the graph whose vertex set is P and whose edge set is $E = \{h \cap P \mid h \in H^+ \text{ and } |h \cap P| = 2\}$. Thus G, contains the skeleton graph of the upper convex hull of P, although G may contain additional edges. The rank of the hypergraph H is 2. Indeed, every subset $P' \in H^+(P)$ such that $|P'| > 2$ must contain an edge of E. To see this, let $h \in H^+$ be a positive halfspace containing at least three points of P. Without loss of generality, assume that the plane π bounding h is in "general position" with respect to the points of P (i.e., no line passing through two points of P is parallel to π). This can be achieved by a proper perturbation of π such that the set of points above the perturbed plane does not change. Then, we can translate π upwards keeping the translated plane π' parallel to π until the positive halfspace bounded by π' contains exactly two points of P'. By definition, these pair of points form an edge in G, so the rank of H is indeed 2. By Lemma 2.4, it is enough to color the vertices of G properly (i.e., such that no color class contains an edge). We will show that G is a planar graph and by the Four-Color Theorem (see, e.g., [2, 3]) it is four colorable. To show that G is planar, we project P onto the plane orthogonally and draw the graph G using straight line segments to represent the edges. We want to show that in this drawing there are no crossings. Assume to the contrary that there are two edges $e_1 = (p_1, q_1), e_2 = (p_2, q_2)$ whose projections cross. Let l be the line parallel to the z-axis that passes through this crossing point. Since $e_1, e_2 \in E$ belong to G, there exists a plane π_1 (respectively, π_2) such that the positive halfspace bounded by π_1 (respectively, π_2) contains only e_1 (respectively, only e_2). l must pass through a point v_1 on the line segment (in $\mathbb{R}^3$) connecting p_1, q_1 and a point v_2 on the line segment connecting p_2, q_2. Assume without loss of generality that v_1 is below v_2. See Figure 1 for an illustration. It is easy to see that π_1 intersect l in a point q that is below v_1. Indeed, since p_1 and q_1 are above π_1 (recall that p_1 and q_1 are the only points of P above π_1) then (by convexity) also the point v_1 is above π_1. Thus q is also below v_2. However, we know that both p_2 and q_2 lie below π_1 (π_1

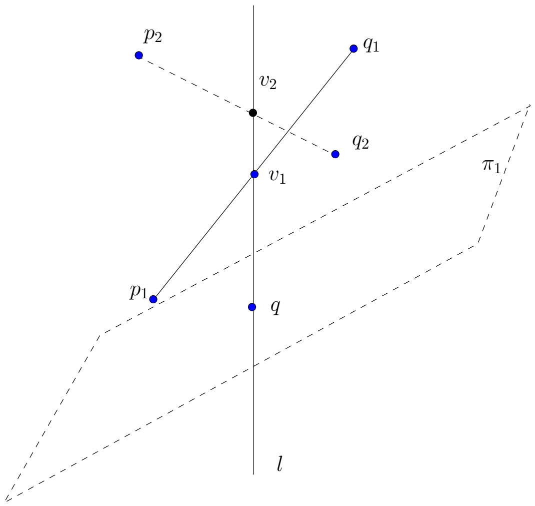


Fig. 1. *Illustration of the proof of Lemma* 3.1 *with the two edges* e_1, e_2*, whose projection cross. The plane* π_1 *must intersect* l *below* v_1 *but above* v_2*, a contradiction.*

separates p_1, q_1 from the rest of the points in P). This means that the line segment connecting p_2, q_2 is also below π_1 which means that v_2 is below q, a contradiction. Thus G is a planar graph and therefore four-colorable. This completes the proof of the lemma. □

Proof of Theorem 1.3. We use the above "lifting transformation" such that the discs are transformed into points in $\mathbb{R}^3$. By Lemma 3.1, there is a coloring of the transformed points with four colors, such that any positive halfspace that contains at least two of these points contains at least two points with distinct colors. We use the same coloring for the preimages of the points and obtain a valid coloring for the hypergraph $H(\mathcal{D})$. □

Remark. It is not clear how to obtain a different proof of Theorem 1.3 without the lifting transformation. The major problem is that $H(\mathcal{D})$ may have rank greater than two. Indeed, if a point p is contained in at least three discs of $\mathcal{D}$, it does not necessarily imply that two of those discs have a point common only to them. This is illustrated in Figure 2. In section 5 we obtain a general upper bound on the chromatic number of regions with low union complexity. Discs are an example of such regions. Therefore, section 5 provides a different way to obtain an upper bound. However, the method we develop in section 5 will only imply an upper bound of six on the chromatic number of discs.

4. Axis-parallel rectangles.

In this section we deal with coloring axis-parallel rectangles. We show that any hypergraph that is induced by a family of n axis-parallel rectangles admits an $O(\log n)$ coloring. This bound is asymptotically tight.

We show that the maximum number of colors $f(n)$ needed to color n axis-parallel rectangles satisfies the recursion $f(n) \leq 8 + f(\frac{n}{2})$, and thus implies the asserted bound. We start with a lemma concerning a restricted case when all rectangles of $\mathcal{R}$ intersect some vertical line.

Lemma 4.1. *Let* $\mathcal{R}$ *be a finite family of axis-parallel rectangles all of which intersect some vertical line* l*. Then* $\chi(H(\mathcal{R})) \leq 8$.

Proof. We assume that the rectangles are in general position in the sense that no

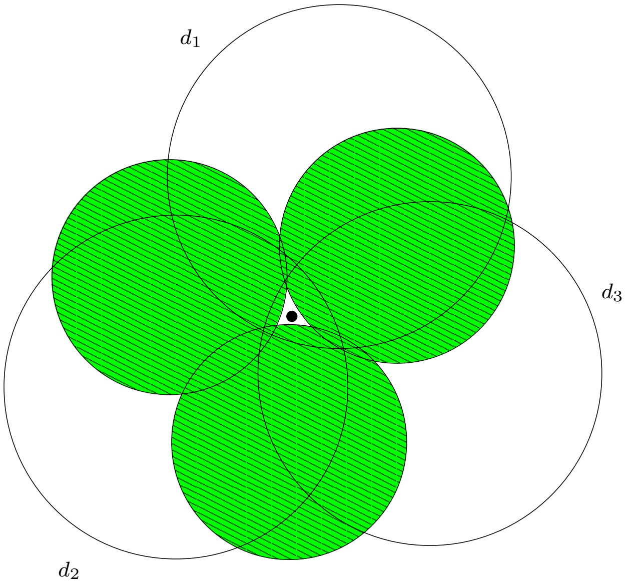


FIG. 2. *An example of a set $\mathcal{D}$ of six discs* (taken from [16]) *whose induced hypergraph $H(\mathcal{D})$ has rank three; there is a point covered only by the three discs d_1, d_2, and d_3. However, there is no point that is covered by two of those discs and no other disc. Note that not every coloring of the Delaunay graph $G(H(\mathcal{D}))$ induces a valid coloring of $H(\mathcal{D})$. Indeed in a coloring of $G(H(\mathcal{D}))$, the discs d_1, d_2, d_3 might get the same color. However, then we would have the hyperedge $\{d_1, d_2, d_3\}$ monochromatic.*

two vertical (or two horizontal) sides share a point. This can be achieved by a small perturbation which does not decrease the number of colors needed. It is easily seen that in this case the hypergraph $H = H(\mathcal{R})$ has rank two. Therefore, by Lemma 2.4, it is enough to show that $\chi(G(H)) \leq 8$. We will show that $G = G(H)$ is 7-degenerate and therefore by Lemma 2.6 is 8-colorable. It is sufficient to argue that the average degree of every (vertex-induced) subgraph of G is less than 8. Let p be a point that is covered by exactly two rectangles $r_1, r_2 \in \mathcal{R}$. Assume without loss of generality that p is to the right of l (see Figure 3). We will charge the pair r_1, r_2 to one of the horizontal sides of one of the rectangles of $\mathcal{R}$ so that each horizontal side is charged at most twice. We translate p to the right until it reaches a vertical side of one of the rectangles r_1, r_2. Assume without loss of generality that this is the side of r_1. Then we move downward until we reach a horizontal side e of some rectangle at a point p'. The important fact is that the horizontal line segment connecting the line l to the point p' is contained in $r_1 \cap r_2$ (we consider the rectangles in $\mathcal{R}$ as closed regions). We charge the pair r_1, r_2 to e. We show that to the right of l at most one charge can occur at such a side. There are two cases to consider: The side e is an upper horizontal side of some rectangle r_3 (note that it cannot be a lower horizontal side of r_3 since then p would have belonged to $r_1 \cap r_2 \cap r_3$). Indeed, assume that e is charged twice to the right of l and that the other charge can occur at a point p''. Assume without loss of generality that p'' is to the left of p'. It is easily seen that p'' belongs to $r_1 \cap r_2$ and therefore could not belong to any other rectangle of $\mathcal{R}$. The second case is when e is a lower horizontal side of either r_1 or r_2. In a similar manner it is easily seen that such a side can be charged at most once to the right of l. Altogether we charge each horizontal side of a rectangle in $\mathcal{R}$ at most once to the right of l. A symmetric argument implies that every horizontal side of a rectangle in $\mathcal{R}$ is charged at most once to the left of l. Altogether we have at most $4n$ charges. We have just shown that $G(H)$ has at most $4n$ edges. As a matter of fact, G has at most $4n - 4$ edges since the uppermost (respectively, the lowermost) horizontal side of the rectangles in $\mathcal{R}$ cannot be charged. Thus, the average degree of G is at most $8 - \frac{8}{n}$ and therefore there must exist a vertex with degree at most 7. Obviously, this charging scheme works for any subgraph of G as well. Thus G is 7-degenerate. By Lemmas 2.6 and 2.4, H is 8-colorable, as asserted. □

Proof of Theorem 1.4. Let l be a vertical line such that at most $n/2$ of the rectangles in $\mathcal{R}$ lie fully to the right of l and at most $n/2$ rectangles of $\mathcal{R}$ lie fully to its left. Let $\mathcal{R}'$ (respectively, $\mathcal{R}''$) denote the subset of rectangles that lie to the right (respectively, to the left) of l. Let $\mathcal{R}_l$ denote the subset of rectangles in $\mathcal{R}$ that intersect the line l. Let $f(n)$ denote the maximum number of colors needed to color a family of n axis-parallel rectangles in the plane. We color (separately) the rectangles in $\mathcal{R}_l$ with eight colors and recursively color the set $\mathcal{R}'$ and $\mathcal{R}''$ using the same set of colors but keeping this set disjoint from the colors used to color $\mathcal{R}_l$. Thus $f(n)$ obeys the recursive relation

$$f(n) \leq 8 + f(n/2),$$

which is easily seen to imply that $f(n) \leq 8 \log n$. This completes the proof of the theorem. □

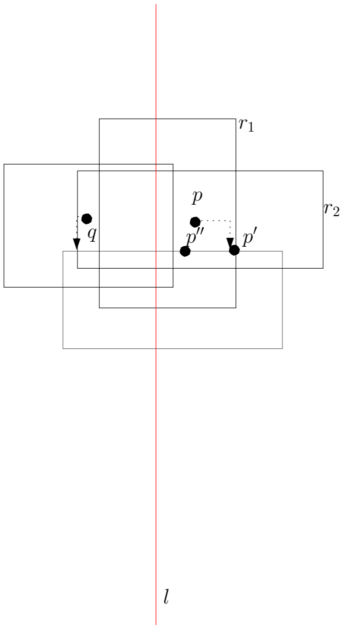


FIG. 3. *Illustration of the charging scheme in the proof of Lemma* 4.1. *Note that here the Delaunay graph of the four rectangles is the clique K_4, so any coloring must use at least four colors.*

5. Chromatic number of regions with low-union complexity. In this section we show a relation between the chromatic number of a hypergraph induced by a finite family of regions $\mathcal{R}$ to the union complexity of $\mathcal{R}$. For example, we show that if $\mathcal{R}$ is a family of planar simple Jordan regions such that any finite subset of $\mathcal{R}$ has linear union complexity, then there exists a constant $c = c(\mathcal{R})$ such that any hypergraph induced by a finite subset of $\mathcal{R}$ is c colorable. Thus, for example, since pseudo-discs have linear union complexity (see, e.g., [11]), there is a constant c such that any family of pseudo-discs can be colored with c colors.

Lemma 5.1. *Let $\mathcal{R}$ be a set of n regions and let $\mathcal{U} : \mathbb{N} \rightarrow \mathbb{N}$ be a function such that $U(m)$ is the maximum complexity of any k regions in $\mathcal{R}$ over all $k \leq m$, for $1 \leq m \leq n$. Then, the Delaunay graph G of the hypergraph $H = H(\mathcal{R})$ has a vertex with degree at most $O(\frac{\mathcal{U}(n)}{n})$.*

Proof. Let $\mathcal{A}(\mathcal{R})$ denote the arrangement of the boundary curves of the regions in $\mathcal{R}$ and let F_2 denote the set of faces of $\mathcal{A}(\mathcal{R})$ that are contained in exactly two regions of $\mathcal{R}$. Obviously, the number of edges in G is bounded by $|F_2|$. We may assume that the regions of $\mathcal{R}$ are in general position, in the sense that no three distinct boundaries pass through a common point. This can be enforced by a slight perturbation of the curves, which does not decrease $|F_2|$. Let $S_{\leq 2}(\mathcal{R})$ be the set of vertices of the arrangement $\mathcal{A}(\mathcal{R})$ that lie in the interior of at most 2 regions of $\mathcal{R}$. By the analysis of Clarkson and Shor [5], we have $|S_{\leq 2}(\mathcal{R})| = O(\mathcal{U}(n))$. We charge a face $f \in F_2$ to one of the vertices on the boundary ∂f, if ∂f has vertices. Thus, the only faces unaccountable for by this charging scheme are the faces that have no vertices on their boundary. However, the number of such faces is only $O(n)$, as we can charge such a face to the region of $\mathcal{R}$ that forms its outer boundary. It is easily seen that in this charging scheme a vertex is charged at most four times, since it can belong to the boundary of at most four faces. Note also that every charged vertex is contained in at most two regions of $\mathcal{R}$ and therefore belongs to $S_{\leq 2}(\mathcal{R})$. Thus $E(G) \leq |F_2| \leq 4 \cdot S_{\leq 2}(\mathcal{R}) + n = O(\mathcal{U}(n) + n)$. Thus, the average degree of G is $O(\frac{\mathcal{U}(n)}{n} + 1)$ and therefore G must contain a vertex with degree at most $O(\frac{\mathcal{U}(n)}{n})$ as asserted. □

We are ready to prove Theorem 1.2.

Proof of Theorem 1.2. By Lemma 5.1 there exists a constant c such that the Delaunay graph G of $H(\mathcal{R})$ has a vertex with degree at most $c \cdot \frac{\mathcal{U}(n)}{n}$. We prove that $\chi(H(\mathcal{R})) \leq c \cdot \frac{\mathcal{U}(n)}{n} + 1$. The proof is by induction on n. Let $r \in \mathcal{R}$ be a region with at most $c \cdot \frac{\mathcal{U}(n)}{n}$ neighbors in G. By the induction hypothesis, the hypergraph $H(\mathcal{R} \setminus \{r\})$ is $c \cdot \frac{\mathcal{U}(n-1)}{n-1} + 1 \leq c \cdot \frac{\mathcal{U}(n)}{n} +$ 1-colorable (by our monotonicity assumption on $\frac{\mathcal{U}(n)}{n}$). We need to choose a color (out of the $c \cdot \frac{\mathcal{U}(n)}{n} + 1$ colors that are available for us) for r such that the coloring of $\mathcal{R}$ is valid. Obviously, points that are not covered by r are not affected by the coloring of r. Note also that any point $p \in r$ that is contained in at least two regions of $\mathcal{R} \setminus r$ is not affected by the color of r since by induction the set of regions in $\mathcal{R} \setminus \{r\}$ containing such points is nonmonochromatic. We thus only need to color r with a color that is different from the colors of all regions $r' \in \mathcal{R} \setminus r$ for which there is a point p that is contained only in $r \cap r'$. However, by our choice of r, there are at most $c \cdot \frac{\mathcal{U}(n)}{n}$ such regions. Thus we can assign to r a color among the $c \cdot \frac{\mathcal{U}(n)}{n} + 1$ colors available to us and keep the coloring of $\mathcal{R}$ proper. This completes the inductive step. As for the algorithmic perspective, we briefly sketch the simple ideas behind it. Here, we do not attempt to optimize the efficiency of the algorithm. Assume a model of computation as in [15] in which computing the intersection points of any pair of regions in $\mathcal{R}$, and a few similar operations, can be performed in constant time. One can compute the arrangement $\mathcal{A}(\mathcal{R})$ using standard methods as in [15]. In addition, one can compute in polynomial time, for each face f of the arrangement

$\mathcal{A}(R)$, its depth $d(f)$ which is the number of ranges in $\mathcal{R}$ containing f. Next, we compute the graph $G(H(\mathcal{R}))$. Note that the edges of $G(H(\mathcal{R}))$ consist of all pairs of regions (r_1, r_2) whose intersection contains a face of depth two. This can be done by checking for each face f of the arrangement and each region $r \in \mathcal{R}$ whether $f \subset r$. This takes time which is proportional to $\mathcal{A}(\mathcal{R}) \cdot n$. Let $r \in \mathcal{R}$ be a vertex of minimum degree in $G(H(\mathcal{R}))$. We update the depth of the faces of the arrangement $\mathcal{A}(\mathcal{R} \setminus \{r\})$ and construct $G(H(\mathcal{R} \setminus \{r\}))$ and color $\mathcal{R} \setminus \{r\}$, recursively. This can be done in total time proportional to the sum $\sum_{i=1}^{n} i f(i)$, where $f(i)$ is the maximum complexity of the arrangement of any i regions in $\mathcal{R}$. Thus if f is polynomial then the total running time is polynomial in n. See Algorithm 1 for a pseudo-code. $\square$

Algorithm 1. Color($\mathcal{R}$): Color the hypergraph $H(\mathcal{R})$.

```
1: Order the elements of $\mathcal{R}$: Compute a permutation $\mathcal{R} = \{r_1, \ldots, r_n\}$ such
   that the degree of $r_i$ in the Delaunay graph $G(H(\{r_1, \ldots, r_i\}))$ is bounded
   by $c \cdot \frac{\mathcal{U}(i)}{i}$, for $i = \{1, \ldots, n\}$.
2: Color the elements: For($i = 2; i \leq n; i++$) $r_i \leftarrow$ A color different from its
   neighbors in $G(H(\{r_1, \ldots, r_i\}))$.
```

6. Application to conflict-free colorings. Among other results, Even et al. [7] proved that any set of n discs in the plane can be CF-colored with $O(\log n)$ colors and that this bound is tight in the worst case. They also provide a deterministic polynomial time algorithm for coloring a given collection of n discs with only $O(\log n)$ colors. Har-Peled and Smorodinsky [9] extended this result to any family of regions with linear union complexity. For example, they provide a randomized algorithm for CF-coloring any family of n pseudo-discs with $O(\log n)$ colors with high probability. In particular, this randomized algorithm serves as a probabilistic proof that a CF-coloring of any family of n pseudo-discs with only $O(\log n)$ colors exists. One of the open problems left in [9] is to obtain a deterministic framework for CF-colorings any family of regions with linear union complexity. As an application of Theorem 1.2 and Algorithm 1, we provide such a framework. Our algorithm outperforms the one used in [9] by being deterministic and conceptually simpler. The number of colors used in our algorithm for CF-coloring the given regions depends on their union complexity.

Algorithm 2 CF-Color($\mathcal{R}$): CF-Color the hypergraph $H(\mathcal{R})$.

```
1: $i \leftarrow 0$: $i$ denotes an unused color
2: while $\mathcal{R} \neq \emptyset$
3:     Increment: $i \leftarrow i + 1$
4:     Color the hypergraph $H(\mathcal{R})$: find a coloring $\chi$ of $H(\mathcal{R})$ with "few" colors,
       using (in most cases) Algorithm 1
5:     $\mathcal{R}' \leftarrow$ Largest color class of $\chi$
6:     Color: $f(x) \leftarrow i$ for all $x \in \mathcal{R}'$
7:     Prune: $\mathcal{R} \leftarrow \mathcal{R} \setminus \mathcal{R}'$
8: end while
```

THEOREM 6.1. *Algorithm* 2 *outputs a valid CF-coloring of* $\mathcal{R}$.

Proof. For a point $p \in \cup_{r \in \mathcal{R}} r$, let i be the maximal index (color) for which there is a region $r \in \mathcal{R}$ that contains p and is colored with i. We claim that there is exactly

Table 1

Summary of relation between the union complexity of the underlying objects $\mathcal{R}$, the chromatic number $H(\mathcal{R})$, and its CF-chromatic number.

Type of regions	$\mathcal{U}(n)$	$\chi(H(\mathcal{R}))$	$\chi_{CF}(H(\mathcal{R}))$
pseudo-discs, etc.	$O(n)$	$O(1)$	$O(\log n)$
convex fat regions, etc.	$O(n^{1+\delta})$	$O(n^\delta)$	$O(n^\delta)$
Axis-parallel rectangles	$\Theta(n^2)$	$O(\log n)$	$O(\log^2 n)$

one such region. Indeed, assume to the contrary that there is another such region r'. Consider the $i'th$ iteration where some of the regions of $\mathcal{R}$ were colored with i (including r and r'). Since r and r' belong to an independent set, there must have been a third region r'' containing p that wasn't colored in the $i'th$ iteration. This means that the color of r'' is greater than i, a contradiction to the maximality of i. This completes the proof of the theorem. □

Remark. Algorithm 2 yields a CF-coloring of regions with "low" union complexity with only "few colors" in the following sense: If $\mathcal{R}$ has union complexity bounded by $\mathcal{U}(n)$, then by Theorem 1.2, $H(\mathcal{R})$ can be colored with $O(\frac{\mathcal{U}(n)}{n})$ colors. So the largest color class is at least $\frac{n^2}{\mathcal{U}(n)}$ by the pigeonhole principle. Thus, in the *prune* step of Algorithm 2 we discard at least this many regions so, in total, Algorithm 2 does only few iterations. This depends on the function $\frac{n^2}{\mathcal{U}(n)}$. Table 1 summarizes the relation between the union complexity of the underlying objects, the chromatic number of the induced hypergraph, and its CF-chromatic number. The bounds given on the chromatic number and the CF-chromatic number are also bounds on the numbers of colors produced by Algorithms 1 and 2, respectively.

Theorem 6.2. *Let $\mathcal{R}$ be a set of n axis-parallel rectangles. Then Algorithm* 2 *applied to $\mathcal{R}$, provides a CF-coloring of $\mathcal{R}$ with $O(\log^2 n)$ colors in polynomial-time.*

Remark. Note that the union complexity of n rectangles can be quadratic. Thus, we cannot apply Theorem 1.2 directly to $\mathcal{R}$, since we would obtain a coloring of $H(\mathcal{R})$ with potentially n colors. Thus, in the prune step of Algorithm 2 we might discard only a constant number of rectangles and the algorithm might use linear number of colors. The bound on the chromatic number of $H(\mathcal{R})$ is asymptotically tight as already mentioned. However, it is not clear that the bound $O(\log^2 n)$ on the CF-chromatic number of $\mathcal{R}$ is asymptotically tight. Maybe one can get better bounds. We leave this as an open problem.

Proof. By Theorem 1.4, we can color $H(\mathcal{R})$ with $O(\log|\mathcal{R}|)$ colors. Thus, in each prune step of Algorithm 2 we discard at least $\Omega(\frac{|\mathcal{R}|}{\log|\mathcal{R}|})$ rectangles. It is easily seen that the total number of iterations (which is the number of colors used by the algorithm) will be $O(\log^2 n)$. □

Definition 6.3 (see [11]). *A family $\mathcal{R}$ of Jordan regions in the plane is called a family of* pseudo-discs *if the boundaries of each pair of them intersect at most twice.*

Theorem 6.4. *Let $\mathcal{R}$ be a family of n pseudo-discs. Then Algorithm* 2 applied to $\mathcal{R}$ provides a CF-coloring with $O(\log n)$ colors in polynomial-time.

Proof. The complexity of the union of any m regions of $\mathcal{R}$ is $O(m)$ (see [11]). By Theorem 1.2, there is a constant c such that Algorithm 1 provides a coloring χ of $H(\mathcal{R})$ with $\leq c$ colors. Such a coloring can be computed in polynomial-time. In the prune step of Algorithm 2 we discard at least $\frac{|\mathcal{R}|}{c}$ regions. Thus, Algorithm 2 provides a CF-coloring of $\mathcal{R}$ with only $\frac{\log n}{\log \frac{c}{c-1}}$ colors in polynomial-time. □

7. Discussion and open problems. Naturally, the problems addressed in this paper have analogous versions in higher dimensions. For example, what is the minimum number of colors that always suffice to color any hypergraph induced by any set $\mathcal{B}$ of n balls in $\mathbb{R}^3$? Unfortunately, the complexity of the union of n balls could be quadratic already in $\mathbb{R}^3$, and we cannot apply the methods developed in this paper directly. Moreover, for $d \geq 4$ and any $n > 1$, there exist families of n balls in $\mathcal{R}^d$ that are pairwise touching and therefore require n distinct colors for any proper coloring of $H(\mathcal{R})$ as any two balls contain a point witnessing the fact that the two balls must be colored with distinct colors. The best upper bound known for CF-coloring any set of n balls in $\mathbb{R}^3$ is the trivial bound n. It is interesting to relax the CF-coloring requirements as follows: what is the minimum number of colors needed to color any hypergraph induced by a set $\mathcal{B}$ of n balls in $\mathbb{R}^3$ such that every hyperedge of cardinality at least 3 is nonmonochromatic. It is easily seen that this number is bounded by $O(\sqrt{n})$ since the maximum degree of any element is bounded by $O(n)$ in the 3-uniform hypergraph consisting of all hyperedges of $H(\mathcal{B})$ with cardinality 3. However, we conjecture that fewer colors are enough. This relates to the notion of 2-CF-coloring studied in [9]. Any improvement over the $O(\sqrt{n})$ bound would imply a better bound on 2-CF-coloring of balls in $\mathbb{R}^3$. Here, we omit the detailed description of this relation.

Another open problem is to bound the chromatic number of any hypergraph induced by n axis-parallel boxes in $\mathbb{R}^d$ (for $d > 2$). We conjecture that $O(\log^{d-1} n)$ colors always suffice.

Acknowledgements. The author wishes to thank Janos Pach and Rados Radoičić for helpful discussions concerning the problems studied in this paper. In particular, the author would like to thank Janos Pach for pointing out that the graph $G(H)$ discussed in Lemma 3.1 might have a simple planar drawing when projected onto the plane. Thanks are also extended to Boris Aronov and Sariel Har-Peled for helpful comments on a preliminary version of this paper.

REFERENCES

[1] N. Alon and S. Smorodinsky, *Conflict-free colorings of shallow discs*, in Proceedings of the 22nd Annual ACM Symposium on Computational Geometry (SoCG 2006) Sedona, AZ, ACM Press, NY, 2006, pp. 41–43.

[2] K. Appel and W. Haken, *Every planar map is four colorable. 1. Discharging*, Illinois J. Math., 21 (1977), pp. 421–490.

[3] K. Appel and W. Haken, *Every planar map is four colorable. 2. Reducibility*, Illinois J. Math., 21 (1977), pp. 491–567.

[4] K. Chen, *On how to play a coloring game against color-blind adversaries*, in Proceedings of the 22nd Annual ACM Symposium on Computational Geometry (SoCG 2006), Sedona, AZ, ACM Press, NY, 2006, pp. 44–51.

[5] K. L. Clarkson and P. W. Shor, *Applications of random sampling in computational geometry*, II. Discrete Comput. Geom., 4 (1989), pp. 387–421.

[6] K. Elbassioni and N. Mustafa, *Conflict-Free Colorings of Rectangles Ranges*, in Proceedings of the Lecture Notes in Comput. Sci. 3884, 2006, pp. 254–263.

[7] G. Even, Z. Lotker, D. Ron and S. Smorodinsky, *Conflict-free colorings of simple geometric regions with applications to frequency assignment in cellular networks*, SIAM J. Comput., 33 (2003), pp. 94–136.

[8] K. Chen, A. Fiat, H. Kaplan, M. Levy, J. Matoušek, E. Mossel, J. Pach, M. Sharir, S. Smorodinsky, U. Wagner, and E. Welzl, *Online conflict-free coloring for intervals*, SIAM J. Comput., 36 (2006), pp. 1342–1359.

[9] S. Har-Peled and S. Smorodinsky, *On conflict-free coloring of points and simple regions in the plane*, Discrete Comput. Geom., 34 (2005), pp. 47–70.

[10] K. Chen, H. Kaplan and M. Sharir, *Online Conflict-Free Coloring for Halfplanes, Congruent Discs, and Axis-Parallel Rectangles*, manuscript, 2005.

[11] K. Kedem, R. Livne, J. Pach, and M. Sharir, *On the union of Jordan regions and collision-free translational motion amidst polygonal obstacles,* Discrete Comput. Geom., 1 (1986), pp. 59–71.

[12] P. Koebe, *Kontaktprobleme der konformen Abbildung*, Ber. Sächs. Akad. Wiss. Leipzig, Math-Phys., KL, 88 (1936), pp. 141–164.

[13] J. Pach and G. Tóth, *Conflict free colorings*, Discrete Comput. Geom. Algorithms Combin. 25, Springer Verlag, Berlin, 2003, pp. 665–671.

[14] J. Pach and G. Tardos, personal communication.

[15] M. Sharir and P. K. Agarwal, *Davenport-Schinzel Sequences and Their Geometric Applications*, Cambridge University Press, Cambridge, UK, 1995.

[16] S. Smorodinsky, *Combinatorial Problems in Computational Geometry*, Ph.D. dissertation, School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel 2003.

SIAM J. DISCRETE MATH.
Vol. 21, No. 3, pp. 688–706

LABELINGS OF GRAPHS WITH FIXED AND VARIABLE EDGE-WEIGHTS*

ROBERT BABILON†, VÍT JELÍNEK†, DANIEL KRÁĽ†, AND PAVEL VALTR†

Abstract. Motivated by $L(p, q)$-labelings of graphs, we introduce a notion of λ-graphs: a λ-graph G is a graph with two types of edges: 1-edges and x-edges. For a parameter $x \in [0, 1]$, a proper labeling of G is a labeling of vertices of G by nonnegative reals such that the labels of the endvertices of a 1-edge differ by at least 1 and the labels of the endvertices of an x-edge differ by at least x; $\lambda_G(x)$ is the smallest real such that G has a proper labeling by labels from the interval $[0, \lambda_G(x)]$. We study properties of the function $\lambda_G(x)$ for finite and infinite λ-graphs and establish the following results: if the function $\lambda_G(x)$ is well defined, then it is a piecewise linear function of x with finitely many linear parts. Surprisingly, the set $\Lambda(\alpha, \beta)$ of all functions λ_G with $\lambda_G(0) = \alpha$ and $\lambda_G(1) = \beta$ is finite for any $\alpha \leq \beta$. We also prove a tight upper bound on the number of segments for finite λ-graphs G with convex functions $\lambda_G(x)$.

Key words. channel assignment problem, graph labeling with distance conditions

AMS subject classification. 05C15

DOI. 10.1137/040619545

1. Introduction.

Several graph theory models for radio frequency assignment were suggested by Hale [16]. One of the most important models is $L(p, q)$-labeling of graphs, which can be traced back to the paper by Griggs and Yeh [15]. An $L(p, q)$-*labeling* of a graph G for $1 \leq q \leq p$ is a labeling of the vertices by nonnegative integers such that the labels of adjacent vertices differ by at least p and the labels of vertices at distance two differ by at least q. The least integer K such that there is a proper labeling using integers between 0 and K is called the *span* and is denoted by $\lambda_{p,q}(G)$.

The case of $L(2, 1)$-labelings attracted a special attention of researchers, in particular with the connection to the conjecture of Griggs and Yeh [15] that $\lambda_{2,1}(G) \leq \Delta^2$ for every graph G with maximum degree Δ. Bounds on the span in terms of the maximum degree have been proved in a series of papers [15, 5, 24, 9], and currently, the best upper bound is $\lambda_{2,1}(G) \leq \Delta^2 + \Delta - 2$. The conjecture itself has been verified for several classes of graphs, including graphs of maximum degree two, chordal graphs [28]; see also [4, 22] and Hamiltonian cubic graphs [19, 20]. However, even the case of general cubic graphs remains open. Because of practical motivation of the problem, $L(p, q)$-labelings are also widely studied from the algorithmic point of view [1, 3, 7, 8, 21, 27].

In this paper, we study how the span $\lambda_{p,q}(G)$ depends on the parameters p and q. This is well motivated from practical point of view since in applications, the parameters p and q are not fixed in advance but rather adjusted ad hoc depending on the level of interference experienced for their different combinations. Our approach is

*Received by the editors November 24, 2004; accepted for publication (in revised form) January 4, 2007; published electronically September 12, 2007. The authors acknowledge the support by the tripartite grant KONTAKT ME-525.

http://www.siam.org/journals/sidma/21-3/61954.html

†Department of Applied Mathematics and Institute for Theoretical Computer Science (ITI), Faculty of Mathematics and Physics, Charles University, Malostranské náměstí 25, 118 00 Prague, Czech Republic (babilon@kam.mff.cuni.cz, jelinek@kam.mff.cuni.cz, kral@kam.mff.cuni.cz, valtr@kam.mff.cuni.cz). Institute for Theoretical Computer Science is supported by the Ministry of Education of the Czech Republic as project LN00A056. The third author was a postdoctoral fellow at Institute for Mathematics of the Technical University, Berlin from October 2004 to July 2005.

similar to that of [25], but we focus on the original notion of $L(p, q)$-labeling rather than its circular coloring version, and we do not determine the behavior for some particular graphs, but rather prove general results. Also, we do not restrict our attention to finite graphs. The inclusion of infinite graphs is motivated by applications, e.g., $L(p, q)$-labelings of infinite triangular, square and hexagonal planar lattices naturally arise in practice and have been addressed from the theoretical point of view as well [18].

$L(p, q)$-labelings are closely related to the channel assignment problem. Our definition of the channel assignment problem is slightly more general than usual: both the weights of edges and the labels of vertices are real numbers rather than just integers. A *channel assignment problem* is determined by a pair (G, w) consisting of a (finite or infinite) graph G and a function $w : E(G) \to \mathbb{R}^+$. A labeling $c : V(G) \to \mathbb{R}_0^+$ of the vertices of G by nonnegative reals is *proper* if $|c(v) - c(v')| \geq w(vv')$ for each edge vv' of G. The *span of a labeling* c is the supremum of the labels used by c and the *span* $\lambda_w(G)$ *of* (G, w) is the infimum of the spans of proper labelings of (G, w). An $L(p, q)$-labeling of a graph G can be viewed as the channel assignment problem for the square of G (the second distance power): the edges of G have weights p and the edges of G^2 not belonging to G have weights q. The reader is also welcome to see the survey [26] on the channel assignment problem.

The alternative view of $L(p, q)$-labelings presented above is a starting point for our work. A *λ-graph* G is a graph with two types of edges: *1-edges* and *x-edges.* For a parameter $x \in [0, 1]$, one forms a channel assignment problem on G by assigning the weight 1 to every 1-edge and the weight x to every x-edge. The span of this channel assignment problem is denoted by $\lambda_G(x)$; the function $\lambda_G(x)$ is called the *λ-function* of G. For a graph H, let G_H be the λ-graph on the same set of vertices as H such that the vertices adjacent in H are joined by 1-edges in G_H, and the vertices at distance two in H are joined by x-edges in G_H. Clearly, the following holds:

$$\lambda_{G_H}\left(\frac{q}{p}\right) = \frac{\lambda_{p,q}(H)}{p}.$$

Therefore, the λ-function of G_H can be viewed as normalized one-dimensional function describing the behavior of the two-parameter function $\lambda_{p,q}(H)$. Note also that $\lambda_G(0) = \chi(G^{(1)}) - 1$ and $\lambda_G(1) = \chi(G) - 1$, where $G^{(1)}$ is the subgraph of G formed by the 1-edges. This approach reflects the practical application of radio frequency assignment: the 1-edges represent the pairs of close transmitters where huge interference occurs, and the x-edges correspond to more distant transmitters where smaller interference may appear. The value of the parameter x is then proportional to the interference experienced and is adjusted according to its level. To get acquainted with the concepts used in this paper, the reader may consult the appendix, where we provide the complete list of λ-graphs with four vertices together with their λ-functions, as well as examples of other interesting λ-graphs.

A similar approach to the study of the span of $L(p, q)$-labeling was developed by Griggs and Jin [11, 12, 13]. They presented their results, e.g., during the SIAM Conference on Discrete Mathematics in Nashville, TN, in June 2004. In particular, they proved (using a different terminology) that if H is a (finite or infinite) graph with bounded maximum degree, then λ_{G_H} is a piecewise linear function of x for $x \in [0, \infty)$ with finitely many linear parts. Moreover, the coefficients of the linear functions forming λ_{G_H} are bounded by a constant that depends solely on the maximum degree of H. The former statement can be derived from our Theorem 3.2 (see Corollary 3.3).

Our Theorem 4.5 yields that there are only finitely many different λ-functions for λ-graphs of the form G_H where H is a graph of bounded maximum degree. Hence, Theorem 4.5 also implies that the coefficients of linear functions forming λ_{G_H} are bounded by a constant depending only on the maximum degree of H.

Our method is different from that in [11]: the arguments in [11] are based on the structure of optimum labelings for a graph H obeying the given distance constraints, whereas we use a close correspondence between orientations of graphs and their labelings, developed in section 2. Still, some of our results and their proofs, e.g., Lemma 3.1, are analogous to those in [11]. Since we prove our results in a more general setting, we decided, for the sake of completeness, to include full arguments even in such cases.

Let us remark that the concept of λ-graphs have been further developed, e.g., in [23]. The reader can check a recent survey [14] for more results in this area.

1.1. Our results. We study general λ-graphs without restricting our attention to those equal to G_H for some H. In section 3, we show that the function $\lambda_G(x)$ is a piecewise linear function with finitely many linear parts, under the assumption that it is well defined for some $x > 0$. The proof of this statement is quite straightforward if G is finite, but it becomes more complex for infinite λ-graphs. In section 4, we study λ-functions with prescribed values for $x = 0, 1$. Let $\Lambda(\alpha, \beta)$ be the set of all λ-functions $\lambda_G(x)$ of finite and infinite λ-graphs G with $\lambda_G(0) = \alpha$ and $\lambda_G(1) = \beta$. One could expect that the set $\Lambda(\alpha, \beta)$ is infinite for $\alpha < \beta$, but the opposite is true: in fact, the set $\Lambda(\alpha, \beta)$ is finite for any integers $\alpha \leq \beta$. In Theorem 4.5, we present the bound $2^{2^{\frac{(2\alpha\beta^2+\alpha\beta+\beta^2+2)^2}{2}}}$ on the size of the set $\Lambda(\alpha, \beta)$. At the end of the paper, we focus on finite λ-graphs whose λ-function is convex and prove an asymptotically tight upper bound on the number of the linear parts of the λ-functions in terms of the order of a λ-graph: if G is a finite λ-graph of order n and the function $\lambda_G(x)$ is convex, then $\lambda_G(x)$ consists of at most $O(n^{2/3})$ linear parts.

2. Gallai–Roy theorem. We establish an analogue of the Gallai–Roy theorem for channel assignment problems with (finite and) infinite underlying graphs. The Gallai–Roy theorem in its original form relates colorings and lengths of paths in acyclic orientations of a graph. Our proof follows the lines of a similar theorem for channel assignment problems with finite graphs by McDiarmid [27], but we include the proof for the sake of completeness.

First, we introduce some additional definitions necessary for stating and proving the theorem. An orientation of a graph is *finitary* if there is a constant $K \geq 0$ such that every oriented walk has length at most K. The *weight* of a path is the sum of the weights of the edges on the path. The channel assignment problem (G, w) is said to be *finitary* if the image set of the function w is finite. If (G, w) is finitary, then there exists a proper labeling c whose span is equal to the span of (G, w), and the span of the optimum labeling c is equal to the maximum label used by c (these claims will be established in the proof of Theorem 2.1).

We now state and prove the announced analogue of the Gallai–Roy theorem.

THEOREM 2.1. *Let (G, w) be a finitary channel assignment problem. The span of (G, w) is finite if and only if G has a finitary orientation. In this case, the span of (G, w) is equal to the minimum of the maximum weight of a path in a finitary orientation of G, where the minimum is taken over all finitary orientations of G.*

Proof. Consider a finitary orientation of G and let w_0 be the maximum weight of a path in the orientation. Label a vertex v of G with the maximum weight of an oriented path which ends at v. Clearly, the span of this labeling does not exceed w_0. Moreover, the labeling is proper: consider two vertices v and v' joined by an edge of G. Assume that the edge between v and v' is oriented from v to v'. Since each path leading to the vertex v can be prolonged to v', the label of v' is greater than the label of v and they differ by at least $w(vv')$. Since there is a finite number of edge weights (recall that both the channel assignment problem and the orientation are finitary), we conclude that the span of (G, w) is at most the minimum of the maximum weight of a path taken over all finitary orientations of G.

On the other hand, if c is a proper labeling of (G, w) with finite span, then there is a finitary orientation of G such that the maximum weight of a path in the orientation is at most the span. Consider the following orientation: an edge between two vertices v and v' is oriented from v to v' if $c(v) < c(v')$, otherwise it is oriented from v' to v. Since the labels of the vertices on an oriented path increase on each edge at least by its weight, the maximum weight of the path in the orientation is bounded by the maximum label assigned to a vertex of G. The statement of the theorem now readily follows. □

The next corollary of Theorem 2.1 on the λ-functions of finite λ-graphs immediately follows.

Corollary 2.2. *If G is a finite λ-graph of order n, then for each $x \in [0, 1]$, there exist nonnegative integers a and b with $a + b \le n - 1$ such that $\lambda_G(x) = a + b \cdot x$.*

Proof. Consider the channel assignment problem (G', w'), where G' is the underlying graph of G, the weight $w'(e)$ of a 1-edge e is one and the weight $w'(e)$ of an x-edge e is x. Since the channel assignment problem (G', w') is finitary, its span is equal to the maximum weight of a finite path of a finitary orientation of G'. Therefore, $\lambda_G(x) = a + b \cdot x$ for some nonnegative integers $a + b \le n - 1$. □

3. Piecewise linearity. In this section, we show that the function $\lambda_G(x)$ of every λ-graph is a piecewise linear function of x. As the first step, we show that the function $\lambda_G(x)$ is a linear function of x on some neighborhood of 0.

Lemma 3.1. *Let G be a (finite or infinite) λ-graph. If the function $\lambda_G(x)$ is finite for some $x > 0$, then the function $\lambda_G(x)$ is a linear function of x on the interval $[0, \varepsilon]$ for some $\varepsilon > 0$.*

Proof. Since $\lambda_G(x)$ is finite for some $x > 0$, there is a finitary orientation $\vec{D_0}$ of G. In particular, the chromatic number $\chi(G^{(1)})$ is finite (recall that $G^{(1)}$ is the spanning subgraph of G whose edges are exactly the 1-edges of G), and $\lambda_G(0) = \chi(G^{(1)}) - 1$.

Next, we construct a finitary orientation of G that does not contain any oriented path with more than $\lambda_G(0)$ 1-edges. Let c be any proper coloring of $G^{(1)}$ with $\chi(G^{(1)})$ colors $0, \dots, \lambda_G(0)$. Consider the orientation $\vec{D}$ of G such that an edge vv' of G is

- oriented from v to v', if $c(v) < c(v')$,
- oriented from v' to v, if $c(v) > c(v')$, and
- oriented as in the orientation $\vec{D_0}$, otherwise.

Since on each oriented path, the colors of the vertices form a nondecreasing sequence that strictly increases on each 1-edge, there is no oriented path with more than $\lambda_G(0)$ 1-edges. It remains to show that the orientation $\vec{D}$ is finitary. Let k be the maximum length of a path in $\vec{D_0}$. As we have observed, the colors assigned by c to the vertices of an oriented path of $\vec{D}$ form a nondecreasing sequence. A subpath formed by the vertices of the same color is also an oriented path in $\vec{D_0}$. Hence, its length is at most k. We conclude that each oriented path in $\vec{D}$ has length at most $\chi(G^{(1)})(k + 1)$. In particular, the orientation $\vec{D}$ is finitary.

Choose $\vec{D}$ to be a finitary orientation of G such that

1. $\vec{D}$ does not contain any oriented path with more than $\lambda_G(0)$ 1-edges, and
2. the maximum length of an oriented path with exactly $\lambda_G(0)$ 1-edges is minimal.

Since the orientation of G constructed in the previous paragraph has the first property, the orientation $\vec{D}$ exists and is well defined.

Let k_D be the maximum length of a path in $\vec{D}$, and let k'_D be the maximum number of x-edges on a path of $\vec{D}$ that has $\lambda_G(0)$ 1-edges. We show the following:

$$\lambda_G(x) = \lambda_G(0) + k'_D x \text{ for every } x \in [0, \varepsilon],$$

where $\varepsilon = \frac{1}{k_D}$.

Assume, for contradiction, that there is $x \in (0, \varepsilon]$ such that $\lambda_G(x) < \lambda_G(0) + k'_D x$. Note that this inequality implies that $\lambda_G(x) < \lambda_G(0) + 1$ since $k'_D < k_D$. By Theorem 2.1, G has a finitary orientation $\vec{D}'$ that does not contain any oriented path with more than $\lambda_G(0)$ 1-edges, and in addition, at least one of the following holds:

- $\vec{D}'$ has no oriented path with $\lambda_G(0)$ 1-edges, or
- any oriented path of $\vec{D}'$ with $\lambda_G(0)$ 1-edges has less than k'_D x-edges.

The former is impossible because every finitary orientation of $G^{(1)}$, and therefore, of G, has a path with $\lambda_G(0)$ edges by Theorem 2.1. The latter contradicts the choice of the orientation $\vec{D}$. We infer that $\lambda_G(x) \geq \lambda_G(0) + k'_D x$ for all $x \in [0, \varepsilon]$.

It remains to establish the opposite inequality, i.e., $\lambda_G(x) \leq \lambda_G(0) + k'_D x$ for $x \in [0, \varepsilon]$. Consider an oriented path P in $\vec{D}$. If P contains $\lambda_G(0)$ 1-edges, then it contains at most k'_D x-edges, and consequently, its weight is at most $\lambda_G(0) + k'_D x$. On the other hand, if P contains less than $\lambda_G(0)$ 1-edges, then its weight is at most $\lambda_G(0) - 1 + k_D x \leq \lambda_G(0)$. We conclude that the maximum weight of an oriented path in $\vec{D}$ is at most $\lambda_G(0) + k'_D x$. Therefore, $\lambda_G(x) \leq \lambda_G(0) + k'_D x$ by Theorem 2.1. □

We are ready to establish the main result of this section. Note that the statement of Theorem 3.2 for finite λ-graphs can be easily derived from Corollary 2.2.

THEOREM 3.2. *Let G be a (finite or infinite) λ-graph. If the function $\lambda_G(x)$ is finite for some $x > 0$, then the function $\lambda_G(x)$ is a piecewise linear function of x on the interval $[0, 1]$ with finitely many linear parts.*

Proof. Since the function $\lambda_G(x)$ is finite for some $x > 0$, G has a finitary orientation and the function $\lambda_G(x)$ is finite for all $x \in [0, 1]$ by Theorem 2.1. Let $\varepsilon > 0$ be a real such that the function $\lambda_G(x)$ is linear for $x \in [0, \varepsilon]$. Such ε exists by Lemma 3.1. We may assume that $\varepsilon \leq 1/4$. Moreover, if $\lambda_G(1) = 0$, then λ_G is identically equal to 0 and the theorem holds. Therefore, we only need to consider the case $\lambda_G(1) \geq 1$. Let $K = \lfloor \lambda_G(1)/\varepsilon \rfloor$. By the previous assumptions, $K \geq 4$. Consider the set $\mathcal{D}$ of finitary orientations $\vec{D}$ of G such that the maximum length of an oriented path in $\vec{D}$ is at most K. Note that the set $\mathcal{D}$ is nonempty since G has a finitary orientation with maximum path length $\lambda_G(1)$ by Theorem 2.1 applied to the graph G with all edge weights equal to one.

For an orientation $\vec{D} \in \mathcal{D}$, let $\mathcal{F}(\vec{D})$ be the set of all the functions $a + bx$ such that $\vec{D}$ contains an oriented path with a 1-edges and b x-edges. Since the maximum length of an oriented path of $\vec{D}$ is at most K, the sum $a + b$ is bounded by K. Therefore, the set $\mathcal{F}(\vec{D})$ is finite for every orientation $\vec{D} \in \mathcal{D}$. Let $f_{\vec{D}}(x) = \max_{f \in \mathcal{F}(\vec{D})} f(x)$. Since the set $\mathcal{F}(\vec{D})$ is finite, the function $f_{\vec{D}}(x)$ is the maximum of a finite number of linear functions. In particular, the function $f_{\vec{D}}(x)$ is piecewise linear and has finitely many linear parts. Let us define

$$f_0(x) := \min_{\vec{D}\in\mathcal{D}} f_{\vec{D}}(x) = \min_{\vec{D}\in\mathcal{D}} \max_{f\in\mathcal{F}(\vec{D})} f(x).$$

Since there are at most $\binom{K+2}{2} \leq K^2$ functions $a + bx$ with $0 \leq a, b$ and $a + b \leq K$, there are at most 2^{K^2} distinct sets $\mathcal{F}(\vec{D})$, and the minimum in the definition of $f_0(x)$ is always attained. Moreover, the function $f_0(x)$ is the minimum of at most 2^{K^2} distinct piecewise linear functions, and thus $f_0(x)$ is also a piecewise linear function. In the rest of this proof, we show that $\lambda_G(x) = f_0(x)$ for all $x \in [\varepsilon, 1]$.

Fix $x \in [\varepsilon, 1]$. Let $\vec{D}$ be an orientation of G such that $f_{\vec{D}}(x) = f_0(x)$. In the orientation $\vec{D}$, the maximum weight of an oriented path is $f_{\vec{D}}(x)$ and $\lambda_G(x) \leq f_0(x)$ by Theorem 2.1. Assume for the sake of contradiction that $\lambda_G(x) < f_0(x)$ for some $x \in [\varepsilon, 1]$. By Theorem 2.1, there exists a finitary orientation $\vec{D}$ of G with the maximum weight of an oriented path equal to $\lambda_G(x)$. If $\vec{D}$ contains a path with more than K edges, then the weight of this path is at least $(K + 1)x > \frac{\lambda_G(1)}{\varepsilon}x \geq \lambda_G(1)$. This is impossible, because $\lambda_G(x) \leq \lambda_G(1)$. Therefore, the length of each oriented path in $\vec{D}$ is at most K and $\vec{D} \in \mathcal{D}$. Since the maximum weight of an oriented path in $\vec{D}$ is $f_{\vec{D}}(x)$, we have $f_0(x) \leq f_{\vec{D}}(x) = \lambda_G(x) < f_0(x)$—a contradiction.

We have shown that $\lambda_G(x) = f_0(x)$ for all $x \in [\varepsilon, 1]$. Since the function $\lambda_G(x)$ is piecewise linear on both the intervals $[0, \varepsilon]$ and $[\varepsilon, 1]$ and it has finitely many linear parts on each of the two intervals, it is a piecewise linear function with finitely many linear parts on the whole interval $[0, 1]$. □

Let us now show how Theorem 3.2 implies the results of Griggs and Jin on λ-functions of λ-graphs of the form G_H.

Corollary 3.3. *Let H be a (finite or infinite) graph with a bounded maximum degree, and let $\ell_H(x) := \frac{1}{p}\lambda_{p,q}(H)$ for $x = q/p$. The function $\ell_H(x)$ is a piecewise linear function for $x \in [0, \infty)$ with finitely many linear parts.*

Proof. For $x \in [0, 1]$, the statement follows from Theorem 3.2 applied to the graph G_H whose definition can be found in section 1. Next, consider the graph G' obtained from G_H by replacing 1-edges by x-edges and x-edges by 1-edges. Observe that $\ell_H(x) = x\cdot\lambda_{G'}(1/x)$. Again, Theorem 3.2 yields that $\lambda_{G'}(x')$ is a piecewise linear function with finitely many linear parts for $x' \in [0, 1]$. Hence, $\ell_H(x)$ is a piecewise linear function with finitely many linear parts for $x \in [1, \infty)$, too. □

Note that if H has bounded maximum degree, then G_H has bounded maximum degree as well, and, in particular, G_H has bounded chromatic number. Our results from section 4, namely Theorem 4.5, imply that for every finite bound K there is a finite set $\mathcal{L}_K$ of piecewise linear functions defined on $[0, \infty)$, with finitely many linear parts, such that for any (finite or infinite) graph H with maximum degree at most K we have $\ell_H \in \mathcal{L}_K$.

Another immediate corollary of Theorem 3.2 is the following.

Corollary 3.4. *If G is a finite λ-graph of order n, then there exist an integer k, $1 \leq k \leq n^2$, real numbers $x_0, \ldots, x_k$, $0 = x_0 < x_1 < \cdots < x_k = 1$, and nonnegative integers $a_1, \ldots, a_k$ and $b_1, \ldots, b_k$ with $a_i + b_i \leq n - 1$, such that $\lambda_G(x) = a_i + b_i x$ for every $x \in [x_{i-1}, x_i]$. Moreover, $x_i = \frac{c_i}{d_i}$ for some integers c_i, d_i, with $0 \leq c_i \leq d_i \leq n - 1$.*

Proof. Since the function $\lambda_G(x)$ is piecewise linear by Theorem 3.2, there exist real numbers $x_0, \ldots, x_k$, $0 = x_0 < x_1 < \cdots < x_k = 1$, such that the function $\lambda_G(x)$ is linear on each interval $[x_{i-1}, x_i]$, $i = 1, \ldots, k$, for some integer k. By Corollary 2.2, the coefficients of these linear functions are nonnegative integers whose sum does not exceed $n-1$.

Furthermore, each of the reals $x_1, \dots, x_{k-1}$ can be expressed as a fraction with both the numerator and denominator between 1 and $n-1$: clearly, x_i is the (unique) solution of the equation $a_i + b_i x_i = a_{i+1} + b_{i+1} x$. Hence, $x_i = \frac{a_i - a_{i+1}}{b_{i+1} - b_i} = \frac{|a_{i+1} - a_i|}{|b_{i+1} - b_i|}$ (the latter equality follows from the fact that x_i is positive). Since there are at most $(n-1)^2$ fractions with both the numerator and denominator between 1 and $n-1$, the bound on the number k follows. □

Let us remark that the bound on the number of linear parts in Corollary 3.4 can be improved to $\frac{3n^2}{\pi^2} + o(n^2)$ using the results on the Farey fractions discussed in section 5. However, we think that the order of the bound from Corollary 3.4 can be asymptotically improved and conjecture the following.

CONJECTURE 3.5. *If G is a finite λ-graph of order n, then the function $\lambda_G(x)$ consists of at most n linear parts.*

4. λ-functions with boundary constraints.

As the first step towards the proof of Theorem 4.5, we establish two bounds on the growth of a λ-function.

LEMMA 4.1. *Let G be a (finite or infinite) λ-graph whose λ-function is finite, and let $\lambda_G(x) = a + bx$ for all $x \in [0, \gamma]$ and some $\gamma > 0$. The following inequality holds:*

$$\lambda_G(x) \geq a + bx$$

for all $x \in [0, 1/b]$, if $b > 0$, and for all $x \in [0, 1]$, otherwise.

Proof. If $b = 0$, the lemma holds trivially, because $\lambda_G(x) \geq \lambda_G(0)$ for all $x \in [0, 1]$. In the rest of the proof, we consider the case $b > 0$. Assume for the sake of contradiction that there exists $x_0 \in [0, 1/b]$ such that $\lambda_G(x_0) < a + bx_0 \leq a + 1$. Note that $x_0 > \gamma$ because $\lambda_G(x)$ is equal to $a + bx$ for $x \in [0, \gamma]$. By Theorem 2.1, there exists an orientation $\vec{D}$ of G with the following property: for every oriented path P in $\vec{D}$ it holds that $a' + b'x_0 \leq \lambda_G(x_0) < a + bx_0$, where a' and b' are the numbers of 1-edges and x-edges of P. Since $a + bx_0 \leq a + 1$, we have $a' \leq a$. Therefore, $a' + b'\gamma < a + b\gamma$ for each such path P. We infer from Theorem 2.1 that $\lambda_G(\gamma) < a + b\gamma$. This contradicts the assumptions of the lemma. □

LEMMA 4.2. *Let G be a (finite or infinite) λ-graph whose λ-function is finite. The following inequality holds:*

$$\lambda_G(x) \leq \alpha + (\alpha + 1)\beta x,$$

where $\alpha = \lambda_G(0)$ and $\beta = \lambda_G(1)$.

Proof. Fix vertex colorings $c^{(1)}$ and c of the graphs $G^{(1)}$ and G with colors $0, \dots, \alpha$ and $0, \dots, \beta$. Let $\vec{D}$ be the following orientation of G: an edge $e = uv$ of G with $c^{(1)}(u) < c^{(1)}(v)$ is oriented from u to v. An edge $e = uv$ with $c^{(1)}(u) = c^{(1)}(v)$ is oriented from u to v if $c(u) < c(v)$ and from v to u, otherwise.

We now bound the maximum weight of a path in $\vec{D}$. Consider an oriented path P in $\vec{D}$. The function $c^{(1)}$ is nondecreasing along the path P. Since the value of $c^{(1)}$ increases on each 1-edge of P, the path P contains at most α 1-edges. There are also at most $\alpha + 1$ subpaths of P formed by the vertices with the same color assigned by $c^{(1)}$. On each such subpath, the function c is strictly increasing, and consequently, such a subpath can consist of at most β x-edges. We conclude that each oriented path in $\vec{D}$ contains at most α 1-edges and at most $(\alpha+1)\beta$ x-edges. By Theorem 2.1, $\lambda_G(x) \leq \alpha + (\alpha + 1)\beta x$. □

A key essence of the proof that the set $\Lambda(\alpha,\beta)$ is finite is the following lower bound on the length of the initial linear part of a λ-function in terms of $\lambda_G(0)$ and $\lambda_G(1)$.

LEMMA 4.3. *Let G be a (finite or infinite) λ-graph whose λ-function is finite. The length of the initial linear part of $\lambda_G(x)$ is at least*

$$\frac{1}{2\alpha\beta+\alpha+\beta+1},$$

where $\alpha=\lambda_G(0)$ and $\beta=\lambda_G(1)$.

Proof. Let a and b be the nonnegative integers such that $\lambda_G(x)=a+bx$ for all $x\in[0,\gamma]$ for some $\gamma>0$. Note that $a=\alpha$. By Lemma 4.2, we have $0\le b\le(\alpha+1)\beta$. We show that $\lambda_G(x)=a+bx$ for all $x\in[0,\frac{1}{2\alpha\beta+\alpha+\beta+1}]$. The inequality $\lambda_G(x)\ge a+bx$ follows from Lemma 4.1. In what follows, we focus on establishing the opposite inequality $\lambda_G(x)\le a+bx$.

By Theorem 2.1 applied to the channel assignment problem derived from G for $x=\min\{1/(b+1),\gamma\}$, there exists a finitary orientation $\vec{D}$ of G with the following properties:

1. $\vec{D}$ contains no oriented path with $a+1$ or more 1-edges, and
2. each oriented path of $\vec{D}$ with a 1-edges contains at most b x-edges.

Let a_v, $v\in V(G)$ be the maximum number of 1-edges on an oriented path of $\vec{D}$ which ends at v, and let b_v be the maximum number of x-edges on an oriented path with a_v 1-edges which ends at v. In addition, let c_β be a coloring of G with colors $0,\dots,\beta$. For $x\in[0,\frac{1}{2\alpha\beta+\alpha+\beta+1}]$, we define a labeling c of G as follows:

$$c(v)=\begin{cases} a_v+b_vx & \text{if } b_v\le b,\\ a_v+a_v(\beta+1)x+(c_\beta(v)+b+1)x & \text{otherwise.}\end{cases}$$

We now prove that c is a proper labeling of G for every $x\in[0,\frac{1}{2\alpha\beta+\alpha+\beta+1}]$.

As the first step towards this goal, we show that if $b_v>b$, then the label $c(v)$ is at most a_v+1-x (note that $b_v>b$ implies $a_v<a=\alpha$):

$$\begin{aligned} c(v)&=a_v+a_v(\beta+1)x+(c_\beta(v)+b+1)x\\ &\le a_v+(\alpha-1)(\beta+1)x+(\beta+(\alpha+1)\beta+1)x\\ &=a_v+(2\alpha\beta+\alpha+\beta+1)x-x\le a_v+1-x. \end{aligned}\tag{4.1}$$

Next, we show that the labeling is proper on each edge of G. Consider an edge uv, oriented from u to v in $\vec{D}$. We distinguish two major cases according to the type of the edge uv:

- ***uv* is an *x*-edge.**
  Clearly, $a_u\le a_v$, and if $a_u=a_v$, then $b_u<b_v$. We verify that $|c(u)-c(v)|\ge x$ by considering the following four subcases:
  - $a_u<a_v$
    If $b_u\le b$, we infer from $x\le 1/(b+1)$ that $c(u)\le a_u+b_ux\le a_u+1-x$. On the other hand, if $b_u>b$, it holds that $c(u)\le a_u+1-x$ by (4.1). Since $a_u+1\le a_v\le c(v)$, we have $c(v)-c(u)\ge x$ as desired.
  - $a_u=a_v$ and $b_u<b_v\le b$
    The inequality $c(v)-c(u)\ge x$ follows from the definition of c.
  - $a_u=a_v$ and $b_u\le b<b_v$
    We have $c(v)-c(u)\ge(c_\beta(v)+b+1)x-b_ux\ge x$.

  - $a_u = a_v$ and $b < b_u < b_v$
    By the definition of c, we have $|c(v) - c(u)| = |c_\beta(v) - c_\beta(u)|x \geq x$.
- ***uv* is a 1-edge.**
  Clearly, $a_u < a_v$. If $a_u = a_v - 1$, then $b_u \leq b_v$. We establish that $|c(u) - c(v)| \geq 1$ by considering the next four subcases:
  - $a_u \leq a_v - 2$
    Observe that $c(u) \leq a_u + 1$ and $a_v \leq c(v)$. Since $a_u \leq a_v - 2$, we can immediately conclude that $c(v) - c(u) \geq 1$.
  - $a_u = a_v - 1$ and $b_u \leq b_v \leq b$
    The definition of c immediately yields that $c(v) - c(u) \geq 1$.
  - $a_u = a_v - 1$ and $b_u \leq b < b_v$
    By the definition of c, we have $c(v) - c(u) \geq 1 + (c_\beta(v) + b + 1)x - b_u x \geq 1$.
  - $a_u = a_v - 1$ and $b < b_u \leq b_v$
    We again inspect the definition of c: $c(v) - c(u) \geq 1 + (\beta + 1)x + [(c_\beta(v) + b + 1) - (c_\beta(u) + b + 1)]\, x \geq 1$.

We have shown that c is a proper labeling of G. Note that the maximum label assigned by c does not exceed $a + bx$. The inequality $\lambda_G(x) \leq a + bx$ for $x \in [0, \frac{1}{2\alpha\beta+\alpha+\beta+1}]$ readily follows. □

Before we prove Theorem 4.5, we observe the following proposition. Its statement can be verified by inspection of the proof of Theorem 3.2.

PROPOSITION 4.4. *Let G be a (finite or infinite) λ-graph whose λ-function is finite. Furthermore, let $\mathcal{F}$ be the set of all linear functions $ax + b$ with integral nonnegative coefficients a and b such that $a + b \leq \beta/\gamma$, where $\beta = \lambda_G(1)$ and γ is a real such that $\lambda_G(x)$ is linear on the interval $[0, \gamma]$. There exist sets $\mathcal{F}_1, \dots, \mathcal{F}_k \subseteq \mathcal{F}$ such that the following equality holds for all $x \in [\gamma, 1]$:*

$$\lambda_G(x) = \min_{i=1,\dots,k} \max_{f \in \mathcal{F}_i} f(x).$$

Finally, we are ready to prove the main result of this section.

THEOREM 4.5. *Let $\alpha \leq \beta$ be any two nonnegative integers. The following estimate on the size of $\Lambda(\alpha, \beta)$ holds:*

$$|\Lambda(\alpha, \beta)| \leq 2^{2^{\frac{(2\alpha\beta^2+\alpha\beta+\beta^2+2)^2}{2}}}.$$

In particular, the set $\Lambda(\alpha, \beta)$ is finite.

Proof. Let $f_0 \in \Lambda(\alpha, \beta)$, i.e., there exists a λ-graph G with $\lambda_G(x) = f_0(x)$ and $f_0(0) = \alpha$ and $f_0(1) = \beta$. By Lemma 4.3, the function f_0 is a linear function of x on the interval $[0, \gamma]$, where $\gamma = \frac{1}{2\alpha\beta+\alpha+\beta+1}$. In particular, the values of f_0 on the interval $[0, \gamma]$ are uniquely determined by the value of $f_0(\gamma)$ (recall that $f_0(0) = \alpha$).

As in Proposition 4.4, let $\mathcal{F}$ be the set of all linear functions $ax + b$ with integral nonnegative coefficients a and b such that $a + b \leq \beta/\gamma$. Let us estimate the cardinality of the set $\mathcal{F}$:

$$|\mathcal{F}| = \sum_{i=0}^{\lfloor \beta/\gamma \rfloor} (i + 1) \leq \frac{(\beta/\gamma + 2)^2}{2} = \frac{(2\alpha\beta^2 + \alpha\beta + \beta^2 + 2)^2}{2}. \tag{4.2}$$

By Proposition 4.4, there exist subsets $\mathcal{F}_1, \dots, \mathcal{F}_k \subseteq \mathcal{F}$ such that $f_0(x)$ is equal to $\min_{i=1,\dots,k} \max_{f \in \mathcal{F}_i} f(x)$ for all $x \in [\gamma, 1]$. Once the sets $\mathcal{F}_1, \dots, \mathcal{F}_k$ are fixed, the value $f_0(\gamma)$ is uniquely determined and thus the function f_0 is uniquely determined by $\mathcal{F}_1, \dots, \mathcal{F}_k$ on the entire interval $[0, 1]$. Since $\mathcal{F}$ contains $2^{|\mathcal{F}|}$ subsets, there are $2^{2^{|\mathcal{F}|}}$ choices of the subsets $\mathcal{F}_1, \dots, \mathcal{F}_k$. The statement of the theorem now follows from the estimate (4.2). □

5. Convex $\boldsymbol{\lambda}$-functions. In this section, we focus on λ-graphs with convex λ-functions. We decided to study convex λ-functions in more detail since it seems that most λ-functions exhibit mixed convex-concave behavior and the first step towards understanding this behavior could be the analysis of λ-graphs with convex λ-functions. Let us start with a simple upper bound on the number of linear parts of convex λ-functions of finite λ-graphs.

THEOREM 5.1. *Let G be a finite λ-graph of order n. If the function $\lambda_G(x)$ is convex, then it consists of at most $3n^{2/3} + 1$ linear parts.*

Proof. Let k be the number of linear parts of $\lambda_G(x)$ and let the reals $x_0, \dots, x_k$ and the integers $a_1, \dots, a_k$ and $b_1, \dots, b_k$ be as in Corollary 3.4. Since the function $\lambda_G(x)$ is convex, $a_i > a_j$ and $b_i < b_j$ for every $1 \le i < j \le k$.

Let $\alpha_i = a_i - a_{i+1} > 0$ and $\beta_i = b_{i+1} - b_i > 0$ for $i = 1, \dots, k-1$. In particular, $a_1 = a_k + \alpha_1 + \cdots + \alpha_{k-1}$ and $b_k = b_1 + \beta_1 + \cdots + \beta_{k-1}$. Note that $x_i = \alpha_i/\beta_i$ for all $i = 1, \dots, k-1$. Let I_A be the set of the indices $i = 1, \dots, k-1$ such that $\alpha_i \ge n^{1/3}$, and let I_B be the set of the indices $i = 1, \dots, k-1$ such that $\beta_i \ge n^{1/3}$. Since $a_1 < n$ by Corollary 3.4, $|I_A| \le n^{2/3}$. Similarly, $|I_B| \le n^{2/3}$.

Let $I = \{1, \dots, k\} \setminus (I_A \cup I_B)$. For $i \in I$, the number $x_i = \alpha_i/\beta_i$ is a fraction with both the numerator and denominator between 1 and $n^{1/3}$. Since there are at most $n^{2/3}$ distinct fractions, we infer that $|I| \le n^{2/3}$. Consequently, $k \le |I| + |I_A| + |I_B| \le 3n^{2/3}$. The statement of the theorem now follows. □

In the rest of this section, we construct λ-graphs whose convex λ-functions have $\Omega(n^{2/3})$ linear parts. The first step towards our construction is the next proposition. We leave its straightforward proof to the reader.

PROPOSITION 5.2. *Let G be the λ-graph which is the disjoint union of a clique of order $k_1 + 1$ with 1-edges and a clique of order $k_x + 1$ with x-edges. If $k_x > k_1$, then the function $\lambda_G(x)$ consists of two linear parts meeting at the point k_1/k_x.*

The second tool is the next lemma on joins of λ-graphs.

LEMMA 5.3. *Let G_1 and G_2 be two disjoint λ-graphs with finite λ-functions, and let $G = G_1 \oplus G_2$ be the λ-graph obtained from G_1 and G_2 by adding 1-edges v_1v_2 between any pair of vertices $v_1 \in V(G_1)$ and $v_2 \in V(G_2)$. The following holds:*

$$\lambda_G(x) = \lambda_{G_1}(x) + \lambda_{G_2}(x) + 1.$$

Proof. Fix the number $x \in [0, 1]$. By Theorem 2.1, G_1 and G_2 have finitary orientations $\vec{D_1}$ and $\vec{D_2}$ with the maximum weights of an oriented path equal to $\lambda_{G_1}(x)$ and $\lambda_{G_2}(x)$. Let $\vec{D}$ be the orientation of G obtained from $\vec{D_1}$ and $\vec{D_2}$ by orienting all the edges between G_1 and G_2 from G_1 to G_2. Clearly, the maximum weight of an oriented path in $\vec{D}$ is $\lambda_{G_1}(x) + \lambda_{G_2}(x) + 1$. By Theorem 2.1, $\lambda_G(x) \le \lambda_{G_1}(x) + \lambda_{G_2}(x) + 1$. In the next paragraph, we finish the proof of the lemma by establishing the opposite inequality.

Assume for contradiction that $\lambda_G(x) < \lambda_{G_1}(x) + \lambda_{G_2}(x) + 1$. By Theorem 2.1, G has a finitary orientation $\vec{D}$ with the maximum weight of an oriented path strictly less than $\lambda_{G_1}(x) + \lambda_{G_2}(x) + 1$. On the other hand, the orientation $\vec{D}$ restricted to G_1 contains an oriented path P_1 with weight at least $\lambda_{G_1}(x)$, and $\vec{D}$ restricted to G_2 contains a path P_2 with weight at least $\lambda_{G_2}(x)$. Let G' be the subgraph of G induced by the vertices of P_1 and P_2, and let $p = |V(G')|$. Note that the orientation $\vec{D}$ is acyclic and any two vertices of G' are connected by an oriented path, which implies that there is a unique way to order the vertices of G' into a sequence $S = (v_1, v_2, v_3, \dots, v_p)$, which is topologically sorted with respect to $\vec{D}$, i.e., if $\vec{D}$ contains an edge oriented from v_i to v_j, then $i < j$. The uniqueness of S implies that for each $i < p$ the vertices v_i and v_{i+1} are connected by an oriented edge $v_i v_{i+1}$. Therefore, G' contains an oriented Hamilton path $P = v_1 v_2, \dots, v_p$. Furthermore, every x-edge of P is also an edge of P_1 or P_2, and thus the weight of P is at least $\lambda_{G_1}(x) + \lambda_{G_2}(x) + 1$. This contradicts our assumption that the weight of every oriented path in $\vec{D}$ is strictly smaller than $\lambda_{G_1}(x) + \lambda_{G_2}(x) + 1$. □

Finally, we recall some results on the Farey fractions. The Farey sequence is formed by sets F_n of rationals, where F_n is the set of all irreducible fractions a/b with $0 \le a \le b \le n$, e.g., $F_4 = \{0, 1/4, 1/3, 1/2, 2/3, 3/4, 1\}$ (note that $1/2 = 2/4$). The Farey fractions appear, e.g., in [2, 6, 17]. For our considerations, the following result [10, 29, 30] on the Farey fractions is of interest:

$$\lim_{n\to\infty} \frac{|F_n|}{n^2} = \frac{3}{\pi^2}. \tag{5.1}$$

We can now construct a λ-graph whose λ-function consists of $\Omega(n^{2/3})$ linear parts.

THEOREM 5.4. *For every positive integer n, there is a λ-graph G of order n whose λ-function consists of $\frac{\sqrt[3]{3}}{(2\pi)^{2/3}} n^{2/3} - o(n^{2/3}) \approx 0.42 n^{2/3}$ linear parts.*

Proof. Fix a positive integer k. We construct a graph G of order at most $(2k+1)|F_k|$ whose λ-function consists of $|F_k|-1$ linear parts. The statement of the theorem will then follow from the limit (5.1).

Let F_k° be the set of the Farey fractions from F_k strictly between 0 and 1. For each fraction $\frac{a}{b} \in F_k^\circ$, consider the graph $G_{a/b}$ from Proposition 5.2 with $k_1 = a$ and $k_x = b$. Note that there are $|F_k| - 2$ choices of a and b (we exclude the fractions 0 and 1). The λ-function of $G_{a/b}$ consists of two linear parts meeting at the point $\frac{a}{b}$.

Let G be the λ-graph obtained from vertex-disjoint copies of $G_{a/b}$, $\frac{a}{b} \in F_k^\circ$, by adding 1-edges between all pairs of vertices from distinct copies, i.e., $G = \bigoplus_{\frac{a}{b} \in F_k^\circ} G_{a/b}$. By Lemma 5.3, the λ-function of G is equal to the following:

$$\lambda_G(x) = |F_k^\circ| - 1 + \sum_{\frac{a}{b} \in F_k^\circ} \lambda_{G_{a/b}}(x).$$

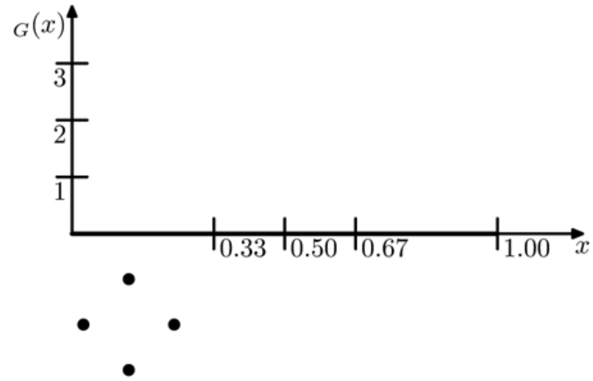
Therefore, the function $\lambda_G(x)$ consists of $|F_k| - 1$ linear parts.

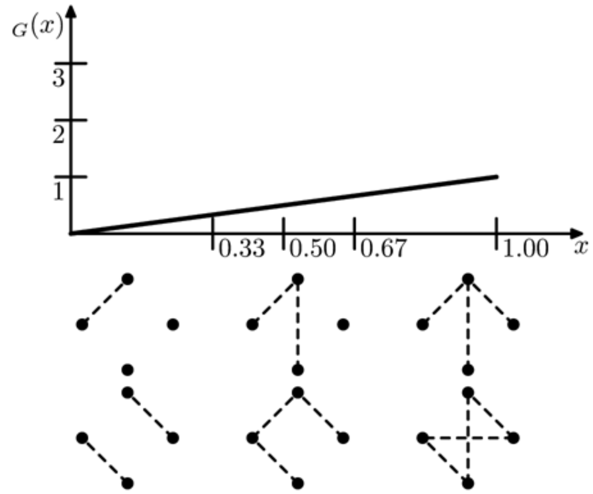
It remains to estimate the order of the λ-graph G. The order of every graph $G_{a/b}$ is at most $2k+1$. Hence, the order of G does not exceed $(2k+1)|F_k|$ as claimed in the beginning. □

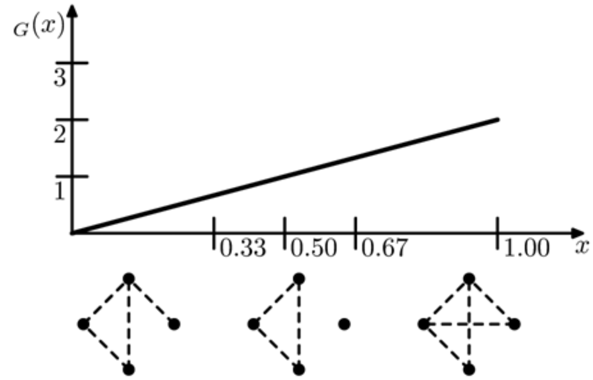
We remark that the multiplicative factors both in Theorems 5.1 and 5.4 can be improved by a finer analysis of the estimates used in the proofs. We decided not to do so in order to keep our arguments simple.

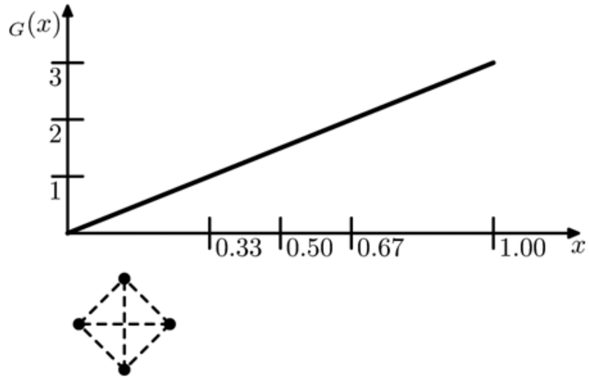
Appendix.

All λ-graphs on four vertices. First, we list all nonisomorphic λ-graphs on four vertices together with their λ-functions. The λ-graphs corresponding to the depicted λ-function can be found under the graph of the function. The 1-edges are depicted as solid segments, while the x-edges are represented by dashed segments.

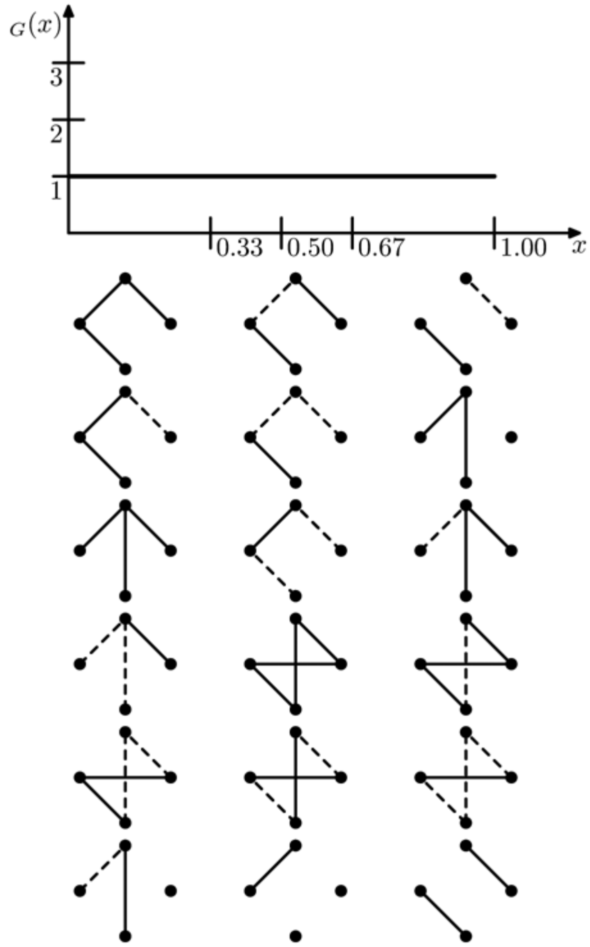

G(x)
3
2
1
0.33
0.50
0.67
1.00
x


G(x)
3
2
1
0.33
0.50
0.67
1.00
x

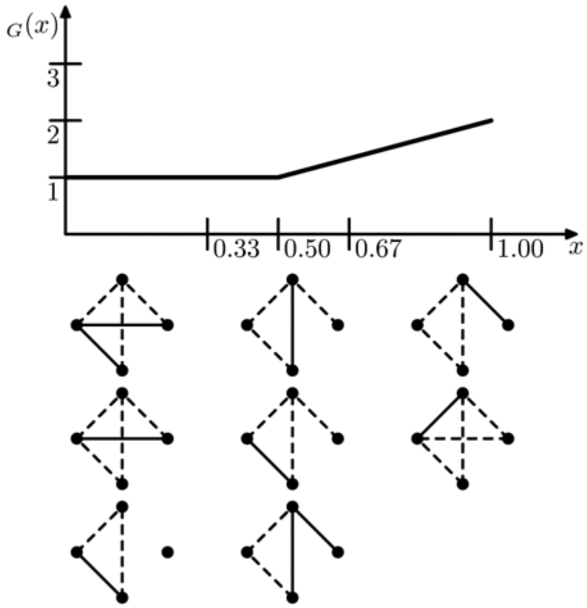

G(x)
3
2
1
0.33
0.50
0.67
1.00
x

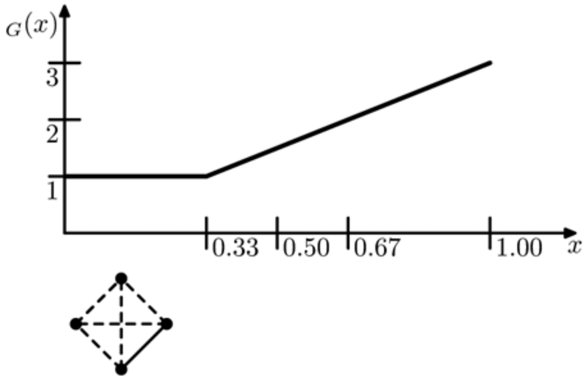


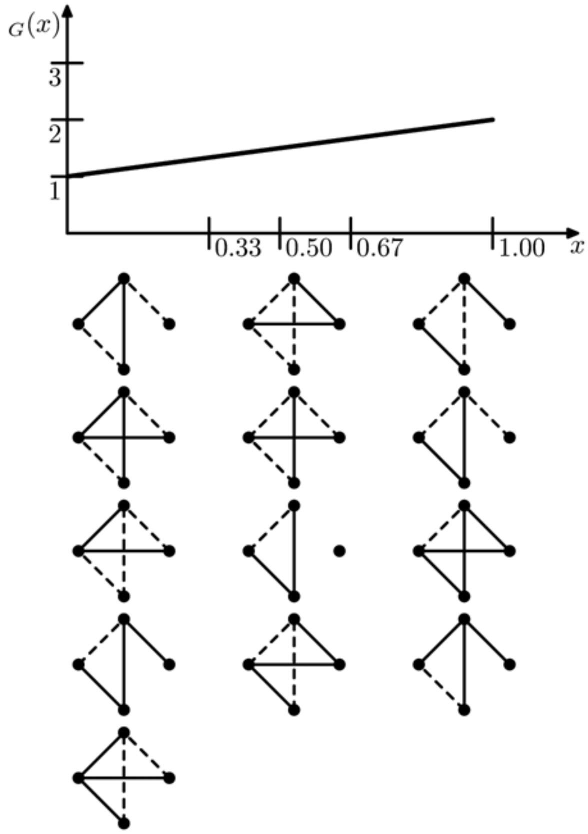
$_G(x)$
3
2
1
0.33
0.50
0.67
1.00
x

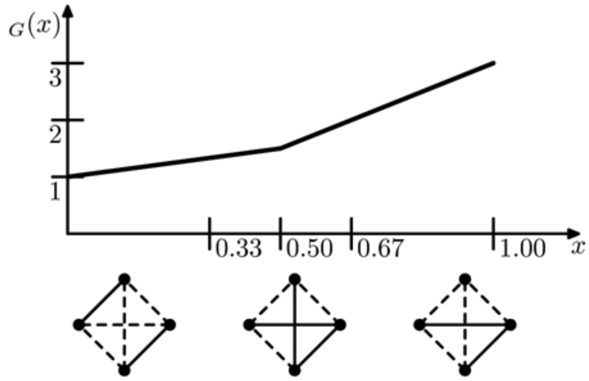


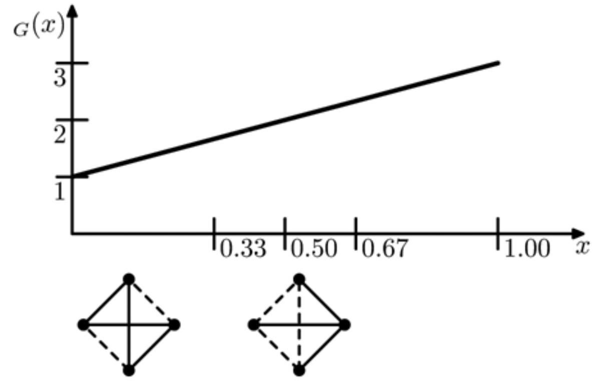
$_G(x)$
3
2
1
0.33
0.50
0.67
1.00
x

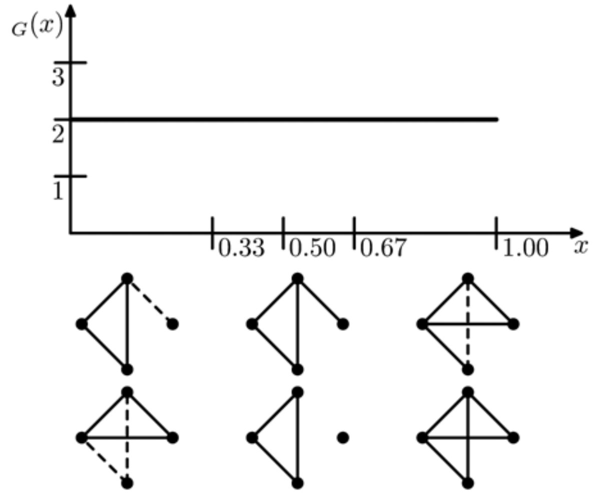

G(x)
3
2
1
0.33
0.50
0.67
1.00
x

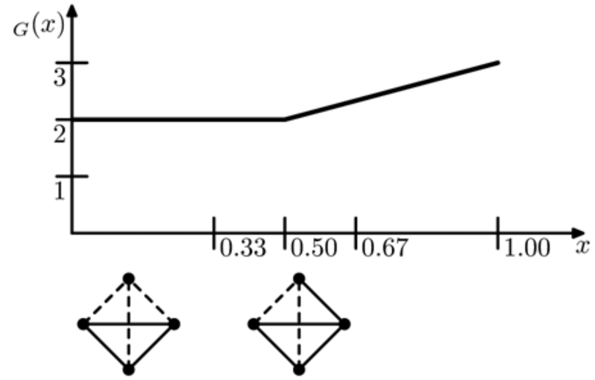

G(x)
3
2
1
0.33
0.50
0.67
1.00
x

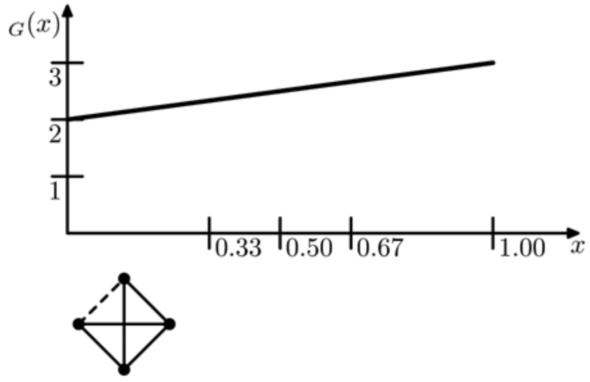

G(x)
3
2
1
0.33
0.50
0.67
1.00
x

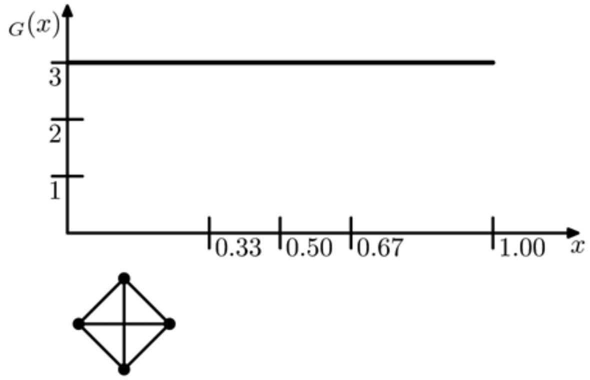

G(x)
3
2
1
0.33
0.50
0.67
1.00
x


$G(x)$
3
2
1
0.33
0.50
0.67
1.00
x

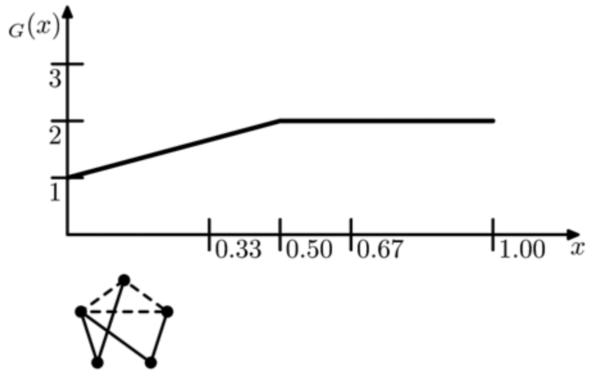

$G(x)$
3
2
1
0.33
0.50
0.67
1.00
x

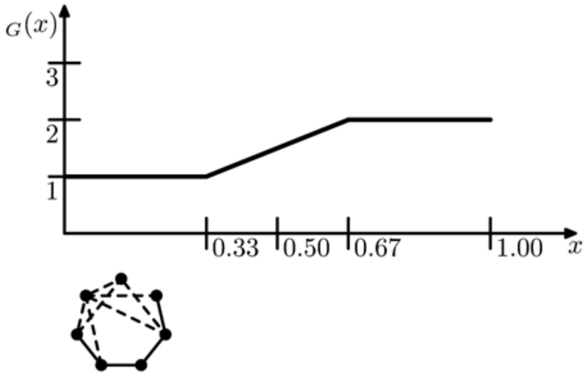

$G(x)$
3
2
1
0.33
0.50
0.67
1.00
x

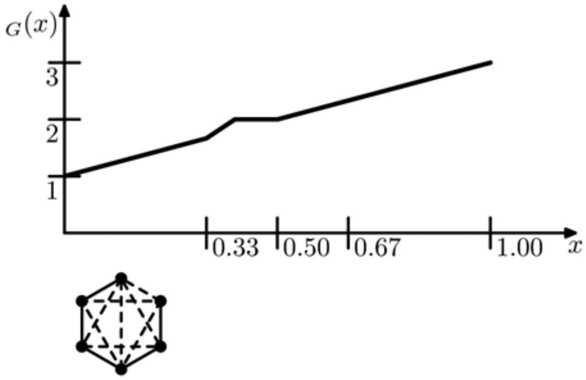




Other selected λ-graphs. We also list some other small λ-graphs with interesting λ-functions: the first one is an example of a λ-graph with a concave λ-function, the second one is an example of a λ-graph whose λ-function is neither convex nor concave (note that it even contains two different constant parts), and the third one is an example of a λ-graph such that two different linear parts of its λ-function correspond to the same linear function.







Acknowledgments. The authors would like to thank Peter Bella for discussions on the Farey fractions and the anonymous referees for their helpful comments.

REFERENCES

[1] G. Agnarsson, R. Greenlaw, and M. M. Halldórsson, *Powers of chordal graphs and their coloring*, Proceedings of the Thirty-First South Eastern International Conference on Combinatorics, Graph Theory, and Computing (Boca Raton, FL, 2000) Congr. Numer., 144 (2000), pp. 41–65.

[2] T. M. Apostol, *Modular Functions and Dirichlet Series in Number Theory*, Springer-Verlag, New York, 1997.

[3] H. L. Bodlaender, T. Kloks, R. B. Tan, and J. van Leeuwen, *λ-coloring of graphs*, in Proc. STACS'00, G. Goos, J. Hartmanis, and J. van Leeuwen, eds., Lecture Notes in Comput. Sci. 1770, Springer-Verlag, Berlin, 2000, pp. 395–406.

[4] G. J. Chang, W.-T. Ke, D. D.-F. Liu, and R. K. Yeh, *On $L(d,1)$-labelings of graphs*, Discrete Math., 220 (2000), pp. 57–66.

[5] G. J. Chang and D. Kuo, *The $L(2,1)$-labeling problem on graphs*, SIAM J. Discrete Math., 9 (1996), pp. 309–316.

[6] J. H. Conway and R. K. Guy, *The Book of Numbers*, Springer-Verlag, Berlin, 1996.

[7] J. Fiala, J. Kratochvíl, and T. Kloks, *Fixed-parameter complexity of λ-labelings*, Discrete Appl. Math., 113 (2001), pp. 59–72.

[8] D. A. Fotakis, S. E. Nikoletseas, V. G. Papadopoulou, and P. G. Spirakis, *NP-completeness results and efficient approximations for radiocoloring in planar graphs*, in Mathematical Foundations of Computer Science 2000 (Bratislava), Lecture Notes in Comput. Sci. 1893, B. Rovan, ed., Springer-Verlag, Berlin, 2000, pp. 363–372.

[9] D. Gonçalves, *On the $L(p,1)$-labelling of graphs*, Discrete Math. Theoret. Comput. Sci., AE (2005), pp. 81–86.

[10] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science*, 2nd ed., Addison-Wesley, Reading, MA, 1994.

[11] J. R. Griggs and T. X. Jin, *Real number graph labellings with distance conditions*, SIAM J. Discrete Math., 20 (2006), pp. 302–327.

[12] J. R. Griggs and T. X. Jin, *Real Number Graph Labellings of Paths and Cycles*, in preparation.

[13] J. R. Griggs and T. X. Jin, *Real Number Graph Labellings of Infinite Graphs*, in preparation.

[14] J. R. Griggs and D. Král', *Graph Labellings with Variable Weights: A Survey*, Discrete Appl. Math., to appear.

[15] J. R. Griggs and R. K. Yeh, *Labelling graphs with a condition at distance* 2, SIAM J. Discrete Math., 5 (1992), pp. 586–595.

[16] W. K. Hale, *Frequency assignment: Theory and applications*, Proceedings of the IEEE, 68 (1980), pp. 1497–1514.

[17] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, Clarendon Press, New York, 1979.

[18] T. X. Jin, Ph.D. thesis, University of South Carolina, Columbia, SC, in preparation.

[19] J.-H. Kang, $L(2,1)$*-labeling of* 3*-regular Hamiltonian graphs*, submitted.

[20] J.-H. Kang, $L(2,1)$*-Labelling of* 3*-Regular Hamiltonian Graphs*, Ph.D. thesis, University of Illinois, Urbana-Champaign, IL, 2004.

[21] D. Král', *An exact algorithm for channel assignment problem*, Discrete Appl. Math., 145 (2005), pp. 326–331.

[22] D. Král', *Coloring powers of chordal graphs*, SIAM J. Discrete Math., 18 (2004), pp. 451–461.

[23] D. Král', *The channel assignment problem with variable weights*, SIAM J. Discrete Math., 20 (2006), pp. 690–704.

[24] D. Král' and R. Škrekovski, *A theorem about channel assignment problem*, SIAM J. Discrete Math., 16 (2003), pp. 426–437.

[25] R. A. Leese and S. D. Noble, *Cyclic labellings with constraints at two distances*, Electron. J. Combin., 11 (2004), Research Paper #16.

[26] C. McDiarmid, *Discrete mathematics and radio channel assignment*, in Recent Advances in Algorithms and Combinatorics CMS Books Math., C. Linhares-Salas and B. Reed, eds., Ouvrages Math. SMC 11, Springer, New York, 2003, pp. 27–63.

[27] C. McDiarmid, *On the span in channel assignment problems: Bounds, computing and counting*, Discrete Math., 266 (2003), pp. 387–397.

[28] D. Sakai, *Labeling chordal graphs: Distance two condition*, SIAM J. Discrete Math., 7 (1994), pp. 133–140.

[29] I. Vardi, *Computational Recreations in Mathematica*, Addison-Wesley, Redwood City, CA, 1991.

[30] E. W. Weisstein, *Farey Sequence*, http://mathworld.wolfram.com/FareySequence.html.

SIAM J. DISCRETE MATH.
Vol. 21, No. 3, pp. 707–725

LARGE COMPLETE BIPARTITE SUBGRAPHS IN INCIDENCE GRAPHS OF POINTS AND HYPERPLANES*

ROEL APFELBAUM† AND MICHA SHARIR‡

Abstract. We show that if the number I of incidences between m points and n planes in $\mathbb{R}^3$ is sufficiently large, then the incidence graph (which connects points to their incident planes) contains a large complete bipartite subgraph involving r points and s planes, so that $rs \geq \frac{I^2}{mn} - a(m+n)$, for some constant $a > 0$. This is shown to be almost tight in the worst case because there are examples of arbitrarily large sets of points and planes where the largest complete bipartite incidence subgraph records only $\frac{I^2}{mn} - \frac{m+n}{16}$ incidences. We also take some steps towards generalizing this result to higher dimensions.

Key words. incidences, hyperplanes, incidence graph

AMS subject classification. 05C35, 52C10, 52C35, 52C45, 68R99

DOI. 10.1137/050641375

1. Introduction.

Let P be a set of m points, and let Π be a set of n hyperplanes in $\mathbb{R}^d$. An *incidence* in this setting is a point-hyperplane pair $(p, \pi) \in P \times \Pi$, such that $p \in \pi$. We denote by $G(P, \Pi) \subseteq P \times \Pi$ the bipartite graph whose edges connect all incident pairs and call it the *incidence graph* of P and Π. We denote by $I(P, \Pi)$ the total number of incidences $|G(P, \Pi)|$. There have been several works on point-hyperplane incidences in the past 15 years [AA, BK, EGS, ET], which we shall review later on. The reader can also consult the recent survey by Pach and Sharir [PS], which reviews some of these results.

As we show, an interesting property of point-hyperplane incidence graphs is that if the number of incidences is large (close to mn), then the incidence graph contains large complete bipartite subgraphs. Such a subgraph is in fact a configuration consisting of many hyperplanes of Π intersecting at a common lower-dimensional affine subspace H, together with many points of P, all incident to H. This property arises, in one way or another, in almost all previous works; see subsection 1.2 for details. In this paper we continue to study this property and ask: Given a point-hyperplane configuration with many incidences, what is the size of the largest complete bipartite incidence subgraph? To state the question more precisely, we define

$$\operatorname{rs}(P, \Pi) = \max \{ rs \mid K_{r,s} \subseteq G(P, \Pi) \},$$

where $K_{r,s}$ denotes the complete bipartite subgraph with r vertices on one side and s vertices on the other, and the notation $K_{r,s} \subseteq G(P, \Pi)$ means that $K_{r,s}$ is a subgraph

*Received by the editors September 28, 2005; accepted for publication (in revised form) February 12, 2007; published electronically September 12, 2007. This work research has been supported by a grant from the Israeli Academy of Sciences for a Center of Excellence in Geometric Computing at Tel Aviv University.

http://www.siam.org/journals/sidma/21-3/64137.html

†School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel (roel6@hotmail.com). This paper is part of this author's M.Sc. and Ph.D. dissertations, prepared under the supervision of the second author.

‡School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel and Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 (michas@post.tau.ac.il).This author's research was also supported by a grant from the U.S.–Israel Binational Science Foundation, by NSF Grants CCR-00-98246 and CCF-05-14079, and by the Hermann Minkowski–MINERVA Center for Geometry at Tel Aviv University.

of $G(P, \Pi)$, such that there are some r points of P and s hyperplanes of Π all incident to one another. We let

$$\mathrm{rs}_d(m, n, I) = \min_{\substack{|P| = m \\ |\Pi| = n \\ I(P, \Pi) \geq I}} \mathrm{rs}(P, \Pi)$$

denote the minimum of $\mathrm{rs}(P, \Pi)$ over all choices of a set $P \subset \mathbb{R}^d$ of m points and a set Π of n hyperplanes in $\mathbb{R}^d$, such that $I(P, \Pi) \geq I$. Note that $\mathrm{rs}_d(m, n, I) \geq \max\left\{\frac{I}{m}, \frac{I}{n}\right\}$, since there always exists a point incident to at least I/m hyperplanes and a hyperplane incident to at least I/n points, which give rise to both subgraphs $K_{I/n,1}$ and $K_{1,I/m}$. We thus have $\mathrm{rs}_d(m, n, I) \geq \frac{I}{\min\{m,n\}} = \Omega\left(\frac{I}{m} + \frac{I}{n}\right)$, and any nontrivial estimate must exceed this lower bound.

For convenience, and to prevent an overflow of constants, we make extensive use of O-notation in our bounds. An expression of the form $y = O(f(x_1, \ldots, x_n))$, is shorthand to the statement that there exist two positive constants b and C such that for each $x_1, \ldots, x_n$, if $f(x_1, \ldots, x_n) > b$, then $y < Cf(x_1, \ldots, x_n)$. Similarly, "$y = \Omega(f(x_1, \ldots, x_n))$" means that there are constants b and c, such that if $f(x_1, \ldots, x_n) > b$, then $y > cf(x_1, \ldots, x_n)$. Finally, "$y = \Theta(f(x_1, \ldots, x_n))$" means that both $y = O(f(x_1, \ldots, x_n))$ and $y = \Omega(f(x_1, \ldots, x_n))$ hold.

1.1. Our results. For the case $d = 3$, we can estimate $\mathrm{rs}_d(m, n, I)$ almost exactly.

Theorem 1.1.

(i) *If $I = \Omega(m\sqrt{n} + n\sqrt{m})$, with a sufficiently large multiplicative constant, then*

$$\mathrm{rs}_3(m, n, I) = \frac{I^2}{mn} - \Theta(m + n).$$

(ii) *If $m \leq n$, $I = O(n\sqrt{m})$, and $I = \Omega((mn)^{3/4})$, for appropriate multiplicative constants, then*

$$\mathrm{rs}_3(m, n, I) = \Theta\left(\frac{I^4}{m^2n^3} + \frac{I}{m}\right).$$

(iii) *Symmetrically, if $m \geq n$, $I = O(m\sqrt{n})$, and $I = \Omega((mn)^{3/4})$, then*

$$\mathrm{rs}_3(m, n, I) = \Theta\left(\frac{I^4}{m^3n^2} + \frac{I}{n}\right).$$

(iv) *If $I = O(m^{3/4}n^{3/4} + m + n)$, then*

$$\mathrm{rs}_3(m, n, I) = \Theta\left(\frac{I}{m} + \frac{I}{n}\right).$$

The interesting case is (i), where the number of incidences is largest. The upper bound construction for this case consists of almost disjoint complete bipartite subconfigurations.

As the dimension d increases beyond 3, the bounds that we are able to derive are less sharp. Moreover, they only apply within certain ranges of the value of I. Extending the analysis to the remaining values of I and tightening the bounds on $rs_d(m, n, I)$ seems to be much harder problems, which we leave open for further research.

We have obtained some nontrivial results for the higher-dimensional case, but not tight ones.

THEOREM 1.2 (lower bound). *If $I = \Omega(mn^{1-\frac{1}{d-1}} + m^{1-\frac{1}{d-1}}n)$, then*

$$\mathrm{rs}_d(m,n,I) = \Omega\left(\left(\frac{I}{mn}\right)^{d-1} mn\right),$$

where the constant of proportionality depends on d.

THEOREM 1.3 (upper bound). *If $I = \Omega((mn)^{1-\frac{1}{d-1}})$, then*

$$\mathrm{rs}_d(m,n,I) = O\left(\left(\frac{I}{mn}\right)^{\frac{d+1}{2}} mn\right),$$

where the constant of proportionality depends on d.

Note that for $d = 3$, both theorems yield the same bound on rs_3, which is also identical (up to multiplicative constants) to that in Theorem 1.1(i). Moreover, all three theorems apply within the same (asymptotic) range $I = \Omega(m\sqrt{n} + n\sqrt{m})$ (Theorem 1.3 applies within a wider range).

It is interesting to compare these bounds to the equivalent bounds for general graphs. There are (m,n)-bipartite graphs with $\frac{1}{2}mn$ edges, such that the largest complete bipartite subgraph has fewer than $2(m+n)$ edges. In fact, a random graph satisfies this property with very high probability. In contrast, Theorems 1.1 and 1.2 assert that a point-hyperplane incidence graph with these many edges has a complete bipartite subgraph with $\Omega(mn)$ edges.

1.2. Previous work.

The problem of bounding the number of incidences between points and curves or surfaces is one of the classical problems in combinatorial geometry and has been studied extensively during the past 20 years; see the recent survey [PS] for a comprehensive review of the state of the art in this area. Most of the study has focused on incidences in the plane, but a considerable amount of work has also been devoted to higher-dimensional problems. The specific problem of analyzing and bounding the number $I(P,\Pi)$ of incidences between a set P of m points and a set Π of n hyperplanes in d dimensions has already been studied in [AA, BK, EGS, ET].

A major issue that arises in the study of point-hyperplane incidences in $d \geq 3$ dimensions is the possible presence of many points of P incident to many hyperplanes of Π. This happens when the intersection of many of the hyperplanes is a nonzero-dimensional affine subspace, and many of the points lie in that subspace. In this case the incidence graph $G(P,\Pi)$ can be a complete bipartite graph, or contain large such subgraphs, and then $I(P,\Pi)$ can be as high as (the trivial upper bound) mn.

Several attempts can be (and have been) made to study this problem in more restricted settings. One is to assume that in $\mathbb{R}^3$ not too many points and/or not too many planes are collinear (or, for hyperplanes in higher dimensions, affinely dependent); see [EGS]. Another is to restrict the problem only to points that are vertices of the arrangement of the hyperplanes [AA, EGS]. Under these assumptions, better (nontrivial) upper bounds on $I(P,\Pi)$ can be obtained. For example:

- The maximum number of incidences between n hyperplanes in $\mathbb{R}^d$ and m vertices of their arrangement is $\Theta(m^{2/3}n^{d/3} + n^{d-1})$, for $m \geq n^{d-2}$, and $\Theta(mn)$ for $m < n^{d-2}$ [AA].

- For m points and n planes in $\mathbb{R}^3$, if no three points are collinear, the number of incidences is $O(m^{3/5}n^{4/5} + m + n)$ (see [EGS][1]). The symmetric bound $O(m^{4/5}n^{3/5} + m + n)$ holds when no three planes are collinear. Brass and Knauer [BK] give a construction from which it follows that the latter bound is tight in the worst case, when no three planes are collinear.[2]
- As Brass and Knauer [BK] show (see also Chazelle [Ch]), for m points and n hyperplanes in $\mathbb{R}^d$, and for any fixed $r, s > 0$, if the incidence graph does not contain $K_{r,s}$ as a subgraph, the number of incidences is $O(((mn)^{1-\frac{1}{d+1}} + m + n)\log(m+n))$. (In fact, using a standard analysis, based on cuttings, one can improve the bound and get rid of the logarithmic factor.)
- Elekes and Tóth [ET] have studied incidences between points and "nondegenerate" hyperplanes, where a hyperplane is considered degenerate if it contains a lower-dimensional affine subspace that contains at least a constant fraction, say β, of its incident points. Elekes and Tóth show that the number of incidences between m points and n nondegenerate hyperplanes in $\mathbb{R}^d$ is $O((mn)^{1-\frac{1}{d+1}} + mn^{1-\frac{1}{d-1}})$, where the constant of proportionality depends on d and β.

Brass and Knauer [BK] considered the general case, where the incidence graph $G(P,\Pi)$ can contain large complete bipartite subgraphs. Rather than bounding $I(P,\Pi)$ itself, they have obtained an upper bound for the overall minimum possible complexity of a representation of $G(P,\Pi)$ as the disjoint union of complete bipartite graphs, that is, $G(P,\Pi) = \bigcup_{i=1}^{s} A_i \times B_i$, where $A_i \subseteq P$ and $B_i \subseteq \Pi$, for all $i = 1, \dots, s$, and each incidence is recorded exactly once in this union. The *complexity* of such a representation of $G(P,\Pi)$ is defined to be $\sum_{i=1}^{s}(|A_i|+|B_i|)$, and the smallest complexity of such a representation, or the *representation complexity* of $G(P,\Pi)$, is denoted by $J(P,\Pi)$. We let $J_d(m,n)$ denote the maximum of $J(P,\Pi)$ over all sets P of m points and Π of n hyperplanes in $\mathbb{R}^d$. Brass and Knauer [BK] have shown that

$$J_d(m,n) = O(((mn)^{1-\frac{1}{d+1}} + m + n)\log(m+n)), \tag{1.1}$$

and that such a decomposition can be computed within the same asymptotic bound.

One way to interpret (1.1) is that if the number of incidences $I(P,\Pi)$ is much larger than $J_d(m,n)$, then $G(P,\Pi)$ should contain large complete bipartite subgraphs (or else the succinct representation would not be possible). This has been one of our main motivations to study how large must these complete bipartite subgraphs be. We strongly suspect, by the way, that the bound (1.1) can be improved by removing the logarithmic factor, which we pose as an interesting technical open problem.

On the flip side of the same coin, one would like to obtain constructions of sets P of m points, and Π of n hyperplanes, so that $G(P,\Pi)$ contains no large complete bipartite subgraphs and $I(P,\Pi)$ is as large as possible. Here too one would hope to obtain such constructions with $I(P,\Pi)$ close to $J_d(m,n)$, or, in three dimensions, to $\Theta(m^{3/4}n^{3/4} + m + n)$. The best three-dimensional construction to date is due to Brass and Knauer [BK], where $G(P,\Pi)$ does not contain any $K_{2,3}$, and $I(P,\Pi) = \Omega(m^{7/10}n^{7/10})$ in the balanced case $m \approx n$. We note that their construction actually

[1] In the original paper, this bound is multiplied by a subpolynomial factor of the form $m^\delta n^\delta$, for any $\delta > 0$. This factor, however, can be eliminated using a more refined analysis.

[2] Brass and Knauer do not derive this specific bound, although it is implicit in their construction; see later in this section and in the appendix. We remark that the symmetric case, where no three points are collinear, is *not* known to be tight in the worst case, because of some subtle aspects of point-hyperplane duality; see the appendix.

yields the bound $I(P, \Pi) = \Omega(m^{4/5}n^{3/5})$ and has the property that no three planes are collinear. Thus, the known upper bound $I(P, \Pi) = O(m^{4/5}n^{3/5} + m + n)$ for this restricted case (which, as noted above, follows from the analysis in [EGS]) is worst-case tight. For the sake of completeness, we present the construction in the appendix.

2. Many incidences yield large complete bipartite incidence subgraphs (in $\mathbb{R}^3$). In this section we prove Theorem 1.1. The main result here is the following lower bound.

THEOREM 2.1 (cf. Theorem 1.1(i)—lower bound). *Let P be a set of m points and Π a set of n planes in $\mathbb{R}^3$, with I incidences between them. Then there exists a line ℓ containing r points of P and contained in s planes of Π, such that*

$$\sqrt{rs} \geq \frac{I}{\sqrt{mn}} - \frac{a(m+n)\sqrt{mn}}{I},$$

where $a > 0$ is some sufficiently large constant.

This inequality, when squared, implies that $rs \geq \frac{I^2}{mn} - 2a(m+n)$. This establishes the lower bound of Theorem 1.1(i). Note that here there is no lower bound requirement on I, as opposed to Theorem 1.1(i), where it is required that $I = \Omega(m\sqrt{n} + n\sqrt{m})$. However, if $I < \sqrt{amn(m+n)}$, then the right-hand side is negative. Thus the theorem is interesting only for point-plane configurations with $I > \sqrt{amn(m+n)} = \Omega(m\sqrt{n} + n\sqrt{m})$.

On the upper bound side, the situation is much simpler, so we first dispose of this case.

LEMMA 2.2 (cf. Theorem 1.1(i)—upper bound). *There exist arbitrarily large configurations of m points and n planes in $\mathbb{R}^3$ with $I = \Omega(m\sqrt{n} + n\sqrt{m})$ incidences, such that every $K_{r,s}$ incidence subgraph satisfies*

$$rs \leq \frac{I^2}{mn} - \frac{1}{16}(m+n).$$

Proof. Without loss of generality, we present the construction for $m \geq n$. Fix three arbitrarily large numbers, $r \geq s \geq k \geq 2$. Take a set L of k parallel lines such that no three lines are coplanar. Then each pair of lines in L determine a distinct plane. We include all these $\binom{k}{2}$ planes in the set Π of n planes. We include in Π additional planes, each of which contains just one of the lines of L, so that each line is incident to exactly s planes. The set P of points consists of $m = rk$ elements, so that each line of L contains r points. The set Π consists of $sk - \binom{k}{2}$ planes, and $I(P, \Pi) = krs$. Put $n_0 = sk = n + \binom{k}{2}$. Note that $\mathrm{rs}(P, \Pi) = rs$, because the corresponding subgraph $K_{r,s}$ cannot contain points from three lines—no plane passes through three lines of L and if it contains points from two lines, then there is only one plane that passes through both lines. This gives

$$\mathrm{rs}(P, \Pi) = rs = \frac{I^2}{mn_0} = \frac{I^2}{m} \cdot \frac{1}{n + \binom{k}{2}} \leq \frac{I^2}{m} \cdot \frac{1}{n + k^2/4}.$$

We now use the inequality

$$\frac{1}{x+h} \leq \frac{1}{x} - \frac{h}{2x^2},$$

which holds for all $x \geq h > 0$, with $x = n$ and $h = k^2/4 \leq n$, to get

$$rs \leq \frac{I^2}{m} \cdot \left(\frac{1}{n} - \frac{k^2/4}{2n^2} \right) = \frac{I^2}{mn} - \frac{(kI)^2}{8mn^2},$$

and since $k = \frac{mn_0}{I}$, we get

$$rs \leq \frac{I^2}{mn} - \frac{(mn_0)^2}{8mn^2} \leq \frac{I^2}{mn} - \frac{m}{8} \leq \frac{I^2}{mn} - \frac{1}{16}(m+n),$$

as claimed. Note that $I = \frac{mn_0}{k} \geq m\left(\frac{n}{k} + \frac{k}{4}\right) \geq m\sqrt{n} = \Omega(m\sqrt{n} + n\sqrt{m})$ is in the required range.

The case $m \leq n$ is handled in a symmetric manner, using duality between points and planes. This completes the proof. □

Remark. A simpler construction, consisting of *disjoint* copies of $K_{r,s}$, yields the lower bound $I^2/(mn)$. Our construction shows that a lower order term proportional to $m + n$ is unavoidable.

Lemma 2.3 (cf. Theorem 1.1(iv)—upper bound). *There exist sufficiently large constants $b, C > 0$ and a sufficiently small constant $c > 0$, such that for any $m, n > b$, and for any I such that $c(m+n) < I < C(mn)^{3/4}$, there exist configurations of at most m points and at most n planes in $\mathbb{R}^3$ with at least I incidences, such that every $K_{r,s}$ incidence subgraph satisfies*

$$rs \leq \frac{6I}{\min\{m,n\}}.$$

Remark. The construction of [BK] (see the appendix), in which there are no $K_{3,2}$ or $K_{2,3}$ incidence subgraphs, provides us with an example where $rs = O(I/\min\{m,n\})$, showing the bound is asymptotically tight. This construction, however, is good only for the range $I = O(m^{3/5}n^{4/5} + m^{4/5}n^{3/5})$ and cannot be used for larger values of I within the assumed larger range $O((mn)^{3/4})$. In contrast, our construction is good for the entire range specified in Lemma 2.3, but may have complete bipartite incidence subgraphs with an arbitrarily large number of elements on both sides.

Proof. This construction resembles similar constructions of Elekes [El]. Put $k = \lfloor \sqrt{2I/n} \rfloor$, $l = \lfloor \sqrt{6I/m} \rfloor$, and $t = \left\lfloor (mn)^{3/2}/(12\sqrt{3}I^2) \right\rfloor$. With an appropriate choice of the constants, we may assume that $k, l, t \geq 100$, say, and so $k^3l^3t \geq I$. Define

$$P = \{ (x_1, x_2, x_3) \mid x_1, x_2 \in \{1, \ldots, k\}, \text{ and } x_3 \in \{1, \ldots, 3klt\} \},$$

and

$$\Pi = \{ x_3 = a_1tx_1 + a_2tx_2 + b \mid a_1, a_2 \in \{1, \ldots, l\}, \text{ and } b \in \{1, \ldots, klt\} \}.$$

The set P consists of $3k^3lt \leq m$ points, and the set Π consists of $kl^3t \leq n$ planes. Each plane is incident to k^2 points, so the number of incidences is $k^3l^3t \geq I$. In addition, each point is incident to at most l^2 planes.

Now there are three types of complete bipartite incidence subgraphs $K_{r,s}$.

1. Between a point and all its incident planes. Then $r = 1$, $s \leq l^2$, and $rs \leq l^2 \leq 6I/m$.
2. Between a plane and all its incident points. Then $r = k^2$, $s = 1$, and $rs = k^2 \leq 2I/n$.

3. Between some r collinear points and s collinear planes, all incident to the same line. Then $r \leq k$, $s \leq l$, and $rs \leq kl = 2\sqrt{3}I/\sqrt{mn} \leq 2\sqrt{3}I/\min\{m,n\}$.

In either case, we have $rs \leq \frac{6I}{\min\{m,n\}}$. This completes the proof. □

To prove Theorem 2.1, we use the results of Szemerédi and Trotter [ST] and of Elekes and Tóth [ET].

Szemerédi and Trotter's celebrated point-line incidence bound states the following theorem.

THEOREM 2.4 (see [ST]). *There exists a constant $C > 0$, such that for any m and n, the number of incidences between any m points and n lines in the plane is upper-bounded by*

$$C(mn)^{2/3} + m + n.$$

Equivalently, for any set P of n points in the plane, and for any $k \leq n$, the number of lines containing at least k points of P is bounded by

$$C'\left(\frac{n^2}{k^3} + \frac{n}{k}\right),$$

and the number of incidences between them is at most

$$C''\left(\frac{n^2}{k^2} + n\right),$$

for corresponding constants C', C''. Furthermore, these bounds are best possible.

Finding the smallest value of C for which Theorem 2.4 is true is an interesting problem in its own right. The best estimates at present are $C \geq 0.42$ [PT], and $C \leq 2.5$ [PRTT]. See also [BMP].

Elekes and Tóth [ET] bound the number of incidences between points and "degenerate" hyperplanes. We define a hyperplane π to be *β-degenerate* with respect to a point set P, if some lower-dimensional flat $F \subset \pi$ contains at least a β-fraction of the points of P incident to π, i.e., if

$$|P \cap F| \geq \beta\,|P \cap \pi|,$$

for some lower-dimensional flat $F \subset \pi$. If no such flat exists, then the hyperplane is said to be *β-nondegenerate*.[3] A hyperplane π is called *k-rich* (with respect to P) if it contains at least k points of P. Elekes and Tóth state the following theorem.

THEOREM 2.5 (see [ET]). *For any integer $d \geq 3$, there is a constant $\beta_d > 0$, such that the number of incidences between any set of m points and any set of n β_d-nondegenerate hyperplanes (with respect to the given point set) in $\mathbb{R}^d$ is*

$$O\left((mn)^{1-\frac{1}{d+1}} + mn^{1-\frac{1}{d-1}}\right).$$

Equivalently, for any set of m points in $\mathbb{R}^d$, the number of k-rich β_d-nondegenerate hyperplanes is

$$O\left(\frac{m^d}{k^{d+1}} + \frac{m^{d-1}}{k^{d-1}}\right).$$

[3] We caution the reader that this notation is the *opposite* to that used in [ET].

Furthermore, these bounds are best possible.

Using this bound, we can prove the following result, which is slightly weaker than Theorem 2.1; see below for a more detailed comparison.

LEMMA 2.6. *Let P be a set of m points and Π a set of n planes in $\mathbb{R}^3$, such that $I = I(P,\Pi) = \Omega((mn)^{3/4} + m\sqrt{n})$, with a sufficiently large multiplicative constant. Then there exists a line ℓ containing r points of P and contained in s planes of Π, such that*

$$rs = \Omega\left(\min\left\{\frac{I^4}{m^2n^3}, \frac{I^2}{mn}\right\}\right).$$

This bound is asymptotically tight in the worst case.

Proof. Applying Theorem 2.5 with $d = 3$, we see that when the constant of proportionality is chosen sufficiently large, most incidences are with planes of Π that are β_3-degenerate, i.e., for each such plane, at least a β_3-fraction of its incident points are contained in a single line.

Put $\beta = \beta_3$, and let $\Pi' \subseteq \Pi$ be the subset of those planes in Π that contain at least $I/(2n)$ points each and are β-degenerate. By the preceding argument, if the constant of proportionality in the assumed lower bound on I is sufficiently large, then $I(P,\Pi') \geq I/3$, say. We replace each plane of Π' with a line that lies on it and contains a β-fraction of its incident points. Thus, each such line contains at least $\beta I/(2n)$ points of P. By projecting these lines and the points of P onto some generic plane and applying Theorem 2.4, the number of incidences between the points of P and these lines is

$$I' = O\left(\frac{m^2}{(\beta I/(2n))^2} + m\right) = O\left(\frac{m^2n^2}{I^2} + m\right).$$

Note that $I(P,\Pi')$ differs from I', because we count in $I(P,\Pi')$ each line with its *multiplicity*, equal to the number of planes of Π' that contain it. The average multiplicity of a line is thus

$$s = \frac{I(P,\Pi')}{I'} = \Omega\left(\frac{I}{I'}\right) = \Omega\left(\min\left\{\frac{I^3}{m^2n^2}, \frac{I}{m}\right\}\right).$$

By the pigeonhole principle, some line ℓ does have at least this multiplicity, i.e., it is contained in at least s planes. By construction, it also contains $r = \Omega(I/n)$ points. Altogether, we get

$$rs = \Omega\left(\min\left\{\frac{I^4}{m^2n^3}, \frac{I^2}{mn}\right\}\right).$$

We have thus found a line ℓ with the asserted properties.

We can obtain a point-plane configuration that has a matching upper bound on $\mathrm{rs}(P,\Pi)$ of the same order of magnitude as the lower bound we have just proved (that is, unless the trivial bound $\mathrm{rs}(P,\Pi) \geq \max\{I/m, I/n\}$ dominates). This is done as follows. We take m points spanning the maximal number of lines incident to $r = \Theta(I/n)$ of the points, which, by Theorem 2.4, is

$$\Theta\left(\frac{m^2}{r^3} + \frac{m}{r}\right) = \Theta\left(\frac{m^2n^3}{I^3} + \frac{mn}{I}\right).$$

We then let each such line occur on $s = \Theta(\min\{I^3/(m^2n^2), I/m\})$ planes. The constants of proportionality are chosen so that the total number of planes is n, and the number of incidences is I. We have thus shown that the bound asserted in the lemma is asymptotically tight in the worst case. □

Proof of Theorem 1.1(ii,iii). If $I = O(n\sqrt{m})$, for an appropriate multiplicative constant, then, in the bound of Lemma 2.6, the first term $I^4/(m^2n^3)$ is smaller than the second term $I^2/(mn)$. Moreover, the lower bound on I that the lemma requires holds under the assumptions $I = \Omega((mn)^{3/4})$ and $m \le n$. Hence, under the assumptions of part (ii) of the theorem, $\mathrm{rs}_3(m,n) = \Omega(I^4/(m^2n^3))$, which clearly implies the lower bound of Theorem 1.1(ii). The upper bound also follows easily from Lemma 2.6. Finally, Theorem 1.1(iii) follows by point-plane duality. □

On the other hand, if $I = \Omega(n\sqrt{m})$, then the second term in Lemma 2.6, namely $I^2/(mn)$, dominates. We thus get the following corollary.

Corollary 2.7. *Let P be a set of m points and Π a set of n planes in $\mathbb{R}^3$, such that $I = I(P,\Pi) = \Omega(m\sqrt{n} + n\sqrt{m})$, with a sufficiently large multiplicative constant. Then there exists a line ℓ containing r points of P and contained in s planes of Π, such that*

$$rs = \Omega\left(\frac{I^2}{mn}\right).$$

(The lemma is applicable since $(mn)^{3/4}$ is always dominated by $m\sqrt{n} + n\sqrt{m}$.)

This is already very close to the bound we are trying to prove, except for the multiplicative constant. We shall now get rid of this constant and finish the proof of Theorem 2.1. Recall that the theorem states that

$$\sqrt{rs} \ge \frac{I}{\sqrt{mn}} - \frac{a(m+n)\sqrt{mn}}{I},$$

for some r points and s planes all incident to one another, and for some constant $a > 0$.

Proof of Theorem 2.1. Let P be a set of m points and Π a set of n planes in $\mathbb{R}^3$, with $I = I(P,\Pi)$ incidences. By Corollary 2.7, there exist positive absolute constants A, k, and β, such that for all $m > k$ and $n > k$, if $I > A(m\sqrt{n} + n\sqrt{m})$, then

$$\sqrt{\mathrm{rs}(P,\Pi)} \ge \frac{\beta I}{\sqrt{mn}}.$$

We choose the constant a so that it satisfies $a \ge \max\left\{4, 2A^2, k, \frac{2}{\beta}\right\}$.

The proof proceeds by induction on m and n. It is easy to see that the theorem holds for sufficiently small values of m, n, or I. More precisely, if $I < \sqrt{amn(m+n)}$, then $\frac{I}{\sqrt{mn}} - \frac{a(m+n)\sqrt{mn}}{I} < 0$, and the theorem holds trivially. Moreover, if $m \le a$ or $n \le a$, then $I \le mn < \sqrt{amn(m+n)}$. Hence, the theorem holds for all m and n such that $m \le a$ or $n \le a$.

Suppose then that $m > a$ and $n > a$ are arbitrary, and that the claim holds for all (m', n') satisfying $m' < m$ and $n' < n$. Let P be a set of m points and Π a set of n planes in $\mathbb{R}^3$ with $I > \sqrt{amn(m+n)}$ incidences between them. Let ℓ be a line that maximizes rs, where $r = |\ell \cap P|$ and $s = |\{\pi \in \Pi \mid \pi \supset \ell\}|$.

Remove from the setting all the points and planes incident to ℓ. We are left with $m-r$ points and $n-s$ planes, and denote by I_1 the number of incidences among them. We note that

$$I_1 \geq I - rs - (m+n) + (r+s). \tag{2.1}$$

Indeed, by removing the elements incident to ℓ, we lose the rs incidences between these elements. We may also lose incidences between the removed points and the surviving planes and between the removed planes and the surviving points. However, each surviving plane can be incident to at most one removed point, and each surviving point can be incident to at most one removed plane. This implies the asserted inequality (2.1).

We next choose a line ℓ_1 incident to r_1 of the remaining points and to s_1 of the remaining planes, such that $r_1 s_1$ is maximized. If $\sqrt{r_1 s_1} \geq \frac{I}{\sqrt{mn}} - \frac{a(m+n)\sqrt{mn}}{I}$, then we are done, since, by construction, $rs \geq r_1 s_1$. Otherwise, we may write

$$\begin{aligned}\frac{I}{\sqrt{mn}} - \frac{a(m+n)\sqrt{mn}}{I} > \sqrt{r_1 s_1} \geq \frac{I_1}{\sqrt{(m-r)(n-s)}} \\ - \frac{a((m+n)-(r+s))\sqrt{(m-r)(n-s)}}{I_1},\end{aligned}$$

where the right inequality follows from the induction hypothesis. Thus,

$$\begin{aligned}\frac{I}{\sqrt{mn}} > \frac{I_1}{\sqrt{(m-r)(n-s)}} \\ + a\left(\frac{(m+n)\sqrt{mn}}{I} - \frac{((m+n)-(r+s))\sqrt{(m-r)(n-s)}}{I_1}\right).\end{aligned}$$

Put

$$h = \frac{(m+n)\sqrt{mn}}{I} - \frac{((m+n)-(r+s))\sqrt{(m-r)(n-s)}}{I_1},$$

so we have

$$\frac{I}{\sqrt{mn}} > \frac{I_1}{\sqrt{(m-r)(n-s)}} + ah.$$

We now distinguish between the two cases $h \geq 0$ and $h < 0$. If $h \geq 0$, then we have

$$\frac{I}{\sqrt{mn}} > \frac{I_1}{\sqrt{(m-r)(n-s)}}, \tag{2.2}$$

or, using the inequality $(m-r)(n-s) \leq (\sqrt{mn} - \sqrt{rs})^2$, and applying (2.1),

$$\frac{I}{\sqrt{mn}} > \frac{I - rs - (m+n)}{\sqrt{mn} - \sqrt{rs}}$$

$$\Longrightarrow \quad I\sqrt{mn} - I\sqrt{rs} > I\sqrt{mn} - \sqrt{mn}rs - (m+n)\sqrt{mn}$$

$$\Longrightarrow \quad rs - \frac{I}{\sqrt{mn}}\sqrt{rs} + (m+n) > 0.$$

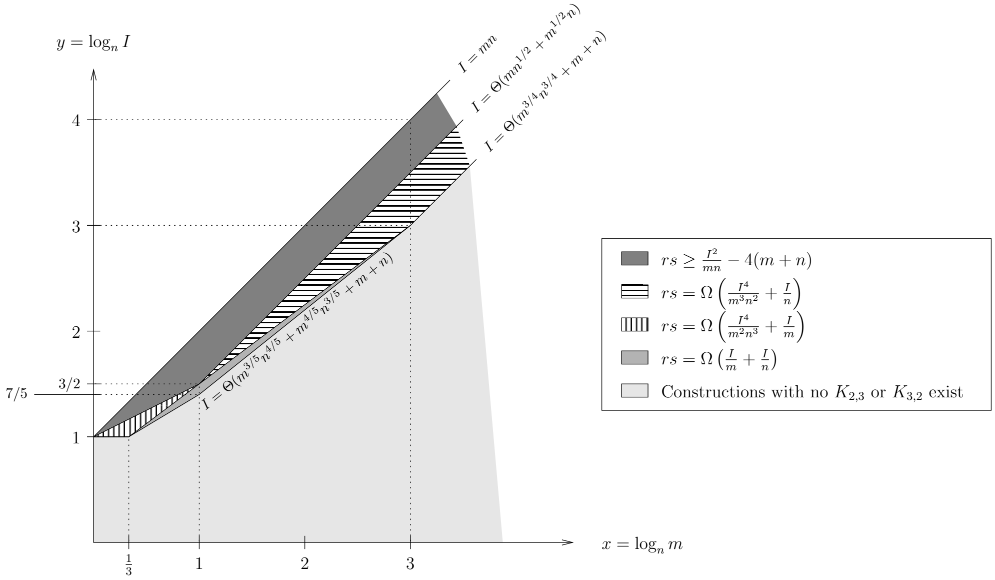


FIG. 2.1. *The known lower bounds for the maximum number of edges in a complete bipartite incidence subgraph in* $\mathbb{R}^3$.

This quadratic inequality in the variable $\sqrt{rs}$ solves to

$$\sqrt{rs} > \frac{\frac{I}{\sqrt{mn}} + \sqrt{\frac{I^2}{mn} - 4(m+n)}}{2}, \quad \text{or}$$
$$\sqrt{rs} < \frac{\frac{I}{\sqrt{mn}} - \sqrt{\frac{I^2}{mn} - 4(m+n)}}{2}.$$

Note that, since $a \geq 4$, it follows that $\frac{I^2}{mn} - 4(m+n) \geq 0$ for the assumed range of I. We can then use the inequality $\sqrt{x - \Delta x} \geq \sqrt{x} - \frac{\Delta x}{\sqrt{x}}$, which holds for $0 \leq \Delta x \leq x$, to obtain

$$\sqrt{rs} > \frac{I}{\sqrt{mn}} - \frac{2(m+n)\sqrt{mn}}{I}, \quad \text{or}$$
$$\sqrt{rs} < \frac{2(m+n)\sqrt{mn}}{I}.$$

Since $a \geq 2A^2$, it is easily checked that Corollary 2.7 is applicable for the assumed range of I and implies that $\sqrt{rs} \geq \frac{\beta I}{\sqrt{mn}}$. Hence, if the second case were possible, we would have $\frac{\beta I}{\sqrt{mn}} < \frac{2(m+n)\sqrt{mn}}{I}$, or $I < \sqrt{\frac{2}{\beta} mn(m+n)}$, which, having chosen $a \geq \frac{2}{\beta}$, would contradict our assumption on I. Hence, only the first inequality is possible, and the theorem holds in this case.

Consider now the case $h < 0$. We have

$$\frac{(m+n)\sqrt{mn}}{I} < \frac{((m+n)-(r+s))\sqrt{(m-r)(n-s)}}{I_1}$$
$$< \frac{(m+n)\sqrt{(m-r)(n-s)}}{I_1}$$

$$\Longrightarrow \quad \frac{I}{\sqrt{mn}} > \frac{I_1}{\sqrt{(m-r)(n-s)}}.$$

But this is exactly inequality (2.2), which, as we have already seen, implies $\sqrt{rs} > \frac{I}{\sqrt{mn}} - \frac{a(m+n)\sqrt{mn}}{I}$, so the theorem holds in this case too.

This completes the induction step, and thus the proof of the theorem. □

Figure 2.1 summarizes our findings. Each differently-shaded region represents certain values of m, n, and I, and has a different lower bound for rs.

3. Large complete bipartite incidence subgraphs in higher dimensions. In Lemma 2.6 we require that $I = \Omega((mn)^{3/4} + m\sqrt{n})$, because we want to ensure that most planes are "degenerate" in the sense that they can be replaced by lines, and the number of incidences will stay roughly the same. However, the lemma holds in a considerably more general setting, involving any family of "degenerate" subsets of points in any dimension. Specifically, we call a finite set of points $S \subset \mathbb{R}^d$ (β, j)-*degenerate*, if some j-flat contains at least a β-fraction of the points of S; in other words, if

$$|F \cap S| \geq \beta |S|$$

for some j-flat F of $\mathbb{R}^d$. If no such j-flat exists, we call S (β, j)-*nondegenerate*.[4] With this notion of degeneracy, Lemma 2.6 becomes a special case of the following lemma (with each plane $\pi \in \Pi$ being mapped to the set $\pi \cap P$, and the entire set of planes Π being mapped to a multiset of subsets of P).

LEMMA 3.1. *Let $P \subset \mathbb{R}^d$ be a set of m points, let $\mathcal{T} \subseteq 2^P$ be a multiset of n subsets of P, and let $0 < \beta < 1$ be some constant, such that all the members of $\mathcal{T}$ are $(\beta, 1)$-degenerate. Then there exists a subset $R \subseteq P$ of $|R| = r$ points and a subfamily $\mathcal{S} \subset \mathcal{T}$ of $|\mathcal{S}| = s$ subsets (counted with multiplicity), such that $R \subseteq S$ for each $S \in \mathcal{S}$, and*

$$rs = \Omega\left(\min\left\{\frac{I^4}{m^2n^3}, \frac{I^2}{mn}\right\}\right),$$

where $I = \sum_{T \in \mathcal{T}} |T|$.

In particular, the multiset $\mathcal{T}$ need not be induced by planes, as in Lemma 2.6, but can be induced by hyperplanes of any dimension. The proof of the lemma is omitted, but it is, essentially, identical to that of Lemma 2.6. We replace each subset $S \in \mathcal{S}$ by a line that contains a fraction of its points and estimate the average multiplicity of the lines using the Szemerédi–Trotter bound within a generic 2-plane onto which we project the points and lines.

We next obtain the following generalization of Lemma 3.1.

LEMMA 3.2. *Let $P \subset \mathbb{R}^d$ be a set of m points, let $\mathcal{T} \subseteq 2^P$ be a multiset of n subsets of P, and let $\beta > 0$ and $j \geq 1$ be some constants, such that all the members of $\mathcal{T}$ are (β, j)-degenerate. Then there exist a subset $R \subseteq P$ of $|R| = r$ points and a subfamily $\mathcal{S} \subset \mathcal{T}$ of $|\mathcal{S}| = s$ subsets (again, counted with multiplicity), such that $R \subseteq S$ for each $S \in \mathcal{S}$, and*

$$rs = \Omega\left(\min\left\{\frac{I^{j+3}}{m^{j+1}n^{j+2}}, \frac{I^{j+1}}{m^j n^j}\right\}\right),$$

[4] Again, this notation is opposite to that of [ET].

where $I = \sum_{T \in \mathcal{T}} |T|$, and the constant of proportionality depends on β and j.

Proof. The proof proceeds by double induction on j and n. The base case $j = 1$ is given by Lemma 3.1 (for any n). Suppose now that the lemma holds for $j - 1 \geq 1$, and also for j and for $n' < n$, and we shall see that it also holds for j and for n. (The base case for n, at any fixed j, is trivial, with an appropriate choice of the constants of proportionality.)

Delete from $\mathcal{T}$ all the members containing fewer than $I/(2n)$ points, and let $\mathcal{T}'$ denote the multiset of the remaining sets. We have $I' = \sum_{T \in \mathcal{T}'} |T| \geq I/2$. If $|\mathcal{T}'| < n/4$, then, by induction on n, we have subsets $R \subseteq P$ and $\mathcal{S} \subseteq \mathcal{T}'$, such that $R \subseteq S$ for each $S \in \mathcal{S}$, and

$$\begin{aligned} |R| \cdot |\mathcal{S}| &= \Omega\left(\min\left\{\frac{(I/2)^{j+3}}{m^{j+1}(n/4)^{j+2}}, \frac{(I/2)^{j+1}}{m^j(n/4)^j}\right\}\right) \\ &= \Omega\left(\min\left\{2^{j+1}\frac{I^{j+3}}{m^{j+1}n^{j+2}}, 2^{j-1}\frac{I^{j+1}}{m^j n^j}\right\}\right). \end{aligned}$$

Since $j \geq 2$, we obtain R and $\mathcal{S}$ that satisfy the asserted lower bound. We can therefore assume that there are at least $n/4$ remaining sets in $\mathcal{T}'$.

For each set $T \in \mathcal{T}'$, let π_T be a j-flat (which exists by assumption) containing at least $\beta|T| \geq \frac{\beta I}{2n}$ points of P. Project these j-flats and the points of P onto some generic $(j+1)$-space Q, and partition $\mathcal{T}'$ into two subfamilies:

$$\mathcal{T}_1 = \{T \in \mathcal{T}' \mid \pi_T \text{ is } \beta_{j+1}\text{-nondegenerate in } Q\}, \quad \text{and} \quad \mathcal{T}_2 = \mathcal{T}' \setminus \mathcal{T}_1.$$

Note that all the members of $\mathcal{T}_2$ are $(\beta\beta_{j+1}, j-1)$-degenerate, that is, informally, they are "more degenerate" than the other members of $\mathcal{T}'$. One of these two families contains at least half of the members of $\mathcal{T}'$. If $|\mathcal{T}_2| \geq |\mathcal{T}'|/2 \geq n/8$, we have, by induction on j,

$$\begin{aligned} rs &= \Omega\left(\min\left\{\frac{I^{j+2}}{m^j n^{j+1}}, \frac{I^j}{m^{j-1}n^{j-1}}\right\}\right) \\ &= \Omega\left(\min\left\{\frac{I^{j+3}}{m^{j+1}n^{j+2}}, \frac{I^{j+1}}{m^j n^j}\right\}\right), \end{aligned}$$

with an appropriate careful choice of the constants of proportionality, and the lemma holds in this case.

Suppose then that $|\mathcal{T}_1| \geq |\mathcal{T}'|/2 \geq n/8$. Put $\Pi = \{\pi_T \mid T \in \mathcal{T}_1\}$. Since the j-flats $\pi \in \Pi$ are $\frac{\beta I}{2n}$-rich and β_{j+1}-nondegenerate with respect to P (in the space Q of projection), Theorem 2.5 implies that the number of these j-flats is upper-bounded by

$$\begin{aligned} |\Pi| &= O\left(\frac{m^{j+1}}{(\beta I/(2n))^{j+2}} + \frac{m^j}{(\beta I/(2n))^j}\right) \\ &= O\left(\frac{m^{j+1}n^{j+2}}{I^{j+2}} + \frac{m^j n^j}{I^j}\right). \end{aligned}$$

Taking into account that $|\mathcal{T}_1| \geq n/8$, the average multiplicity of an element of Π is

$$\frac{|\mathcal{T}_1|}{|\Pi|} = \Omega\left(\min\left\{\frac{I^{j+2}}{m^{j+1}n^{j+1}}, \frac{I^j}{m^j n^{j-1}}\right\}\right).$$

Let $\pi \in \Pi$ be a j-flat with at least this multiplicity. Define $R = \pi \cap P$, and $\mathcal{S} = \{T \in \mathcal{T}_1 \mid \pi_T = \pi\}$. We have (i) $r = |R| \geq \frac{\beta I}{2n} = \Omega(I/n)$, (ii) $s = |\mathcal{S}| \geq |\mathcal{T}_1|/|\Pi|$, (iii) R is contained in every member of $\mathcal{S}$, and

$$rs = \Omega\left(\frac{I}{n} \cdot \frac{|\mathcal{T}_1|}{|\Pi|}\right) = \Omega\left(\min\left\{\frac{I^{j+3}}{m^{j+1}n^{j+2}}, \frac{I^{j+1}}{m^j n^j}\right\}\right),$$

as asserted by the lemma. □

As a corollary, we obtain the following theorem.

THEOREM 3.3. *If $I = \Omega((mn)^{1-\frac{1}{d+1}} + mn^{1-\frac{1}{d-1}})$, with a sufficiently large multiplicative constant, then*

$$\mathrm{rs}_d(m, n, I) = \Omega\left(\min\left\{\frac{I^{d+1}}{m^{d-1}n^d}, \frac{I^{d-1}}{m^{d-2}n^{d-2}}\right\}\right).$$

Proof. Let P be a set of m points in $\mathbb{R}^d$ and Π a set of n hyperplanes in $\mathbb{R}^d$, with $I = I(P, \Pi)$ in the assumed range. By Theorem 2.5, an appropriate choice of constants implies that most incidences are with hyperplanes of Π that are β_d-degenerate with respect to P. We map each hyperplane $\pi \in \Pi$, which is β_d-degenerate, to the set $T_\pi = P \cap \pi$, and let $\mathcal{T}$ be the multiset of all those T_π's. This multiset has $n' < n$ elements, all of which are $(\beta_d, d-2)$-degenerate and $I' \geq I/2$ incidences. By Lemma 3.2, there are subsets $R \subseteq P$ and $\mathcal{S} \subseteq \mathcal{T}$, such that $R \subseteq S$ for each $S \in \mathcal{S}$ and

$$\begin{aligned} rs &= \Omega\left(\min\left\{\frac{(I')^{d+1}}{m^{d-1}(n')^d}, \frac{(I')^{d-1}}{m^{d-2}(n')^{d-2}}\right\}\right) \\ &= \Omega\left(\min\left\{\frac{I^{d+1}}{m^{d-1}n^d}, \frac{I^{d-1}}{m^{d-2}n^{d-2}}\right\}\right), \end{aligned}$$

where $r = |R|$ and $s = |\mathcal{S}|$.

We map each member $S \in \mathcal{S}$ back to the hyperplane $\pi \in \Pi$ that satisfies $S = T_\pi$ (by the multiset structure of $\mathcal{S}$, this inverse mapping can be assumed to be well defined). We denote the resulting set of hyperplanes by Σ. Then $G(R, \Sigma) = K_{r,s}$ and rs has the asserted lower bound. This completes the proof. □

We can now prove Theorem 1.2, which states that in the range $I = \Omega(mn^{1-\frac{1}{d-1}} + nm^{1-\frac{1}{d-1}})$, we have the lower bound

$$\mathrm{rs}_d(m, n, I) = \Omega\left(\left(\frac{I}{mn}\right)^{d-1} mn\right).$$

That is, in this range the minimum in the expression provided by Theorem 3.3 is attained by the second term.

Proof of Theorem 1.2. Let P be a set of m points and Π a set of n hyperplanes in $\mathbb{R}^d$, with $I = I(P, \Pi) = \Omega(mn^{1-\frac{1}{d-1}} + nm^{1-\frac{1}{d-1}})$ incidences. This lower bound is larger than the one required in Theorem 3.3. Indeed, we have $(mn)^{1-\frac{1}{d+1}} \leq mn^{1-\frac{1}{d-1}}$ when $n \leq m^{(d-1)/2}$, and, symmetrically, $(mn)^{1-\frac{1}{d+1}} \leq nm^{1-\frac{1}{d-1}}$ when $m \leq n^{(d-1)/2}$; since $(d-1)/2 \geq 1$, one of the latter inequalities must hold. Therefore, we have in this range

$$\mathrm{rs}_d(m, n, I) = \Omega\left(\min\left\{\frac{I^{d+1}}{m^{d-1}n^d}, \frac{I^{d-1}}{m^{d-2}n^{d-2}}\right\}\right).$$

However, the minimum is attained by the second term when $I = \Omega(mn^{1/2})$, which certainly holds for $I = \Omega(mn^{1-\frac{1}{d-1}} + nm^{1-\frac{1}{d-1}})$, which therefore yields

$$\mathrm{rs}_d(m,n,I) = \Omega\left(\frac{I^{d-1}}{m^{d-2}n^{d-2}}\right) = \Omega\left(\left(\frac{I}{mn}\right)^{d-1} mn\right)$$

as claimed. □

Next, we give an upper bound construction showing that

$$\mathrm{rs}_d(m,n) = O\left(\left(\frac{I}{mn}\right)^{\frac{d+1}{2}} mn\right),$$

as asserted in Theorem 1.3.

Proof of Theorem 1.3. We start with the following d-dimensional structure, which is similar to constructions of Elekes [El]. For arbitrary integers $k, l > 0$, let $P_{d,k,l}$ and $\Pi_{d,k,l}$ denote the following respective sets of points and hyperplanes in $\mathbb{R}^d$:

$$P_{d,k,l} = \{\,(x_1,\ldots,x_d) \mid x_1,\ldots,x_{d-1} \in \{1,\ldots,k\}\,,\ \text{and } x_d \in \{1,\ldots,dkl\}\}\,,$$

$$\Pi_{d,k,l} = \left\{\, x_d = \sum_{i=1}^{d-1} a_i x_i + b \;\middle|\; a_1,\ldots,a_{d-1} \in \{1,\ldots,l\}\,,\ \text{and } b \in \{1,\ldots,kl\} \right\}.$$

Note that $|P_{d,k.l}| = dk^d l$ and $|\Pi_{d,k,l}| = kl^d$. For any hyperplane $\pi \in \Pi_{d,k,l}$, and for each choice of $x_1,\ldots,x_{d-1} \in \{1,\ldots k\}$, there is a point $(x_1,\ldots,x_{d-1},x_d) \in P_{d,k,l} \cap \pi$. The set $P_{d,k,l} \cap \pi$ is thus a $(d-1)$-lattice isomorphic to the hypercube $\{1,\ldots,k\}^{d-1}$, and contains k^{d-1} points. Hence the number of incidences between $P_{d,k,l}$ and $\Pi_{d,k,l}$ is $I = k^d l^d$.

Each j-flat $F \subset \mathbb{R}^d$, which is the intersection of some $d-j$ or more hyperplanes of $\Pi_{d,k,l}$, is the image of some j-flat of the hypercube, as embedded into any of the hyperplanes $\pi \in \Pi_{d,k,l}$ that contain F. Since any j-flat of the hypercube contains at most k^j points, we have $|F \cap P_{d,k,l}| \le k^j$. Furthermore, we have the following observation.

OBSERVATION 3.4. *Any j-flat $F \subset \mathbb{R}^d$ (for $j < d$) is contained in at most l^{d-j-1} hyperplanes of $\Pi_{d,k,l}$.*

Proof. F is the image of some affine mapping $T : \mathbb{R}^j \to \mathbb{R}^d$, that is, $T(y) = My+v$, for some matrix $M \in \mathbb{R}^{d\times j}$, with rank $\rho(M) = j$, and vector $v \in \mathbb{R}^d$.

Let $\pi \in \Pi_{d,k,l}$ be a hyperplane containing F, given by the linear equation $x_d = \sum_{i=1}^{d-1} a_i x_i + b$, for some $a_1,\ldots,a_{d-1} \in \{1,\ldots,l\}$ and $b \in \{1,\ldots,kl\}$. Put $a_d = -1$, and $a = (a_1,\ldots,a_d) \in \mathbb{R}^d$. Thus we can write $\pi = \left\{x \in \mathbb{R}^d \;\middle|\; a^{\mathrm{T}}x + b = 0\right\}$.

Since $\pi \supset F$, we have $a^{\mathrm{T}}(My+v) + b = 0$ for all $y \in \mathbb{R}^j$. In particular, for $y = 0$, we have

$$a^{\mathrm{T}}v + b = 0.$$

This gives $a^{\mathrm{T}}My = 0$, for all $y \in \mathbb{R}^j$, which is equivalent to

$$M^{\mathrm{T}}a = 0.$$

Thus, a is in the kernel of $M^{\mathrm{T}} \in \mathbb{R}^{j\times d}$. We have

$$\dim \operatorname{Ker}(M^{\mathrm{T}}) = d - \dim \operatorname{Im}(M^{\mathrm{T}}) = d - \rho(M) = d - j.$$

Hence a lies in the $(d-j)$-flat $K = \operatorname{Ker}(M^{\mathrm{T}})$. In addition, the requirement $a_d = -1$ constrains a to a hyperplane H. Note that $H \not\supset K$, since $0 \in K$, but $0 \notin H$. Hence a lies in the $(d-j-1)$-flat $K \cap H$. This flat can contain at most l^{d-j-1} points of the $l \times \cdots \times l \times 1$ lattice section. Hence there are at most l^{d-j-1} possible values of a. Once a has been determined, $b = -a^{\mathrm{T}}v$ is also uniquely determined. Thus, there are at most l^{d-j-1} possible hyperplanes $\pi \in \Pi_{d,k,l}$ containing F, and the observation is established. □

By adding another dimension to the construction, an x_{d+1}-axis, we turn every point of $P_{d,k,l}$ into a line parallel to the x_{d+1}-axis, and every $(d-1)$-hyperplane of $\Pi_{d,k,l}$ into a d-hyperplane parallel to the x_{d+1}-axis. We denote the resulting set of lines by $P'_{d,k,l}$ and the set of d-hyperplanes by $\Pi'_{d,k,l}$. These sets have the same incidence relations as the original sets of points and $(d-1)$-hyperplanes. In particular, every j-flat in $\mathbb{R}^{d+1}$, which is the intersection of some $d-j+1$ or more d-hyperplanes of $\Pi'_{d,k,l}$, contains at most k^{j-1} lines of $P'_{d,k,l}$ (all parallel to the x_{d+1}-axis), and is contained in at most l^{d-j} d-hyperplanes of $\Pi'_{d,k,l}$.

To construct an example that attains the asserted bound $rs = O((\frac{I}{mn})^{\frac{d+1}{2}} mn)$, we proceed as follows. Let $P' = P'_{d-2,k,k}$ and $\Pi' = \Pi'_{d-2,k,k}$ be sets of $(d-2)k^{d-1}$ lines and k^{d-1} $(d-2)$-flats in $\mathbb{R}^d$ (these lines and flats are constructed in $\mathbb{R}^{d-1}$, but we embed them in a natural way in $\mathbb{R}^d$). For every line $\ell \in P'$, choose μ arbitrary points on ℓ, and let P denote the overall resulting set of points, and put $m = |P| = (d-2)\mu k^{d-1}$. For every $(d-2)$-flat $\pi' \in \Pi'$, choose ν distinct arbitrary hyperplanes, i.e., $(d-1)$-flats, containing π', and let Π denote the overall resulting set of hyperplanes. The hyperplanes are chosen so that no two hyperplanes containing two different flats from Π' coincide. Put $n = |\Pi| = \nu k^{d-1}$.

Now every hyperplane $\pi \in \Pi$ contains one flat $\pi' \in \Pi'$, which contains k^{d-3} lines of P', yielding a total of μk^{d-3} points of P incident to π. The number of incidences between P and Π is thus $I = \mu\nu k^{2d-4} = \Theta(k^{-2}mn)$, or $\frac{I}{mn} = \Theta(k^{-2})$. Note that the freedom of choice of the parameters k, μ, and ν allows I to have almost any asymptotic value from $\Theta((mn)^{1-\frac{1}{d-1}})$ (choose $\mu = \nu = 1$) up to $\Theta(mn)$ (choose $k = 1$). In particular, we may assume $I = \Omega(mn^{1-\frac{2}{d+1}} + m^{1-\frac{2}{d+1}}n)$. Suppose now that $G(P, \Pi)$ contains a $K_{r,s}$ subgraph, that is, there exists some j-flat F (for some $j = 1, \ldots, d-2$) containing r points of P, and contained in s hyperplanes of Π. Without loss of generality, we may take F to be the intersection of these s hyperplanes. Thus, F is parallel to the x_{d+1}-axis, so any line of P' that meets F is fully contained in F. F contains at most k^{j-1} lines of P', hence, $r \le \mu k^{j-1}$. Also, F is contained in at most k^{d-j-2} flats of Π', hence, $s \le \nu k^{d-j-2}$. Altogether,

$$rs \le \mu\nu k^{d-3} = \underbrace{\mu k^{d-1}}_{\approx m} \cdot \underbrace{\nu k^{d-1}}_{=n} \cdot \underbrace{k^{-d-1}}_{\approx \left(\frac{I}{mn}\right)^{\frac{d+1}{2}}} = O\left(\left(\frac{I}{mn}\right)^{\frac{d+1}{2}} mn\right),$$

as claimed. □

We leave it as an open problem to close the gap between the bounds in Theorems 1.2 and 1.3, for $d \ge 4$.

4. Conclusion. We have studied the structure of point-hyperplane incidence graphs and have shown that whenever the number of incidences is large, the incidence graph contains large complete bipartite subgraphs. Specifically,

1. We have derived lower bounds on the number of edges in the largest complete bipartite incidence subgraph in three dimensions (Theorems 1.1) and in higher dimensions (Theorem 1.2).
2. We have obtained matching upper bound constructions for these lower bounds. The three-dimensional constructions (Lemmas 2.2, 2.3, and 2.6) are worst-case tight, whereas the higher-dimensional one (Theorem 1.3) is not known to be tight.
3. For each of these bounds, we have provided an estimate of how many incidences there must be in order to ensure the existence of large complete bipartite incidence subgraphs that attain the asserted lower bounds. The three-dimensional estimates are tight, whereas the higher-dimensional ones are not known to be tight.

We leave as an open problem to close the gap between the higher-dimensional bounds on the number of edges in the largest complete bipartite point-hyperplane incidence subgraph.

Appendix. Incidences between points and planes in $\mathbb{R}^3$ with no three collinear planes. An upper bound on the number of incidences between m points and n planes in $\mathbb{R}^3$ with no three collinear planes (and a symmetric bound for the dual problem, where no three points are collinear) has been known for a while. As discussed in the introduction, we attribute the result to [EGS]; the bound there is slightly weaker, but can be cleaned-up using a more careful analysis.

THEOREM A.1 (see Edelsbrunner, Guibas, and Sharir [EGS]). *Let $P \subset \mathbb{R}^3$ be a set of m points and let Π be a set of n planes in $\mathbb{R}^3$, such that no three planes of Π are collinear. Then the number of incidences is bounded by*

$$I(P, \Pi) = O(m^{4/5}n^{3/5} + m + n).$$

The symmetric bound $I(P, \Pi) = O(m^{3/5}n^{4/5} + m + n)$ holds in the dual case, where no three points of P are collinear.

The proof of the first bound uses the fact that if no three planes are collinear, then the incidence graph does not contain a $K_{2,3}$, i.e., every two distinct points lying in the intersection of three distinct planes. Note that the converse is not true, i.e., we can construct point-plane configurations with no $K_{2,3}$, but with (many) triples of collinear planes. Thus, Edelsbrunner, Guibas, and Sharir have implicitly proved a slightly stronger statement, whose proof follows the one in [EGS] almost verbatim.

THEOREM A.2. *Let $P \subset \mathbb{R}^3$ be a set of m points and let Π be a set of n planes in $\mathbb{R}^3$, such that $G(P, \Pi)$ does not contain a $K_{2,3}$ subgraph. Then the number of incidences is bounded by*

$$I(P, \Pi) = O(m^{4/5}n^{3/5} + m + n).$$

The symmetric bound $I(P, \Pi) = O(m^{3/5}n^{4/5} + m + n)$ holds in the dual case, where $G(P, \Pi)$ does not contain a $K_{3,2}$ subgraph.

Recently, Brass and Knauer [BK] constructed an example that effectively shows that these bounds are worst-case tight. For the sake of completeness, we repeat (and slightly modify) their construction here. It relies on the following result.

THEOREM A.3 (see Bárány et al. [BHPT]). *Let Q be a subset of the integer lattice in $\mathbb{R}^3$ contained in the ball of radius r centered at the origin. Assume further that* *every three distinct vectors of Q are linearly independent, and that Q is a maximal set satisfying this property. Then*

$$|Q| = \Theta\left(r^{3/2}\right).$$

Theorem A.4 (see Brass and Knauer [BK]). *For any m and n, such that $m = O(n^3)$, there exist a set P of m points and a set Π of n planes in $\mathbb{R}^3$, with no three collinear planes, such that*

$$I(P, \Pi) = \Omega(m^{4/5}n^{3/5}).$$

Proof. Let $P = \left\{1, \ldots, m^{1/3}\right\}^3$ be an $m^{1/3} \times m^{1/3} \times m^{1/3}$ lattice section. Put $r = \Theta(n^{2/5}m^{-2/15})$, and let Q be a maximal lattice subset of the ball of radius r about the origin that satisfies the property in Theorem A.3, i.e., every three vectors of Q are linearly independent, and $|Q| = \Theta(r^{3/2})$. Note that for our assumed range of m and n, we have $r > 1$, with an appropriate choice of the constants of proportionality. For each point $p \in P$ and for each vector $q \in Q$, we construct a plane through p normal to q; its equation is $x \cdot q = p \cdot q$. Let Π denote the resulting set of planes. Since each coordinate of p is an integer of magnitude at most $m^{1/3}$, and each coordinate of q is an integer of magnitude at most r, there are $O(m^{1/3}r)$ distinct values of $p \cdot q$, and the number of planes is thus $|\Pi| = |Q| \cdot O(m^{1/3}r) = O(n)$. The number of incidences between P and Π is $I(P, \Pi) = |P| \cdot |Q| = \Theta(mr^{3/2}) = \Theta(m^{4/5}n^{3/5})$, and no three planes are collinear. Indeed, suppose there were three collinear planes in Π with normals $q_1, q_2, q_3 \in Q$. These normals are all distinct and lie in the plane through the origin normal to the intersection line of the three planes and are thus linearly dependent—a contradiction. $\square$

Interestingly, this construction, when transformed to dual space, does not have the dual property that no three points are collinear. This is because the duals of three parallel planes are three collinear points, and the construction does contain many triples of parallel planes. Thus, the problem of obtaining a tight bound on the number of incidences between m points, no three of which are collinear, and n planes in $\mathbb{R}^3$, remains open. Nevertheless, the following somewhat weaker result, which follows from the dual construction, holds.

Corollary A.5. *The maximum number of incidences between m points and n planes in $\mathbb{R}^3$, such that no three points lie in two or more common planes, is $\Theta(m^{3/5}n^{4/5} + m + n)$.*

In other words, both primal and dual versions of Theorem A.2 yield bounds that are worst-case tight. In contrast, the bound in the primal version of Theorem A.1 is worst-case tight, but the bound in the dual version is not known to be tight.

Acknowledgment. We wish to thank Csaba D. Tóth, one of our referees, for his careful and very helpful review, in which he has made several important comments and noticed some inaccuracies in the original version.

REFERENCES

[AA] P. K. Agarwal and B. Aronov, *Counting facets and incidences*, Discrete Comput. Geom., 7 (1992), 359–369.

[BHPT] I. Bárány, G. Harcos, J. Pach and G. Tardos, *Covering lattice points by subspaces*, Period. Math. Hungarica, 43 (2001), 93–103.

[BK] P. Brass and C. Knauer, *On counting point-hyperplane incidences*, Comput. Geom., 25 (2003), 13–20.

[BMP] P. Brass, W. Moser, and J. Pach, *Research Problems in Discrete Geometry*, Springer, New York, 2005.

[Ch] B. Chazelle, *Cutting hyperplanes for divide-and-conquer*, Discrete Comput. Geom., 9 (1993), 145–158.

[EGS] H. Edelsbrunner, L. Guibas, and M. Sharir, *The complexity of many cells in arrangements of planes and related problems*, Discrete Comput. Geom., 5 (1990), 197–216.

[El] G. Elekes, *Sums versus products in number theory, algebra and Erdős geometry*, Paul Erdős and his Mathematics II, Bolyai Soc. Math. Stud. 11, Budapest, 2002, 241–290.

[ET] G. Elekes and C. D. Tóth, *Incidences of not too degenerate hyperplanes*, Proc. 21st Annu. ACM Sympos. Comput. Geom., Pisa, Italy, 2005, 16–21.

[PRTT] J. Pach, R. Radoicic, G. Tardos, and G. Tóth, *Improving the crossing lemma by finding more crossings in sparse graphs*, Discrete Comput. Geom., 36 (2006), 527–552.

[PS] J. Pach and M. Sharir, *Geometric incidences*, Towards a Theory of Geometric Graphs (J. Pach, ed.), Contemp. Math., Vol. 342, Amer. Math. Soc., Providence, RI, 2004, 185–223.

[PT] J. Pach and G. Tóth, *Graphs drawns with few crossings per edge*, Combinatorica, 17 (1997), 427–439.

[ST] E. Szemerédi and W. Trotter, *Extremal problems in discrete geometry*, Combinatorica, 3 (1983), 381–392.

SIAM J. Discrete Math.
Vol. 21, No. 3, pp. 726–730

CONSTRUCTING FINITE FIELD EXTENSIONS WITH LARGE ORDER ELEMENTS[*]

QI CHENG[†]

Abstract. In this paper, we present an algorithm that, given a fixed prime power q and a positive integer N, finds an integer $n \in [N, 2qN]$ and an element $\alpha \in \mathbf{F}_{q^n}$ of order greater than $5.8^{n/\log_q n}$, in time polynomial in N. We present another algorithm that finds an integer $n \in [N, N + O(N^{0.77})]$ and an element $\alpha \in \mathbf{F}_{q^n}$ of order at least $5.8^{\sqrt{n}}$, in time polynomial in N. Our result is inspired by the recent AKS primality testing algorithm [M. Agrawal, N. Kayal, and N. Saxena, *Ann. of Math.* (2), 160 (2004), pp. 781–793] and the subsequent improvements [P. Berrizbeitia, *Math. Comp.*, 74 (2005), pp. 2043–2059, Q. Cheng, in *Proceedings of the 23rd Annual International Cryptology Conference* (CRYPTO 2003), D. Boneh, ed., Lecture Notes in Comput. Sci. 2729, Springer-Verlag, Berlin, 2003, pp. 338–348, D. J. Bernstein, *Math. Comp.*, 76 (2007), pp. 389–403].

Key words. finite fields, high order elements

AMS subject classifications. 11Y16, 68Q25, 11T71

DOI. 10.1137/S0895480104445514

1. Introduction.

It is well known that every finite field has multiplicative generators, which are sometimes called primitive elements. An important open problem in computational number theory is to construct a multiplicative generator for a given finite field. Although there are plenty of generators in a finite field [7, Chapter 1, Theorem 5.1], finding one is notoriously difficult, since we do not know how to test whether an element is a generator or not without factoring integers or finding discrete logarithms. Assuming the generalized Riemann hypothesis does not seem to help.

In practice, small characteristic fields are particularly useful. In this context, one can ask a relevant but less restrictive question: for a fixed prime power q, can we find an element in $\mathbf{F}_{q^n}$ with large order in time polynomial in n? Note in the question that we are not required to give the exact order of the element. Instead, we only need to give a proof that the element has high order. Besides the apparent connection to the generator problem, the problem is interesting in its own regard [12]. However, it does not seem easier than finding a primitive element if we require the order to be greater than q^{n^c} for a constant c. A weak solution was given in [6], which presented a polynomial time algorithm producing an element with order at least $n^{\log_q n}$. Another relevant question asks one to find a number n greater than a given number N, and an element of order at least q^{n^c} in $\mathbf{F}_{q^n}$ for some constant c. The rationale of this question, which we call *the special finite field high order element problem*, is to deal with special finite fields first, and then try to increase the density of the sequence of n so that the high order element problem can be eventually solved. Von zur Gathen and Shparlinski [12, 11] have obtained the following results.

Proposition 1.1. *Let q be a fixed prime power. For any positive integer N, an integer $n \geq N$ with $n = O(N \log N)$ and an element $\alpha \in \mathbf{F}_{q^n}$ of order at least $2^{(2n)^{1/2}-2}$ can be computed in time polynomial in N.*

[*] Received by the editors August 4, 2004; accepted for publication (in revised form) March 10, 2007; published electronically September 12, 2007. Part of this paper, in its preliminary form, appeared in the Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA) 2004.
http://www.siam.org/journals/sidma/21-3/44551.html

[†] School of Computer Science, University of Oklahoma, Norman, OK 73019 (qcheng@cs.ou.edu). This research was partially supported by NSF Career Award CCR-0237845.

PROPOSITION 1.2. *Let q be a fixed prime power. For any positive integer N, an integer $n \geq N$ with $n = N + O(N/\log^c N)$ and an element $\alpha \in \mathbf{F}_{q^n}$ of order at least $2^{10q^{-12}n^{1/2}-25}$ can be computed in time polynomial in N.*

All of the previous results are based on the properties of Gauss periods. For a survey, see [12].

2. Our results.

A novel technique in the celebrated Agrawal–Kayal–Saxena primality testing algorithm and its subsequent improvements is to use polynomials of degree one to generate a large multiplicative subgroup modulo, an integer, and a polynomial. In this paper, we apply this idea to obtain a new solution to the special finite field high order element problem. Our result, which can be summarized in the following theorems, features a denser sequence of n and/or a much higher order.

THEOREM 2.1. *Let q be a fixed prime power. For a sufficiently large positive integer N we can compute in time polynomial in N an integer $n \in [N, 2qN]$ and an element $\alpha \in \mathbf{F}_{q^n}$ with an order greater than $5.8^{n/\log_q n}$.*

THEOREM 2.2. *Let q be a fixed prime power. We can compute in time polynomial in N an integer $n \in [N, N + O(N^{0.77})]$ and an element $\alpha \in \mathbf{F}_{q^n}$ with an order greater than $5.8^{\sqrt{n}}$.*

The previous theorems are based on the following result.

LEMMA 2.3. *Let r be a prime power. Let m be a positive divisor of $r-1$. Let $x^m - g$, $g \in \mathbf{F}_r$, be an irreducible polynomial over $\mathbf{F}_r$ and α be one of its roots in the extension field $\mathbf{F}_{r^m}$. Then for any $a \in \mathbf{F}_r^*$, $\alpha + a$ has an order greater than*

$$\max_{0 \leq d_- \leq d \leq m} \binom{m}{d_-}\binom{d-1}{d_- - 1}\binom{2m - d_- - d - 2}{m - d_- - 1}.$$

The finite field $\mathbf{F}_{r^m}$ is a Kummer extension of $\mathbf{F}_r$. By a numerical search [3], it can be shown that, asymptotically, $\max_{0 \leq d_- \leq d \leq m} \binom{m}{d_-}\binom{d-1}{d_- - 1}\binom{2m-d_- - d-2}{m-d_- - 1}$ is $\Omega(5.8^m)$ when we take $d_- = 0.292m$ and $d = m/2$.

Proof. Without loss of generality, suppose that $\mathbf{F}_{r^m} = \mathbf{F}_r[x]/(x^m - g)$, and $\alpha = x \pmod{x^m - g}$. Denote the order of $\alpha + a$ by s. Then $\alpha + a$ is one of the roots of $X^s = 1$. We want to estimate the number of roots of $X^s = 1$. For any $c \in (\mathbf{F}_r^*)^{(r-1)/m}$, $c\alpha + a$ is one of the roots as well, since $c\alpha + a$ is a conjugate of $\alpha + a$ over $\mathbf{F}_r$. If A is a solution and B is a solution, then AB and A/B are solutions as well. We use this fact to find more solutions. Let $c_1, c_2, \ldots, c_m$ be a list of all the elements in $(\mathbf{F}_r^*)^{(r-1)/m}$. If $(e_1, e_2, \ldots, e_m)$ and $(e_1', e_2', \ldots, e_m')$ are two different sequences of integers, suppose that $\sum_{1 \leq i \leq r-1} |e_i| = m-1$, $\sum_{1 \leq i \leq r-1} |e_i'| = m-1$, $|\{i : e_i < 0\}| = |\{i : e_i' < 0\}| = d_-$, and $\sum_{e_i < 0} |e_i| = \sum_{e_i' < 0} |e_i| = d$; we claim that $\prod_{1 \leq i \leq m}(c_i\alpha + a)^{e_i} \neq \prod_{1 \leq i \leq m}(c_i\alpha + a)^{e_i'}$. Assume that these two elements are equal, we then have

$$\prod_{1 \leq i \leq m, e_i \geq 0} (c_i\alpha + a)^{e_i} \prod_{1 \leq i \leq m, e_i' < 0} (c_i\alpha + a)^{-e_i'} = \prod_{1 \leq i \leq m, e_i < 0} (c_i\alpha + a)^{-e_i} \prod_{1 \leq i \leq m, e_i' \geq 0} (c_i\alpha + a)^{e_i'}.$$

Since $\sum_{1 \leq i \leq m, e_i \geq 0} e_i + \sum_{1 \leq i \leq m, e_i' < 0}(-e_i') = \sum_{1 \leq i \leq m, e_i < 0}(-e_i) + \sum_{1 \leq i \leq m, e_i' \geq 0} e_i' = m-1$, we obtain that

$$\prod_{1\le i\le m, e_i\ge 0}(c_ix+a)^{e_i}\prod_{1\le i\le m, e_i'<0}(c_ix+a)^{-e_i'}$$
$$=\prod_{1\le i\le m, e_i<0}(c_ix+a)^{-e_i}\prod_{1\le i\le m, e_i'\ge 0}(c_ix+a)^{e_i'}$$

in the ring $\mathbf{F}_r[x]$, contradicting the unique factorization of the ring.

Now consider the subset of $\mathbf{F}_{r^m}$:

$$S=\left\{\prod_{1\le i\le m}(c_i\alpha+a)^{e_i}\,\middle|\,\sum_{1\le i\le m}|e_i|=m-1,\,\middle|\,\{i:e_i<0\}|=d_-,\sum_{e_i<0}|e_i|=d\right\}.$$

All of the elements in S are roots of $X^s=1$. Thus $s\ge |S|$. The cardinality of S is $\binom{m}{d_-}\binom{d-1}{d_--1}\binom{2m-d_--d-2}{m-d_--1}$. The exponential size of the group generated by linear factors in a polynomial ring was known before. Using negative exponents to obtain a better bound was suggested by Voloch [10] recently. □

Does there exist an irreducible polynomial of form x^m-g over $\mathbf{F}_r$? The following lemma answers this question.

LEMMA 2.4. *The polynomial x^m-g is an irreducible polynomial over $\mathbf{F}_r$ if $m|r-1$ and g is not a lth power in $\mathbf{F}_r$ for any $l|m$ $(l>1)$; in particular, if g is a multiplicative generator of $\mathbf{F}_r$.*

Proof. Let α be a root of x^m-g over some extension of $\mathbf{F}_r$. Denote $[\mathbf{F}_r(\alpha):\mathbf{F}_r]$ by d. We have $[\mathbf{F}_r(a\alpha):\mathbf{F}_r]=d$ for any $a\in(\mathbf{F}_r^*)^{(r-1)/m}$, and $a\alpha$ is also a root of x^m-g. This implies that x^m-g can be factored into irreducible polynomials of degree d over $\mathbf{F}_q$ and $d|m$. Take the factor $f(x)$ satisfying $f(\alpha)=0$. Assume that $f(x)$ is monic and the constant coefficient of $f(x)$ is f_0. The roots of $f(x)$ have the form $\alpha, a_1\alpha,\dots,a_{d-1}\alpha$. We have $f_0=(\prod_{i=1}^{d-1}a_i)\alpha^d$. So $\alpha^d=\frac{m}{(\prod_{i=1}^{d-1}a_i)}\in\mathbf{F}_r^*$, and $(\alpha^d)^{m/d}=g$. This contradicts the condition in the lemma. □

3. The algorithms and proofs. Now we are ready to describe the algorithms. Let q be a fixed prime power. The input of the algorithm is a positive integer $N>0$. The first algorithm is designed to prove Theorem 2.1.

1. Find the smallest positive integer t such that $t(q^t-1)\ge N$. Let $n=t(q^t-1)$.
2. Find a generator in $\mathbf{F}_{q^t}$, denote it by g.
3. Solve the equation $x^{q^t-1}-g=0$ in $\mathbf{F}_{q^n}$, let α be one of the roots.
4. Output $\alpha+1$ (or $\alpha+a$ for any $a\in\mathbf{F}_{q^t}^*$).

From Step 1, we see that $N\le n\le 2qN$. Steps 2 and 3 together take time $(q^t)^{O(1)}=N^{O(1)}$. Hence the algorithm takes time $N^{O(1)}$. Applying Theorem 2.3 with $r=q^t\ge n/\log_q n$ and $m=r-1$, we get that the order of the output element is greater than 5.8^{q^t} for a sufficiently large n, which is greater than $5.8^{n/\log_q n}$. This proves Theorem 2.1.

The second algorithm is designed to prove Theorem 2.2

1. Find the smallest prime t greater than $\sqrt{N}+1$.
2. Use the algorithm described in [9, Theorem 2.4] and [8] to construct a small set $G\subseteq\mathbf{F}_{q^t}$ such that at least one of the elements in the subset is a primitive element.
3. For $g\in G$, test the irreducibility of x^t-g. Stop if x^t-g is irreducible over $\mathbf{F}_{q^{t-1}}$.

4. Solve the equation $x^t - g = 0$ in $\mathbf{F}_{q^{(t-1)t}}$, let α be one of the roots.
5. Output $\alpha + 1$ (or $\alpha + a$ for any $a \in \mathbf{F}_{q^t}^*$).

From Step 1, we see that $\sqrt{N} + 1 \leq t \leq \sqrt{N} + O\left(\sqrt{N}^{0.525}\right)$ [2]. Hence $N \leq t(t-1) = N + O(N^{0.77})$. Testing irreducibility and factoring polynomials can be solved in polynomial time if the characteristic of the field is small—and there is at least one primitive element in G. (However, the fact that $x^t - g$ is irreducible does not imply that g is a primitive element in $\mathbf{F}_{q^{t-1}}$.) Hence steps 3 and 4 together take time $(t \log q)^{O(1)} = N^{O(1)}$. The whole algorithm takes time $N^{O(1)}$. Applying Theorem 2.3 with $r = q^{t-1}$ and $m = t$, the order of the output element is greater than 5.8^t for sufficiently large n, which is greater than $5.8^{\sqrt{n}}$.

4. Concluding remarks.

A few comments are in order.

1. A similar idea can be applied to solve the problem of constructing extensions of $\mathbf{F}_{q^r}$ (q is a fixed prime power) with an element of provable high order.
2. Numerical evidences suggest that the order of g is often equal to the group order $q^n - 1$ and is close to the group order otherwise. However, it seems hard to prove that. In fact, this is one of the main obstacles in improving the space efficiency of AKS-style primality testing algorithm [1]. We make the following conjecture.
   CONJECTURE 1. *Let q be a prime power and n be a positive factor of $q - 1$. Assume that $n \geq \log q$. Let $x^n - g$ ($g \in \mathbf{F}_q$) be an irreducible polynomial over $\mathbf{F}_q$ and let α be one of its roots. Then the order of $\alpha + 1$ is greater than $q^{n/c}$ for an absolute constant c.*
3. Let p be a prime. The Artin–Schreier extension of a finite field $\mathbf{F}_p$ is $\mathbf{F}_{p^p}$. It is easy to show that $x^p - x - a = 0$ is an irreducible polynomial in $\mathbf{F}_p$ for any $a \in \mathbf{F}_p^*$. Therefore, we may take $\mathbf{F}_{p^p} = \mathbf{F}_p[x]/(x^p - x - a)$. Let $\alpha = x \pmod{x^p - x - a}$. It can be shown similarly that the order of $\alpha + b$ for any $b \in \mathbf{F}_p$ is asymptotically greater than 5.8^p.

Acknowledgments. We thank Dr. Igor Shparlinski for his suggestion to derandomize the second algorithm. We are also grateful to Dr. Pedro Berrizbeitia and Mr. Yu-Hsin Li for helpful discussions.

REFERENCES

[1] M. Agrawal, N. Kayal, and N. Saxena, *Primes is in P*, Ann. of Math. (2), 160 (2004), pp. 781–793.

[2] R. C. Baker, G. Harman, and J. Pintz, *The difference between consecutive primes* II, Proc. London Math. Soc. (3), 83 (2001), pp. 532–562.

[3] D. J. Bernstein, *Proving primality in essentially quartic random time*, Math. Comp., 76 (2007), pp. 389–403.

[4] P. Berrizbeitia, *Sharpening "primes is in p" for a large family of numbers*, Math. Comp., 74 (2005), pp. 2043–2059.

[5] Q. Cheng, *Primality proving via one round in ECPP and one iteration in AKS*, in Proceedings of the 23rd Annual International Cryptology Conference (CRYPTO 2003), CA, Santa Barbara D. Boneh, ed., Lecture Notes in Comput. Sci. 2729, Springer-Verlag, Berlin, 2003, pp. 338–348.

[6] S. Gao, *Elements of provable high orders in finite fields*, Proc. Amer. Math. Soc., 127 (1999), pp. 1615–1623.

[7] K. Prachar, *Primzahlverteilung*, Springer-Verlag, Berlin, 1957.

[8] V. Shoup, *Searching for primitive roots in finite fields*, Math. Comp., 58 (1992), pp. 369–380.

[9] I. E. Shparlinski, *Computational and Algorithmic Problems in Finite Fields*, Kluwer Academic Publishers, Dordrecht, 1992.

[10] J. F. Voloch, *On some subgroups of the multiplicative group of finite rings*, J. Théor Nombres. Bordeaux, 16 (2004), pp. 233–239.

[11] J. von zur Gathen and I. Shparlinski, *Orders of Gauss periods in finite fields*, in Proceedings of the 6th International Symposium on Algorithms and Computation, Lecture Notes in Comput. Sci. 1004, Springer-Verlag, 1995. Also appeared as *Orders of Gauss periods in finite fields*, Appl. Algebra Eng., Commun. Comput., 9 (1998), pp. 15–24.

[12] J. von zur Gathen and I. Shparlinski, *Gauss periods in finite fields*, in Proceedings of the 5th Conference of Finite Fields and their Applications, Ausgsburg, Germany, 1999, Springer-Verlag, Berlin, 2001, pp. 162–177.

SIAM J. DISCRETE MATH.
Vol. 21, No. 3, pp. 731–736

A TWO-SET PROBLEM ON COLORING THE INTEGERS[*]

JEFFREY A. RYAN[†]

Abstract. For positive integers m, r and a system of inequalities $\Re$, define $f(m, r, \Re)$ to be the minimum integer n such that for every coloring of $\{1, 2, \ldots, n\}$ with r colors, there exist two monochromatic subsets $X, Y \subseteq [1, n]$ (but not necessarily of the same color) which satisfy: (i) $\Re$, (ii) the largest number in X is less than the smallest number in Y, (iii) $|X| = |Y| = m$. Let $L_X = -2x_1 + x_{m-1} + x_m$ for $x_1, x_{m-1}, x_m \in X$, $L_Y = -2y_1 + y_{m-1} + y_m$ for $y_1, y_{m-1}, y_m \in Y$, and let $\Re := L_X \le L_Y$. In this paper we prove that $f(m, r, \Re) = 5m - 3$ and consider the corresponding question for zero-sum sets and generalize our result in the sense of the Erdős–Ginzburg–Ziv theorem.

Key words. combinatorics, Ramsey theory, number theory, sequences mod m

AMS subject classification. 05D10, 11B50

DOI. 10.1137/S0895480103429320

1. Introduction.

Following the Erdős–Ginzburg–Ziv theorem [1], several theorems of Ramsey-type have been generalized by considering $\mathbb{Z}_m$-colorings, where $\mathbb{Z}_m$ is the cyclic group of order m and zero-sum configurations, rather than two-colorings and monochromatic configurations. We call such theorems generalizations to be in the sense of Erdős–Ginzburg–Ziv (EGZ). Surveys on zero-sum problems appear in [2] and [3].

First, we introduce some notation. We denote $X <_p Y$ if and only if max $(x) <$ min (y). A mapping $\Delta : X \to C$ is called a *coloring* and we refer to C as the set of *colors*. For a nonempty subset $S \subseteq X$ let $\Delta(S)$ denote the set $\{\Delta(s) \mid s \in S\}$. We say that a subset $S \subseteq X$ is *monochromatic* if and only if $\Delta(s) = \Delta(s^*)$ for all $s, s^* \in S$. Let $L_X = -2x_1 + x_{m-1} + x_m$ for $x_1, x_{m-1}, x_m \in X$ and $L_Y = -2y_1 + y_{m-1} + y_m$ for $y_1, y_{m-1}, y_m \in Y$. For $c \in C$, where $\Delta^{-1}(c)$ is nonempty, we denote $first(c) = \min\{x \in X \mid \Delta(x) = c\}$, $last(c) = \max\{x \in X \mid \Delta(x) = c\}$, and *second to last*$(c) = \max\{x \in X \mid \Delta(x) = c \text{ and } x \neq last(c)\}$. Moreover, colorings $\Delta : \{1, 2, \ldots, n\} \to C$ will be identified with the strings $\Delta(1)\Delta(2)\cdots\Delta(n)$, and we use x^i to denote the string $xx\cdots x$ of length i. Finally, let $[a, b]$ be the set of integers $\{n \in N \mid a \le n \le b\}$.

DEFINITION 1. *Let m, r be positive integers. Let $\Re$ be a system of inequalities. Define $n = f(m, r, \Re)$ to be the least positive integer such that for every coloring $\Delta : [1, n] \to [1, r]$ there exist two monochromatic subsets $X, Y \subseteq [1, n]$ (but elements from X, Y may be colored differently) which satisfy*

(i) *$\Re$,*

(ii) *$|X| = |Y| = m$,*

(iii) *$X <_p Y$.*

In this paper we consider $\Re := L_X \le L_Y$.

At present there is no general theory which addresses generalizations of systems of inequalities and equations $\Re$ in the sense of the EGZ theorem. Partial results, which motivated this research, can be found in [6] and [7]. As of this writing these

[*] Received by the editors June 7, 2003; accepted for publication (in revised form) February 23, 2007; published electronically September 19, 2007. Many thanks to the National Science Foundation for supporting this research.

http://www.siam.org/journals/sidma/21-3/42932.html

[†] Courant Institute of Mathematical Sciences, New York University, New York, NY (JRyan@Courant.nyu.edu).

are the first two-set results concerning more than two variables x_i or y_i in a system $\Re$.

2. Two-colorings. In this section we will study $f(m, r, \Re)$. To do so, we use the following lemma.

LEMMA 2. *Let m be a positive integer, $m \geq 3$. Let $\Delta : [1, 3m-2] \to \{1, 2\}$ be a coloring. Then the following holds:*

(i) *either there exists a monochromatic m-element subset $Y \subseteq [1, 3m-2]$ with $L_Y \geq 4m-5$,*

(ii) *or there exist monochromatic m-element subsets $X, Y \subseteq [1, 3m-2]$ with $X <_p Y$ and $L_X \leq L_Y$,*

(iii) *or* $\Delta([1, 3m-2]) = 1^{m-1}2^{2m-3}12$.

Proof. Let $\Delta : [1, 3m-2] \to \{1, 2\}$ be a coloring. Without loss of generality, let $\Delta(1) = 1$. If $|\Delta^{-1}(1)| < m$, then $|\Delta^{-1}(2)| \geq 2m-1$; hence there exists a set Y with $L_Y \geq 4m-5$, and (i) follows. Therefore, we can assume that $|\Delta^{-1}(1)| \geq m$. Similarly, $|\Delta^{-1}(2)| \geq m$. We have the following two cases.

Case 1. $\Delta(3m-2) = 1$. Assume there is some coloring $\Delta : [1, 3m-2] \to \{1, 2\}$ for which the lemma does not hold. Then we must have $\Delta([m-1, 3m-3]) = 2^{2m-1}$ as otherwise (i) follows; however, this contradicts our assumption that $|\Delta^{-1}(1)| \geq m$.

Case 2. $\Delta(3m-2) = 2$. First, suppose $\Delta([1, m]) = 1^m$. If we have at least m of some color in $[m+1, 3m-2]$, then we can take $X = [1, m]$ and Y as some monochromatic m-element subset of $[m+1, 3m-2]$ so that (ii) is satisfied. Therefore, to avoid satisfying the lemma we must have no more than $m-1$ elements of either color in $[m+1, 3m-2]$; however, this contradicts our assumption that $|\Delta^{-1}(2)| \geq m$.

Therefore, suppose that there exists some $x \in [1, m]$ with $\Delta(x) = 2$. If $\Delta(3m-3) = 2$, then since $|\Delta^{-1}(2)| \geq m$ we can take $y_1 = \mathit{first}\ (2)$, $y_{m-1} = 3m-3$, $y_m = 3m-2$ so that $L_Y \geq 4m-5$, and (i) holds. So, assume $\Delta(3m-3) = 1$, and we can assume $\mathit{secondtolast}(1) < m$ as otherwise (i) would follow. This implies that $\Delta([m, 3m-4]) = 2^{2m-3}$. Since $|\Delta^{-1}(1)| \geq m$, we must have $\Delta([1, 3m-2]) = 1^{m-1}2^{2m-3}12$, satisfying (iii). In this case, we can take $y_1 = 1, y_{m-1} = m-1, y_m = 3m-3$ for $L_Y = 4m-6$. □

THEOREM 3. *For each positive integer $m \geq 3$,*

$$f(m, 2, \Re) = 5m-3.$$

Proof. The coloring $\Delta : [1, 5m-4] \to \{1, 2\}$ given by the string

$$1^{m-2}2^{m-2}121^{2m-1}2^{m-1}$$

shows the lower bound $f(m, 2, \Re) \geq 5m-3$.

To see that $f(m, 2, \Re) \leq 5m-3$, consider an arbitrary coloring $\Delta : [1, 5m-3] \to \{1, 2\}$. Without loss of generality let $\Delta(1) = 1$. By the pigeonhole principle the set $[1, 2m-1]$ contains a monochromatic m-element subset X with $L_X \leq 4m-5$. Next, consider the restriction of the coloring Δ to the set $[2m, 5m-3]$ which is isometric to the set $[1, 3m-2]$ and apply Lemma 2. The only coloring $\Delta : [1, 2m-1] \to \{1, 2\}$ that avoids some monochromatic m-element subset X with $L_X \leq 4m-6$ is $\Delta([1, 2m-1]) = 12^{m-1}1^{m-1}$, so we have two strings left to consider:

(a) $12^{m-1}1^{2m-2}2^{2m-3}12$,

(b) $12^{m-1}1^{m-1}2^{m-1}1^{2m-3}21$.

In (a), we can take $X = [m+1, 2m]$ and $Y = [3m-1, 4m]$ so that we have $L_X \leq L_Y$. In (b), we can take $x_1 = 2, x_{m-1} = m, x_m = 2m$ for $L_X = 3m-4$, and $y_1 = 3m-1, y_{m-1} = 5m-5, y_m = 5m-3$ for $L_Y = 4m-6$, and we have $L_X \leq L_Y$. □

3. Zero-sum sets.

THEOREM 4. (see Erdős, Ginzburg, and Ziv [1]). *If $A = a_1, a_2, \ldots, a_{2m-1}$ is a sequence of integers, then there are m indices $i_1, i_2, \ldots i_m \in \{1, 2, \ldots, 2m-1\}$ such that*

$$a_{i_1} + a_{i_2} + \cdots + a_{i_m} \equiv 0 \pmod{m}.$$

THEOREM 5 (see Caro [2]). *Let S be a finite set with $|S| = 2m-2$. If $\Delta : S \to \mathbb{Z}_m$ is a coloring and S does not contain a zero-sum m-element subset T, then $\Delta(S) = \{a, b\}$ and $|\Delta^{-1}(a)| = |\Delta^{-1}(b)| = m-1$.*

Let $A = a_1, a_2, \ldots, a_t$ be a sequence. If b is an element of A, which belongs to a residue class X, then $|x|_A$ denotes the cardinality of the set $\{i \mid a_i \in x$, where a_i is in the sequence $A\}$. We say that the sequence A is arranged normally with parameters $v_1, v_2, \ldots, v_s$, if there are s distinct residue classes modulo m, say $x_1, x_2, \ldots, x_s$, where $|x_1|_A \geq |x_2|_A \geq \ldots \geq |x_s|_A$ and $|x_i|_A = v_i$ for $i = 1, 2, \ldots, s$ such that the first v_1 elements A belong to x_1, the next v_2 of the elements of A belong to x_2, and so on up to the last v_s of the elements of A which belong to x_s.

THEOREM 6. (see Bialostocki and Lotspeich [5]). *Let m be an integer, $m \geq 3$, and let $A = a_1, a_2, \ldots, a_{2m-3}$ be a sequence of integers. Suppose that A is arranged normally with parameters $v_1, v_2, v_3 (v_1$ and v_2 if $m = 3$.) If there are no m indices $i_1, i_2, \ldots, i_m \in \{1, 2, \ldots, 2m-3\}$ such that*

$$a_{i_1} + a_{i_2} + \cdots + a_{i_m} \equiv 0 \pmod{m},$$

then $v_1 = m-1, v_2 = m-3$, and $v_3 = 1 (v_1 = 2$ and $v_2 = 1$, if $m = 3)$.

DEFINITION 7. *Let m, r be positive integers. Let $\Re$ be a system of inequalities. Define $n = f(m, \mathbb{Z}_m, \Re)$ to be the least positive integer such that for every coloring $\Delta : [1, n] \to \mathbb{Z}_m$ there exist two zero-sum subsets $X, Y \subseteq [1, n]$ which satisfy*

(i) *$\Re$,*

(ii) *$|X| = |Y| = m$,*

(iii) *$X <_p Y$.*

Again, we consider $\Re := L_X \leq L_Y$.

4. Zero-sum generalizations. We now determine the value of $f(m, \mathbb{Z}_m, \Re)$. We use the following lemma.

LEMMA 8. *Let m be an integer, $m \geq 3$, and let $\Delta : [1, 3m-2] \to \mathbb{Z}_m$ be a coloring that uses at least three colors. Then the following holds:*

(i) *either there exists a zero-sum m-element subset $Y \subseteq [1, 3m-2]$ with $L_Y \geq 4m-5$,*

(ii) *or there exist two zero-sum m-element subsets $X, Y \subseteq [1, 3m-2]$ with $X <_p Y$ and $L_X \leq L_Y$.*

Proof. Let $[1, 3m-2]$ be the disjoint union of P, Q, and R, where $P = [1, m-2]$, $Q = [m-1, 2m-1]$, and $R = [2m, 3m-2]$. If there is a monochromatic m-element subset $B \subseteq P \cup R$, then we have $L_B \geq 4m-5$, and (i) holds. Therefore, assume there is no such set B. Then as $|P \cup R| = 2m-3$, we have by Theorem 6 that $P \cup R$ contains $m-3$ elements of one color, say a, $m-1$ elements of another color (different from a), say b, and one element of a third color that may or may not be a and cannot be b, as otherwise we would have a zero-sum m-element subset of color b . Let us call this third color c. Now, either $P \cup R$ is colored by two colors or it is colored by three according to c. We have the following two cases.

Case 1. $c \neq a$. Recall that we also must have $c \neq b$. Let $\Delta(m-1) = \alpha$. By Theorem 5 we must have a zero-sum m-element subset $X \subseteq P \cup \{m-1\} \cup R$. We have three subcases according to α.

Subcase (1a). $\alpha = a$. If X has less than $m-2$ elements of color a, then we would have a zero-sum m-element subset of $P \cup R$, a contradiction. Therefore, assume that X has $m-2$ elements of color a. We know that X has at least one element of color b and there must be at least one element $z \in R$ such that $\Delta(z) = b$. If $|\{\Delta(x) = a | x \in R\}| \geq 1$, then we can take x_{m-1} and x_m to be elements from R and (i) holds. Therefore, assume that $|\{\Delta(x) = a | x \in R\}| = 0$. If $\Delta^{-1}(c)$ is not an element of X, then there are two elements of color b in X, and so we can choose x_{m-1}, x_m as two b-colored elements of R and (i) follows. Now, assume $\Delta^{-1}(c) \in X$. If $\Delta^{-1}(c) \in R$, then we can choose x_{m-1}, x_m as z and c in R and (i) holds. If $\Delta^{-1}(c) \in P$, then $\Delta(R) = b^{m-1}$; therefore, we can take $x_1 = 1, x_{m-1} = m-1$, and $x_m = 3m-2$ so that $L_X = 4m-5$ and (i) holds.

Subcase (1b). $\alpha = b$. Then all of the elements of X must have color b. If $|\{\Delta(x) = b | x \in R\}| \geq 2$, then (i) holds. So, as we must have $|\{\Delta(x) = b | x \in R\}| \geq 1$, assume $|\{\Delta(x) = b | x \in R\}| = 1$, which implies that $\Delta([1, m-1]) = b^{m-1}$. Suppose $\Delta(m) = b$. Then we have $\Delta([1, m]) = b^m$, and so $[1, m]$ is zero-sum. If there is another zero-sum m-element subset $Y \subseteq [m+1, 3m-2]$, then (ii) follows. So, assume there is no such set Y. In this case we have by Theorem 4 that there is a zero-sum m-element subset $Z \subseteq \{1\} \cup [m+1, 3m-2]$, and since there is no m-element zero-sum set $Y \subseteq [m+1, 3m-2]$, we must have $z_1 = 1$ and (i) follows.

Therefore, assume $\Delta(m) \neq b$. The set $[1, m-3] \cup \{m\} \cup R$ contains at most $(m-2)$ elements of one color and by Theorem 6 must contain a zero-sum m-element subset and (i) follows.

Subcase (1c). $\alpha \neq a$ and $\alpha \neq b$. In this case we must have $(m-1) \in X$ as otherwise $P \cup Q$ would contain a zero-sum m-element subset B, contradicting our assumption that there is no such B. Let us assume that there is some coloring $\varphi : [1, 3m-2] \to \mathbb{Z}_m$ that uses at least three colors with $c \neq a$, $\varphi(m-1) \neq a$, $\varphi(m-1) \neq b$, and no monochromatic m-element subset $B \subseteq P \cup R$ for which the lemma does not hold.

Suppose $|\{\varphi^{-1}(b) \in X\}| \geq 2$. Then we must have $\varphi(P) = b^{m-2}$, as otherwise we could take x_{m-1} and x_m from R and (i) follows. This implies that $|\{\varphi^{-1}(b) \in X \backslash (m-1)\}| = m-1$, which implies that $\varphi(m-1) = b$, a contradiction.

Therefore, we must have $|\{\varphi^{-1}(b) \in X\}| = 1$. This implies that we must have $m-3$ elements of color a in X, and that X must include the element of color c. Since we must have at least one element of color b in R, we must have no elements of colors a or c in R, as otherwise (i) would follow. Therefore, $\varphi(R) = b^{m-1}$, but in this case we can take $x_1 = 1, x_{m-1} = m-1$, and $x_m = 3m-2$ so that $L_X = 4m-5$, satisfying (i) and contradicting our assumption that there exists a coloring φ with the above assumptions for which the lemma does not hold.

Case 2. $c = a$. Then in $P \cup R$ we have $(m-2)$ elements of color a and $(m-1)$ elements of color b. By assumption, there is some $\beta \in Q$ such that $\Delta(\beta) \neq a$ and $\Delta(\beta) \neq b$, and let $\Delta(\beta) = c$. By Theorem 5 we must have a zero-sum m-element subset $X \subseteq P \cup \{\beta\} \cup R$. Since $\beta \in X$, X must contain elements of both colors a and b. We have $\Delta(1) \in \{a, b\}$, and so we can take $x_1 = 1$ in each of the following three subcases.

Subcase (2a). X has exactly one element of color a. If $\Delta(1) = a$, then $|\{\Delta(x) = b | x \in R\}| \geq 2$ and we can take x_{m-1} and x_m from R and (i) follows. If $\Delta(1) = b$, then we can assume $|\{\Delta(x) = b | x \in R\}| = 1$, as otherwise (i) holds. This implies that R contains $m-2$ elements of color a. In this case, we can take x_{m-1} and x_m to be two elements of colors a and b from R, which implies (i).

Subcase (2b). X has exactly one element of color b. If $\Delta(1) = a$, we can assume that $|\{\Delta(x) = a | x \in R\}| = 0$, as otherwise (i) follows. Then, we can take $x_{m-1} = \beta$ and $x_m = 3m-2$ so that $L_X \geq -2 + (m-1) + (3m-2) = 4m-5$. Therefore, assume that $\Delta(1) = b$. If $|\{\Delta(x) = a | x \in R\}| \geq 2$ (i) follows, so we must have exactly one element of color a in R. Therefore $\Delta(P) = ba^{m-3}$, and in this case we can take $x_1 = 2$ and x_{m-1} and x_m to be elements of colors a and b in $\Re$, satisfying (i).

Subcase (2c). X has more than one element of color a and more than one element of color b. In this case, we can take x_{m-1} and x_m from R, implying (i), and the lemma follows. □

COROLLARY 9. *Any coloring of* $[1, 3m-2]$ *that avoids satisfying either condition* (i) *or condition* (ii) *of the lemma must be a two-coloring.*

THEOREM 10. *Let* m *be an integer,* $m \geq 3$. *Then,*

$$f(m, \mathbb{Z}_m, \Re) = 5m - 3.$$

Proof. The string $1^{m-2}0^{m-2}101^{2m-1}0^{m-1}$, which corresponds to a coloring $\Delta : [1, 5m-4] \to \{0, 1\}$, implies the lower bound $f(m, \mathbb{Z}_m, \Re) \geq 5m-3$. Next, we show the upper bound $f(m, \mathbb{Z}_m, \Re) \leq 5m-3$. By the pigeonhole principle we are guaranteed a monochromatic m-element subset $X \subseteq [1, 2m-1]$ with $L_X \leq 2m-1$. By Lemma 8 we are guaranteed some monochromatic m-element subset $Y \subseteq [2m, 5m-3]$ with $L_Y \geq 4m-5$ unless $[2m, 5m-3]$ is two-colored, and then applying Lemma 2 shows that $\Delta([2m, 5m-3]) = c^{m-1}d^{2m-3}cd$ for some $c, d \in \mathbb{Z}_m$. Since the only colorings of $[1, 2m-1]$ that avoid a zero-sum m-element subset X with $L_X \leq 4m-6$ are those of the form $ab^{m-1}a^{m-1}$, the only strings that we have left to check are those of the form $ab^{m-1}a^{m-1}c^{m-1}d^{2m-3}cd$. First, consider the case where $\Delta(2m) = b$. Then we have a set X with $L_X = 3m-4$, and we can take Y from among the last $2m-1$ elements so that $L_Y \geq 3m-4$, and we have $L_X \leq L_Y$. Second, consider the case where $\Delta(2m) \neq b$. Then in $[2, 2m]$ we must have a set X with $L_X \leq 4m-6$, and we can take Y from the last $(2m-1)$ elements so that $L_Y = 4m-6$ and we have $L_X \leq L_Y$; hence, the theorem is proven. □

Acknowledgment. Many thanks to Arie Bialostocki for his guidance.

REFERENCES

[1] P. Erdős, A. Ginzburg, and A. Ziv, *Theorem in additive number theory*, Bull. Research Council Israel, 10 (1961), pp. 41–43.

[2] Y. Caro, *Zero-sum problems—A survey*, Discrete Math., 152 (1996), pp. 93–113.

[3] A. Bialostocki, *Zero sum trees: A survey of results and open problems*, in Finite and Infinite Combinatorics in Sets and Logic (Banff, AB, 1991), 19-29, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 411, Kluwer Acad. Publ., Dordrecht, 1993.

[4] A. Bialostocki and P. Dierker, *On the Erdős-Ginzburg-Ziv theorem and the Ramsey numbers for stars and matchings*, Discrete Math., 110 (1992), pp. 1–8.

[5] A. Bialostocki and M. Lotspeich, *Some Developments of the Erdős-Ginzburg-Ziv Theorem* I, Sets, Graphs and Numbers, Colloq. Math. Soc. János Bolyai 60, (1991), Budapest, North-Holland, Amsterdam, 1992, pp. 97–117.

[6] A. Bialostocki, G. Bialostocki, and D. Schaal, *A zero-sum theorem*, J. Combin. Theory Ser. A, 101 (2003), pp. 147–152.

[7] A. Bialostocki, R. Sabar, and D. Schaal, *On a Generalization of a Variation of Schur's Equation*, preprint, 2003.

[8] Y. Caro, *On zero-sum Ramsey numbers—Stars*, Discrete Math., 104 (1992), pp. 1–6.

SIAM J. DISCRETE MATH.
Vol. 21, No. 3, pp. 737–749

CROSSING STARS IN TOPOLOGICAL GRAPHS[*]

GÁBOR TARDOS[†] AND GÉZA TÓTH[‡]

Abstract. Let G be a graph without loops or multiple edges drawn in the plane. It is shown that, for any k, if G has at least $C_k n$ edges and n vertices, then it contains three sets of k edges, such that every edge in any of the sets crosses all edges in the other two sets. Furthermore, two of the three sets can be chosen such that all k edges in the set have a common vertex.

Key words. topological graph, crossing edges

AMS subject classifications. 05C35, 52C45, 52C10

DOI. 10.1137/050623693

1. Introduction.

A *topological graph* is a graph drawn in the plane with no loops or multiple edges so that its vertices are represented by points and its edges are represented by Jordan curves connecting the corresponding points. We do not distinguish these points and curves of the topological graph from the vertices and edges of the underlying abstract graph they represent. We assume that (i) the edges of a topological graph do not pass through any vertex, (ii) two edges share a finite number of interior points and properly cross each other, and (iii) no three edges cross at the same point. Conditions (ii) and (iii) are simplifying assumptions only; graph drawings violating them can be modified to satisfy them without affecting which pairs of edges cross. A topological graph is called *simple* if any pair of its edges have at most one point in common (either a common endpoint or a crossing).

It is well known that every planar graph with n vertices has at most $3n - 6$ edges. Equivalently, every topological graph G with n vertices and more than $3n - 6$ edges has a pair of crossing edges. This simple statement was generalized in several directions. For a survey, see [P99].

Pach and Tóth [PT97] and Pach et al. [PRTT04] proved that a topological graph of n vertices and more than $(k + 2)(n - 2)$ edges must have k edges that cross the same edge. This bound is tight for $k = 1, 2, 3$ but can be substantially improved for large values of k.

For $k \geq 2$, let $f_k(n)$ (resp., $f_k^s(n)$) be the maximum number of edges of a topological graph (resp., simple topological graph) on n vertices and no k pairwise crossing edges. Agarwal et al. [AAPPS97] proved (for simple topological graphs), and Pach, Radoičić, and Tóth [PRT03] proved, with a shorter and more general argument, that for some $c > 0$, every topological graph with n vertices and more than cn edges has *three* pairwise crossing edges. That is, $f_3^s(n) \leq f_3(n) \leq cn$. Very recently, Ackerman and Tardos [AT07] proved that $7n - O(1) \leq f_3(n) \leq 8n$ and that $f_3^s(n) = 6.5n + \Theta(1)$. Moreover, Ackerman [A06] managed to prove that $f_4(n) \leq 36n$. For $m \geq 5$ the

[*] Received by the editors February 1, 2005; accepted for publication (in revised form) March 5, 2007; published electronically September 26, 2007. A preliminary version of this work appeared in [TT05].

http://www.siam.org/journals/sidma/21-3/62369.html

[†] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada, V5A 1S6, and Rényi Institute, Hungarian Academy of Sciences, Budapest, Hungary (tardos@cs.sfu.ca). The work of this author was supported by NSERC grant 611470 and by Hungarian Foundation for Scientific Research (OTKA) grants T037846, T046234, AT048826, and NK62321.

[‡] Rényi Institute, Hungarian Academy of Sciences, Budapest, Hungary (geza@renyi.hu). The work of this author was supported by OTKA grants T038397 and T046246.

best known upper bounds are $f^s_m(n) \leq cn\log^{2m-8}$ and $f_m(n) \leq c'n\log^{4m-16} n$ (see [PSS96], [PT05], [A06]), while the best known lower bounds are all linear functions of n, and it is conjectured that they are much closer to the truth.

CONJECTURE 1. *For every $k \geq 3$ there is a $c_k > 0$ such that every topological graph with n vertices and $C_k n$ edges contains k pairwise crossing edges.*

In [PRT04], the results about *three* pairwise crossing edges were further generalized: for every integer $k > 0$, there exists a constant $c_k > 0$, such that every topological graph with n vertices and more than $c_k n$ edges has $k+2$ edges such that the first two cross each other and both of them cross the remaining k edges (see Figure 1(a)).

In [PPST05] another generalization was shown. For any k and l there is a constant $c_{k,l}$ with the following property: Every topological graph with n vertices and more than $c_{k,l}n$ edges has $k+l$ edges such that the first k have a common vertex, and each of them crosses all of the remaining l edges (see Figure 1(b)).

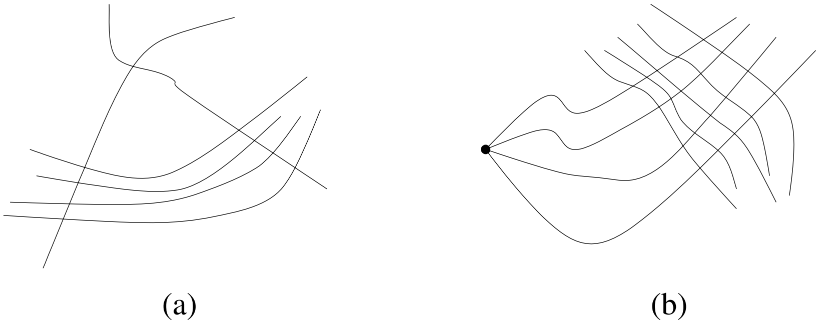


FIG. 1. *A topological graph without either configuration has only a linear number of edges.*

In this paper we prove a common generalization of the above results.

Let k be a positive integer. The edges $A \cup B \cup X$ of a topological graph form a *k-star grid* if A is a set of k edges incident to a common endpoint x, B is a set of k edges incident to a common endpoint y, and any edge from A crosses any edge from B; furthermore, X also contains k edges, and any edge in X crosses all edges in $A \cup B$. See Figure 2. In this definition we allow the case $x = y$ and we also allow the edges of X to be incident to x or y. These pathological cases are not possible in a simple topological graph.

THEOREM 1.1. *For any $k \geq 1$, there is a constant C_k such that every topological graph with n vertices and at least $C_k n$ edges contains a k-star grid.*

We did not attempt to optimize our proof for C_k, but we note that this proof gives C_k that is triply exponential in k. The condition that all edges in the set A (resp., B) have a common endpoint is essential; our proof does not work if we want to have *independent* edges in the set A (or B). The situation is very similar with the previous results [PRT04], [PPST05], and we cannot even prove the following well-established conjecture.

CONJECTURE 2. *For every $k \geq 2$, there is a $C_k > 0$ such that every topological graph with n vertices and $C_k n$ edges contains $k+2$ independent edges such that the first two of them cross each of the last k.*

2. Proof of the theorem. The proof of Theorem 1.1 is rather technical and consists of several steps. We give an overview first and then indicate which steps of the proof can be eliminated if we consider only simple topological graphs. Note that

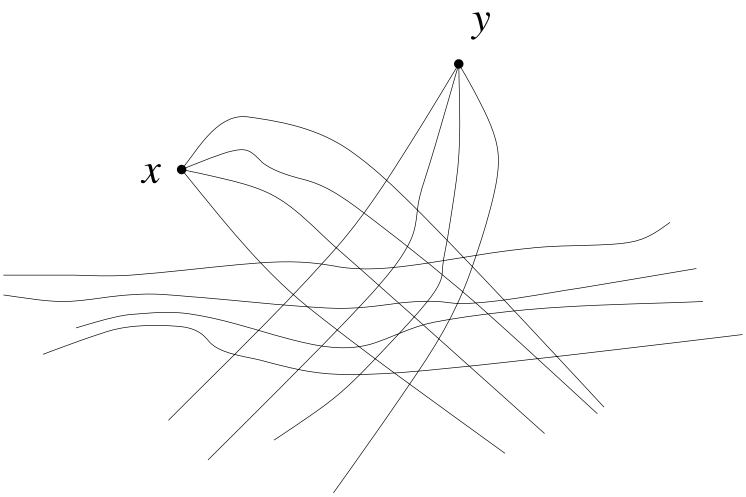


FIG. 2. *A 4-star grid.*

we do not strive for absolute preciosity in this overview. The reader will find the precise definitions later in the proof.

For the proof, we fix k and take an arbitrary topological graph F. We let $C = |E(F)|/|V(F)|$. Our goal is to prove that if C is large enough (as a function of k), then we find a k-star grid in F. This clearly establishes Theorem 1.1.

First we take a *densest subgraph* F_0 of F and concentrate on F_0 only.

Next we *redraw* F_0, i.e., we take another topological graph G_0 which has the same underlying abstract graph as F_0 but eliminates certain unnecessary crossings. This step of the proof is not needed if F is a simple topological graph, i.e., we may take $G_0 = F_0$.

We then use *subdivisions*, i.e., we introduce vertices at certain edge-crossings. We obtain a subdivision G_1 of G_0 with a crossing-free spanning tree T. This step is taken from [PRT04] and [PPST05].

We further subdivide G_1 to obtain G_2 and its crossing-free spanning subgraph H with no *proper cut*. This means that any two consecutive crossing points of any edge e in $G_2 \setminus H$ with H are with "close-by" edges of H. This step is taken from [PPST05]. In this and the previous step we make sure that the size (number of vertices) of the graph increases by only a constant factor. Note also that subdivisions in these two steps can create k-star grids. This does not happen for simple topological graphs for $k > 2$.

The next step represents the new idea in this paper. For many vertices we find a large number of edges emanating from that vertex with the property that they go "parallel" (with respect to H) for a long time and then one by one "depart" from the rest of the edges. All these "departures" take place in separate cells of H. We call these sets of edges *bundles*.

Using the fact that C is large enough, we find a *cross-track configuration* in G_2, i.e., k edges of a bundle, another k edges of a (perhaps different) bundle such that these $2k$ edges go parallel through $l - 1$ cells of H but such that, eventually, the first k edges cross the second k edges. For simple topological graphs we can choose $l = k$, and the proof ends here. Indeed, the $2k$ edges in the cross-track configuration plus k edges of H form a k-star grid. In the general case, however, some of the edges of H

crossed by the edges in the cross-track configuration may coincide or may be parts of the same edge of G_0, separated only by our subdivision process. We take l to be an exponential function of k and use the following result of Schaefer and Stefankovič [SS04] to take care of the technical difficulties mentioned above.

THEOREM A (Schaefer and Stefankovič [SS04]). *Let T be a topological graph. Redraw T so that the resulting topological graph T' satisfies the following two conditions:*

(i) *If two edges of T' cross each other, then the corresponding edges also cross in T;*

(ii) *T' has the minimum number of crossings among all drawings with property* (i).

Now for any $i > 0$ and any edge e, any 2^i consecutive crossings on e arise from at least i different edges.

By an application of Theorem A, we show that if l is large enough, then out of the l edges of H crossed by the parallel track of the edges of the cross-track configuration, at least k must come from distinct edges of G_0.

We continue with the detailed execution of the above plan.

Let $k \geq 1$ fixed, and let F be a topological graph with n' vertices and Cn' edges. Our goal is to prove that F contains a k-star grid if C is large enough. The bound on C depends on k but not on n'. This will establish the validity of Theorem 1.1.

Let F_0 be the densest nonempty connected subgraph of F; that is, $F_0 \subseteq F$ connected and $|E(F_0)|/|V(F_0)|$ is maximal. Clearly, the requirement that F_0 has to be connected does not change the value of the maximum, so we have $|E(F_0)|/|V(F_0)| \geq |E(F)|/|V(F)| = C$. Removing a vertex of F_0 of degree d increases the ratio if $d < C$; therefore each vertex in F_0 has degree at least C. Let n denote the number of vertices of F_0. Clearly, $n > C$ so we may assume $n \geq 5$.

Redraw F_0 so that the resulting topological graph G_0 satisfies the following two conditions:

(i) If two edges of G_0 cross each other, then the corresponding edges also cross in F_0;

(ii) G_0 has the minimum number of crossings among all drawings with property (i).

It is enough to find a k-star grid in G_0, as property (i) shows that the corresponding edges form a k-star grid in F_0, and thus in F too.

We will apply a *subdivision* to G_0, i.e., we declare a certain intersection point of two edges as a *new vertex* and replace each of the two edges by their two segments up to and from that new vertex. Notice that in this way we may create two edges connecting the same pair of vertices, and thus we have to extend our definition of topological graph to allow for this. No pair of vertices will ever be connected by more than two edges. The graph obtained from G_0 by several subdivisions is called a *subdivision of* G_0. To distinguish this from the new vertices of the subdivision, vertices of G_0 are called *old vertices.*

Notice that for $k > 2$, a subdivision does not introduce a k-star grid in a simple topological graph, so if G_0 is simple, it is enough to find a k-star grid in a subdivision of G_0. The situation is somewhat more complex if G_0 is not simple. If G_0 contains two k-edge stars A and B such that each edge of A is crossed by each edge of B and another edge e_0 crosses every edge in $A \cup B$ k times, then the repeated subdivision of e_0 may result in a k-star grid.

Obviously, no edge of G_0 intersects itself; otherwise we could reduce the number of crossings by removing the loop. Suppose that G_0 has two distinct edges, e and

f, that meet at least twice (including their common endpoints, in the case that they have a common endpoint). A simply connected region whose boundary is composed of an arc of e and an arc of f is called a *lens*.

CLAIM 1. *Every lens in G_0 has a vertex in its interior.*

Proof. Suppose, for a contradiction, that there is a lens ℓ that contains no vertex of G in its interior. Consider a *minimal* lens $\ell' \subseteq \ell$, by containment. Notice that by swapping the two sides of ℓ', we could reduce the number of crossings without creating any new pair of crossing edges, contradicting property (ii) above. □

Clearly, the property of having no self-intersecting edge and the property stated in Claim 1 are both inherited from G_0 to its subdivisions.

Let G be a topological graph and H a subgraph of G. Let e be an edge of G not contained in H. We always consider e with an orientation. Each edge can be considered with either orientation. The edge e has a finite number of intersection points with edges of H, and these points split the Jordan curve e into a finite number of shorter curves. We call these shorter curves the *segments* of the edge e determined by H and denote them by $s_1(e), s_2(e), \ldots$ in the order they appear on e. The dependence on H is not explicit in the notation, but H will always be clear from the context. If e does not cross the edges of H, the entire edge is a single segment.

We consider a crossing-free subgraph H of a topological graph G that is connected and contains all vertices. Such a graph H subdivides the plane into *cells*. The boundary of a cell is a closed walk in H that may visit vertices several times and may even pass through an edge twice. The *size* of a cell is the length of the corresponding walk, that is, the number of edges in the walk, with multiplicity. A segment s of an edge e not in H inherits its orientation from e. It is contained in a single cell α, and the endpoints of s are on the boundary of α. We call the cell α, and the vertex or edge of the boundary walk of α where s starts, the *origin* of s. Similarly, α and the vertex or edge of this walk where s ends is the *destination* of s. Notice that in the case when the boundary of α visits the relevant vertex or edge more than once, the origin or destination of e contains more information than the vertex or edge itself and tells us "which side" of the vertex or edge is involved. If two segments have the same origin and the same destination, we call them *parallel* and say that their *type* is the same. If two segments s and s' have the same origin but different destinations, then they are contained in the same cell. We say that *s turns left from s'* if the common origin, the destination of s, and the destination of s' appear in this order in the clockwise tour of the boundary of the cell. Notice that the common origin must differ from both of the destinations. A segment with equal origin and destination would define an "empty lens," thus contradicting Claim 1. As a consequence, for segments s and s' with a common origin, either s and s' are parallel, or s turns left from s', or s' turns left from s.

As in [PRT04] and [PPST05], first we construct a subdivision G_1 of G_0 that contains a crossing-free spanning tree T.

Since the abstract underlying graph of G_0 is connected, we can choose a sequence of edges $e_1, e_2, \ldots, e_{n-1} \in E(G_0)$ such that $e_1, e_2, \ldots, e_i$ form a tree T_i, for every $1 \leq i \leq n-1$. In particular, $e_1, e_2, \ldots, e_{n-1}$ form a spanning tree T_{n-1} of G.

Construct the crossing-free topological graphs $\tilde{T}_1, \tilde{T}_2, \ldots, \tilde{T}_{n-1}$, as follows. Each is a subtree of a subdivision of G_0. Let $\tilde{T}_1$ be defined as a topological graph of two vertices consisting of the single edge e_1. Suppose that $\tilde{T}_i$ has already been defined for some $1 \leq i < n-1$, and let v denote the endpoint of e_{i+1} that does not belong to T_i. Then we define $\tilde{T}_{i+1}$ as follows. Add to $\tilde{T}_i$ the piece of e_{i+1} between v and its first crossing with $\tilde{T}_i$. More precisely, follow the edge e_{i+1} from v up to the point v'

where it hits $\tilde{T}_i$ for the first time. If this is a vertex of $\tilde{T}_i$, simply add e_{i+1} to $\tilde{T}_i$ to get $\tilde{T}_{i+1}$. If v' is in the interior of an edge e, then we apply subdivision: we introduce v' as a new vertex. We replace the edge e of $\tilde{T}_i$ with the two resulting parts and add the segment of e_{i+1} between v and v' to obtain $\tilde{T}_{i+1}$. See Figure 3.

We let $T = \tilde{T}_{n-1}$ and G_1 be the subdivision of G_0 obtained in the process. Note that G_1 has n old and at most $n-2$ new vertices.

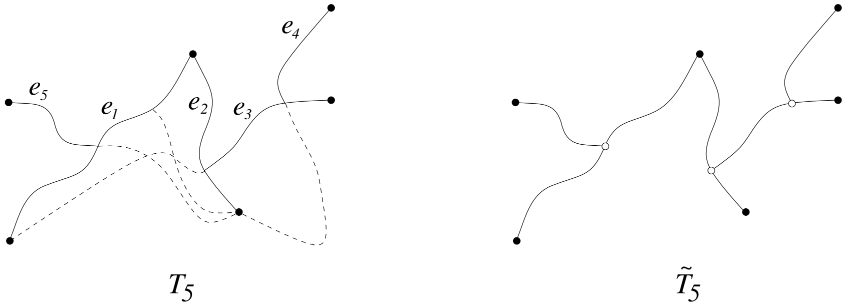


FIG. 3. *Constructing $\tilde{T}_5$ from T_5.*

Next, just like in [PPST05], we further subdivide G_1 to obtain G_2 and a crossing-free subgraph H of G_2.

Start with $H_0 = T$ and $\tilde{G}_0 = G_1$. Define $H_1, \dots, H_u$ and $\tilde{G}_1, \dots, \tilde{G}_u$ recursively, maintaining that H_i is a crossing-free connected subgraph of a subdivision $\tilde{G}_i$ of G_0. Furthermore, H_i is connected and contains all vertices of $\tilde{G}_i$, and all the cells of H_i are of size at least 8. This clearly holds for H_0 and $\tilde{G}_0$ if $n \geq 5$.

Having defined H_i and $\tilde{G}_i$, consider the segments of the edges of $\tilde{G}_i$ as determined by H_i. Let s be such a segment. By *adding s to H_i* we mean constructing a subdivision of $\tilde{G}_i$ by inserting new vertices for the endpoints of s, if necessary, and defining a subgraph H_i^s of the subdivision by adding s to H_i. More precisely, we also have to replace any edge of H_i that contains in its interior an endpoint of s by the two new edges resulting from the subdivision. Notice that s itself is an edge after the subdivision. The resulting graph H_i^s is a crossing-free connected spanning subgraph of the resulting subdivision of $\tilde{G}_i$. The cell of H_i containing s is now subdivided into two cells, and the other cells remain intact (but their size may increase). We call s a *proper cut of H_i* if both new cells of H_i^s are of size at least 8. See Figure 4.

If there exists a proper cut of H_i, then we choose one such segment s, set $H_{i+1} = H_i^s$, and let $\tilde{G}_{i+1}$ be the resulting subdivision of $\tilde{G}_i$. If there is no proper cut of H_i, we set $u = i$, $H = H_u$, and $G_2 = \tilde{G}_u$.

The number of cells starts at 1 cell, at $H_0 = T$, and increases by 1 in every step, so H_i contains $i+1$ cells. Each of these cells is of size at least 8, so we have at least $4i+4$ edges in H_i. From the Euler formula, the number of vertices v_i of H_i is at least $3i+5$. As $H_0 = T$ contains at most $2n-2$ vertices and we introduce at most 2 new vertices in every step, so we also have $v_i \leq 2i + 2n - 2$. The upper and lower bounds on v_i imply $i \leq 2n-7$. So the above process terminates in $u \leq 2n-7$ steps. This proves the following.

CLAIM 2. *G_2 is a subdivision of G_0 with at most $6n-16$ vertices. H is a connected, spanning, crossing-free subgraph of G_2 with no proper cut. H has at most $8n-24$ edges.*

We call an old vertex of G_2 *important* if its degree in H is less than 32. By Claim 2, H has less than $n/2$ vertices of degree 32 or more. Out of the n old vertices,

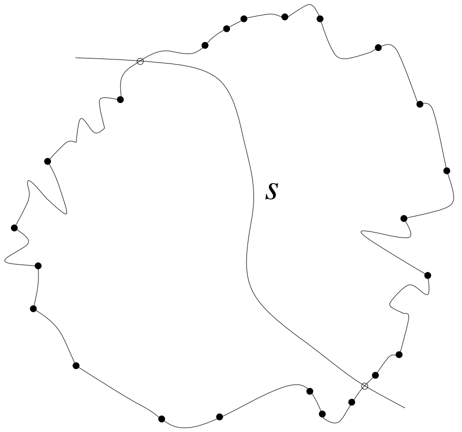


Fig. 4. *A proper cut.*

we must have more than $n/2$ important vertices.

Let $l = 2^{k+1}k^2 + 1$. Consider an edge e of G_2 not in H. Call any l consecutive of the segments $s_1(e), s_2(e), \ldots$ a *track* of e. The *type* of a track is simply the sequence of the types of the l segments, $s_i(e), \ldots, s_{i+l-1}(e)$. Tracks (of possibly different edges) of the same type are called *parallel*. Consider two edges e and f of G_2 that are not in H. Let $d(e, f)$ be the largest index $i \geq 1$ such that for all $1 \leq j < i$ the segments $s_j(e)$ and $s_j(f)$ exist and are parallel. For example, if e and f start at different vertices or in different cells, we have $d(e, f) = 1$.

Notice that for any origin of a segment, at most 24 destinations are possible. For large cells of H, more choices would be possible but they yield proper cuts of H which do not exist by Claim 2. By the same claim, there are less than $32n$ possible origins and therefore less than $768n$ types of segments. The destination of a segment determines the origin of the next segment; therefore there are less than $32 \cdot 24^l n$ different types of tracks.

Let $m = 300k \cdot 24^l$. We call the sequence $e_1, \ldots, e_{2m}$ of $2m$ edges of G_2 but not in H a *bundle* if $l \leq d(e_1, e_{2m}) < d(e_2, e_{2m}) < \cdots < d(e_{2m-1}, e_{2m})$. Notice that the edges of a bundle start at a common vertex. We say that the bundle *emanates* from this common starting vertex.

Claim 3. *If* $C \geq C_k := 31 \cdot 24^{2m+l} + 31$*, then there exists a bundle emanating from every important vertex.*

Proof. Consider an important vertex x. Let S_0 be the set of edges of G_2 not in H that start at x. The vertex x has degree at least C in G_0 and has the same degree in its subdivision G_2. Its degree in H is at most 31, so $|S_0| \geq C - 31$. For $i \geq 1$ we define S_i to be a subset of maximal size of S_{i-1} with $s_i(e)$ existing and having equal type for each $e \in S_i$. The number of possible origins for the type of segment $s_1(e)$ of an edge $e \in S_0$ is the degree of x in H. Since x is important, at most 31 origins and at most 744 types of $s_1(e)$ may exist for $e \in S_0$. Thus, $|S_1| \geq |S_0|/744$. Notice that the type of $s_i(e)$ determines if e ends with the segment $s_i(e)$, and if so, then it determines the ending vertex. So if one of the edges $e \in S_i$ ends with its ith segment, then all do, and they all connect the same pair of vertices. Thus, as long as $|S_i| > 2$, $s_{i+1}(e)$ exists for all $e \in S_i$. Furthermore, the type of $s_i(e)$ determines the origin of $s_{i+1}(e)$. So if $|S_i| > 2$, then $|S_{i+1}| \geq |S_i|/24$.

The finiteness of the entire topological graph G_2 implies that $S_i = \emptyset$ for large enough i. Let $l \le d_1 < d_2 < \cdots < d_v$ be all the indices $d \ge l$ such that $|S_{d+1}| < |S_d|$. The above calculations yield that $|S_{d_1}| \ge 24^{2m}$ and $S_{d_i+1} = S_{d_{i+1}} \ge 24^{2m-i}$ for $i \le 2m$. We choose e_i to be an arbitrary element of $S_{d_i} \setminus S_{d_i+1}$. We have $d(e_i, e_{2m}) = d_i + 1$ for $i < 2m$. This establishes that $(e_1, \ldots, e_{2m})$ form a bundle. □

Fix a bundle $B^x = \{e_1^x, \ldots, e_{2m}^x\}$ from every important vertex x. The existence is given by Claim 3. These will be all the bundles, and in fact all the edges of $G_2 \setminus H$ we consider from now on.

The segments $s_1(e_{2m}^x), s_2(e_{2m}^x), \ldots, s_{d^x}(e_{2m}^x)$ for $d^x = d(e_m^x, e_{2m}^x)$ form the *backbone* of the bundle B^x. The tracks of e_{2m}^x contained in the backbone are called the *vertebras*. We denote the vertebra starting with the segment $s_i(e_{2m}^x)$ by t_i^x. Notice that the vertebras interleave: the last $l-1$ segments of a vertebra are the first $l-1$ segments of the next vertebra. With any vertebra t_i^x we find $m-1$ parallel tracks: the tracks starting with the segments $s_i(e_{m+1}^x), \ldots, s_i(e_{2m-1}^x)$.

Let $e = t_i^x$ and $f = t_j^y$ be two distinct parallel vertebras. Notice that $i > 1$ and $j > 1$ must hold, since we consider only a single bundle from any (important) vertex. Let e' and f' be the inverse orientation of the "previous" segments $s_{i-1}(e_{2m}^x)$ and $s_{j-1}(e_{2m}^y)$, respectively. Notice that e' and f' have the same origin. We say that $e < f$ if e' turns left from f'. We also say that $e < f$ if t_{i-1}^x and t_{j-1}^y are parallel, and $t_{i-1}^x < t_{j-1}^y$. Notice that the recursive definition is well founded and defines a linear order among parallel vertebras. We call a vertebra *extremal* if it is smallest or largest among the vertebras of its type. If e is a nonextremal vertebra, we let e^+ stand for the next larger vertebra of the same type, while e^- stands for the next smaller vertebra. We say that a vertebra e is *special* either if it is extremal or if one of e^+ or e^- is the last vertebra in a backbone.

CLAIM 4. *The number of special vertebras is at most* $65 \cdot 24^l n$.

Proof. We have at most two extremal vertebras for every type, that is, at most $64 \cdot 24^l n$ extremal vertebras. We have one last vertebra in every backbone, that is, at most n last vertebras. Each last vertebra makes its at most two neighbors special, so the claimed bound holds. □

We define a *cross-track configuration* as two sets of k edges such that every edge from the first set crosses every edge from the second set, and all $2k$ edges go parallel for a long time. More precisely, let A and B both be sets of k edges. We say that $A \cup B$ is a *cross-track configuration* if the following conditions are satisfied:

(i) Every $a \in A$ crosses every $b \in B$.

(ii) Every $a \in A$ is incident to an old vertex x, and every $b \in B$ is incident to an old vertex y.

(iii) There are $\alpha, \beta > 0$ such that for every $a \in A$, $b \in B$, and $0 \le i < l-1$, $s_{\alpha+i}(a)$ and $s_{\beta+i}(b)$ exist and are parallel.

Notice that for simple topological graphs, a cross-track configuration $A \cup B$ can be appended with the set $X \subseteq E(H)$ consisting of k of the origins of the segments in the parallel tracks of the edges in $A \cup B$. These edges cross every edge in $A \cup B$; therefore $A \cup B \cup X$ form a k-star grid. Unfortunately, if G_2 is not simple, then X may contain fewer than k edges; in extreme situations X might consist of a single edge (the edges in $A \cup B$ go around and around, crossing this single edge many times). Also, finding a k-star grid in G_2 is not enough in this case.

Our immediate goal is to find a cross-track configuration in G_2; see Claim 6. As explained above, this leads immediately to a k-star grid in G_2, and also in G_0 if G_0 is simple. For nonsimple topological graphs, we will also use the cross-track configuration to find k-star grids in G_0, but the argument is more involved.

The following claim is based on a similar observation in [AAPPS97].

CLAIM 5. *Let e and f be two consecutive vertebras of the bundle B^x, neither of which is special. Then either e^+ and f^+ are also consecutive vertebras of a backbone, or there exists a cross-track configuration in G_2. The same holds for e^- and f^-.*

Proof. Assume f follows e in B^x, and let $e^+ = t_i^y$. We have to show that $f^+ = t_{i+1}^y$. Suppose that $f^+ = t_j^z$.

Since e is not special, $e^* = s_{i+l}(e_{2m}^y)$ is still in the backbone of B^y. Let f^* be the last segment of f. These two segments have a common origin. We distinguish three cases. See Figure 5.

Case 1. e^* and f^* are parallel. Then, by the definition of the order of vertebras, t_{i+1}^y must be f^+.

Case 2. f^* turns left from e^*. In this case all edges e_a^x intersect all edges e_b^y for $m < a, b \leq 2m$. This provides a cross-track configuration. See Figure 5(a).

Case 3. e^* turns left from f^*. Now the edges e_a^y and e_b^z must cross for $m < a, b \leq 2m$, and this also provides a cross-track configuration. See Figure 5(b).

The proof for e^- and f^- is similar. □

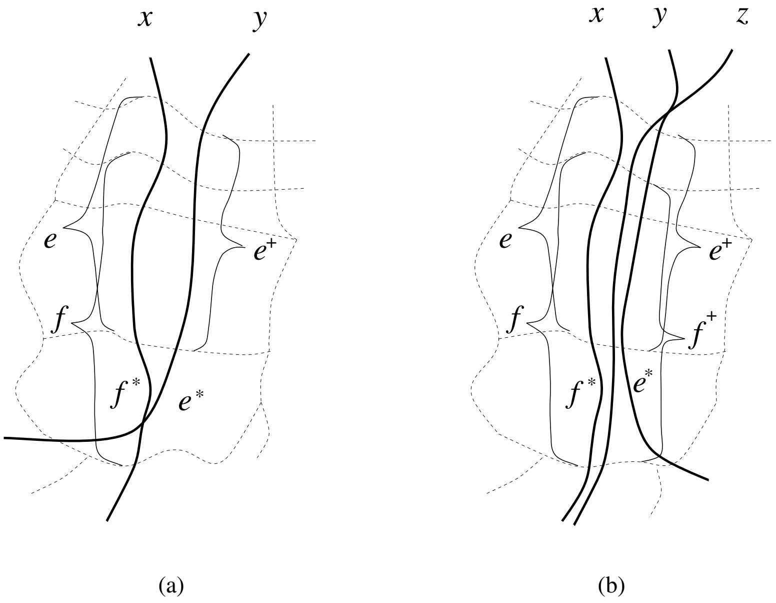


FIG. 5. *e^+ and f^+ are consecutive vertebras.*

We considered at least $n/2$ bundles. By Claim 4 we have at most $65 \cdot 24^l n$ special vertebras, so the pigeonhole principle gives the existence of a bundle B^x with at most $130 \cdot 24^l$ special vertebras. We fix such a bundle B^x and let e_i stand for the ith segment in the backbone of B^x: $e_i = s_i(e_{2m}^x)$ for $1 \leq i \leq d(e_m^x, e_{2m}^x)$. We call e_i a *departure point* if $i = d(e_j^x, e_{2m}^x)$ for some $1 \leq j \leq m$. We look for an interval of the backbone of B^x without special vertebras but with the largest number of departure points. There are m departure points, so at least $\lfloor m/(130 \cdot 24^l + 1) \rfloor$ of them are in an interval that has no special vertebra. Formally, we have $1 \leq i < j \leq d(e_m^x, e_{2m}^x) - l + 1$, such that none of the vertebras $t_i^x, t_{i+1}^x, \ldots, t_j^x$ are special, but for some indices $1 \leq i' < j' \leq m$ we have $i + l \leq d(e_{i'}^x, e_{2m}^x) < d(e_{j'}^x, e_{2m}^x) \leq j + l - 1$ and $j' - i' + 1 \geq \lfloor m/(130 \cdot 24^l + 1) \rfloor$.

By Claim 5 either we have a cross-track configuration, or the vertebras $(t_i^x)^+$, $(t_{i+1}^x)^+, \ldots, (t_j^x)^+$ are consecutive tracks of some bundle B^y, while $(t_i^x)^-, (t_{i+1}^x)^-, \ldots, (t_j^x)^-$ are also consecutive tracks of some bundle B^z. In the latter case, for any $i' \leq v \leq j'$ the edge e_v^x crosses all edges e_w^y with $m < w \leq 2m$, or it crosses all edges e_w^z with $m < w \leq 2m$. One of the options must occur with at least $\lfloor m/(260 \cdot 24^l + 2) \rfloor \geq k$ edges. This provides us with a set A of k edges of the bundle B^x and another set B of k edges of a bundle such that the properties of cross-track configuration are satisfied. Thus, a cross-track configuration must exist. See Figure 6. This proves the following.

Claim 6. *For $C \geq C_k$ there exists a cross-track configuration in G_2.*

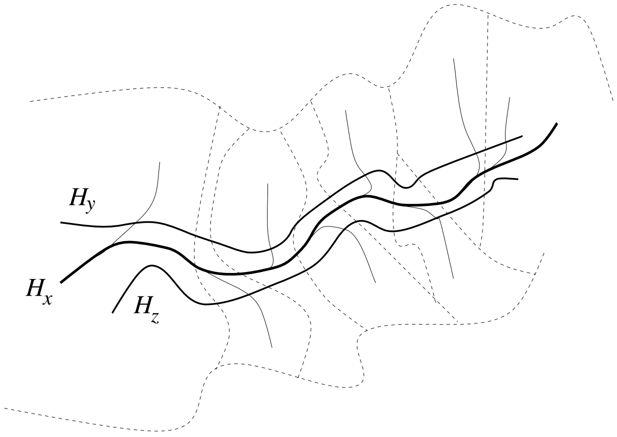


Fig. 6. *H_y and H_z envelop a vertebra of H_x.*

Let $A \cup B$ be a cross-track configuration in G_2. We use it to find a k-star grid in G_0. There are $\alpha, \beta > 0$ such that for every $a \in A$, $b \in B$, and $0 \leq i < l - 1$, the segments $s_{\alpha+i}(a)$ and $s_{\beta+i}(b)$ are parallel. Let $s_i^*(e) = s_{\alpha+i}(e)$ for $e \in A$ and $s_i^*(e) = s_{\beta+i}(e)$ for $e \in B$. We say that $0 \leq i < l - 1$ is *bad* if two distinct segments from the set $\{s_i^*(e) \mid e \in A \cup B\}$ intersect.

Observe that we counted at most one crossing for each pair of edges in $A \cup B$; otherwise we would get an "empty lens." Therefore, there are at most $\binom{2k}{2}$ bad values of i. So there are $0 \leq i_0 < i_1 \leq l - 1$, with $i_1 - i_0 + 2 > l/(\binom{2k}{2} + 1) > 2^k + 1$ such that there is no bad i with $i_0 \leq i \leq i_1$. For $i_0 \leq i \leq i_1$, let h_i be the edge of H that is the common origin of the segments $s_i^*(e)$ for $e \in A \cup B$. Order the edges $e \in A \cup B$ according to the order in which the starting points of $s_i^*(e)$ appear on h_i. Notice that we get the same order for each i. Let a and b be the first and last edges in this order. Let p_i and q_i be the starting points of $s_i^*(a)$ and $s_i^*(b)$, respectively. Let a^* be the "relevant" part of a; that is, a^* is the interval of a between p_{i_0} and p_{i_1}.

At this point we shift our attention from G_2 and H to the original graph G_0 and modify its drawing in the plane. Let S be the set of edges of G_0 containing the edges $A \cup B$ of G_2. Note that the edges in A are incident to the same old vertex; therefore they cannot be different segments of an edge of G_0. The same holds for the edges in B. Moreover, any edge of A and any edge of B intersect, so they are not different segments of the same edge. Consequently, S contains $2k$ distinct edges. We do not redraw the edge containing a but redraw some segments of other edges making sure that conditions (i) and (ii) of the definition of G_0 are maintained, and furthermore, every edge that intersects a^* also intersects all edges in S.

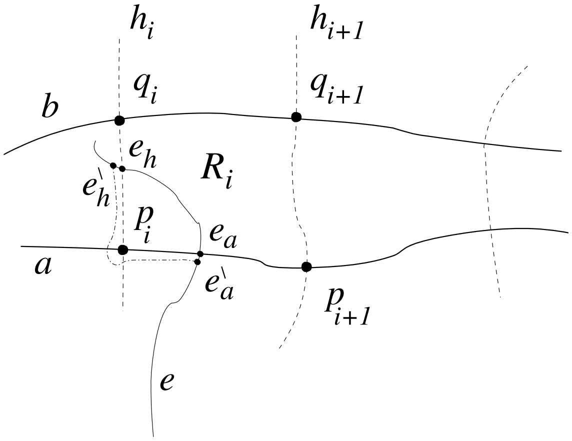


FIG. 7. *Procedure* REDRAW.

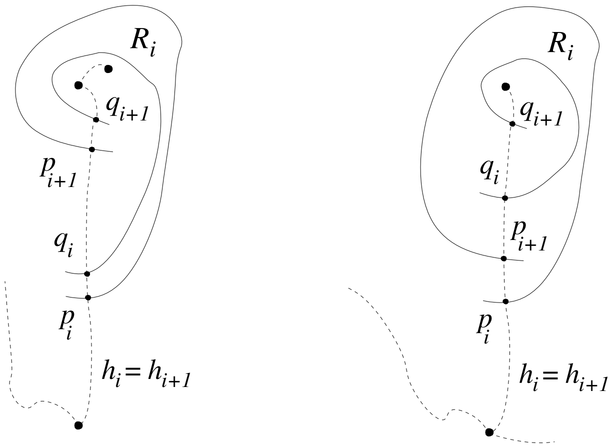


FIG. 8. *We cannot even rule out that* $h_i = h_{i+1}$.

Let $i_0 \leq i < i_1$ and consider the following four intervals: (1) the interval of h_i between p_i and q_i, (2) $s_i^*(a)$, (3) the interval of h_{i+1} between p_{i+1} and q_{i+1}, and (4) $s_i^*(b)$. These segments bound a quadrilateral shaped region R_i, with "vertices" p_i, q_i, p_{i+1}, and q_{i+1}. See Figure 7. We cannot rule out that some of the regions R_i are not disjoint and, in fact, we cannot even rule out that $h_i = h_{i+1}$ (see Figure 8), but it does not affect the argument to be presented.

The region R_i does not contain vertices; therefore no edge of G_0 entering R_i through $s_i^*(a)$ may leave R_i through $s_i^*(a)$ again, as that would contradict Claim 1. We distinguish three types of edges of G_0 entering R_i through $s_i^*(a)$. Note that an edge can cross $s_i^*(a)$ several times; in this case we consider separately each of the segments of e inside R_i.

Type 1. The edge e enters R_i through $s_i^*(a)$ and leaves R_i through $s_i^*(b)$. In this case, e crosses each edge in S.

Type 2. The edge e enters R_i through $s_i^*(a)$ and leaves R_i through h_i.

Type 3. The edge e enters R_i through $s_i^*(a)$ and leaves R_i through h_{i+1}.

We describe procedure REDRAW. If there exists $i_0 \leq i < i_1$ with an edge of

Type 2 crossing $s_i^*(a)$, then we choose an arbitrary such i and the edge e of Type 2 crossing $s_i^*(a)$ closest to p_i. Let e_a be the point of e where it enters R_i, and let e_h be the point where it leaves R_i. Let e'_a and e'_h be points on e outside R_i but close to e_a and e_h, respectively. Replace the interval $e'_a e'_h$ of e by a curve outside R_i, which follows very closely the interval of a between e_a and p_i, and then follows the interval of h_i between p_i and e_h. In the case when $h_i = h_{i+1}$, the new curve is drawn similarly, but it does not go outside the region R_i. It is easy to verify that if the new segment of e follows the boundary of R_i close enough, then no new crossings are created, and therefore the modified topological graph satisfies properties (i) and (ii). See Figure 7.

If there exists $i_0 \leq i < i_1$ and an edge of Type 3 crossing $s_i^*(a)$, then we proceed analogously. We choose such an i arbitrarily, choose a Type 3 edge that crosses $s_i^*(a)$ closest to p_{i+1}, and redraw the segment of the edge in R_i taking a detour around p_{i+1}.

As long as there is an i, $i_0 \leq i < i_1$, with a Type 2 or Type 3 edge, execute Redraw.

If a^* enters the region R_i (we cannot rule out this possibility), then Redraw choosing this i affects other regions R_j. In the extreme case when p_{i+1} is on h_i between p_i and q_i, by redrawing edges of Type 2 we create another crossing with $s_i^*(a)$ itself, possibly another Type 2 crossing. Nevertheless, it can be shown that the procedure terminates after finitely many steps. To see this, consider an edge e. The set $\bigcup_{i=i_0}^{i_1-1} R_i$ divides e into several intervals. Let e^* be one of them. For each crossing p of e^* and a^* let $r(p) = i$ if and only if p is on $s_i^*(a)$. Let $r(e^*, a^*)$ be the sum of all $r(p)$ over all crossings. This sum will either always decrease or always increase when we execute Redraw involving e^*; therefore e^* is involved in only finitely many steps. To see this "monotonicity condition" notice that each segment of a^* entering R_i has the "same orientation"; that is, it enters R_i through h_i and leaves through h_{i+1}.

Let G'_0 be the topological graph obtained in the process. All edges of G'_0 crossing the curve a^* cross all edges in S. We did not create any additional crossing, so the graph G'_0 satisfies properties (i) and (ii) in the definition of G_0. These properties and a result of Schaefer and Stefankovič [SS04] imply the following.

Claim 7. *For any edge e of G'_0 and for any $i > 0$, any 2^i consecutive crossings on e arise from at least i different edges.*

The interval a^* of a crosses H at least 2^k times, and we did not "redraw" these segments of edges of G_0. We can therefore take 2^k consecutive crossings of a^* in G'_0, and by Claim 7 they are from at least k edges. Let X be a set of k edges of G'_0 crossing a^*. Clearly, $S \cup X$ is a k-star grid in G'_0.

Clearly, the corresponding edges form a k-star grid in F too. This finishes our proof of Theorem 1.1.

REFERENCES

[A06] E. Ackerman, *On the maximum number of edges in topological graphs with no four pairwise crossing edges*, in Proceedings of the 22nd Annual ACM Symposium on Computational Geometry (SoCG), ACM, New York, 2006, pp. 259–263.

[AT07] E. Ackerman and G. Tardos, *The maximum number of edges in quasi-planar graphs*, J. Combin. Theory Ser. A, 114 (2007), pp. 563–571.

[AAPPS97] P. K. Agarwal, B. Aronov, J. Pach, R. Pollack, and M. Sharir, *Quasi-planar graphs have a linear number of edges*, Combinatorica, 17 (1997), pp. 1–9.

[P99] J. Pach, *Geometric graph theory*, in Surveys in Combinatorics, J. D. Lamb and D. A. Preece, eds., London Math. Soc. Lecture Note Ser. 267, Cambridge University Press, Cambridge, UK, 1999, pp. 167–200.

[PPST05] J. Pach, R. Pinchasi, M. Sharir, and G. Tóth, *Topological graphs with no large grids*, Graphs Combin., 21 (2005), pp. 355–364.

[PRTT04] J. Pach, R. Radoičić, G. Tardos, and G. Tóth, *Improving the crossing lemma by finding more crossings in sparse graphs*, in Proceedings of the 20th Annual ACM Symposium on Computational Geometry (SoCG), ACM, New York, 2004, pp. 68–75.

[PPTT02] J. Pach, R. Pinchasi, G. Tardos, and G. Tóth, *Geometric graphs with no self-intersecting path of length three*, in Graph Drawing, M. T. Goodrich and S. G. Kobourov, eds., Lecture Notes in Comput. Sci. 2528, Springer-Verlag, Berlin, 2002, pp. 295–311.

[PRT03] J. Pach, R. Radoičić, and G. Tóth, *Relaxing planarity for topological graphs*, in Discrete and Computational Geometry, J. Akiyama and M. Kano, eds., Lecture Notes in Comput. Sci. 2866, Springer-Verlag, Berlin, 2003, pp. 221–232.

[PRT04] J. Pach, R. Radoičić, and G. Tóth, *A generalization of quasi-planarity*, in Towards a Theory of Geometric Graphs, J. Pach, ed., Contemp. Math. 342, AMS, Providence, RI, 2004, pp. 177–183.

[PSS96] J. Pach, F. Shahrokhi, and M. Szegedy, *Applications of the crossing number*, Algorithmica, 16 (1996), pp. 111–117. Also in Proceedings of the 10th Annual ACM Symposium on Computational Geometry (SoCG), ACM, New York, 1994, pp. 198–202.

[PT97] J. Pach and G. Tóth, *Graphs drawn with few crossings per edge*, Combinatorica, 17 (1997), pp. 427–439.

[PT05] J. Pach and G. Tóth, *Disjoint edges in topological graphs*, in Combinatorial Geometry and Graph Theory: Indonesia-Japan Joint Conference (Bandung, Indonesia, 2003), Revised Selected Papers, J. Akiyama, E. T. Baskoro, and M. Kano, eds., Lecture Notes in Comput. Sci. 3330, Springer-Verlag, Berlin, 2005, pp. 133–140.

[SS04] M. Schaefer and D. Stefankovič, *Decidability of string graphs*, J. Comput. System Sci., 68 (2004), pp. 319–334. Also in Proceedings of the 33rd Annual ACM Symposium on Theory of Computing (STOC), ACM, New York, 2001, pp. 241–246.

[TT05] G. Tardos and G. Tóth, *Crossing stars in topological graphs*, in Proceedings of the Japan Conference on Discrete and Computational Geometry 2004, in Honor of János Pach on His 50th Year, Lecture Notes in Comput. Sci. 3742, Springer-Verlag, Berlin, 2005, pp. 184–197.

SIAM J. Discrete Math.
Vol. 21, No. 3, pp. 750–762

OPTIMAL LEE-TYPE LOCAL STRUCTURES IN CARTESIAN PRODUCTS OF CYCLES AND PATHS[*]

SIMON ŠPACAPAN[†]

Abstract. We define the neighborhood of an r-ball $B(u, r)$ centered on $u \in G$ as the set of all vertices x, such that $d(u, x) = r + 1$, and denote it by $N(u, r)$. We call a set X of pairwise disjoint r-balls an optimal local structure for $B(u, r)$ if $N(u, r) \subset \bigcup X$ and no r-ball from X intersects $B(u, r)$. We prove the nonexistence of an optimal local structure in $G = C_{q_1} \Box C_{q_2} \Box \cdots \Box C_{q_n}$ for any $B(u, r) \subset G$, where $n \geq 3$, $r \geq n$, and $q_i \geq 2r + 1$ for $i = 1, \ldots, n$. In particular, this confirms the nonexistence of perfect Lee codes with parameters $n \geq 3$, $e \geq n$, and $q \geq 2e + 1$. We also prove that if q_i is even for $i = 1, \ldots, n$, $\sum_{i=1}^{n} q_i/2$ is odd, and $2r + 1 = \sum_{i=1}^{n} q_i/2$, then for every r-ball in G there is an optimal local structure.

Key words. perfect Lee codes, optimal local structure

AMS subject classifications. 05C69, 05C12

DOI. 10.1137/040621387

1. Introduction.

The question of existence of perfect Lee codes was raised in 1968 by Golomb and Welch, when they proved in [8] the existence of perfect Lee codes with parameters $(n, e, q) = (2, e, 2e^2 + 2e + 1)$ for $e \geq 1$ and $(n, e, q) = (n, 1, 2n + 1)$ for $n \geq 1$, where n denotes the length of the codewords, e denotes the number of errors this code corrects, and q is the number of symbols in the alphabet. They also proved the nonexistence in some special cases and conjectured [8] the nonexistence of perfect Lee codes for all values of n and e, except $n \in \{1, 2\}$ or $e = 1$.

In 1975 Post proved in [20] the nonexistence of perfect Lee codes over large alphabets ($q \geq 2e + 1$) for $3 \leq n \leq 5$ and $e \geq n - 1$, and for $n \geq 6$ and $e \geq n\sqrt{2}/2 - 3\sqrt{2}/4 - 1/2$. There have also been numerous results given in [2, 3, 4, 16, 17]. For example, in [2] Astola gives an infinite family of new perfect Lee codes, called the repetition codes. These are codes with parameters (n, e, q), where n is odd, q congruent to 2 (mod 4), and $e = (nq - 2)/4$. However, the question of the existence of perfect Lee codes has not been completely solved and remains open for large alphabets and small values of e and is also not completely solved for small alphabets ($q < 2e + 1$).

Lee codes are known to be useful in transmitting information over channels with error patterns suitable for the Lee metric. Some recent results from [6, 21] give (de)coding algorithms for a certain class of BCH (Bose, Ray-Chaudhuri, Hocquenghem) Lee codes. BCH Lee codes are also considered in [1], where the construction of BCH Lee codes is addressed. Some recent results from [19] improve certain known bounds on the cardinality of Lee codes. We also mention that in [10] an upper bound for the domination number and r-domination number of the Cartesian product of paths is obtained by using known results of Golomb and Welch on perfect Lee codes. For more information about perfect codes in graphs we refer the reader to [5] and [15].

Note that the existence of a perfect Lee code with parameters (n, e, q) is equivalent

[*] Received by the editors December 23, 2004; accepted for publication (in revised form) March 13, 2007; published electronically September 26, 2007.
http://www.siam.org/journals/sidma/21-3/62138.html

[†] FME, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia (simon.spacapan@uni-mb.si).

to the existence of an e-perfect code in the Cartesian product of n copies of C_q, $C_q \Box C_q \Box \cdots \Box C_q$ (n times).

There are basically two possible ways to prove the nonexistence of perfect Lee codes. One way, such as in the previously mentioned papers, is by dealing with a sphere packing condition and some other related conditions. There have also been some proofs in which certain counting arguments were given to obtain nonexistence theorems (see, for example, [20]); however, the sphere packing condition and similar conditions do not provide an understanding of the local structure of a code. In this paper we are concerned with the local structure of the Cartesian product of cycles, and we show the nonexistence of an optimal local structure for a large class of products. Clearly, the nonexistence of an optimal local structure for $B(u,e)$ in $C_q \Box C_q \Box \cdots \Box C_q$ (n times) implies the nonexistence of a perfect Lee code with parameters (n,e,q); moreover, this result gives even better insight into the structure of the Lee code, in particular the local structure of the code. In this way Golomb and Welch [9] proved nonexistence for $n = 3, e = 2$, and $q \geq 2e+1$. Gravier, Mollard, and Payan proved in [11] the nonexistence of an optimal local structure in $C_{q_1} \Box C_{q_2} \Box C_{q_3}$ for any $B(u,e)$, where $e \geq 2$ and $q_i \geq 2e+1$ for $i = 1,2,3$. The main goal of this paper is to prove the nonexistence of an optimal local structure for any $B(u,r) \subset C_{q_1} \Box \cdots \Box C_{q_n}$, where $n \geq 3$, $r \geq n$, and $q_i \geq 2r+1$ for $i = 1, \ldots, n$.

2. Preliminaries. For a graph $G = (V(G), E(G))$, the *distance* $d(u,v)$ between vertices u and v is defined as the number of edges on a shortest u,v-path. We define the r-ball $B(u,r)$ centered on $u \in V(G)$ by

$$B(u,r) = \{x \in V(G) \,|\, d(u,x) \leq r\}\,.$$

A set $C \subseteq V(G)$ is an *r-code* in G if $B(c_1,r) \cap B(c_2,r) = \emptyset$ for any $c_1, c_2 \in C, c_1 \neq c_2$, and an r-code C is called an *r-perfect code* if $\cup_{c \in C} B(c,r) = V(G)$.

The distance between two r-balls L_1 and L_2 is defined by

$$d(L_1, L_2) = \min\{d(l_1, l_2) \,|\, l_1 \in L_1, l_2 \in L_2\},$$

and L_1 and L_2 are said to be *neighboring* r-balls if $d(L_1, L_2) = 1$. The *neighborhood* $N(u,r)$ of r-ball $B(u,r)$ is defined by

$$N(u,r) = B(u, r+1) \setminus B(u,r)\,.$$

We now formally define an optimal local structure. Let $L_0 = B(u,r)$ be an r-ball and suppose that r-balls $L_1, \ldots, L_k$ are neighboring L_0. If $L_i \cap L_j = \emptyset$ for $i,j \geq 0, i \neq j$, and $N(u,r) \subset \bigcup_{i=1}^{k} L_i$, then $L_1, \ldots, L_k$ is called an *optimal local structure* for L_0.

For $i = 1, \ldots, n$ let $G_i = (V(G_i), E(G_i))$. The *Cartesian product* of graphs $G_1, G_2, \ldots, G_n$ is the graph $G_1 \Box G_2 \Box \cdots \Box G_n$ with the vertex set $V_1 \times V_2 \times \cdots \times V_n$ (where $\times$ denotes the Cartesian product of sets), and the vertices $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ are adjacent in $G_1 \Box G_2 \Box \cdots \Box G_n$ if for some $j \in \{1, \ldots, n\}$ u_j is adjacent to v_j in G_j and for any $i \neq j$, $u_i = v_i$.

We will denote the cycle of length q by C_q, and we set $V(C_q) = \mathbb{Z}_q = \{0, 1, \ldots, q-1\}$. The vertices of C_q will be calculated modulo q whenever applicable. A perfect Lee code with parameters n, e, and q is an e-perfect code in the graph $C_q^n = C_q \Box C_q \Box \cdots \Box C_q$ (n times).

Let $G = C_{q_1} \Box C_{q_2} \Box \cdots \Box C_{q_n}$; then we will denote the vertices of G with bold letters. For $\mathbf{u} \in G$ we reserve the symbols $u_1, \ldots, u_n$ for the components of $\mathbf{u}$, and thus $\mathbf{u} = (u_1, \dots, u_n)$. The distance function in the graph G will be denoted by d, and the distance function in C_{q_i} will be denoted by d_i, so that for $\mathbf{x}, \mathbf{y} \in G$ we have

$$d_i(x_i, y_i) = \min\{|x_i - y_i|, q_i - |x_i - y_i|\}$$

and

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} d_i(x_i, y_i)\,.$$

Note also that $d(\mathbf{x}, \mathbf{y}) \le \sum_{i=1}^{n} |x_i - y_i|$ and $d(\mathbf{x}, \mathbf{0}) = \sum_{i=1}^{n} d(x_i, 0) \le \sum_{i=1}^{n} |x_i|$.

3. Some existence results. The following theorem describes an optimal local structure for an r-ball in the graph $G = C_{q_1} \Box C_{q_2} \Box \cdots \Box C_{q_n}$, where $n \ge 3$ and $r \ge 2$ (and $q_1, \dots, q_n$ are appropriate).

THEOREM 3.1. *Let $n \ge 3$, $G = C_{q_1} \Box C_{q_2} \Box \cdots \Box C_{q_n}$, where q_i is even for $i = 1, \dots, n$. If $\sum_{i=1}^{n} q_i/2$ is odd and $2r + 1 = \sum_{i=1}^{n} q_i/2$, then for any r-ball in G there exist an optimal local structure.*

We prove the existence of an optimal local structure for $B(\mathbf{0}, r)$. Let $\mathbf{u} = (q_1/2, \dots, q_n/2)$. We claim that $B(\mathbf{u}, r)$ is an optimal local structure for $B(\mathbf{0}, r)$ (and vice versa). We have $d(\mathbf{0}, \mathbf{u}) = \sum_{i=1}^{n} q_i/2 = 2r + 1$, and hence $B(\mathbf{0}, r) \cap B(\mathbf{u}, r) = \emptyset$. If $\mathbf{x} \in G \setminus B(\mathbf{0}, r)$, then $\sum_{i=1}^{n} d_i(x_i, 0) \ge r + 1$, and therefore $d(\mathbf{u}, \mathbf{x}) = \sum_{i=1}^{n} d_i(q_i/2, x_i) = \sum_{i=1}^{n} q_i/2 - d_i(x_i, 0) \le r$. Thus $C = \{\mathbf{0}, \mathbf{u}\}$ is an r-perfect code in G, and hence $B(\mathbf{u}, r)$ is an optimal local structure for $B(\mathbf{0}, r)$ (and vice versa).

Let $n \ge 3, r \ge 2$, and $G = C_{q_1} \Box C_{q_2} \Box \cdots \Box C_{q_n}$. We conjecture that the only existing optimal local structures in G are the optimal local structures of Theorem 3.1; that is, an optimal local structure for an r-ball exists if and only if q_i is even for $i = 1, \dots, n$, $\sum_{i=1}^{n} q_i/2$ is odd, and $2r + 1 = \sum_{i=1}^{n} q_i/2$.

4. Nonexistence of an optimal local structure for $n = 3$. We first show the nonexistence of an optimal local structure for $B(\mathbf{0}, r) \subset C_{q_1} \Box C_{q_2} \Box C_{q_3}$, where $r \ge 3$ and $q_1, q_2, q_3 \ge 2r + 1$. Clearly, the proof of the nonexistence of an optimal local structure for $B(\mathbf{0}, r)$ implies nonexistence for $B(\mathbf{u}, r)$ for any $\mathbf{u} \in C_{q_1} \Box C_{q_2} \Box C_{q_3}$. The general idea of the proof is to find a large enough set $\mathcal{U} \subset N(\mathbf{0}, r)$, such that if $\mathbf{x}, \mathbf{y} \in \mathcal{U}, \mathbf{x} \ne \mathbf{y}$, then every r-ball containing $\mathbf{x}$ and $\mathbf{y}$ intersects $B(\mathbf{0}, r)$. Thus the number of r-balls constituting an optimal local structure is at least $|\mathcal{U}|$. Since the set $\mathcal{U}$ is large, any family of r-balls neighboring $B(\mathbf{0}, r)$ and containing the set $\mathcal{U}$ will contain intersecting r-balls. In this section we assume $r \ge 3$ and $G = C_{q_1} \Box C_{q_2} \Box C_{q_3}$, where $q_1, q_2, q_3 \ge 2r + 1$. Since the results of this section are more or less special cases of the results in the next section, we refer the reader to the proofs of Lemma 4.1–4.6.

LEMMA 4.1. *Let*

$A_1 = \{(2, 1, r - 2), (1, 2, r - 2), (1, 1, r - 1)\};$
$A_2 = \{(-2, 1, r - 2), (-1, 2, r - 2), (-1, 1, r - 1)\};$
$A_3 = \{(2, -1, r - 2), (1, -2, r - 2), (1, -1, r - 1)\};$
$A_4 = \{(-2, -1, r - 2), (-1, -2, r - 2), (-1, -1, r - 1)\}\,.$

Then $A_i \subset N(\mathbf{0}, r)$ for $i = 1, \dots, 4$.

Proof. Since $q_1, q_2, q_3 \ge 2r+1$, we have for $\mathbf{t} \in \bigcup_{i=1}^{4} A_i$ that $d(\mathbf{t}, \mathbf{0}) = \sum_{k=1}^{3} |t_k| = r + 1$. □

The vertices of A_1 and A_2 are depicted in Figure 4.1 for the case $r = n = 3$.

LEMMA 4.2. *Let $\mathbf{t} \in A_i$ for some $i = 1, \dots, 4$. If $\mathbf{t} \in B(\mathbf{u}, r)$, $B(\mathbf{u}, r) \cap B(\mathbf{0}, r) = \emptyset$, then the following hold.*

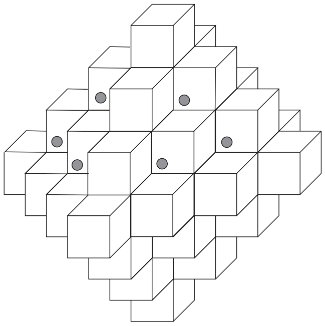


FIG. 4.1. *The vertices of A_1 and A_2.*

(i) *If $i \in \{1,2\}$ and $t_i > 0$, then $u_i = t_i + w_i$, where $0 \le w_i \le r$.*

(ii) *If $i \in \{1,2\}$ and $t_i < 0$, then $u_i = t_i - w_i$, where $0 \le w_i \le r$.*

(iii) *$u_3 = t_3 + r - |u_1 - t_1| - |u_2 - t_2|$, where $|u_1 - t_1| + |u_2 - t_2| \le r$.*

Conversely, for every $\mathbf{u}$ satisfying (i), (ii), *and* (iii), $\mathbf{t} \in B(\mathbf{u}, r)$.

Proof. See the proof of Lemma 5.2. □

Lemma 4.2 is describing all potential centers $\mathbf{u}$ of an r-ball containing $\mathbf{t}$ and not intersecting $B(\mathbf{0}, r)$. Let $\mathbf{t} \in A_i$ and denote by $U_{\mathbf{t}}$ the set of all vertices $\mathbf{u}$ satisfying (i), (ii), and (iii) of Lemma 4.2. With this notation we have the following corollary.

COROLLARY 4.3. *If $B(\mathbf{x^{(1)}}, r), B(\mathbf{x^{(2)}}, r), \ldots, B(\mathbf{x^{(m)}}, r)$ is an optimal local structure for $B(\mathbf{0}, r)$, then for every $\mathbf{t} \in \bigcup_{i=1}^{4} A_i$ there is exactly one $k \in \{1, \ldots, m\}$ such that $\mathbf{x^{(k)}} \in U_{\mathbf{t}}$.*

For $i = 1, \ldots, 4$ we define the sets $\mathcal{A}_i$ by

$$\mathcal{A}_i = \bigcup_{\mathbf{t} \in A_i} U_{\mathbf{t}} \,.$$

The following lemma gives an explicit description of the sets $\mathcal{A}_i$ (see Figure 4.2).

LEMMA 4.4. *For $i = 1, \ldots, 4$ the sets $\mathcal{A}_i$ are the following:*

$\mathcal{A}_1 = \{(1 + v_1, 1 + v_2, 2r - 1 - v_1 - v_2) \,|\, v_1, v_2 \ge 0, v_1 + v_2 \le r + 1\}$;
$\mathcal{A}_2 = \{(-1 - v_1, 1 + v_2, 2r - 1 - v_1 - v_2) \,|\, v_1, v_2 \ge 0, v_1 + v_2 \le r + 1\}$;
$\mathcal{A}_3 = \{(1 + v_1, -1 - v_2, 2r - 1 - v_1 - v_2) \,|\, v_1, v_2 \ge 0, v_1 + v_2 \le r + 1\}$;
$\mathcal{A}_4 = \{(-1 - v_1, -1 - v_2, 2r - 1 - v_1 - v_2) \,|\, v_1, v_2 \ge 0, v_1 + v_2 \le r + 1\}$.

Moreover, if $q_i \ge 2r + 5$ for $i = 1, 2$, then $\mathcal{A}_j \cap \mathcal{A}_k = \emptyset$ for $j \ne k$.

Proof. See the proof of Lemma 5.4. □

By the definition of the set $\mathcal{A}_i$, for any vertex $\mathbf{x} \in \mathcal{A}_i$ we have $B(\mathbf{x}, r) \cap A_i \ne \emptyset$, and conversely, for any $\mathbf{t} \in A_i$ and any $\mathbf{u}$ such that $B(\mathbf{u}, r) \cap B(\mathbf{0}, r) = \emptyset$ we have $\mathbf{u} \in \mathcal{A}_i$. In what follows we will use the following notation for $\mathbf{x} \in \mathcal{A}_1$. Instead of writing

$$\mathbf{x} = (1 + v_1, 1 + v_2, 2r - 1 - v_1 - v_2),$$

we write

$$\mathbf{x} = (1 + v_1(\mathbf{x}), 1 + v_2(\mathbf{x}), 2r - 1 - k(\mathbf{x}))\,,$$

where $v_1(\mathbf{x}) = v_1, v_2(\mathbf{x}) = v_2$, and $k(\mathbf{x}) = v_1(\mathbf{x}) + v_2(\mathbf{x})$. Similar notation is used for any $\mathbf{x} \in \bigcup \mathcal{A}_i$.

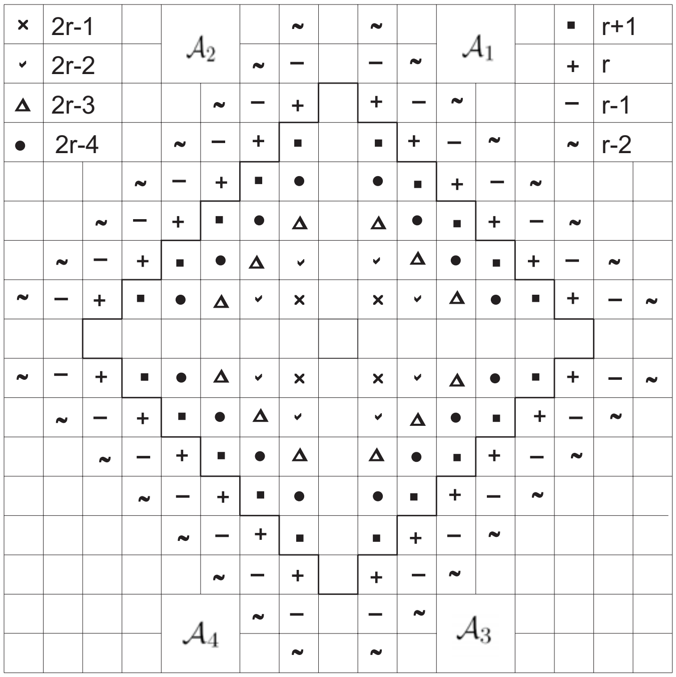


FIG. 4.2. *The sets $\mathcal{A}_i, i = 1, 2, 3, 4$, for $r = 6$.*

LEMMA 4.5. *Suppose $B(\mathbf{x^{(1)}}, r), B(\mathbf{x^{(2)}}, r), \ldots, B(\mathbf{x^{(m)}}, r)$ is an optimal local structure for $B(\mathbf{0}, r)$, and let $\mathcal{X} = \{\mathbf{x^{(1)}}, \mathbf{x^{(2)}}, \ldots, \mathbf{x^{(m)}}\}, \mathcal{X}_j = \mathcal{A}_j \cap \mathcal{X}$ for $j = 1, \ldots, 4$. Then for every $j \in \{1, \ldots, 4\}$ there is exactly one $\mathbf{x} \in \mathcal{X}_j$ such that $k(\mathbf{x}) \leq r$. Additionally, for every $s \in \{1, 2\}$ and every $j \in \{1, \ldots, 4\}$, there is exactly one $\mathbf{y} \in \mathcal{X}_j$, such that $v_s(\mathbf{y}) > 0$.*

Proof. See the proof of Lemma 5.5. □

LEMMA 4.6. *Let $\mathbf{u} \in \mathcal{A}_1, \mathbf{w} \in \mathcal{A}_2$, $\mathbf{x} \in \mathcal{A}_3$, and $\mathbf{y} \in \mathcal{A}_4$. If*

$\mathbf{u} = (1 + v_1(\mathbf{u}), 1 + v_2(\mathbf{u}), 2r - 1 - k(\mathbf{u}))$;
$\mathbf{w} = (-1 - v_1(\mathbf{w}), 1 + v_2(\mathbf{w}), 2r - 1 - k(\mathbf{w}))$;
$\mathbf{x} = (1 + v_1(\mathbf{x}), -1 - v_2(\mathbf{x}), 2r - 1 - k(\mathbf{x}))$;
$\mathbf{y} = (-1 - v_1(\mathbf{y}), -1 - v_2(\mathbf{y}), 2r - 1 - k(\mathbf{y}))$,

then

$d(\mathbf{u}, \mathbf{w}) \leq \max\{2k(\mathbf{u}), 2k(\mathbf{w})\} + 2 - 2\min\{v_2(\mathbf{u}), v_2(\mathbf{w})\}$;
$d(\mathbf{u}, \mathbf{x}) \leq \max\{2k(\mathbf{u}), 2k(\mathbf{x})\} + 2 - 2\min\{v_1(\mathbf{u}), v_1(\mathbf{x})\}$; *and*
$d(\mathbf{u}, \mathbf{y}) \leq \max\{2k(\mathbf{u}), 2k(\mathbf{y})\} + 4$.

Proof. See the proof of Proposition 5.7. □

The above lemma gives us the distance function defined on $\bigcup_{i=1}^{4} \mathcal{A}_i \times \bigcup_{i=1}^{4} \mathcal{A}_i$. We are now ready to prove the nonexistence of an optimal local structure for $B(\mathbf{0}, r)$ in the case when $n = 3$.

THEOREM 4.7. *Let $r \geq 3$ and $G = C_{q_1} \Box C_{q_2} \Box C_{q_3}$, where $q_i \geq 2r+1$ for $i = 1, 2, 3$. Then there does not exist an optimal local structure for $B(\mathbf{u}, r)$ for any $\mathbf{u} \in G$.*

Proof. Assume first $q_1, q_2 \geq 2r + 5$; thus $\mathcal{A}_j \cap \mathcal{A}_k = \emptyset$ for $j \neq k$. Suppose

$$B(\mathbf{x^{(1)}}, r), B(\mathbf{x^{(2)}}, r), \ldots, B(\mathbf{x^{(m)}}, r)$$

is an optimal local structure for $B(\mathbf{0}, r)$, and denote $\mathcal{X} = \{\mathbf{x^{(1)}}, \mathbf{x^{(2)}}, \ldots, \mathbf{x^{(m)}}\}$ and

$\mathcal{X}_j = \mathcal{A}_j \cap \mathcal{X}$ for $j = 1, \ldots, 4$. By Lemma 4.5 we have exactly one $\mathbf{x} \in \mathcal{X}_1$ such that $k(\mathbf{x}) \leq r$. Let

$$\mathbf{x} = (1 + v_1(\mathbf{x}), 1 + v_2(\mathbf{x}), 2r - 1 - k(\mathbf{x}))$$

and suppose $v_1(\mathbf{x}) \geq 2$. Then for every $\mathbf{u} \in \mathcal{A}_3$,

$$\mathbf{u} = (-1 - v_1(\mathbf{u}), 1 + v_2(\mathbf{u}), 2r - 1 - k(\mathbf{u}))\,,$$

such that $v_1(\mathbf{u}) \geq 1$ and $\mathbf{u} \neq (2, -r-1, r-2)$, we have

$$d(\mathbf{x}, \mathbf{u}) \leq \max\{2k(\mathbf{x}), 2k(\mathbf{u})\} + 2 - 2\min\{v_1(\mathbf{x}), v_1(\mathbf{u})\} \leq 2r\,.$$

Thus $(2, -r-1, r-2)$ is the only element from $U_{\mathbf{h}}$, where $\mathbf{h} = (2, -1, r-2)$, at a distance of more than $2r$ from $\mathbf{x}$. Thus by Corollary 4.3 and Lemma 4.5,

$$\mathcal{X}_3 = \{(2, -r-1, r-2), (1, -1, 2r-1)\},$$

and therefore for every $\mathbf{u} \in \mathcal{A}_4$ we get $\mathbf{u} \in U_{\mathbf{w}}$, where $\mathbf{w} = (-2, -1, r-2)$, and therefore $|\mathcal{X}_4| = 1$. Thus $\mathcal{X}_4 = \{(-r, -2, r-1)\}$. If $\mathbf{z} = (-r, -2, r-1)$ and $\mathbf{y} \in \bigcap_{\mathbf{t} \in A_2} U_{\mathbf{t}}$, then $v_1(\mathbf{y}), v_2(\mathbf{y}) \geq 1$ and $k(\mathbf{y}) \leq r$, and hence

$$d(\mathbf{y}, \mathbf{z}) \leq \max\{2k(\mathbf{y}), 2k(\mathbf{z})\} + 2 - 2\min\{v_1(\mathbf{y}), v_1(\mathbf{z})\} \leq 2r\,.$$

Thus $|\mathcal{X}_2| \geq 2$, but this is impossible since $\mathbf{u} \in \mathcal{X}_2$ implies $v_2(\mathbf{u}) \geq 1$ (because $\mathbf{z}, (1, -1, 2r-1) \in \mathcal{X}$). Analogously we get a contradiction in the case when $v_2(\mathbf{x}) \geq 2$. Suppose then $v_1(\mathbf{x}), v_2(\mathbf{x}) \leq 1$.

Case 1. If $v_1(\mathbf{x}), v_2(\mathbf{x}) = 1$, then $k(\mathbf{x}) = 2$ and $\mathcal{X}_1 = \{(2, 2, 2r-3)\}$. Since $k(\mathbf{y}) \leq r$ for some $\mathbf{y} \in \mathcal{X}_3$, we have

$$\mathbf{y} = (1, -r-1, r-1) \in \mathcal{X}_3\,,$$

since this is the only vertex of $\mathcal{A}_3$, such that $k(\mathbf{y}) \leq r$ and $d(\mathbf{x}, \mathbf{y}) > 2r$. Analogously

$$\mathbf{u} = (-1-r, 1, r-1) \in \mathcal{X}_2\,.$$

But then for every $\mathbf{z} \in \mathcal{A}_4$ we have that $d(\mathbf{z}, \mathbf{x})$ or $d(\mathbf{z}, \mathbf{u})$ or $d(\mathbf{z}, \mathbf{y})$ is less than $2r$.

Case 2. If $v_1(\mathbf{x}) = 1$ and $v_2(\mathbf{x}) = 0$, then $k(\mathbf{x}) = 1$ and

$$\mathcal{X}_1 = \{(2, 1, 2r-2), (1, r+2, r-2)\}\,.$$

In this case $(-r, 2, r-1), (1, -r-1, r-1), (r+2, -1, r-2) \in \mathcal{X}$. Thus if $d(\mathbf{x}, \mathbf{y}) > 2r$ for $\mathbf{x}, \mathbf{y} \in \mathcal{X}, \mathbf{x} \neq \mathbf{y}$, then $\mathcal{X}_4 = \emptyset$, which is a contradiction.

Case 3. If $v_1(\mathbf{x}) = 0$ and $v_2(\mathbf{x}) = 0$, then $k(\mathbf{x}) = 0$ and $|\mathcal{X}_1| \geq 2$. If $|\mathcal{X}_1| = 2$, then

$$\mathcal{X}_1 = \{(1, 1, 2r-1), (1 + v_1, 1 + v_2, r-2)\}$$

and $v_1, v_2 \geq 1$. Without loss of generality, $v_2 \geq 2$, and then for all $\mathbf{u} \in \mathcal{X}_2$, we have $v_2(\mathbf{u}) \leq 1$ and $k(\mathbf{u}) \geq r$, and hence $\mathcal{X}_2 = \{(-r, 2, r-1)\}$. Now let $\mathbf{y} \in \mathcal{X}_3$ be the element with $k(\mathbf{y}) = r$. If $v_2(\mathbf{y}) = 0$, then $\mathbf{y} = (r+1, -1, r-1)$, and therefore $\mathbf{t} = (1, -r-2, r-2) \in \mathcal{X}_3$ (otherwise, $v_2(\mathbf{y}) \geq 1$). Hence if $\mathbf{z} \in \mathcal{X}_4$, such that

$k(\mathbf{z}) \le r$, then $d(\mathbf{z},\mathbf{w}) \le 2r$ for some $\mathbf{w} \in \mathcal{X}$, which by Lemma 4.5 is a contradiction. If $|\mathcal{X}_1| = 3$, then

$$\mathcal{X}_1 = \{(1,1,2r-1),(r+2,1,r-2),(1,r+2,r-2)\}$$

and

$$\mathcal{X}_2 = \{(-r,2,r-1)\} \text{ and } \mathcal{X}_3 = \{(2,-r,r-1)\}\,,$$

and hence $\mathcal{X}_4 = \emptyset$.

The same proof works also for $q_1, q_2 \ge 2r+1$; one only has to be aware that a vertex $\mathbf{x}$ from $\mathcal{A}_i$ could be contained in $\mathcal{A}_j$, and in this case one coordinate of $\mathbf{x}$ is greater than r. $\square$

Note that in the case when one cycle is of length less than $2r+1$, an optimal local structures might exist. Take the example of $G = C_6 \Box C_6 \Box C_{14}$. If $\mathbf{x} = (0,0,0)$ and $\mathbf{y} = (3,3,7)$, then $B(\mathbf{x},6)$ is an optimal local structure for $B(\mathbf{y},6)$, and vice versa. Moreover, $C = \{\mathbf{x},\mathbf{y}\}$ is a 6-perfect code in G.

5. The general case. In this section we consider products $C_{q_1} \Box C_{q_2} \Box \cdots \Box C_{q_n}$, where $n \ge 4, r \ge n$, and $q_i \ge 2r+1$ for $i = 1,\dots,n$, and we wish to prove that they possess no optimal local structure.

LEMMA 5.1. *Let $\mathcal{U}$ be the set of all vertices $(t_1,\dots,t_{n-1},t_n)$, such that $|t_i| \ge 1$ for $i = 1,\dots,n-1$, $t_n = r+1-n$, or $t_n = r+2-n$, and such that*

(i) *if $t_n = r+1-n$, then there is exactly one $i \le n-1$, such that $|t_i| = 2$ and $|t_j| = 1$ for $j \ne i, n$;*

(ii) *if $t_n = r+2-n$, then $|t_i| = 1$ for $i = 1,\dots,n-1$.*

Then $\mathcal{U} \subset N(\mathbf{0},r)$.

Proof. Since for every $\mathbf{t} \in \mathcal{U}$, we have $d(\mathbf{t},\mathbf{0}) = \sum_{i=1}^{n} |t_i| = r+1$, hence we have $\mathcal{U} \subset N(\mathbf{0},r)$. $\square$

LEMMA 5.2. *Let $\mathbf{t} \in \mathcal{U}$ and suppose $\mathbf{u}$ is a vertex such that $\mathbf{t} \in B(\mathbf{u},r)$ and $B(\mathbf{u},r) \cap B(\mathbf{0},r) = \emptyset$. Then*

(i) *if $i \in \{1,\dots,n-1\}$ and $t_i > 0$, then $u_i = t_i + w_i$, where $0 \le w_i \le r$;*

(ii) *if $i \in \{1,\dots,n-1\}$ and $t_i < 0$, then $u_i = t_i - w_i$, where $0 \le w_i \le r$;*

(iii) *$u_n = t_n + r - \sum_{i=1}^{n-1} |u_i - t_i|$, where $\sum_{i=1}^{n-1} |u_i - t_i| \le r$.*

Conversely, for every $\mathbf{u}$ satisfying (i), (ii), *and* (iii), $\mathbf{t} \in B(\mathbf{u},r)$.

Proof. Let $\mathbf{t} \in \mathcal{U}$ and suppose $\mathbf{u}$ is a vertex, such that $\mathbf{t} \in B(\mathbf{u},r)$ and $B(\mathbf{u},r) \cap B(\mathbf{0},r) = \emptyset$. Suppose there is a coordinate $i_0 \le n-1$ for which (i) does not hold. Then $u_{i_0} = t_{i_0} + w_{i_0}$, and since $d_{i_0}(u_{i_0},t_{i_0}) \le r$ we have $-r < w_{i_0} < 0$. Thus $\mathbf{t}' = (t_1,\dots,t_{i_0}-1,\dots,t_n) \in B(\mathbf{u},r)$, contradicting the fact that $B(\mathbf{u},r) \cap B(\mathbf{0},r) = \emptyset$. With analogous argument we prove (ii).

Let $u_n = t_n + r - k$ for some k. Since $d_n(u_n,t_n) \le r$ we have $0 \le k \le 2r$. If $k > r$, then $\mathbf{t}' = (t_1,\dots,t_{n-1},t_n - 1) \in B(\mathbf{u},r)$, which is a contradiction. Thus $u_n = t_n + r - k$ for some $k \in \{0,\dots,r\}$. Since $B(\mathbf{u},r) \cap B(\mathbf{0},r) = \emptyset$ and $\mathbf{t} \in N(\mathbf{0},r)$ we have $d(\mathbf{u},\mathbf{t}) = r$, and thus $\sum_{i=1}^{n-1} |u_i - t_i| = k$.

Conversely, if $\mathbf{u}$ satisfies (i), (ii), and (iii), then $d(\mathbf{t},\mathbf{u}) = r$, and thus $\mathbf{t} \in B(\mathbf{u},r)$. Note that if $q_n > 4r-2$ and $q_i > 4r+4$ for $i = 1,\dots,n-1$, then (i), (ii), and (iii) together give $B(\mathbf{u},r) \cap B(\mathbf{0},r) = \emptyset$. $\square$

Lemma 5.2 is an analog of Lemma 4.2, describing all potential centers $\mathbf{u}$ of an r-ball containing $\mathbf{t}$ and not intersecting $B(\mathbf{0},r)$. Let $\mathbf{t} \in \mathcal{U}$ and denote by $U_{\mathbf{t}}$ the set of all vertices $\mathbf{u}$ satisfying (i), (ii), and (iii) of Lemma 5.2. With this notation we have the following corollary.

COROLLARY 5.3. *If $B(\mathbf{x^{(1)}}, r), B(\mathbf{x^{(2)}}, r), \ldots, B(\mathbf{x^{(m)}}, r)$ is an optimal local structure for $B(\mathbf{0}, r)$, then for every $\mathbf{t} \in \mathcal{U}$ there is exactly one $i \in \{1, \ldots, m\}$ such that $\mathbf{x^{(i)}} \in U_{\mathbf{t}}$.*

For $i = 0, \ldots, n-1$ we define the sets $\mathcal{U}_i$ as follows:
$\mathcal{U}_0 = \{(t_1, \ldots, t_n) \in \mathcal{U} \,|\, t_i > 0 \text{ for } i = 1, \ldots, n\}$ and
$\mathcal{U}_i = \{(t_1, \ldots, t_n) \in \mathcal{U} \,|\, t_i < 0, t_j > 0 \text{ for } j \neq i\}$ for $i \neq 0$.
Note that $\mathcal{U}_0$, $\mathcal{U}_1$, and $\mathcal{U}_2$ are the sets A_1, A_2, and A_3, if $n = 3$. Finally, we define for $i = 0, \ldots, n-1$ the set V_i as follows:

$$V_i = \bigcup_{\mathbf{t} \in \mathcal{U}_i} U_{\mathbf{t}}\,.$$

Note that the sets V_i are analogous to the sets $\mathcal{A}_i$ and that these sets appear symmetrically. The following lemma gives an explicit description of the sets V_i.

LEMMA 5.4. *Let $\mathbf{x} \in G$. Then $\mathbf{x} \ \in V_0$ if and only if*

$$\mathbf{x} = \left(1 + v_1, \ldots, 1 + v_{n-1}, 2r + 2 - n - \sum_{i=1}^{n-1} v_i\right)$$

for some $v_1, \ldots, v_{n-1}$ such that $\sum_{i=1}^{n-1} v_i \leq r+1$ and $0 \leq v_i \leq r+1$ for $i = 1, \ldots, n-1$, and $\mathbf{x} \ \in V_i$ if and only if

$$\mathbf{x} = \left(1 + v_1, \ldots, -1 - v_i, \ldots, 1 + v_{n-1}, 2r + 2 - n - \sum_{i=1}^{n-1} v_i\right) \tag{5.1}$$

for some $v_1, \ldots, v_{n-1}$ such that $\sum_{i=1}^{n-1} v_i \leq r+1$ and $0 \leq v_i \leq r+1$ for $i = 1, \ldots, n-1$.

Proof. We use the notation of Lemma 5.2. Suppose $\mathbf{u} \in V_i$ for some $i \geq 1$. We shall prove $\mathbf{u}$ is of the form (5.1). Since $\mathbf{u} \in V_i$, there is a $\mathbf{t} \in \mathcal{U}_i$ such that $\mathbf{u} \in U_{\mathbf{t}}$. Thus $\mathbf{u}$ satisfies (i), (ii), and (iii) of Proposition 5.2. Consider the equation

$$\begin{aligned}\mathbf{u} &= \left(t_1 + w_1, \ldots, t_i - w_i, \ldots, t_{n-1} + w_{n-1}, t_n + r - \sum_{i=1}^{n-1} w_i\right)\\ &= (1 + v_1, \ldots, -1 - v_i, \ldots, 1 + v_{n-1}, 2r + 2 - n - k)\end{aligned}$$

and define $k = r + 2 - n - t_n + \sum_{i=1}^{n-1} w_i$. Since $t_n = r + 1 - n$ or $t_n = r + 2 - n$ we infer $k \in \{0, \ldots, r+1\}$. We have $0 \leq v_i \leq r + 1$ and $0 \leq v_j \leq r + 1$ for $j \neq i$ by (ii) and (i), respectively. Finally, we have for $t_n = r + 2 - n$

$$\sum_{i=1}^{n-1} v_i = \sum_{i=1}^{n-1} w_i = k,$$

and for $t_n = r + 1 - n$

$$\sum_{i=1}^{n-1} v_i = \sum_{i=1}^{n-1} w_i + 1 = k\,.$$

Now let's prove the converse. Suppose $\mathbf{u}$ is of the form (5.1) and let $\sum_{i=1}^{n-1} v_i = k$.

Case 1. If $k \leq r$, then let $t_n = r + 2 - n$, $t_i = -1$, and $t_j = 1$ for $j \neq i$. It is easy to check that then $\mathbf{t} \in \mathcal{U}_i$ and $\mathbf{u} \in U_{\mathbf{t}}$.

Case 2. If $k = r+1$, then let $t_n = r+1-n$. Since $\sum_{i=1}^{n-1} v_i = r+1$, at least one $v_i > 0$. Let i_0 be any index such that $v_{i_0} > 0$, and let $t_{i_0} = 2, t_j = 1$ for $j \neq i$ and $t_i = -1$ (if $i_0 \neq i$) or $t_{i_0} = -2$ and $t_j = 1$ for $j \neq i$ (if $i_0 = i$). With these settings we again have $\mathbf{t} \in \mathcal{U}_i$ and $\mathbf{u} \in U_{\mathbf{t}}$. □

Note that in the case when $q_i \geq 2r+5$ for $i = 1, \ldots, n-1$, we have $V_j \cap V_k = \emptyset$ for $j \neq k$. In what follows we will use the following notation for $\mathbf{x} \in V_i$. Instead of writing

$$\mathbf{x} = \left(1+v_1, \ldots, -1-v_i, \ldots, 1+v_{n-1}, 2r+2-n-\sum_{i=1}^{n-1} v_i\right),$$

we write

$$\mathbf{x} = (1+v_1(\mathbf{x}), \ldots, -1-v_i(\mathbf{x}), \ldots, 1+v_{n-1}(\mathbf{x}), 2r+2-n-k(\mathbf{x}))\,,$$

where $v_i = v_i(\mathbf{x})$ for $i = 1, \ldots, n-1$ and $k(\mathbf{x}) = \sum_{i=1}^{n-1} v_i(\mathbf{x})$.

Lemma 5.5. *Suppose $B(\mathbf{x^{(1)}}, r), B(\mathbf{x^{(2)}}, r), \ldots, B(\mathbf{x^{(m)}}, r)$ is an optimal local structure for $B(\mathbf{0}, r)$, and let $\mathcal{X} = \{\mathbf{x^{(1)}}, \mathbf{x^{(2)}}, \ldots, \mathbf{x^{(m)}}\}, \mathcal{X}_j = \mathcal{X} \cap V_j$ for $j = 0, \ldots, n-1$. Then for every j there is exactly one $\mathbf{x} \in \mathcal{X}_j$ such that $k(\mathbf{x}) \leq r$, and for every $s \in \{1, \ldots, n-1\}$, there is exactly one $\mathbf{y} \in \mathcal{X}_j$ such that $v_s(\mathbf{y}) > 0$.*

Proof. Both assertions follow from Corollary 5.3 and the proof of Lemma 5.4. □

Example. We give an illustration of Lemma 5.5. Suppose $n = 7$ and $r = 9$; then the set $\mathcal{U}_0$ can be covered by r-balls centered on the vertices $\mathbf{x}, \mathbf{y}, \mathbf{z}$, and $\mathbf{w}$:

$\mathbf{x} = (2,4,1,1,1,1,9)$ and $(2,1,1,1,1,1,3), (1,2,1,1,1,1,3), (1,1,1,1,1,1,4) \in B(\mathbf{x}, 9)$,
$\mathbf{y} = (1,1,11,1,1,1,3)$ and $(1,1,2,1,1,1,3) \in B(\mathbf{y}, 9)$,
$\mathbf{z} = (1,1,1,11,1,1,3)$ and $(1,1,1,2,1,1,3) \in B(\mathbf{z}, 9)$,
$\mathbf{w} = (1,1,1,1,4,8,3)$ and $(1,1,1,1,1,2,3), (1,1,1,1,2,1,3) \in B(\mathbf{w}, 9)$.

Note that $k(\mathbf{x}) = 4 \leq r$ and $k(\mathbf{y}) = k(\mathbf{z}) = k(\mathbf{w}) = 10 = r+1$ and that the vertices $\mathbf{x}, \mathbf{y}, \mathbf{z}$, and $\mathbf{w}$ are pairwise on a distance $2r+2 = 20$.

Definition 5.6. *Let $\mathbf{u} \in V_0, \mathbf{x} \in V_i$, and $\mathbf{y} \in V_j$ for some $i, j \neq 0, i \neq j$. If*

$$\begin{aligned}
\mathbf{u} &= (1+v_1(\mathbf{u}), \ldots, 1+v_{n-1}(\mathbf{u}), 2r+2-n-k(\mathbf{u}))\,,\\
\mathbf{x} &= (1+v_1(\mathbf{x}), \ldots, -1-v_i(\mathbf{x}), \ldots, 1+v_{n-1}(\mathbf{x}), 2r+2-n-k(\mathbf{x}))\,,\\
\mathbf{y} &= (1+v_1(\mathbf{y}), \ldots, -1-v_j(\mathbf{y}), \ldots, 1+v_{n-1}(\mathbf{y}), 2r+2-n-k(\mathbf{y}))\,,
\end{aligned}$$

then we define the function $\chi : (\bigcup_{i=0}^{n-1} V_i) \times (\bigcup_{i=0}^{n-1} V_i) \to \mathbb{N}_0$ by

$$\begin{aligned}
&\chi(\mathbf{u}, \mathbf{x}) = \textstyle\sum_{k \neq i} \min\{v_k(\mathbf{u}), v_k(\mathbf{x})\}\,,\\
&\chi(\mathbf{u}, \mathbf{y}) = \textstyle\sum_{k \neq j} \min\{v_k(\mathbf{u}), v_k(\mathbf{y})\}, \text{ and}\\
&\chi(\mathbf{x}, \mathbf{y}) = \textstyle\sum_{k \neq i,j} \min\{v_k(\mathbf{x}), v_k(\mathbf{y})\}\,.
\end{aligned}$$

Proposition 5.7. *Let $\mathbf{u}, \mathbf{x}$, and $\mathbf{y}$ be as in Definition 5.6. Then*

$$\begin{aligned}
&d(\mathbf{u}, \mathbf{x}) \leq \max\{2k(\mathbf{u}), 2k(\mathbf{x})\} + 2 - 2\chi(\mathbf{u}, \mathbf{x}),\\
&d(\mathbf{u}, \mathbf{y}) \leq \max\{2k(\mathbf{u}), 2k(\mathbf{y})\} + 2 - 2\chi(\mathbf{u}, \mathbf{y}), \text{ and}\\
&d(\mathbf{x}, \mathbf{y}) \leq \max\{2k(\mathbf{x}), 2k(\mathbf{y})\} + 4 - 2\chi(\mathbf{x}, \mathbf{y})\,.
\end{aligned}$$

Proof. Since $|v_k(\mathbf{u}) - v_k(\mathbf{x})| = v_k(\mathbf{u}) + v_k(\mathbf{x}) - 2\min\{v_k(\mathbf{u}), v_k(\mathbf{x})\}$ for $k \neq i$, we have

$$\begin{aligned} d(\mathbf{u},\mathbf{x}) &\leq \sum_{k=1}^{n} |u_k - x_k| = \sum_{k\neq i,n} |v_k(\mathbf{u}) - v_k(\mathbf{x})| + v_i(\mathbf{u}) + v_i(\mathbf{x}) + 2 + |k(\mathbf{u}) - k(\mathbf{x})| \\ &= \sum_{k=1}^{n-1} v_k(\mathbf{u}) + \sum_{k=1}^{n-1} v_k(\mathbf{x}) + 2 - 2\chi(\mathbf{u},\mathbf{x}) + |k(\mathbf{u}) - k(\mathbf{x})| \\ &= k(\mathbf{u}) + k(\mathbf{x}) + 2 - 2\chi(\mathbf{u},\mathbf{x}) + |k(\mathbf{u}) - k(\mathbf{x})|, \end{aligned}$$

and therefore $d(\mathbf{u},\mathbf{x}) \leq \max\{2k(\mathbf{u}), 2k(\mathbf{x})\} + 2 - 2\chi(\mathbf{u},\mathbf{x})$. The rest we prove analogously. □

COROLLARY 5.8. *Suppose $\mathbf{x},\mathbf{y} \in \bigcup V_i$ and $d(\mathbf{x},\mathbf{y}) \geq 2r+1$. Then*

(i) *if $\mathbf{x} \in V_0, \mathbf{y} \in V_i, i \neq 0$, then $\chi(\mathbf{x},\mathbf{y}) \leq 1$;*
(ii) *if $\mathbf{x} \in V_0, \mathbf{y} \in V_i, i \neq 0$, and $k(\mathbf{x}), k(\mathbf{y}) \leq r$, then $\chi(\mathbf{x},\mathbf{y}) = 0$;*
(iii) *if $\mathbf{x} \in V_i, \mathbf{y} \in V_j, i \neq j$, then $\chi(\mathbf{x},\mathbf{y}) \leq 2$;*
(iv) *If $\mathbf{x} \in V_i, \mathbf{y} \in V_j, i \neq j$, and $k(\mathbf{x}), k(\mathbf{y}) \leq r$ then, $\chi(\mathbf{x},\mathbf{y}) \leq 1$.*

Now suppose that $B(\mathbf{x^{(1)}},r), B(\mathbf{x^{(2)}},r),\ldots,B(\mathbf{x^{(m)}},r)$ is an optimal local structure for $B(\mathbf{0},r)$ and $\mathcal{X} = \{\mathbf{x^{(1)}},\mathbf{x^{(2)}},\ldots,\mathbf{x^{(m)}}\}$. We already know by Corollary 4.3 and the definition of V_j that $\mathcal{X}_j = \mathcal{X} \cap V_j \neq \emptyset$ for $j = 0,\ldots,n-1$. Since $d(\mathbf{x},\mathbf{y}) \geq 2r+1$ for $\mathbf{x},\mathbf{y} \in \mathcal{X}, \mathbf{x} \neq \mathbf{y}$, we can use Proposition 5.7 to give us some further requirements on the set $\mathcal{X}$. Also, Lemma 5.5 gives some restrictions on the sets $\mathcal{X}_j$. In the following lemmas we will first show some additional properties of the sets $\mathcal{X}_j$ and then show that either $d(\mathbf{x},\mathbf{y}) \leq 2r$ for some $\mathbf{x},\mathbf{y} \in \mathcal{X}, \mathbf{x} \neq \mathbf{y}$, or for some j the set $\mathcal{X}_j$ does not fulfill Lemma 5.5, contradicting the fact that r-balls centered on vertices from $\mathcal{X}$ constitute an optimal local structure.

LEMMA 5.9. *Suppose $B(\mathbf{x^{(1)}},r), B(\mathbf{x^{(2)}},r),\ldots,B(\mathbf{x^{(m)}},r)$ is an optimal local structure for $B(\mathbf{0},r)$. Then $|\mathcal{X}_j| \geq 2$ for $j = 0,\ldots,n-1$.*

Proof. Assume on the contrary that $|\mathcal{X}_j| = 1$ for some $j \leq n-1$. Without loss of generality we can assume $j = 0$. Thus $\mathcal{X}_0 = \{\mathbf{x}\}$, where

$$\mathbf{x} = (1 + v_1(\mathbf{x}),\ldots,1 + v_{n-1}(\mathbf{x}), 2r + 2 - n - k(\mathbf{x}))\,,$$

$k(\mathbf{x}) \leq r$, and $v_i(\mathbf{x}) \geq 1$ for $i = 1,\ldots,n-1$. For $i = 1,\ldots,n-1$ let $\mathbf{y_i} \in \mathcal{X}_i$ be the element such that $k(\mathbf{y_i}) \leq r$ (see Lemma 5.5). Since there is at most one $i \in \{1,\ldots,n-1\}$ such that $k(\mathbf{y_i}) = 0$ (see Proposition 5.7) and $n-1 \geq 3$, there exist $a, b \leq n-1, a \neq b$, such that $k(\mathbf{y_a}), k(\mathbf{y_b}) \geq 1$. Since $v_i(\mathbf{x}) \geq 1$ for $i = 1,\ldots,n-1$ we have by Corollary 5.8(ii) that

$$\mathbf{y_a} = (1,\ldots,1,-1-k(\mathbf{y_a}),1,\ldots,1,2r+2-n-k(\mathbf{y_a})) \in \mathcal{X}_a$$

and

$$\mathbf{y_b} = (1,\ldots,1,-1-k(\mathbf{y_b}),1,\ldots,1,2r+2-n-k(\mathbf{y_b})) \in \mathcal{X}_b$$

(recall that the elements of $\mathcal{X}_a$ and $\mathcal{X}_b$ have, respectively, an ath and a bth coordinate negative). For $i = 1,\ldots,n-1$ and $i \neq a$ let $\mathbf{w_i} \in \mathcal{X}_a$ be the element with $v_i(\mathbf{w_i}) \geq 1$; then by Corollary 5.8(i), we have

$$\mathbf{w_i} = (1,\ldots,-1,\ldots,r+2,\ldots,r+1-n),$$

where the ith coordinate of $\mathbf{w_i}$ is $r+2$. Analogously for $i = 1, \dots, n-1$ and $i \neq b$ let $\mathbf{z_i} \in \mathcal{X}_b$ be the element with $v_i(\mathbf{z_i}) \geq 1$; then by Corollary 5.8(i) we have

$$\mathbf{z_i} = (1, \dots, -1, \dots, r+2, \dots, r+1-n),$$

where the ith coordinate of $\mathbf{z_i}$ is $r+2$. Thus

$$\mathcal{X}_a = \{\mathbf{y_a}, \mathbf{w_i} \,|\, i = 1, \dots, n-1, i \neq a\} \text{ and } \mathcal{X}_b = \{\mathbf{y_b}, \mathbf{z_i} \,|\, i = 1, \dots, n-1, i \neq b\}\,.$$

Since $n \geq 4$ we have $i \neq a, b$ for some $i \in \{1, \dots, n-1\}$ and $d(\mathbf{z_i}, \mathbf{w_i}) = 4 \leq 2r$. □

THEOREM 5.10. *Let $r \geq n \geq 4$ and $G = C_{q_1} \Box C_{q_2} \Box \cdots \Box C_{q_n}, q_i \geq 2r+1$ for $i = 1, \dots, n$. Then there does not exist an optimal local structure for $B(\mathbf{u}, r)$ for any $\mathbf{u} \in G$.*

Proof. It is enough to prove this for $\mathbf{u} = \mathbf{0}$. Assume on the contrary that $B(\mathbf{x^{(1)}}, r), B(\mathbf{x^{(2)}}, r), \dots, B(\mathbf{x^{(m)}}, r)$ is an optimal local structure for $B(\mathbf{0}, r)$. Suppose that there is an $\mathbf{x} \in \cup_{i=0}^{n-1} \mathcal{X}_i$ such that $k(\mathbf{x}) \leq r-1$. Since the sets V_i are symmetric we can assume, without loss of generality, that $\mathbf{x} \in \mathcal{X}_0$. (If there is no $\mathbf{x} \in \cup_{i=0}^{n-1} \mathcal{X}_i$ such that $k(\mathbf{x}) \leq r-1$, then let $\mathbf{x} \in \mathcal{X}_0$ be the vertex with $k(\mathbf{x}) \leq r$.) Therefore, by Proposition 5.7, $k(\mathbf{y}) \geq r$ for every $\mathbf{y} \in \bigcup_{i=0}^{n-1} \mathcal{X}_i, \mathbf{y} \neq \mathbf{x}$. Let s be the number of coordinates of $\mathbf{x}$, with $v_i(\mathbf{x}) \neq 0$. Thus

$$s = |\{i \,|\, v_i(\mathbf{x}) > 0\}|\,,$$

and we can assume, without loss of generality, that $v_1(\mathbf{x}), \dots, v_s(\mathbf{x}) > 0$.

By Lemma 5.9, $0 \leq s \leq n-2$ and $|\mathcal{X}_i| \geq 2$ for $i = 1, \dots, n-1$. We claim that for every $i \neq 0$ there is a $\mathbf{y} \in \mathcal{X}_i$, such that $v_k(\mathbf{y}) \geq 3$ for some $k \in \{1, \dots, n-1\}, k \neq i$. Assume on the contrary that for some $i_0 \in \{1, \dots, n-1\}$ and for every $\mathbf{y} \in \mathcal{X}_{i_0}, v_k(\mathbf{y}) \leq 2$ for $k \neq i_0$. Since $|\mathcal{X}_{i_0}| \geq 2$, there is a $\mathbf{u} \in \mathcal{X}_{i_0}$ such that $v_{i_0}(\mathbf{u}) = 0$. Thus $\sum_{i \neq i_0} v_i(\mathbf{u}) = r$ if $k(\mathbf{u}) = r$ and $\sum_{i \neq i_0} v_i(\mathbf{u}) = r+1$ if $k(\mathbf{u}) = r+1$. Assume $k(\mathbf{u}) = r+1$. Let

$$Q_1 = \{i \,|\, v_i(\mathbf{u}) = 1\} \text{ and } Q_2 = \{i \,|\, v_i(\mathbf{u}) = 2\}\,,$$

and denote

$$q_1 = |Q_1| \text{ and } q_2 = |Q_2|\,.$$

Thus

$$q_1 + 2q_2 = r+1 \geq n+1\,.$$

Let $\mathbf{z} \in V_0$ and $\mathbf{z} \neq \mathbf{x}$. Since $k(\mathbf{x}) \leq r$, we have $k(\mathbf{z}) = r+1$ and $v_i(\mathbf{z}) = 0$ for $i \leq s$. For any $i \in Q_2$, we have $v_i(\mathbf{z}) \leq 1$, since otherwise $v_i(\mathbf{z}), v_i(\mathbf{u}) \geq 2$, and $\chi(\mathbf{u}, \mathbf{z}) \geq 2$, and therefore $d(\mathbf{u}, \mathbf{z}) = 2k(\mathbf{z}) + 2 - 2\chi(\mathbf{u}, \mathbf{z}) \leq 2r$. Note that $q_2 \geq 3$, and by Corollary 5.8(i) there is at most one coordinate $i \leq s$ with $v_i(\mathbf{u}) \geq 1$. For every $i \in Q_2, i \geq s+1$, there is $\mathbf{z} \in V_0$ such that $v_i(\mathbf{z}) = 1$. Since $\sum_{i=s+1}^{n-1} v_i(\mathbf{z}) = r+1 \geq 5$, there is a coordinate $\ell \neq i$ such that $v_\ell(\mathbf{z}) \geq 1$. Moreover, $\ell \geq s+1$ and $\ell \notin Q_1 \cup Q_2$. This reasoning gives the following inequality:

$$(5.2) \quad n-1-s-(q_1+q_2-1) \geq q_2-1 \text{ for } s \geq 1, \qquad n-1-(q_1+q_2) \geq q_2 \text{ for } s = 0.$$

The left side of the inequality is (at most) the number of the coordinates $i \geq s+1$ with $v_i(\mathbf{u}) = 0$, and the right side is (at least) the number of the coordinates $i \geq s+1, i \neq i_0$, with $v_i(\mathbf{u}) = 2$. But since $q_1 + 2q_2 = r + 1 \geq n + 1$, the inequalities (5.2) yield a contradiction. Assume now that $k(\mathbf{u}) = r$; then

$$n - 1 - s - (q_1 + q_2) \geq q_2 \quad \text{for } s \geq 0,$$

and since $q_1 + 2q_2 = r \geq n$ we have a contradiction again. So in any case, we ran into contradiction, and thus for every $i \neq 0$ there is a $\mathbf{y} \in \mathcal{X}_i$, such that $v_k(\mathbf{y}) \geq 3$ for some $k \in \{1, \dots, n-1\}, k \neq i$. But if $\mathbf{y_1}, \mathbf{y_2}$ are from $\mathcal{X}_{i_1}, \mathcal{X}_{i_2}$, respectively, such that $v_k(\mathbf{y_1}), v_k(\mathbf{y_2}) \geq 3$ for some $k \neq i_1, i_2$, then $\chi(\mathbf{y_1}, \mathbf{y_2}) \geq 3$, and hence $d(\mathbf{y_1}, \mathbf{y_2}) \leq 2r$. Thus for every $k \in \{1, \dots, n-1\}$, there is a $\mathbf{y} \in \mathcal{X}_i, i \neq k$, such that $v_k(\mathbf{y}) \geq 3$. Since there is a $\mathbf{z} \in \mathcal{X}_0$ such that $v_k(\mathbf{z}) \geq 2$ for some $k \geq s + 1$ and a $\mathbf{y} \in \mathcal{X}_i, i \neq k$, such that $v_k(\mathbf{y}) \geq 3$, we have $\chi(\mathbf{y}, \mathbf{z}) \geq 2$, and therefore $d(\mathbf{y}, \mathbf{z}) \leq 2r$, which is a contradiction. □

Note that in the proof of the above theorem the sets V_i do not need to be disjoint, all arguments work also if $2r + 1 \leq q_i \leq 2r + 4$, and the sets V_i are not disjoint.

Since the local structure of Cartesian products of paths is the same as the local structure of Cartesian products of cycles, we can state the following theorem (we denote the path of length k by P_k and set $V(P_k) = \{0, 1, \dots, k\}$).

THEOREM 5.11. *Let $r \geq n \geq 3$. If $G = P_{q_1} \Box P_{q_2} \Box \cdots \Box P_{q_n}$, $q_i \geq 4$ for $i = 1, \dots, n$, and $q_j \geq r+2-n$ for some $j \in \{1, \dots, n\}$, then there does not exist an optimal local structure for $B(\mathbf{u}, r)$ for any $\mathbf{u}$ such that $u_j \geq r + 2 - n$ (or $u_j \leq q_j - r - 2 + n$) and $2 \leq u_i \leq q_i - 2$ for $i \neq j$.*

The position of $\mathbf{u}$ makes it possible to define a set analogous to the set $\mathcal{U}$ of Lemma 5.1 and proceed with the proof of the above theorem, as done in the proof for products of cycles.

6. Conclusion.

Combining sections 4 and 5, we have the following theorem.

THEOREM 6.1. *Let $r \geq n \geq 3$ and $G = C_{q_1} \Box C_{q_2} \Box \cdots \Box C_{q_n}, q_i \geq 2r + 1$ for $i = 1, \dots, n$. Then there does not exist an optimal local structure for $B(\mathbf{u}, r)$ for any $\mathbf{u} \in G$.*

The results of this paper and the fact that there does not exist an optimal local structure for $n = 3, r \geq 2$ (cf. [9, 11]) and for $n = 4, r \geq 2$ (cf. [22]) justify the following conjecture.

CONJECTURE 6.2. *Let $n \geq 3$ and $2 \leq r \leq n - 1$. If $G = C_{q_1} \Box C_{q_2} \Box \cdots \Box C_{q_n}$ and $q_i \geq 2r + 1$ for $i = 1, \dots, n$, then there does not exist an optimal local structure for $B(\mathbf{u}, r)$ for any $\mathbf{u} \in G$.*

Lee codes have been a subject of intense research, and numerous results related to Lee codes and perfect Lee codes have been given by different authors. However, the underlying question of the existence of an optimal local structure remained unsolved (except for some special cases). The results of this paper provide an understanding of the local structure of Lee codes and give further suggestions for the study of Lee codes and related topics such as domination and packing in Cartesian product of cycles and paths.

Acknowledgment. Many thanks to Sandi Klavžar for many valuable suggestions regarding this manuscript.

REFERENCES

[1] A. A. ANDRADE, J. C. INTERLANDO, AND R. PALAZZO, JR., *Alternant and BCH codes over certain rings,* Comput. Appl. Math., 22 (2003), pp. 233–247.

[2] J. T. Astola, *On perfect codes in the Lee-metric,* Ann. Univ. Turku. Ser. A I, no. 176 (1978).

[3] J. T. Astola, *An Elias-type of bound for Lee codes over large alphabets and its application to perfect codes,* IEEE Trans. Inform. Theory, 28 (1982), pp. 111–113.

[4] J. T. Astola, *On the asymptotic behaviour of Lee-codes,* Discrete Appl. Math., 8 (1984), pp. 13–23.

[5] N. Biggs, *Perfect codes in graphs*, J. Combin. Theory Ser. B, 15 (1973), pp. 289–296.

[6] E. Byrne, *Decoding a class of Lee metric codes over Galois Ring*, IEEE Trans. Inform. Theory, 48 (2002), pp. 966–975.

[7] P. Cull and I. Nelson, *Error-correcting codes on the Towers of Hanoi graphs*, Discrete Math., 208/209 (1999), pp. 157–175.

[8] S. W. Golomb and L. R. Welch, *Algebraic coding and the Lee metric,* in Error Correcting Codes (Proc. Sympos. Math. Res. Center, Madison, WI), John Wiley, New York, 1968, pp. 175–194.

[9] S. W. Golomb and L. R. Welch, *Perfect codes in the Lee metric and the packing of polyominoes,* SIAM J. Appl. Math., 18 (1970), pp. 302–317.

[10] S. Gravier, M. Mollard, *On domination numbers of Cartesian product of paths,* Discrete Appl. Math., 80 (1997), pp. 247–250.

[11] S. Gravier, M. Mollard, and C. Payan, *On the nonexistence of* 3*-dimensional tilting in the Lee metric*, European J. Combin., 19 (1998), pp. 567–572.

[12] S. Gravier, M. Mollard, and C. Payan, *Variations on tilings in the Manhattan metric*, Geom. Dedicata, 76 (1999), pp. 265–273.

[13] W. Imrich and S. Klavžar, *Product Graphs: Structure and Recognition,* John Wiley, New York, 2000.

[14] P. K. Jha, *Smallest independent dominating sets in Kronecker products of cycles*, Discrete Appl. Math., 113 (2001), pp. 303–306.

[15] J. Kratochvíl, *Perfect Codes in General Graphs*, Rozpravy Československé Akad. Věd Řada Mat. Přírod. Věd, no. 7, Akademia, Praha, 1991.

[16] T. Lepistö, *A modification of the Elias-bound and nonexistence theorems for perfect codes in the Lee-metric*, Inform. Control, 49 (1981), pp. 109–124.

[17] T. Lepistö, *Bounds for perfect Lee-codes over small alphabets*, Ann. Univ. Turku. Ser. A I, no. 186, (1984), pp. 59–63.

[18] M. Livingstone and Q. F. Stout, *Perfect dominating sets*, Congr. Numer., 79 (1990), pp. 187–203.

[19] J. Quistorff, *Some remarks on the Plotkin bound,* Electron. J. Combin., 10 (2003), Note 6.

[20] K. A. Post, *Nonexistence theorems on perfect Lee codes over large alphabets,* Inform. Control, 29 (1975), pp. 369–380.

[21] R. M. Roth and P. H. Siegel, *Lee-Metric BCH codes and their application to constrained and partial-response channels*, IEEE Trans. Inform. Theory, 40 (1994), pp. 1083–1096.

[22] S. Špacapan, *A Complete Proof of the Nonexistence of Regular Four Dimensional Tilings in the Lee Metric*, preprint 933, IMFM Preprint Series, Vol. 42, Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia, 2004.

SIAM J. Discrete Math.
Vol. 21, No. 3, pp. 763–784

CAYLEY DIGRAPHS OF FINITE ABELIAN GROUPS AND MONOMIAL IDEALS*

DOMINGO GÓMEZ†, JAIME GUTIERREZ†, AND ÁLVAR IBEAS†

Abstract. In the study of double-loop computer networks, the diagrams known as *L-shapes* arise as a graphical representation of an optimal routing for every graph's node. The description of these diagrams provides an efficient method for computing the diameter and the average minimum distance of the corresponding graphs. We extend these diagrams to multiloop computer networks. For each Cayley digraph with a finite abelian group as vertex set, we define a monomial ideal and consider its representations via its minimal system of generators or its irredundant irreducible decomposition. From this last piece of information, we can compute the graph's diameter and average minimum distance. That monomial ideal is the initial ideal of a certain lattice with respect to a graded monomial ordering. This result permits the use of Gröbner bases for computing the ideal and finding an optimal routing. Finally, we present a family of Cayley digraphs parametrized by their diameter d, all of them associated to irreducible monomial ideals.

Key words. monomial ideals, Cayley digraph, Gröbner bases, multiloop networks

AMS subject classifications. 13P10, 05C25, 68M10

DOI. 10.1137/050646056

1. Introduction.

Let Γ be a group and $S \subseteq \Gamma$ a subset. The *Cayley digraph* associated to (Γ, S) is a directed graph whose vertex set is Γ and whose edge set is $\{(g,h) \in \Gamma^2 \mid g^{-1}h \in S\}$. Every Cayley digraph is vertex-symmetric and its degree equals the number of elements in S. These graphs are connected if and only if the set S generates the group. We are dealing with digraphs associated to finite abelian groups, but we are mainly interested in those associated to cyclic groups. Let N be a positive integer and $\mathbb{Z}_N$ the integers modulo N. For any subset $S = \{j_1, \ldots, j_r\}$ of this abelian group we denote by $C_N(S) = C_N(j_1, \ldots, j_r)$ the corresponding Cayley digraph (see Figure 1.1), which is called the *circulant digraph* or *multiloop computer network* of jumps $j_1, \ldots, j_r$. It is connected if and only if $\gcd(j_1, \ldots, j_r, N) = 1$. If S is a subset of $\mathbb{Z}_N$ such that for every element in S its inverse also lies in S, then $C_N(S)$ is an undirected graph called a *circulant graph* or *distributed multiloop computer network*.

Multiloop networks were first proposed in [32] for organizing multimodule memory services and have a vast number of applications in telecommunication networking, VLSI design, and distributed computation. Their properties, such as diameter and reliability, have been the focus of much research in computer network design; see, for instance, [5, 7, 12, 13, 19, 21, 25, 33].

The *single-loop network* or *ring network* is mathematically trivial. Digraphs with $r = 2$ or *double-loop networks* and their corresponding undirected graphs (*distributed double-loop networks*, with degree four) have been extensively studied; see the surveys [3, 20] and the references therein. When $C_N(j_1, j_2)$ is connected, one can define a *minimum distance diagram* (MDD) as an array with vertex 0 in cell $(0, 0)$ and vertex c

*Received by the editors November 25, 2005; accepted for publication (in revised form) March 22, 2007; published electronically September 26, 2007. This research was partially supported by the Research Project MTM2004-07086 of the Spanish Ministry of Science. An extended abstract of part of this work appeared in [14].

http://www.siam.org/journals/sidma/21-3/64605.html

†University of Cantabria, E–39071 Santander, Spain (domingo.gomez@unican.es, jaime.gutierrez@unican.es, alvar.ibeas@unican.es).

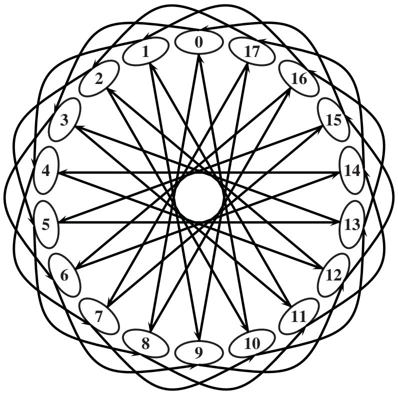


Fig. 1.1. $C_{18}(3, 8)$.

8											
27											
13											
32											
18											
4											
23											
9											
28											
14	19	24	29	1	6	11	16	21	26	31	3
0	5	10	15	20	25	30	2	7	12	17	22

Fig. 1.2. *MDD of* $C_{33}(5, 14)$.

in cell (x, y) (x is the column index and y the row index), for a particular choice satisfying $j_1x + j_2y \equiv c \bmod N$, and $x + y$ minimum. One example is shown in Figure 1.2.

The classical work of Wong and Coppersmith [32] presents an algorithm for constructing an MDD of $C_N(j_1, j_2)$ in $O(N^2)$ steps and shows it has an "L" shape. Several characterizations and applications of this idea for describing circulants with desirable properties appear in [1, 8, 9, 13]. However, they do not focus on higher degree digraphs.

Two notable parameters in a graph are the diameter d and the average minimum distance $\bar{d}$. The former represents the worst delay in the communication between two nodes, and the latter represents the average delay. Given an L-shape, it is easy to compute d and $\bar{d}$.

On the other hand, let $d_r(N) := \min\{d(C_N(j_1, \dots, j_r) \mid j_1, \dots, j_r \in \mathbb{Z}_N\}$. An important problem is to determine this value and find a specific $C_N(j_1, \dots, j_r)$ attaining this minimum. The network $C_N(j_1, \dots, j_r)$ is said to be *optimal* if its diameter equals $d_r(N)$. In some cases, it is difficult to obtain optimal networks; however, one can find general simple functions serving as upper and lower bounds for $d_r(N)$; see [3]. The paper [32] shows $d_2(N) \leq \sqrt{3N} - 2$ and presents a family of circulant digraphs with diameter $2\sqrt{N} - 2$.

In this article we present monomial ideals as a natural tool for studying the MDDs of arbitrary Cayley digraphs, provided that the vertex group is finite and abelian. Given a graded monomial ordering and a Cayley digraph (Γ, S), we build a monomial ideal in the polynomial ring $\mathbb{K}[X_1, \dots, X_r]$, where $\mathbb{K}$ is an arbitrary field and $r = \#S$. We obtain some properties of this monomial ideal: in particular, a certain generalization of the two-dimensional L-shape is shown. On the other side, it is the initial ideal of a certain lattice. This result permits the use of Gröbner bases for computing the ideal and finding an optimal routing for each pair of nodes. Given the representation of the monomial ideal via its irreducible decomposition, we provide formulae to compute d and $\bar{d}$. We also show a family of circulant digraphs of degree two which coincides with the family obtained in paper [32]. Finally, we present a new and attractive family of circulant digraphs of arbitrary degree parametrized by the diameter d, with average minimum distance $d/2$, and whose associated monomial ideals are irreducible.

The paper is divided into nine sections. In section 2 we collect several known facts about monomial ideals, presenting examples and fixing notation for later use. Section 3 presents the key idea of associating monomial ideals to digraphs in order to obtain an MDD, and it also provides an algorithm to construct an MDD for Cayley digraphs with a finite abelian group as vertex set. Section 4 is devoted to presenting the relation between MDDs and the ideal of a lattice. In section 5 we present an algorithm to compute a shortest path between two vertices by means of Gröbner bases. Section 6 presents an algorithm specifically tailored for degree three circulants. It computes the minimal system of generators in $O(s \log N)$ arithmetic operations, where s is the number of generators and N is the number of nodes. Section 7 is dedicated to providing formulae to find the diameter and the average minimum distance. Then section 8 presents a family of multiloop computer networks with an arbitrary number of jumps, parametrized by the diameter d, and all of them associated to irreducible monomial ideals. We conclude with a short summary and a discussion of open questions.

2. Monomial ideals.

2. Monomial ideals. Monomial ideals form an important link between commutative algebra and combinatorics. Here we review several basic related results and definitions concerning monomial ideals; see, for instance, [2, 30].

Let $\mathbb{K}$ be an arbitrary field and $\mathbb{K}[X_1, \dots, X_r]$ the polynomial ring in the variables $X_1, \dots, X_r$. Throughout the paper, we very often identify monomials of $\mathbb{K}[X_1, \dots, X_r]$ with vectors of $\mathbb{N}^r$ and use the following notation:

$$\mathbf{x}^{\mathbf{a}} = X_1^{a_1} \cdots X_r^{a_r} \longleftrightarrow \mathbf{a} = (a_1, \dots, a_r),$$

$$\mathbf{x}^{\mathbf{a}} | \mathbf{x}^{\mathbf{b}} \iff \mathbf{a} = (a_1, \dots, a_r) \leq \mathbf{b} = (b_1, \dots, b_r) \overset{\text{def}}{\iff} a_i \leq b_i \quad \forall i = 1, \dots, r,$$

$$\mathbf{a} = (a_1, \dots, a_r) \sqsubset \mathbf{b} = (b_1, \dots, b_r) \overset{\text{def}}{\iff} (b_i > 0 \Rightarrow a_i < b_i)\,,$$

$$\mathbf{e}_i := (0, \dots, \overset{\overset{i}{\smile}}{1}, \dots, 0), \quad \mathfrak{m}^{\mathbf{a}} := (X_i^{a_i} \mid a_i > 0), \quad \mathbf{1} := (1, \dots, 1).$$

The definition of $\sqsubset$ suits the characterization in (2.2), and when it is employed (in expressions like $\mathbf{a} \sqsubset \mathbf{b}$), we usually have $\mathbf{1} \leq \mathbf{b}$.

A *monomial ideal* is an ideal generated by monomials, i.e., $I \subset \mathbb{K}[X_1, \dots, X_r]$ is a monomial ideal if there is a subset $A \subseteq \mathbb{N}^r$ such that

$$I = (\mathbf{x}^{\mathbf{a}} \mid \mathbf{a} \in A) = (A).$$

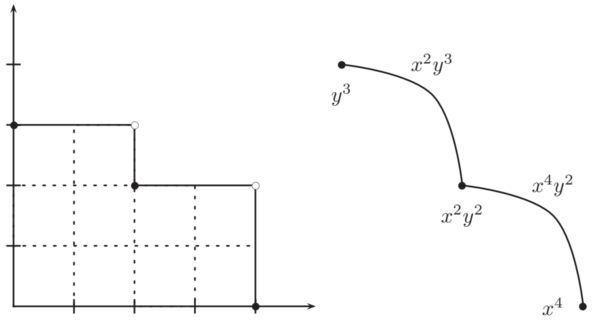


FIG. 2.1. *Staircase diagram and Buchberger's graph.*

There are two standard ways of describing a nontrivial monomial ideal:

- Via the (unique) minimal system of monomial generators $I = (\mathbf{x}^{\mathbf{a}_1}, \ldots, \mathbf{x}^{\mathbf{a}_s})$, we have

$$\mathbf{x}^{\mathbf{u}} \in I \iff \exists\ i \in \{1, \ldots, s\} \mid \mathbf{a}_i \leq \mathbf{u}. \tag{2.1}$$

- Via the (unique) irredundant decomposition by irreducible monomial ideals $I = \mathfrak{m}^{\mathbf{b}_1} \cap \cdots \cap \mathfrak{m}^{\mathbf{b}_n}$, we have

$$\mathbf{x}^{\mathbf{u}} \notin I \iff \exists\ i \in \{1, \ldots, n\} \mid \mathbf{u} \sqsubset \mathbf{b}_i. \tag{2.2}$$

The so-called staircase diagram is a useful graphical representation of monomial ideals.

Example 2.1. The monomial ideal $I_1 := (x^4, x^2y^2, y^3) = (x^2, y^3) \cap (x^4, y^2)$ is represented on the left in Figure 2.1.

There is an algorithm for finding the irredundant irreducible decomposition of a monomial ideal based on Alexander duality; see [27]. An irreducible component $\mathfrak{m}^{\mathbf{a}}$ can be associated to $\operatorname{lcm}(X_1^{a_1}, \ldots, X_r^{a_r}) = \mathbf{x}^{\mathbf{a}}$. On the other hand, if $\mathbb{K}[X_1, \ldots, X_r]/I$ is an artinian ring, then the monomial $\mathbf{x}^{\mathbf{a}}$ associated to the irreducible component $\mathfrak{m}^{\mathbf{a}}$ must coincide with the least common multiple of a subset of the minimal generators of I. In the above Example 2.1 we have

$$x^2y^3 = \operatorname{lcm}(x^2y^2, y^3),\ x^4y^2 = \operatorname{lcm}(x^4, x^2y^2).$$

The diagram on the right in Figure 2.1 is called *Buchberger's graph* of the monomial ideal I_1; see [28]. At any stage in Buchberger's algorithm for computing Gröbner bases, one considers the S-pairs among the current polynomials and removes those which are redundant; the minimal S-pairs define a graph on the generators of any monomial ideal.

THEOREM 2.2. *Let I be a nontrivial monomial ideal given by a minimal system of generators $I = (\mathbf{x}^{\mathbf{a}_1}, \ldots, \mathbf{x}^{\mathbf{a}_s})$ and by the irredundant irreducible decomposition $I = \mathfrak{m}^{\mathbf{b}_1} \cap \cdots \cap \mathfrak{m}^{\mathbf{b}_n}$. The following are equivalent:*

1. *$\mathbb{K}[X_1, \ldots, X_r]/I$ is an artinian ring.*
2. *$\forall i = 1, \ldots, r$, one of the generators' exponents is $\mathbf{a}_j = \alpha_i \mathbf{e}_i$ for some $\alpha_i \in \mathbb{N}$.*
3. *$\forall i = 1, \ldots, n,\ \forall j = 1, \ldots, r,\ b_{i,j} > 0$.*

Proof. We need to prove that the number of monomials outside I is finite if and only if either of the two last items is satisfied. We do that using the characterizations in (2.1) and (2.2).

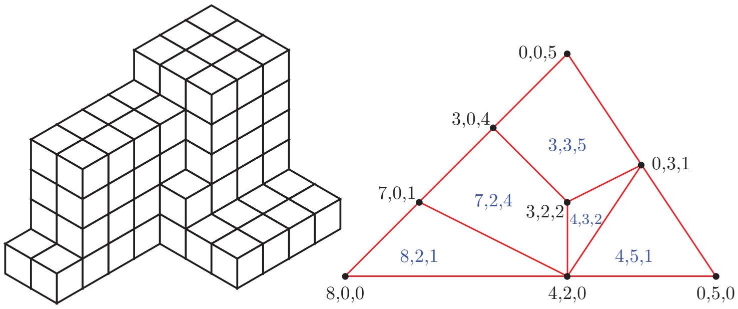


FIG. 2.2. *Planar graph associated to I_2.*

If the second item is true, then the number of monomials which do not lie in the ideal is bounded by the product $\prod \alpha_i$. Conversely, if that item is false, there exists an index $i \in \{1, \ldots, r\}$ such that $X_i^\alpha \notin I \; \forall \alpha \in \mathbb{N}$.

The third item is obviously equivalent to $\#\{\mathbf{u} \in \mathbb{N}^r \mid \mathbf{u} \sqsubset \mathbf{b}_i \text{ for some } i \in \{1, \ldots, r\}\} < \infty$. $\square$

We conclude this section by illustrating those facts in the following example.

Example 2.3. In [28], a planar graph is associated to every monomial ideal in three variables satisfying the conditions in Theorem 2.2. The monomial $\mathbf{x}^{\mathbf{b}}$ associated to an irreducible component $\mathfrak{m}^{\mathbf{b}}$ is identified with a connected component in the graph's complement and can be obtained as the least common multiple of the generators in its boundary. In Figure 2.2 we show this construction for the ideal:

$$\begin{aligned} I_2 &:= (x^8, x^4y^2, y^5, y^3z, z^5, x^3z^4, x^7z, x^3y^2z^2) \\ &= (x^8, y^2, z) \cap (x^7, y^2, z^4) \cap (x^4, y^3, z^2) \cap (x^4, y^5, z) \cap (x^3, y^3, z^5). \end{aligned}$$

The description of those relations permits the simplification of some computations on Cayley digraphs, as pointed out in section 7.

3. Minimum distance diagrams.

There are different ways to relate monomial ideals with graphs (see, for instance, [30]). In this section we propose a new approach to studying Cayley digraphs in which we associate a graph with a monomial ideal. The routing problem for Cayley digraphs reduces to studying paths originating at a fixed vertex, as these graphs are vertex-symmetric. Given a graph associated to $(\Gamma, \{s_1, \ldots, s_r\})$, where Γ is finite and abelian, we are looking for the shortest path from node 0_Γ to node $c \; \forall c \in \Gamma$, i.e., a *minimum distance diagram* (MDD). We can construct the *routing mapping* R:

$$\begin{array}{rccc} R: & \mathbb{N}^r & \longrightarrow & \Gamma \\ & \mathbf{a} & \mapsto & a_1s_1 + \cdots + a_rs_r. \end{array} \tag{3.1}$$

Thus, we need to find a right inverse map of R:

$$D: \; \Gamma \; \longrightarrow \; \mathbb{N}^r,$$

such that

$$R(D(c)) = c \quad \forall c \in \Gamma \quad \text{and} \quad \|D(c)\|_1 = \min\{\|\mathbf{x}\|_1 \mid \mathbf{x} \in R^{-1}(c)\}.$$

In general, map D is not unique; see Figure 3.1. This happens when the set $R^{-1}(c)$ contains two or more elements with minimum ℓ_1-norm for some $c \in \Gamma$.

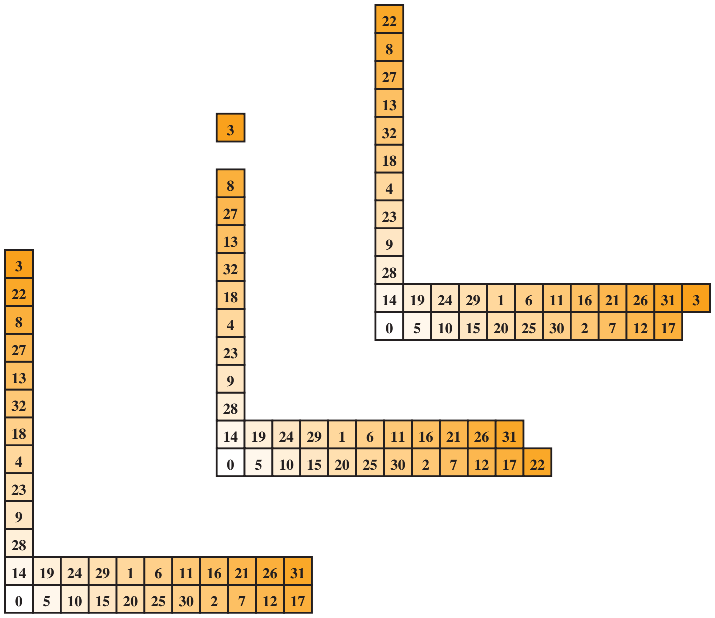


FIG. 3.1. *Different MDDs for* $C_{33}(5,14)$.

In digraphs of degree two, we can characterize this situation in terms of lattices. Let $\bar{R}$ be the extended map of R from $\mathbb{N}^r$ to $\mathbb{Z}^r$, and $\mathcal{L}$ the kernel of $\bar{R}$.

PROPOSITION 3.1. *Let D be an MDD for $(\Gamma, \{s_1, s_2\})$, where Γ is finite and abelian. Then there is a different MDD for the same graph if and only if there exists a vector $(T, -T) \in \mathcal{L}$ with $T > 0$ and $T \leq \max\{a_1, a_2\}$ for some $\mathbf{a} = (a_1, a_2) \in D(\Gamma)$.*

In the example $C_{33}(5, 14)$ from Figure 1.2, the associated lattice is generated by $\{(-16, 1), (-1, -2)\}$:

$$(T, -T) = \alpha(-16, 1) + \beta(-1, -2) \in \mathcal{L} \iff \alpha = \frac{-T}{11},\ \beta = \frac{5T}{11} \in \mathbb{Z} \iff T \in (11).$$

In consequence, this graph admits exactly four MDDs: the L-shape one given in the introduction and the three shown in Figure 3.1. However, only two of them have an “L” shape. These correspond with the only two graded monomial orderings in $\mathbb{K}[X, Y]$.

In accordance with the previous discussion, a well-ordering in $\mathbb{N}^r$ compatible with the norm ℓ_1 determines a unique MDD. Then, fixing a graded monomial ordering $\prec$, the obtained MDD is

$$\begin{array}{cccc} D: & \Gamma & \longrightarrow & \mathbb{N}^r \\ & c & \mapsto & \min(R^{-1}(c)). \end{array} \tag{3.2}$$

For each graded monomial ordering we can associate the bijective map $p : \mathbb{N} \longrightarrow \mathbb{N}^r$, such that $n < m \Rightarrow p(n) \prec p(m)$, that is, satisfying

$$p(i) = \min\left(\mathbb{N}^r \backslash \{p(j) \mid j < i\}\right).$$

This map provides a method of constructing the MDD with respect to a fixed monomial ordering. The procedure visits (through p) the elements in $\mathbb{N}^r$ corresponding with vertices (elements in Γ) until all of them are completed.

Algorithm 3.1: MDD construction.

```
Input: $\Gamma = \{c_i \mid 0 \leq i < N\}$, abelian group, $\{s_1, \ldots, s_r\}$, generating set; $s$.
Output: $D(c_i),\ i = 0, \ldots, N-1$.
$D[c_0, \ldots, c_{N-1}] := \bar{\emptyset},\ S := 0,\ \mathbf{a} := 0$;
while $S < N$ do
    $c := R(\mathbf{a})$;
    if $D(c) = \emptyset$ then
        $D(c) := \mathbf{a}$;
        $S := S + 1$;
    end
    $\mathbf{a} := s(\mathbf{a})$;
end
```

We include in the MDD building method's input the mapping s, such that

$$\begin{array}{rccc} s: & \mathbb{N}^r & \longrightarrow & \mathbb{N}^r \\ & \mathbf{a} & \mapsto & p(p^{-1}(\mathbf{a}) + 1), \end{array}$$

$$\mathbf{a} \prec s(\mathbf{a}),\ (\mathbf{a} \prec \mathbf{b} \Rightarrow s(\mathbf{a}) \preceq \mathbf{b}).$$

Of course, computing the whole diagram $D[0, \ldots, N-1]$ of a circulant cannot be computationally efficient, its size being exponential in the input size. Furthermore, Algorithm 3.1 performs an exhaustive search that can last at most for $\binom{d+r}{d}$ loops until reaching its ending, where d is the graph's diameter. When $r \ll d$, that bound is approximately $\frac{1}{r!}d^r$. The examples in Figure 3.2 illustrate the algorithm's output.

Definition 3.2. *Let Γ be a finite abelian group and (Γ, S) an associated connected digraph. Let $\prec$ be a graded monomial ordering. The monomial ideal*

$$I_S := (\mathbb{N}^r \backslash D(\Gamma))$$

is the ideal associated with the graph (Γ, S) and the monomial ordering $\prec$.

In the examples of Figure 3.2 we have two monomial ideals (J_1 and J_2) associated with $C_{104}(1, 5, 31)$ and with graded lex $x \prec y \prec z$ and $x \prec z \prec y$, respectively:

$$J_1 = (x^5, xy^6, y^7, y^3z^3, z^4, xy^2z^3) = (x^5, y^2, z^4) \cap (x^5, y^6, z^3) \cap (x, y^7, z^3) \cap (x, y^3, z^4),$$

$$J_2 = (x^5, y^4, y^3z^3, z^7, xy^2z^3) = (x^5, y^2, z^7) \cap (x^5, y^4, z^3) \cap (x, y^3, z^7).$$

Proposition 3.3. *With the above notation, we have that $\mathbb{N}^r \backslash D(\Gamma)$ is an ideal of the semigroup $\mathbb{N}^r$.*

Proof. Let $\mathbf{a}$ be an element in the ideal generated by $\mathbb{N}^r \backslash D(\Gamma)$. Then $\exists \mathbf{b} \in \mathbb{N}^r$, $\exists \mathbf{z} \in \mathbb{N}^r \backslash D(\Gamma)$ such that $\mathbf{a} = \mathbf{b} + \mathbf{z}$. Now, $\mathbf{z} \notin D(\Gamma)$. Then $\exists \mathbf{u} \in \mathbb{N}^r$, with $R(\mathbf{u}) = R(\mathbf{z})$, $\mathbf{u} \prec \mathbf{z}$. Since $\mathbf{u} + \mathbf{b} \prec \mathbf{z} + \mathbf{b}$ and R is a linear map, $R(\mathbf{u} + \mathbf{b}) = R(\mathbf{a})$ and $\mathbf{a} \notin D(\Gamma)$. □

Obviously, D is an injective map and $\#\big(D(\Gamma)\big) = \#\Gamma < \infty$. So, the monomial ideal I_S always contains generators of the form $X_1^{a_1}, \ldots, X_r^{a_r}$; that is, the quotient ring $\mathbb{K}[X_1 \ldots, X_r]/I_S$ is artinian (see Theorem 2.2). We say that an MDD built from a graded monomial ordering is *degenerated* if I_S is an irreducible ideal, that is, when the minimal system of generators of I_S contains only as many generators as the cardinal of S. In general, it is not the case as illustrated in the above examples. The paper [32] constructed MDDs in L-shape from circulant digraphs of degree two (i.e., $r = 2$). The following concept is the generalization of L-shapes to arbitrary dimension.

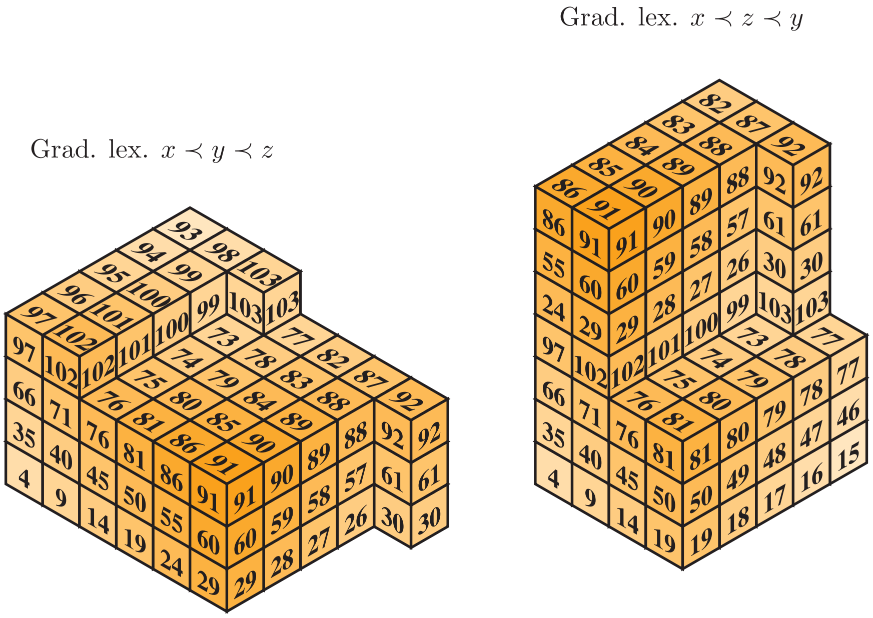

Grad. lex. $x \prec z \prec y$
Grad. lex. $x \prec y \prec z$

Fig. 3.2. *MDD of* $C_{104}(1, 5, 31)$.

Definition 3.4. *Let I be a monomial ideal and let A be the minimal system of generators of I. We say that I is an L-shape if there exists at most one element $\mathbf{x}^{\mathbf{a}} = X_1^{a_1} \cdots X_r^{a_r} \in A$ such that $a_i > 0$ $\forall i = 1, \ldots, r$.*

We say that an MDD built following Algorithm 3.1 *is an L-shape if the associated monomial ideal is an L-shape.*

In the examples of Figure 3.2 the generator involving every variable is xy^2z^3. We will prove that any MDD built with Algorithm 3.1 is an L-shape. First we need the following technical result.

Lemma 3.5. *Let A be the minimal system of generators of I_S. If the exponent of $\mathbf{x}^{\mathbf{a}} \in A$ has some component a_i positive, then $\mathbf{b} = (b_1, \ldots, b_r) := D(R(\mathbf{a}))$ satisfies $b_i = 0$.*

Proof. Since $\mathbf{a}$ is an element of A, then $\mathbf{a} \notin D(\Gamma)$. We must have $\mathbf{a} - \mathbf{e}_i \in D(\Gamma)$, because otherwise $\mathbf{a}$ would not be a minimal generator. Now, $\mathbf{b} \prec \mathbf{a}$ and $R(\mathbf{b}) = R(\mathbf{a})$. If we suppose $b_i > 0$, then

$$R(\mathbf{b} - \mathbf{e}_i) = R(\mathbf{a} - \mathbf{e}_i),\ \mathbf{b} - \mathbf{e}_i \prec \mathbf{a} - \mathbf{e}_i,$$

which contradicts $\mathbf{a} - \mathbf{e}_i \in D(\Gamma)$. □

Now, we state the main result in this section.

Proposition 3.6. *The output of Algorithm* 3.1 *is an L-shape.*

Proof. Let A be the minimal system of generators of I_S. If $\mathbf{a} \in A$ is such that $a_i > 0$ $\forall i$, then by Lemma 3.5 we have $R(\mathbf{a}) = R(\mathbf{0}) = 0_\Gamma$.

Moreover, $\mathbf{a} - \mathbf{e}_1 \in D(\Gamma)$, and $R(\mathbf{a} - \mathbf{e}_1) = -s_1$. So, if $\mathbf{a} \in A$ and $\mathbf{b} \in A$ are two generators with every component positive, then $\mathbf{a} - \mathbf{e}_1 = D(-s_1) = \mathbf{b} - \mathbf{e}_1$. That completes the proof. □

Now, the problem is to find the list of generators describing the ideal associated to a circulant digraph in a convenient way. The following section answers this question.

4. Lattice ideals and L-shapes. In this section we study the initial ideal of the lattice defined by the kernel of the extended routing map $\bar{R}$ and the monomial ideal associated to a circulant graph.

An integral lattice $\mathcal{L}$ of $\mathbb{Z}^r$ is the set of integer linear combinations of some integral vectors; in other words, an integral lattice is a $\mathbb{Z}$-submodule of $\mathbb{Z}^r$. This object has been used to solve many problems in mathematics and computer science (see, for instance, [4, 16, 24, 26]).

For any integral lattice $\mathcal{L} \subset \mathbb{Z}^r$ there is an associated binomial ideal (see [10, 31]):

$$I_{\mathcal{L}} := (\mathbf{x}^{\mathbf{a}^+} - \mathbf{x}^{\mathbf{a}^-} \mid \mathbf{a} \in \mathcal{L}) \subseteq \mathbb{K}[X_1, \ldots, X_r],$$

where $\mathbf{a}^+$ and $\mathbf{a}^-$ are the *positive* and *negative parts* of vector $\mathbf{a}$, that is, the unique vectors with no negative component and such that $\mathbf{a} = \mathbf{a}^+ - \mathbf{a}^-$.

The interesting articles [10, 31] study the combinatorics, geometry, and complexity of Gröbner bases for the ideals $I_{\mathcal{L}}$. In particular, they show that

$$\mathbf{x}^{\mathbf{a}} - \mathbf{x}^{\mathbf{b}} \in I_{\mathcal{L}} \iff \mathbf{a} - \mathbf{b} \in \mathcal{L}. \tag{4.1}$$

Given a Cayley digraph $(\Gamma, S = \{s_1, \ldots, s_r\})$ associated to a finite abelian group, we can extend the routing map R defined in (3.1) from $\mathbb{N}^r$ to $\mathbb{Z}^r$:

$$\begin{array}{cccc} \bar{R}: & \mathbb{Z}^r & \longrightarrow & \Gamma \\ & \mathbf{a} & \mapsto & a_1 s_1 + \cdots + a_r s_r. \end{array}$$

The kernel $\mathcal{L}_S$ of the map $\bar{R}$, i.e.,

$$\mathcal{L}_S := \{(a_1, \ldots, a_r) \in \mathbb{Z}^r \mid a_1 s_1 + \cdots + a_r s_r = 0_\Gamma\},$$

is the lattice associated to $(\Gamma, \{s_1, \ldots, s_r\})$.

Given an integral lattice $\mathcal{L}$ and a monomial ordering $\prec$, for every nonzero binomial $\mathbf{x}^{\mathbf{a}} - \mathbf{x}^{\mathbf{b}} \in I_{\mathcal{L}}$, the *leading* or *initial monomial* with respect to $\prec$ is given by

$$LM(\mathbf{x}^{\mathbf{a}} - \mathbf{x}^{\mathbf{b}}) := \begin{cases} \mathbf{x}^{\mathbf{a}} & \text{if } \mathbf{x}^{\mathbf{a}} \succ \mathbf{x}^{\mathbf{b}}, \\ \mathbf{x}^{\mathbf{b}} & \text{otherwise.} \end{cases}$$

As usual, given a polynomial ideal J in $\mathbb{K}[X_1, \ldots, X_r]$ we denote by $LM(J)$ the monomial ideal generated by the leading monomials of all nonzero elements of J, that is,

$$LM(J) := (LM(f) \mid f \in J^*).$$

The following is one of the main results in this section.

PROPOSITION 4.1. *For every graded monomial ordering $\prec$, we have that*

$$LM(I_{\mathcal{L}_S}) = I_S.$$

Proof. The ideal $I_{\mathcal{L}_S}$ is generated by binomials of the form $\mathbf{x}^{\mathbf{a}} - \mathbf{x}^{\mathbf{b}}$. Then it has a Gröbner basis G also consisting of that kind of binomial. Let $\mathbf{x}^{\mathbf{a}}$ be a monomial in $LM(I_{\mathcal{L}_S})$. There exists a binomial $\mathbf{x}^{a} - \mathbf{x}^{b}$ in the basis G, and by (4.1), $\mathbf{a} - \mathbf{b} \in \mathcal{L}_S$. Now, since $\mathbf{a} \succ \mathbf{b}$ and both paths have the same image by R, then $\mathbf{a} \notin D(\mathbb{Z}_N)$. Conversely, let $\mathbf{x}^{\mathbf{a}} \in I_S$. We take $\mathbf{b} := D(R(\mathbf{a})) \prec \mathbf{a}$. It is clear that $\mathbf{a} - \mathbf{b} \in \mathcal{L}_S$, and so $\mathbf{x}^{\mathbf{a}} - \mathbf{x}^{\mathbf{b}}$ is a binomial in $I_{\mathcal{L}_S}$, whose leading monomial is $\mathbf{x}^{\mathbf{a}}$. □

Gröbner bases were introduced by Buchberger in his thesis [6] and their use has become widespread in commutative algebra and algebraic geometry. The theory of Gröbner bases is related to several areas in mathematics and computer science; see, for instance, [2, 17, 30]. As a consequence of the previous result we have that if G is a minimal or reduced Gröbner basis of the ideal $I_{\mathcal{L}_S}$, then the leading monomials of the elements of G constitute a minimal system of generators of our MDD. In order to apply Buchberger's algorithm for computing a finite Gröbner basis of an ideal, we need to start with a finite set of generators. In this sense, we must point out that every generating set of binomials for $I_{\mathcal{L}_S}$ corresponds to a generating set of $\mathcal{L}_S$ (see (4.1)), but the converse is not true. Lemma 2.1 of [31] provides a sufficient condition for this converse result to be true.

Proposition 4.2. *Let $C_N(j_1, \ldots, j_r)$ be a connected circulant graph with associated lattice $\mathcal{L}$. We have $I_{\mathcal{L}} = (X_1^N X_2^N \cdots X_r^N - 1, \quad \mathbf{x}^{\mathbf{a}^+} - \mathbf{x}^{\mathbf{a}^-} \mid \mathbf{a} \in U)$, where*

$$U := \{(N\alpha_1, \ldots, N\alpha_r), (\alpha_1 j_1 - 1, \alpha_2 j_2, \ldots, \alpha_r j_r), (\alpha_1 j_1, \alpha_2 j_2 - 1, \ldots, \alpha_r j_r), \ldots, \\ (\alpha_1 j_1, \ldots, \alpha_{r-1} j_{r-1}, \alpha_r j_r - 1)\},$$

and $\beta, \alpha_i \in \mathbb{Z}, (i = 1, \ldots, r)$, satisfying $1 = \alpha_1 j_1 + \cdots + \alpha_r j_r + \beta N$.

Proof. The proof follows from [31, Lemma 2.1] and a simple linear algebra exercise. ∎

Using Propositions 4.1 and 4.2 we can compute a minimal system of generators of I_S for circulant digraphs. The paper [31] also contains results on the complexity of computing the reduced Gröbner basis of lattice ideals and on its size. In particular, it provides an upper bound for the number of elements and shows an example lattice $\mathcal{L}$ with exponential size in the bit complexity of a basis of $\mathcal{L}$. Nevertheless, we must cite program 4*ti*2, which is extremely efficient in computing the reduced Gröbner basis of binomials ideals. That software is available at http://www.4ti2.de; see [18].

5. Optimal routing. In this section we show an algorithm for computing a shortest path between two vertices for any Cayley digraph with a finite abelian group as vertex set using a finite Gröbner basis of $I_{\mathcal{L}_S}$.

Message routing is a basic function in communication networks. The problem is to find a route along which messages should be sent. The routing algorithm dictates token passing strategies in communication networks.

Given a pair of nodes (t, s) in a graph, there are several paths which join the origin t and the destination s. We are interested in *optimal paths*, i.e., those with minimum length. For general graphs, finding a shortest path between two vertices is a well-known and important problem. Efficient polynomial time algorithms have been developed for various routing problems. However, for the family of circulant graphs, there is an important distinction to be made, and that concerns the natural input size to a problem. For an arbitrary graph it is common to consider the input size to be $O(N^2)$, which is the number of bits in its adjacency matrix. However, any circulant graph can be described by only r integers. In this representation the input size is $O(r \log N)$. Thus, polynomial time algorithms for general graphs may exhibit exponential complexity in the special case of circulant graphs for this compact input representation. In [7] it is shown that the shortest path problem is NP-hard for this concise representation. The paper [15] presents very efficient algorithms for computing a shortest path for circulants with two jumps.

As we have already pointed out, in our case the routing problem is reduced to pairs of nodes $(0_\Gamma, j)$ where the starting point is fixed. Using the well-known extended Euclidean algorithm we compute a path $\mathbf{c}$ from vertex 0_Γ to vertex j if it exists.

We can apply the general integer programming techniques (see [29]) to find a shortest path for circulant digraphs as follows.

LEMMA 5.1. *Any shortest path from* 0 *to* j *in* $C_N(j_1, j_2, \ldots, j_r)$ *is a solution to the following integer program:* $\min\{\mathbf{d} \cdot \mathbf{x} | A\mathbf{x} \geq \mathbf{b}, \mathbf{x} \in \mathbb{Z}^{r+1}\}$*, where* $\mathbf{x} = (x_1, x_2, \ldots, x_r, y) \in \mathbb{Z}^{r+1}$*,* $\mathbf{d} = (1, 1, \ldots, 1, 0) \in \mathbb{Z}^{r+1}$*,* $\mathbf{b} = (j, -j, 0, \ldots, 0) \in \mathbb{Z}^{r+2}$*, and*

$$A = \begin{pmatrix} j_1 & j_2 & \ldots & j_r & N \\ -j_1 & -j_2 & \ldots & -j_r & -N \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix} \in \mathbb{Z}^{(r+2)\times(r+1)},$$

and conversely.

So, with the number of jumps r fixed, we can derive an algorithm to compute a shortest path in circulant digraphs requiring $O(r + \log r \log N)$ arithmetic operations on rational numbers of size $O(\log N)$; see [11, 15, 22, 23].

The main result of this section is the following.

PROPOSITION 5.2. *Let* G *be a Gröbner basis of the ideal* $I_{\mathcal{L}_S}$ *with respect to any graded monomial ordering* $\prec$*, and let* $\mathbf{c}$ *be a path (not necessarily a shortest one) in* $R^{-1}(j)$*. Then the normal form of* $\mathbf{x^c} - 1$ *with respect to* G *is* $\mathbf{x^d} - 1$*, where* $\mathbf{d}$ *is the shortest path from vertex* 0_Γ *to vertex* j *with respect to the monomial ordering* $\prec$*.*

Proof. We have $\mathbf{c} - \mathbf{d} \in I_{\mathcal{L}}$, which implies $(\mathbf{x^c} - 1) - (\mathbf{x^d} - 1) \in I_{\mathcal{L}_S}$. Clearly, $\mathbf{x^d} - 1$ is a normal form, because $\mathbf{x^d} \notin I_S$. □

This result provides a convenient algorithm to compute a shortest path and then to design optimal routings.

6. An algorithm of MDD for triple-loop computer networks. In this section we provide an algorithm specifically tailored for computing the minimal system of generators for a triple-loop computer network, which requires $O(s \log N)$ arithmetic operations, where s is the number of generators of the minimal system.

The case of degree two circulants is very simple. We always have two generators of the form x^a, y^b, and there are two possibilities: there is one other generator $x^c y^d$ $(c < a,\ d < b)$ or those two are the only generators (irreducible ideal case). We can obtain this representation in an efficient way, for instance, using the algorithm in [8].

We present Algorithm 6.1 to compute the minimal generators of the ideal I_S associated to a circulant digraph of degree three. Once we have fixed a graded monomial ordering, we need as an intermediate step a procedure to decide, given a path $\mathbf{b}$, whether or not it lies in the MDD. For $\mathbf{b} \in \mathbb{N}^3$, we define the Boolean function $P(\mathbf{b})$ to be the truth value of $D(R(\mathbf{b})) = \mathbf{b}$.

Algorithm 6.1 works by computing, one by one, every generator in the ideal's minimal system. For each generator we use one or two binary searches. So, its complexity is $O(s\ \log N)$ steps, where s is the number of generators. In the worst case, an upper bound for s is $2N + 1$; see [31]. In practice, most of the time consumed in each step is used calling up the boolean function P, which will be proved to be computable in polynomial time.

PROPOSITION 6.1. *Algorithm* 6.1 *is correct.*

Proof. By Theorem 2.2, among the generators of I_S are monomials of the form x^a, y^b, and z^c. These are computed in lines 2–14 (part I). Lines 15–44 (part II) find every

Algorithm 6.1: MDD description. The three jumps case. (I)

```
   Input: $j_1, j_2, j_3, N \in \mathbb{N},\ \gcd(j_1, j_2, j_3, N) = 1,\ P$.
   Output: $\mathbf{a}_1, \ldots, \mathbf{a}_s \in \mathbb{N}^3 \mid (\mathbf{x}^{\mathbf{a}_1}, \ldots, \mathbf{x}^{\mathbf{a}_s}) = (\mathbb{N}^3 \backslash D(\mathbb{Z}_N))$; $a_i \not\leq a_j$ if $i \neq j$.
1  $k := 1$;
2  for $i = 1, 2, 3$ do
3      $m := 0,\ M := N.$;
4      while $M - m > 1$ do
5          $l := \left\lfloor \frac{m+M}{2} \right\rfloor$;
6          if $P(l\mathbf{e}_i)$ then
7              $m := l$;
8          else
9              $M := l$;
10         end
11     end
12     $\mathbf{a}_k := M\mathbf{e}_i$;
13     $k := k + 1$;
14 end
```

generator involving two variables, and lines 45–54 (part III) work for the (possibly missing) generator with all three variables.

The key fact is that if $(a, 0, 0)$ is one of the generators we are looking for in the first part, then for any $l \in \mathbb{N}$, $P(l, 0, 0) \iff l < a$. We can perform a binary search to obtain the three generators.

In the second part, we start with generators involving the first two variables, continue with the one without the y, and so on. For instance, for the first case, we look at the generator $(0, a, 0)$ found in the previous step. Then if $(q, *, 0)$ is the generator with lowest first component involving the first two variables, we can use $P(l, a-1, 0) \iff l < q$ to find q by a binary search. Once this is done, we fix the generator's second component $*$, aided by $P(q, l, 0) \iff l < *$. In a similar way, we continue to discover all the generators in this form.

Finally, there is only one generator possibly missing, which must satisfy $R(\mathbf{b}) = 0$. So, steps 45–47 find a candidate. This possible generator is checked for possible irredundancy in the remaining lines. □

To finish the method, we need a way to decide $P(\mathbf{b})$. In fact, we can use integer programming to solve the problem of finding a shortest path; see Lemma 5.1.

However, we need to find the minimum element according to the ordering $\prec$. We can follow Algorithm 6.2, which takes as input a matrix A to represent the monomial ordering (see [2]) in this way:

$$\mathbf{x} \prec \mathbf{y} \iff A\mathbf{x} \underset{\text{lex}}{<} A\mathbf{y}.$$

We represent the matrix rows with subindices: $A_1, \ldots, A_m$. Then we obtain Algorithm 6.2.

Proposition 6.2. *Algorithm* 6.2 *is correct.*

Proof. Steps 1–6 are clear. The only trouble arises when the vector that we get as result of the integer programming–type search $\mathbf{c}$ has the same ℓ_1-norm as $\mathbf{b}$, Algorithm 6.1: MDD description. The three jumps case. (II)

```
for $i = \{(1,2),(1,3),(2,3)\}$ do
    $T := a_{i[2]}[i[2]] - 1$;
    $Q := 0$;
    repeat
        $m := Q,\ M := a_{i[1]}[i[1]]$;
        while $M - m > 1$ do
            $l := \left\lfloor \frac{m+M}{2} \right\rfloor$;
            if $P(l\mathbf{e}_{i[1]} + T\mathbf{e}_{i[2]})$ then
                $m := l$;
            else
                $M := l$;
            end
        end
        $Q := M$;
        if $Q < a_{i[1]}[i[1]]$ then
            $m := 0,\ M := T$;
            while $M - m > 1$ do
                $l := \left\lfloor \frac{m+M}{2} \right\rfloor$;
                if $P(Q\mathbf{e}_{i[1]} + l\mathbf{e}_{j[2]})$ then
                    $m := l$;
                else
                    $M := l$;
                end
            end
            $\mathbf{a}_k := Q\mathbf{e}_{i[1]} + l\mathbf{e}_{i[2]}$;
            $k := k + 1$;
            $T := l - 1$;
        end
    until $Q = a_{i[1]}[i[1]]$ ;
end
```

Algorithm 6.1: MDD description. The three jumps case. (III)

```
$c := N - j_1 \bmod N$;
$\mathbf{b} := D(c)$;
$b[1] := b[1] + 1$;
for $i = 1, \ldots, k - 1$ do
    if $\mathbf{a}_i \leq \mathbf{b}$ then
        $k := k - 1$;
        STOP;
    end
    $\mathbf{a}_k := \mathbf{b}$;
end
```

Algorithm 6.2: Deciding if a given path lies in an MDD.

```
   Input: $j_1, \dots, j_r, N \in \mathbb{N}$, $\gcd(j_1, \dots, j_r, N) = 1$, $A \in \mathbb{R}^{m \times r}$, $\mathbf{b} \in \mathbb{N}^r$.
   Output: Boolean value $P(\mathbf{b}) := (\mathbf{b} = D(R(\mathbf{b})))$.
1  Execute an integer programming–type algorithm to get $\mathbf{c}$, an element with
   minimum $\ell_1$-norm in $R^{-1}(R(\mathbf{b}))$;
2  if $\|\mathbf{c}\|_1 < \|\mathbf{b}\|_1$ then
3      OUTPUT false;
4  else
5      if $\mathbf{c} \prec \mathbf{b}$ then
6          OUTPUT false;
7      else
8          Compute a basis for the lattice
           $\mathcal{L} := \{\mathbf{c} \in \mathbb{N}^r \mid \ < \mathbf{c}, (j_1, \dots, j_r) >= 0, \ < \mathbf{c}, (1, \dots, 1) >= 0\}$;
9          for $i = 1, \dots, m$ do
10             Set $* = (< A_i, \mathbf{b} > -(\min \ A)/2)$;
11             Set the boolean value $\alpha$, depending on whether there is a point in
               the set $(\mathbf{b} + \mathcal{L}) \cap \mathbb{N}^r \cap \{\mathbf{c} \in \mathbb{N}^r \mid \ < \mathbf{c}, A_1 >=< \mathbf{b}, A_1 >, \dots, <$
               $\mathbf{c}, A_{i-1} >=< \mathbf{b}, A_{i-1} >, < \mathbf{c}, A_i > \leq *\}$;
12             if $\alpha$ then
13                 OUTPUT false;
14             end
15         end
16         OUTPUT true;
17     end
18 end
```

and $\mathbf{b} \preceq \mathbf{c}$. In this case, we have to decide whether there is another vector $\mathbf{d} \in \mathbb{N}^r$, satisfying

$$\|\mathbf{d}\|_1 = \|\mathbf{b}\|_1 = \|\mathbf{c}\|_1, \ \mathbf{d} \prec \mathbf{b}.$$

Obviously, if such a vector $\mathbf{d}$ does exist, it lies in the set $(\mathbf{b} + \mathcal{L}) \cap \mathbb{N}^r$. So, we check in steps 9–14 if there is another path $\mathbf{c}$ such that $A\mathbf{c} \underset{\text{lex}}{<} A\mathbf{b}$. □

7. Diameter and average minimum distance. Two notable parameters in a digraph are the diameter and the average minimum distance. The former represents the worst delay in the communication between two nodes, and the latter represents the average delay. In this section we show formulae to compute those parameters in a circulant digraph given by the irredundant irreducible decomposition of the monomial ideal I_S.

7.1. Diameter. Given an MDD of a digraph (Γ, S), it is easy to obtain the diameter

$$d = \max\{\|\mathbf{a}\|_1 \mid \mathbf{a} \in D(\Gamma)\}.$$

The description of the monomial ideal I_S in terms of its irreducible components permits a simplification.

Proposition 7.1. *Let $\mathfrak{m}^{\mathbf{b}_1} \cap \cdots \cap \mathfrak{m}^{\mathbf{b}_n}$ be the irredundant irreducible decomposition of the ideal I_S. Then*

$$d = \max\{\|\mathbf{b}_i\|_1 - r \mid i = 1, \dots, r\}.$$

Proof. If we define the *corners* of I_S as

$$E(D) := \{\mathbf{a} \in D(\Gamma) \mid \mathbf{a} + \mathbf{e}_i \notin D(\Gamma)\ \forall i = 1, \dots, r\},$$

then it is clear that $d = \max\{\|\mathbf{a}\|_1 \mid \mathbf{a} \in E(D)\}$. We will prove that $\{\mathbf{a} + \mathbf{1} \mid \mathbf{a} \in E(D)\} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$.

Let $i \in \{1, \dots, n\}$. By Theorem 2.2, we have $\mathbf{b}_i \geq \mathbf{1}$. Let us check that $\mathbf{a} := \mathbf{b}_i - \mathbf{1} \in E(D)$. If $\mathbf{x}^{\mathbf{a}} \in I_S$, we would have $\mathbf{x}^{\mathbf{a}} \in \mathfrak{m}^{\mathbf{b}_i} \Rightarrow \exists j \in \{1, \dots, r\} \mid a_j \geq b_{ij} = a_j + 1$. So, $\mathbf{x}^{\mathbf{a}} \notin I_S$. Further, if $\exists j \in \{1, \dots, r\}$ such that $\mathbf{x}^{\mathbf{a}+\mathbf{e}_j} \notin I_S$, then $\exists k \in \{1, \dots, n\}$, $k \neq i \mid \mathbf{x}^{\mathbf{a}+\mathbf{e}_j} \notin \mathfrak{m}^{\mathbf{b}_k} \Rightarrow \mathbf{a} + \mathbf{e}_j \sqsubset \mathbf{b}_k \Rightarrow \mathbf{b}_i \leq \mathbf{b}_k \Rightarrow \mathfrak{m}^{\mathbf{b}_k} \subseteq \mathfrak{m}^{\mathbf{b}_i}$. So, $\mathbf{x}^{\mathbf{a}+\mathbf{e}_j} \in I_S$ and $\mathbf{a} \in E(D)$.

On the other hand, let $\mathbf{a} \in E(D)$. First we will see that $I_S \subseteq \mathfrak{m}^{\mathbf{a}+\mathbf{1}}$. Suppose that $\mathbf{x}^{\mathbf{u}} \in I_S \backslash \mathfrak{m}^{\mathbf{a}+\mathbf{1}}$. Then $\mathbf{a} + \mathbf{1} > \mathbf{u} \Rightarrow \mathbf{a} \geq \mathbf{u}$. Since $\mathbf{x}^{\mathbf{u}} \in I_S$, then $\mathbf{x}^{\mathbf{a}} \in I_S$; this is a contradiction because $\mathbf{a} \in E(D)$. If $\mathfrak{m}^{\mathbf{a}+\mathbf{1}}$ were not an irreducible component in the decomposition of I_S, it would be satisfied:

$$\begin{aligned} &\exists j \in \{1, \dots, n\} \mid \mathfrak{m}^{\mathbf{b}_j} \subsetneq \mathfrak{m}^{\mathbf{a}+\mathbf{1}} \Rightarrow \left\{ \begin{array}{c} \mathbf{a} + \mathbf{1} \leq \mathbf{b}_j \\ \mathbf{a} + \mathbf{1} \neq \mathbf{b}_j \end{array} \right\} \\ &\quad \Rightarrow \exists i \in \{1, \dots, r\} \mid \mathbf{x}^{\mathbf{a}+\mathbf{e}_i} \in D(\Gamma). \qquad \square \end{aligned}$$

7.2. Average minimum distance. Again, given an MDD of a digraph, it is easy to obtain the average minimum distance. Let N be the number of nodes.

$$\bar{d} = \frac{\sum_{c \in \Gamma} \|D(c)\|_1}{N} = \frac{\sum_{\mathbf{x}^{\mathbf{u}} \notin I_S} \|\mathbf{u}\|_1}{N}.$$

The following result provides a formula for computing $\bar{d}$ in digraphs with a degenerated MDD.

Lemma 7.2. *Let* $I_S = \mathfrak{m}^{\mathbf{a}+\mathbf{1}} = \mathfrak{m}^{\mathbf{b}}$. *Then*

$$\sum_{\mathbf{x}^{\mathbf{u}} \notin I_S} \|\mathbf{u}\|_1 = \frac{b_1 \cdots b_r}{2}(b_1 + \cdots + b_r - r) = \frac{a_1 + \cdots + a_r}{2} \prod_{i=1}^{r} (a_i + 1).$$

Proof. By Proposition 7.1, we have

$$d = \|\mathbf{b}\|_1 - r = \|\mathbf{a}\|_1.$$

On the other hand,

$$\mathbf{u} = (u_1, \dots, u_r) \in D(\Gamma) \iff \forall i \in \{1, \dots, r\},\ u_i < b_i.$$

We define the following relation in $D(\Gamma)$: $(u_1, \dots, u_r) \equiv (a_1 - u_1, \dots, a_r - u_r)$. So, every equivalence class contains two elements (whose degrees add up to $\|\mathbf{a}\|_1$) or only one. This last happens if and only if $\forall i \in \{1, \dots, r\}$, $u_i = a_i - u_i \Rightarrow 2\|\mathbf{u}\|_1 = \|\mathbf{a}\|_1$. We can state the following:

$$\sum_{\mathbf{x}^{\mathbf{u}} \notin I} \|\mathbf{u}\|_1 = \frac{N}{2} d,$$

and the proof is complete. $\square$

Note 7.3. In the above case, that is, when I_S is an irreducible ideal, we have $\bar{d} = d/2$.

To discuss the general case we introduce some new notation. Let $\mathfrak{m}^{\mathbf{b}_1} \cap \cdots \cap \mathfrak{m}^{\mathbf{b}_n}$ be the irreducible decomposition of the monomial ideal I_S; so we define

$$\mathbf{d}_\Delta := \text{exponent}\big(\gcd(\mathbf{x}^{\mathbf{b}_i} \mid i \in \Delta)\big) \ \forall \Delta \subseteq \{1, \ldots, n\}, \ \Delta \neq \emptyset.$$

$$\sigma(\mathbf{u}) := \frac{u_1 \cdots u_r}{2}(u_1 + \cdots + u_r - r).$$

Our next goal is to find a formula for the average minimum distance. We will apply the general inclusion-exclusion principle as follows.

PROPOSITION 7.4. *Let $\mathfrak{m}^{\mathbf{b}_1} \cap \cdots \cap \mathfrak{m}^{\mathbf{b}_n}$ be the irreducible decomposition of the ideal I_S. We have*

$$\sum_{\mathbf{x}^{\mathbf{u}} \notin I_S} \|\mathbf{u}\|_1 = \sum_{\emptyset \subsetneq \Delta \subseteq \{1,\ldots,n\}} (-1)^{\#\Delta+1} \sigma(\mathbf{d}_\Delta).$$

Proof. Applying Lemma 7.2, we obtain

$$\sum_{\emptyset \subsetneq \Delta \subseteq \{1,\ldots,n\}} (-1)^{\#\Delta+1} \sigma(\mathbf{d}_\Delta) = \sum_{\Delta} (-1)^{\#\Delta+1} \sum_{\mathbf{x}^{\mathbf{u}} \notin \mathfrak{m}^{\mathbf{b}_i} \ \forall i \in \Delta} \|\mathbf{u}\|_1.$$

If $\mathbf{x}^{\mathbf{u}} \notin I_S$, that is, if $\exists i \in \{1, \ldots, n\} \mid \mathbf{x}^{\mathbf{u}} \notin \mathfrak{m}^{\mathbf{b}_i}$, then the above sum includes $\|\mathbf{u}\|_1$ exactly once, as seen in the following equation, where $j = \#\{j \in 1, \ldots, n\} \mid \mathbf{x}^{\mathbf{u}} \notin \mathfrak{m}^{\mathbf{b}_i}$:

$$\binom{j}{1} - \binom{j}{2} + \cdots + (-1)^{j+1}\binom{j}{j} = 1.$$

This completes the proof. $\square$

Considering the ideal $I_1 = (x^4, x^2y^2, y^3)$ from Example 2.1, the sum of the degrees of the monomials outside this ideal is (see Figure 7.1)

$$\sum_{\mathbf{x}^{\mathbf{u}} \notin I_1} \|\mathbf{u}\|_1 = \sigma(2,3) + \sigma(4,2) - \sigma(2,2) = 9 + 16 - 4 = 21.$$

The several results introduced in section 2 permit a strong reduction in the number of sum terms we need to consider in the expression of Proposition 7.4. For instance, if we consider Example 2.3, Proposition 7.4 solves (see Figure 7.2)

$$\begin{aligned}
\sum_{\mathbf{x}^{\mathbf{u}} \notin I} \|\mathbf{u}\|_1 &= \sigma(8,2,1) + \sigma(7,2,4) + \sigma(3,3,5) + \sigma(4,3,2) + \sigma(4,5,1) \\
&\quad - [\sigma(7,2,1) + \sigma(3,2,1) + \sigma(4,2,1) + \sigma(4,2,1) + \sigma(3,2,4) \\
&\quad + \sigma(4,2,2) + \sigma(4,2,1) + \sigma(3,3,2) + \sigma(3,3,1) + \sigma(4,3,1)] \\
&\quad + \sigma(3,2,1) + \sigma(4,2,1) + \sigma(4,2,1) + \sigma(3,2,1) + \sigma(3,2,1) \\
&\quad + \sigma(4,2,1) + \sigma(3,2,2) + \sigma(3,2,1) + \sigma(4,2,1) + \sigma(3,3,1) \\
&\quad - [\sigma(3,2,1) + \sigma(3,2,1) + \sigma(4,2,1) + \sigma(3,2,1) + \sigma(3,2,1)] + \sigma(3,2,1) \\
&= \sigma(8,2,1) + \sigma(7,2,4) + \sigma(3,3,5) + \sigma(4,3,2) + \sigma(4,5,1) \\
&\quad - [\sigma(7,2,1) + \sigma(3,2,4) + \sigma(4,2,2) + \sigma(3,3,2) + \sigma(4,3,1)] \\
&\quad + s(3,2,2) = 454.
\end{aligned}$$

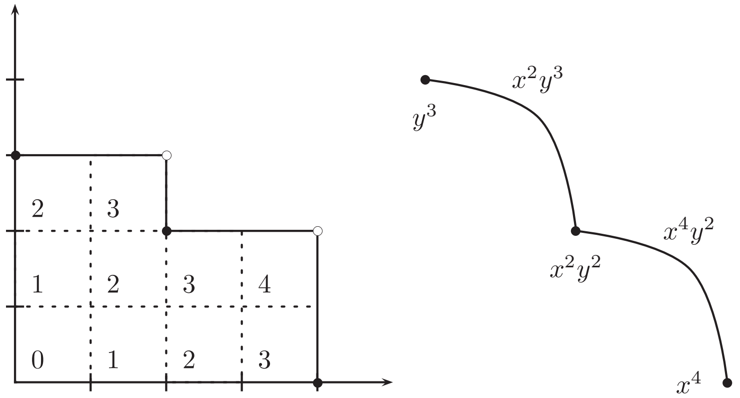


FIG. 7.1.

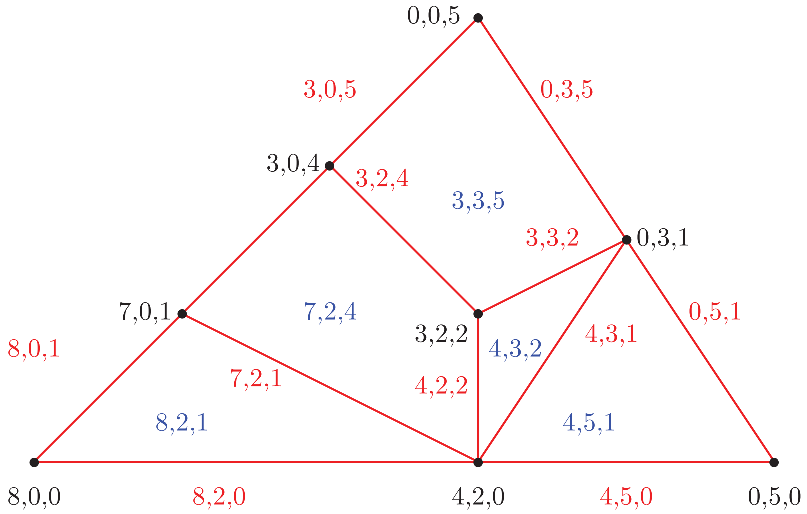


FIG. 7.2.

Clearly, if $\mathbf{b} \in \mathbb{N}^r$ has a zero coordinate, then $\sigma(\mathbf{b}) = 0$. This fact produces several cancellations in the formula of Proposition 7.4. We end up with a sum of the simplex labels, affected with the sign: + for faces, − for edges, and + for nodes.

In Cayley digraphs of degree two the associated monomial ideal has only one or two irreducible components (see Proposition 3.6). Then the computation of the average minimum distance is immediate. For digraphs of degree three we can follow this strategy:

- Construct the Miller–Sturmfels graph G as in the previous examples such that each irreducible component corresponds with the least common multiple of some generators of the minimal system.
- Let E be the set of all edges, F the set of faces, and N the set of vertices of G:

$$\bar{d} = \frac{1}{N}\left(\sum_{e\in F}\sigma(e) - \sum_{e\in E}\sigma(e) + \sum_{e\in N}\sigma(e)\right).$$

8. Degenerated L-shapes. We recall that an MDD is degenerated if the associated monomial ideal is irreducible, that is, of the form $\left(X_1^{\alpha_1}, \ldots, X_r^{\alpha_r}\right)$. In general, the family of graphs having this property does not have optimal properties according to the ratio nodes/diameter. In this section we present families of circulant digraphs having a degenerated MDD and with a relatively small diameter.

PROPOSITION 8.1. *Let a, s, k be natural numbers such that* $\gcd(a, s) = 1$ *and* $a < s$. *The monomial ideal associated with $C_{sk}(a, s)$ is $I_S = (x^s, y^k)$ for any monomial ordering.*

Proof. Since $\mathbb{K}[x, y]/I_S$ is an artinian ring (see Theorem 2.2), the minimal system of generators of I_S contains monomials of the form x^α, y^β. We claim that $\beta = k$. In order to prove it, we note that $D(si) = (0, i)$ $\forall i = 0, \ldots, k-1$. In fact, let $i \in \{0, \ldots, k-1\}$ and suppose that $\exists (u, v) \in \mathbb{N}^2$, $\mid u + v \leq i$ with $R(u, v) = si$. Then

$$si \equiv_{sk} au + sv \Rightarrow \exists h \in \mathbb{N} \ / \ si = au + sv + hsk \Rightarrow s|au \Rightarrow s|u$$

$$\Rightarrow \begin{cases} u = 0 \\ \vee \\ \exists t \in \mathbb{N}^* \ / \ u = st. \end{cases}$$

In the first case, we have

$$i = v + kh \leq k - 1 \Rightarrow h = 0 \Rightarrow v = i.$$

In the second,

$$i = at + v + kh \leq k - 1 \Rightarrow h = 0, \qquad i = at + v \geq u + v \Rightarrow at \geq u,$$

but this a contradiction because $a < s$. So, $\beta \geq k$. On the other hand, $D(0, k) = 0 = D(0, 0)$ implies $\beta = k$. Finally, suppose that (see Figure 8.1)

$$I_S = (x^\alpha, x^\gamma y^\delta, y^k), \quad \gamma < \alpha, \ \delta < k, \quad R(\gamma, \delta) = R(0, k) = 0.$$

Thus,

$$R(\gamma, k) = R(\gamma, \delta) + R(0, k - \delta) = R(0, k - \delta),$$

$$R(\gamma, k) = R(\gamma, 0) + R(0, k) = R(\gamma, 0).$$

Therefore, one of the two vectors $(0, k - \delta)$, $(\gamma, 0)$ should be in I_S, but this is false.

Consequently, I_S is degenerated and

$$sk = \dim\left(\mathbb{K}[x,y]/(x^\alpha, y^k)\right) = \alpha k \Rightarrow s = \alpha. \qquad \square$$

The following example shows that we cannot omit from the above result hypotheses the requirement $a < s$.

Example 8.2. The monomial ideal I_S associated with $C_{60}(7, 6)$ and any graded monomial ordering in $\mathbb{K}[x, y]$ is not degenerated: $I_S = (x^{12}, x^6 y^3, y^7)$.

Using the Gröbner bases theory and previous results we can generalize Proposition 8.1 from two jumps to an arbitrary number of them.

PROPOSITION 8.3. *Let $\alpha_1, \ldots, \alpha_r$ be positive integers, neither of them equal to one. Setting $N := \alpha_1 \cdots \alpha_r$, the circulant digraph $\mathcal{C}_N(1, \alpha_1, \alpha_1\alpha_2, \ldots, \alpha_1 \cdots \alpha_{r-1})$*

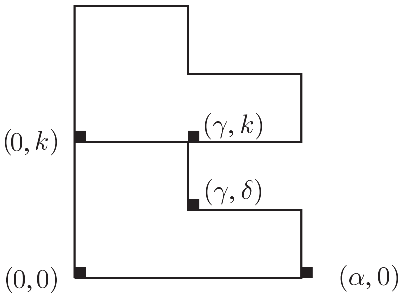


FIG. 8.1.

is associated to the—incidentally, irreducible—monomial ideal $(X_1^{\alpha_1},\dots,X_r^{\alpha_r})$*, with any graded monomial ordering. The Gröbner basis of the associated binomial ideal is* $\{X_i^{\alpha_i}-X_{i+1} \mid\ i=1,\dots,r-1\}\cup\{X_r^{\alpha_r}-1\}$*.*

Proof. First of all, every element of the proposed basis lies in the binomial ideal. This is because their associated lattice points,

$$\{(0,\dots,\overset{i}{\smile}{\alpha_i},-1,\dots,0)\mid i=1,\dots,r-1\}\cup\{(0,\dots,0,\alpha_r)\},$$

are paths for node 0. Then the initial ideal of this lattice ideal must contain the following one:

$$(X_1^{\alpha_1},\dots,X_r^{\alpha_r})\subseteq I_S.$$

We know that the dimension of the quotient vector space $\mathbb{K}[\mathbf{x}]/I_S$ equals the number of nodes $N=\alpha_1\cdots\alpha_r$. Moreover, the dimension of $\mathbb{K}[\mathbf{x}]/(X_i^{\alpha_i}\mid i=1,\dots,r)$ is N, so both ideals must coincide. In order to obtain a reduced Gröbner basis, we must have one binomial for each generator in the initial ideal. That is, the reduced Gröbner basis is

$$\{X_i^{\alpha_i}-m_i(\mathbf{x})\mid i=1,\dots,r\},$$

where m_i is a monomial satisfying $m_i\notin(X_1^{\alpha_1},\dots,X_r^{\alpha_r})$. Then $m_i=\mathbf{x}^{\mathbf{a}}$, with $a_i<\alpha_i, i=1,\dots,r$. Set $X_{r+1}:=1$. Then $(X_i^{\alpha_i}-X_{i+1})-(X_i^{\alpha_i}-m_i)=m_i-X_{i+1}$ belongs to the ideal. If $m_i\neq X_{i+1}$, we would have $\alpha_{i+1}=1$, which is a contradiction. □

The following result is an immediate consequence.

COROLLARY 8.4. *Let* d,r *be two positive integers. Let* k *be the residue class of* d *modulo* r*. Then, if we fix*

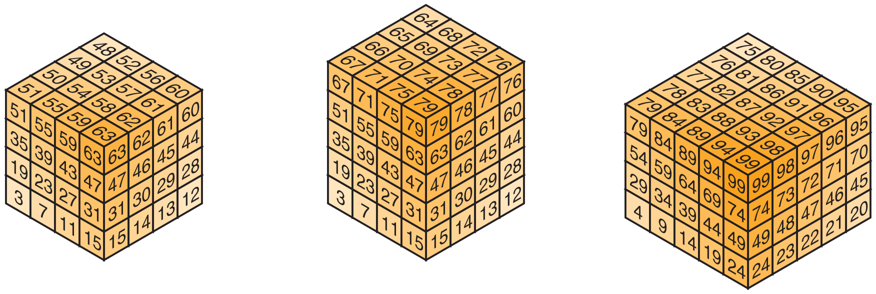
$$\alpha_1=\cdots=\alpha_k=\frac{d-k}{r}+2,\quad \alpha_{k+1}=\cdots=\alpha_r=\frac{d-k}{r}+1,$$

the following is a directed circulant graph with r *jumps,* $N:=\alpha_1\cdots\alpha_r$ *nodes, and diameter* d*:*

$$\mathcal{C}_N(1,\alpha_1,\alpha_1\alpha_2,\dots,\alpha_1\cdots\alpha_{r-1}).$$

We note that the number of vertices is

$$N=\left(\frac{d-k}{r}+2\right)^k\left(\frac{d-k}{r}+1\right)^{r-k}.$$



d=9=4+4+4- 3

d =10=4+4+5-3

d =11=5+5+4-3

FIG. 8.2. *Family of circulant digraphs.*

Once r is fixed, increasing the diameter d makes the number of nodes in this graph family increase as $O(d^r)$.

Proposition 8.1 provides a family with diameter $2\sqrt{N}-2$ and average minimum distance $\sqrt{N}-1$. Let $d>1$ be a natural number:

$$C_{\left(\frac{d+2}{2}\right)^2}\left(1,\frac{d+2}{2}\right) \text{ if } d\equiv 0 \bmod 2 \quad \text{and} \quad C_{\frac{(d+1)(d+3)}{4}}\left(1,\frac{d+1}{2}\right) \text{ if } d\equiv 1 \bmod 2.$$

Basically, this family was discovered in the paper [32]. However, determining $d_2(N)$ and finding the optimal $C_N(j_1,j_2)$ is an open problem.

In the case of undirected circulant graphs of degree four, i.e., $C_N(j_1,-j_1,j_2,-j_2)$, several papers have shown that the lower bound $\frac{1}{2}\left(\sqrt{2N-1}-1\right)$ can be achieved by taking $j_1=\frac{1}{2}\left(\sqrt{2N-1}-1\right)$ and $j_2=\frac{1}{2}\left(\sqrt{2N-1}-1\right)+1$; see the survey [3]. In the middle, that is, between circulant digraphs of degree two and circulant graphs of degree four, Proposition 8.3 and the above corollary provide a very attractive family of circulant graph of degree three; see Figure 8.2. Let $d>2$ be a natural number:

$$C_{\left(\frac{d+3}{3}\right)^3}\left(1,\frac{d+3}{3},\left(\frac{d+3}{3}\right)^2\right) \text{ if } d\equiv 0 \bmod 3,$$

$$C_{\frac{(d+2)^2(d+5)}{27}}\left(1,\frac{d+2}{3},\left(\frac{d+2}{3}\right)^2\right) \text{ if } d\equiv 1 \bmod 3,$$

$$C_{\frac{(d+4)^2(d+1)}{27}}\left(1,\frac{d+4}{3},\left(\frac{d+4}{3}\right)^2\right) \text{ if } d\equiv 2 \bmod 3.$$

Graphs in this family have diameter d and average minimum distance $d/2$.

9. Conclusions.

In this paper we have proposed monomial ideals as a natural tool for studying Cayley digraphs with a finite abelian group as vertex set. We have generalized the L-shape concept in the plane to L-shape in the r-dimensional affine space. We think that this new point of view may shed light on problems in multiloop computer networks. We also have introduced the Gröbner bases theory in this context, which seems very useful. Many interesting questions remain unsolved. We would like to provide fault tolerance routing algorithms. From a more practical point of view, it would be interesting to investigate the implementation in computer networks of the family of circulant graphs of degree three under parameters such as routing, fault tolerance, etc.

Acknowledgment. We would like to thank Prof. B. Sturmfels, who gave us the ideal of a lattice point of view.

REFERENCES

[1] F. Aguiló and M. A. Fiol, *An efficient algorithm to find optimal double loop networks*, Discrete Math., 138 (1995), pp. 15–29.

[2] T. Beker and V. Weispfenning, *Gröbner Bases. A Computational Approach to Commutative Algebra*, Grad. Texts in Math. 141, Springer-Verlag, New York, 1993.

[3] J.-C. Bermond, F. Comellas, and D. F. Hsu, *Distributed loop computer networks: A survey*, J. Parallel Distrib. Comput., 24 (1995), pp. 2–10.

[4] S. R. Blackburn, D. Gomez-Perez, J. Gutierrez, and I. E. Shparlinski, *Predicting nonlinear pseudorandom number generators*, Math. Comp., 74 (2005), pp. 1471–1494.

[5] F. T. Boesch and R. Tindell, *Circulants and their connectivity*, J. Graph Theory, 8 (1984), pp. 487–499.

[6] B. Buchberger, *Ein Algorithmus zum Auffinden der Basiselemente des Restklassenringes nach einem nulldimensionalen Polynomideal*, Ph.D. thesis, University of Innsbruck, Austria, 1965.

[7] J.-Y. Cai, G. Havas, B. Mans, A. Nerurkar, J.-P. Seifert, and I. Shparlinski, *On routing in circulant graphs*, in Computing and Combinatorics (Tokyo, 1999), Lecture Notes in Comput. Sci. 1627, Springer-Verlag, Berlin, 1999, pp. 360–369.

[8] Y. Chen and F. K. Hwang, *Diameters of weighted double-loop networks*, J. Algorithms, 9 (1988), pp. 401–410.

[9] D. Z. Du, D. F. Hsu, and F. K. Hwang, *Double-linked ring networks*, IEEE Trans. Comput., 34 (1985), pp. 853–877.

[10] D. Eisenbud and B. Sturmfels, *Binomial ideals*, Duke Math. J., 84 (1996), pp. 1–45.

[11] F. Eisenbrand, *Fast integer programming in fixed dimension*, in Algorithms—ESA 2003, Lecture Notes in Comput. Sci. 2832, Springer-Verlag, Berlin, 2003, pp. 196–207.

[12] P. Erdös and D. F. Hsu, *Distributed loop networks with minimum transmission delay*, Theoret. Comput. Sci., 100 (1992), pp. 223–241.

[13] M. Fiol, J. L. Yebra, I. Alegre, and M. Valero, *A discrete optimization problem in local networks and data alignment*, IEEE Trans. Comput., C-36 (1987), pp. 702–713.

[14] D. Gomez, J. Gutierrez, and A. Ibeas, *Circulant digraphs and monomial ideals*, in Computer Algebra in Scientific Computing, Lecture Notes in Comput. Sci. 3718, Springer-Verlag, Berlin, 2005, pp. 196–207.

[15] D. Gomez, J. Gutierrez, and A. Ibeas, *Optimal routing in double loop networks*, Theoret. Comput. Sci., 381 (2007), pp. 68–85.

[16] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag, Berlin, 1993.

[17] J. Gutierrez and R. Rubio, *Reduced Groebner bases under composition*, J. Symbolic Comput., 26 (1999), pp. 433–444.

[18] R. Hemmecke, R. Hemmecke, and P. Malkin, *4ti2 Version 1.2—Computation of Hilbert Bases, Graver Bases, Toric Gröbner Bases, and More*, 2005; available online from www.4ti2.de.

[19] D. F. Hsu and X.-D. Jia, *Extremal problems in the construction of distributed loop networks*, SIAM J. Discrete Math, 7 (1994), pp. 57–71.

[20] F. K. Hwang, *A complementary survey on double-loop networks*, Theoret. Comput. Sci., 263 (2001), pp. 211–229.

[21] F. K. Hwang, *A survey on multi-loop networks*, Theoret. Comput. Sci., 299 (2003), pp. 107–121.

[22] R. Kannan, *Minkoswski's convex body theorem and integer programing*, Math. Oper. Res., 12 (1987), pp. 415–440.

[23] H. W. Lenstra, *Integer programming with a fixed number of variables*, Math. Oper. Res., 8 (1983), pp. 538–548.

[24] A. K. Lenstra, H. W. Lenstra, and L. Lovász, *Factoring polynomials with rational coefficients*, Math. Ann., 261 (1982), pp. 515–534.

[25] B. Mans, *Optimal distributed algorithms in unlabeled tori and chordal rings*, J. Parallel Distrib. Comput., 46 (1997), pp. 80–90.

[26] D. Micciancio and S. Goldwasser, *Complexity of Lattice Problems*, Kluwer Internat. Ser. Engrg. Comput. Sci. 671, Kluwer Academic, Boston, MA, 2002.

[27] E. Miller, *Resolutions and Duality for Monomial Ideals*, Ph.D. Thesis, University of California, Berkeley, 2000.

[28] E. Miller and B. Sturmfels, *Monomial ideal and planar graphs*, in Proceedings of AAECC-13, Applied Algebra, Algebraic Algorithms and Error-Correcting Codes (Honolulu, HI, 1999), Lecture Notes in Comput. Sci. 1719, Springer-Verlag, Berlin, 1999, pp. 19–28.

[29] A. Schrijver, *Theory of Linear and Integer Programming*, Wiley-Intersci. Ser. Discrete Math, Wiley-Interscience, Chichester, UK, 1986.

[30] B. Sturmfels, *Gröbner Bases and Convex Polytopes*, Univ. Lecture Ser. 8, AMS, Providence, RI, 1996.

[31] B. Sturmfels, R. Weismantel, and G. M. Ziegler, *Gröbner bases of lattices, corner polyhedra, and integer programming*, Beiträge Algebra Geom., 36 (1995), pp. 281–298.

[32] C. K. Wong and D. Coppersmith, *A combinatorial problem related to multimodule memory organizations*, J. ACM, 21 (1974), pp. 392–402.

[33] J. Žerovnik and T. Pisanski, *Computing the diameter in multiple-loop networks*, J. Algorithms, 14 (1993), pp. 226–243.

SIAM J. Discrete Math.
Vol. 21, No. 3, pp. 785–793

CONSTRUCTIONS OF OPTICAL ORTHOGONAL CODES FROM FINITE GEOMETRY*

T. L. ALDERSON† and KEITH E. MELLINGER‡

Abstract. The link between finite geometry and various classes of error-correcting codes is well known. Arcs in projective spaces, for instance, have a close tie to linear MDS codes as well as the high-performing low-density parity-check codes. In this article, we demonstrate a connection between arcs and optical orthogonal codes (OOCs), a class of nonlinear binary codes used for many modern communication applications. Using arcs and Baer subspaces of finite projective spaces, we construct some infinite classes of OOCs with auto-correlation and cross-correlation both larger than 1.

Key words. optical orthogonal codes, arcs, Baer subplanes

AMS subject classifications. 94B27, 51E20

DOI. 10.1137/050632257

1. Introduction.

An $(n, w, \lambda_a, \lambda_c)$-optical orthogonal code (OOC) is a family of binary sequences (codewords) of length n, with constant hamming weight w satisfying the following two conditions:

- (auto-correlation property) for any codeword $c = (c_0, c_1, \ldots, c_{n-1})$ and for any integer $1 \le t \le n-1$, there holds $\sum_{i=0}^{n-1} c_i c_{i+t} \le \lambda_a$,
- (cross-correlation property) for any two distinct codewords c, c' and for any integer $0 \le t \le n-1$, there holds $\sum_{i=0}^{n-1} c_i c'_{i+t} \le \lambda_c$,

where each subscript is reduced modulo n.

One of the first proposed applications of OOCs was to optical code-division multiple access communication systems where binary sequences with strong correlation properties are required [1, 3, 5]. Subsequently, OOCs have found application for multimedia transmissions in fiber-optic LANs [9]. OOCs have also been called cyclically permutable constant weight codes in the construction of protocol sequences for multiuser collision channels without feedback [11].

An $(n, w, \lambda_a, \lambda_c)$-OOC with $\lambda_a = \lambda_c$ is denoted (n, w, λ)-OOC. The number of codewords is the size of the code. For fixed values of n, w, and λ, the largest size of an (n, w, λ)-OOC is denoted $\Phi(n, w, \lambda)$. An (n, w, λ)-OOC of size $\Phi(n, w, \lambda)$ is said to be *optimal*. From the Johnson bound for constant weight codes it follows that [3]

$$\Phi(n, w, \lambda) \le \left\lfloor \frac{1}{w} \left\lfloor \frac{n-1}{w-1} \left\lfloor \frac{n-2}{w-2} \left\lfloor \cdots \left\lfloor \frac{n-\lambda}{w-\lambda} \right\rfloor \right\rfloor \cdots \right\rfloor \right\rfloor \right\rfloor . \qquad (1.1)$$

Much of the literature is restricted to (n, w, λ)-OOCs. If C is an $(n, w, \lambda_a, \lambda_c)$-OOC with $\lambda_a \ne \lambda_c$, then we obtain a (perhaps naive) bound on the size of C by taking $\lambda = \max\{\lambda_a, \lambda_c\}$ in (1.1). In [17], Yang and Fuja discuss OOCs with $\lambda_a > \lambda_c$

*Received by the editors May 24, 2005; accepted for publication (in revised form) April 9, 2007; published electronically September 28, 2007.
http://www.siam.org/journals/sidma/21-3/63225.html

†Department of Mathematical Sciences, University of New Brunswick, Saint John, NB, Canada, E2L 4L5 (tim@unbsj.ca). This author's research was supported by the NSERC of Canada.

‡Department of Mathematics, University of Mary Washington, 1301 College Avenue, Trinkle Hall, Fredericksburg, VA 22401 (kmelling@umw.edu). This author's research was supported by a faculty development grant from the University of Mary Washington.

and the following bound is established:

$$\Phi(n, w, \lambda + m, \lambda) \leq \Phi(n, w, \lambda) \cdot (\lambda + m). \tag{1.2}$$

The codes we construct in sections 4.1, 4.2, and 5 have $\lambda_a < \lambda_c$. As such, (1.1) seems the only applicable bound. We do, however, offer some analysis regarding the possible optimality of our codes.

Let F be an infinite family of OOCs with $\lambda_a = \lambda_c$. For any (n, w, λ)-OOC $C \in F$ containing at least one codeword, the number of codewords in C is denoted by $M(n, w, \lambda)$ and the corresponding Johnson bound is denoted by $J(n, w, \lambda)$. F is called asymptotically optimal if

$$\lim_{n\to\infty} \frac{M(n, w, \lambda)}{J(n, w, \lambda)} = 1. \tag{1.3}$$

For $\lambda = 1, 2$, optimal OOCs are known to exist (see, e.g., [3, 4, 13]). There are very limited examples of such optimal OOCs with $\lambda > 2$ (in [12, 13] optimal OOCs consisting of a single codeword are shown to exist). Our constructions were originally motivated by the results in [10] where certain families of conics in $PG(2, q)$ are used to construct $(n, q+1, 2)$-OOCs that are close to optimal. We build on the ideas in [10] and construct several new classes of OOCs based on arcs in finite projective spaces.

2. Preliminaries.

As our work relies heavily on the structure of finite projective spaces, we start with a short overview of the fundamentals of finite projective geometry. We let $PG(k, q)$ represent the finite projective geometry of dimension k and order q. Due to a result of Veblen and Young [16], all finite projective spaces of dimension greater than two are isomorphic up to the order q. The space $PG(k, q)$ can be modeled most easily with the vector space of dimension $k + 1$ over the finite field $GF(q)$. In this model, the one-dimensional subspaces represent the points, two-dimensional subspaces represent lines, etc. Using this model, it is not hard to show by elementary counting that the number of points of $PG(k, q)$ is given by $\theta_{k,q} = \frac{q^{k+1}-1}{q-1}$.

The fundamental theorem of projective geometry states that the full automorphism group of $PG(k, q)$ is the group $P\Gamma L(k + 1, q)$ of semilinear transformations acting on the underlying vector space. The subgroup $PGL(k+1, q) \cong GL(k+1, q)/Z_0$ (where Z_0 represents the center of the group $GL(k+1, q)$) of projective linear transformations is easily modeled by matrices and will be useful in our constructions. Another property that we rely on is the principle of duality. For any space $S = PG(k, q)$, there is a *dual* space S^* whose points and hyperplanes (subspaces of dimension $k - 1$) are, respectively, the hyperplanes and points of S. For any result about points of S, there is always a corresponding result about hyperplanes of S^*. More generally, for any result dealing with subspaces of S, replacing each reference to a subspace $PG(m, q)$, $m < k$, with a reference to the subspace $PG(k - m - 1, q)$ yields a corresponding *dual* statement of S^* that has the same truth value. For instance, a result about a set of points of $PG(k, q)$, no three of which are collinear, could be rewritten dually about a set of hyperplanes of $PG(k, q)$, no three of which meet in a common subspace of dimension $k - 2$.

3. OOCs from lines of $PG(k, q)$.

In [3], Chung, Salehi, and Wei provide a method for constructing $(n, w, 1)$-OOCs using lines of the projective geometry $PG(k, q)$. Briefly, let ω be a primitive element of $GF(q^{k+1})$. The points of $\Sigma = PG(k, q)$ can be represented as $\omega^0 = 1, \omega, \omega^2, \ldots, \omega^{n-1}$, where $n = \frac{q^{k+1}-1}{q-1}$. Hence, in a natural way a point set A of $PG(k,q)$ corresponds to binary n-tuple (or codeword) $(a_0, a_1, \ldots, a_{n-1})$, where $a_i = 1$ if and only if $\omega^i \in A$.

Denote by ϕ the collineation of Σ defined by $\omega^i \mapsto \omega^{i+1}$, a singer cycle acting on Σ. The map ϕ acts transitively on the points (and dually on the hyperplanes) of Σ. If A is a point set of Σ corresponding to the codeword $c = (a_0, a_1, \ldots, a_{n-1})$, then ϕ induces a cyclic shift on the entries of c.

For each line ℓ of Σ, consider the orbit $\mathcal{O}_\ell$ under ϕ. If $\mathcal{O}_\ell$ is a full orbit (has size n), then a representative line and corresponding codeword are chosen. Short orbits are discarded. Let $\mathcal{L}(k,q)$ represent the cardinality of this set of chosen lines. Two lines of Σ intersect in at most one point and each line contains $q+1$ points. It follows that the codewords satisfy both $\lambda_a \leq 1$ and $\lambda_c \leq 1$, and by counting the number of full orbits under ϕ, the following theorem is obtained.

THEOREM 3.1. *For any prime power q and any positive integer k, there exists an (optimal) $(\theta_{k,q}, q+1, 1)$-OOC consisting of $\mathcal{L}(k,q) = \left\lfloor \frac{q^k-1}{q^2-1} \right\rfloor$ codewords.*

4. OOCs from arcs in $PG(k,q)$.

An *n-arc* in $PG(k,q)$ is a collection of $n > k$ points such that no $k+1$ are incident with a common hyperplane. It follows that if $\mathcal{K}$ is an n-arc in $PG(k,q)$, then no $k+1$ points of $\mathcal{K}$ lie on a hyperplane, no k lie on a $(k-2)$-flat,..., and no 3 lie on a line. An n-arc is called *complete* if it is not contained in an $(n+1)$-arc.

For given k and q, let $m(k,q)$ denote the maximum value of n for which an n-arc exists in $PG(k,q)$. Then $m(k,q) = k+2$ for $q \leq k+1$. In homogeneous coordinates, the points (1, 0, ..., 0), (0, 1, 0, ..., 0), ...,(0, ..., 0, 1), and (1, 1, ..., 1) constitute such an arc. Hence, for $q \leq k+1$, every point in $PG(k,q)$ is a linear combination of at most k of these $k+2$ points. In $PG(2,q)$, a (nondegenerate) conic is a $(q+1)$-arc and elementary counting shows that this arc is complete when q is odd.

When q is even, one can add one additional point to each conic, the so-called *knot* where all of the tangent lines intersect. The resulting $(q+2)$-arc is called a hyperoval and is necessarily complete. Conics are a special case of the so-called normal rational curves. A *rational curve $\mathcal{C}_n$* of order n in $PG(d,q)$ is a set of points

$$\{P(t) = P\left(g_0(t_0,t_1), \ldots, g_d(t_0,t_1)\right) \,|\, t_0, t_1 \in GF(q)\},$$

where each g_i is a binary form of degree n and the highest common factor of $g_0, g_1, \ldots, g_d$ is 1. The curve $\mathcal{C}_n$ may also be written

$$\{P(t) = P\left(f_0(t), \ldots, f_d(t)\right) \,|\, t \in GF(q) \cup \{\infty\}\}, \tag{4.1}$$

where $f_i(t) = g_i(1,t)$.

DEFINITION 4.1. *A normal rational curve (NRC) in $PG(d,q)$, $2 \leq d \leq q-2$, is a rational curve (of order d) projectively equivalent to the set of points*

$$\{(1, t, \ldots, t^d) \,|\, t \in GF(q)\} \cup \{(0, \ldots, 0, 1)\}.$$

It is well known that an NRC is, in fact, a $(q+1)$-arc. When stated in terms of arcs, the *main conjecture* (MC) for MDS codes, always taking $q > k+1$, is $m(k,q) = q+2$ for $k = 2$ and $k = q-2$, both with q even, and $q+1$ in all other cases. The main conjecture has its roots in a problem first posed over 50 years ago by B. Segre. The MC has not been proved in general, but it has been verified in many cases. See [6] for a recent survey of results relating to the MC.

DEFINITION 4.2. *Let $\pi = PG(k,q)$. A t-family $\mathcal{F}$ of m-arcs in π is a collection of m-arcs mutually meeting in at most t points.*

Theorem 4.3. *Let $\mathcal{F}$ be a t-family of m-arcs in $\pi = PG(k,q)$. Let $\mu = \max\{k,t\}$. Then there exists a $(\frac{q^{k+2}-1}{q-1}, m, k, \mu)$-OOC C consisting of $|\mathcal{F}|$ codewords.*

Proof. Consider $\pi = PG(k,q)$ as embedded in $\Sigma = PG(k+1,q)$, and let ω be a primitive element of $GF(q^{k+2})$. Let C be t-family of m-arcs π. Identify each arc in C with the corresponding codeword of length $\frac{q^{k+2}-1}{q-1}$ and weight m. As in section 3, let $\phi : \omega^i \mapsto \omega^{i+1}$ be a singer group acting on Σ. Let $\mathcal{K}$ be an arc in C. The auto-correlation λ_a is the maximum number of points in the intersection of $\phi^i(\mathcal{K})$ and $\phi^j(\mathcal{K})$, where $i \neq j$. Since ϕ is a collineation of Σ, $\phi^i(\mathcal{K}) \cap \phi^j(\mathcal{K}) \subset \phi^i(\pi) \cap \phi^j(\pi)$. As ϕ acts regularly on the hyperplanes of Σ, $\phi^i(\pi) \neq \phi^j(\pi)$ and $\phi^i(\pi) \cap \phi^j(\pi)$ is necessarily a $(k-1)$-flat. It follows that λ_a is bounded above by the maximum intersection of an arc in $PG(k,q)$ and a $(k-1)$-flat; hence $\lambda_a \leq k$. Now let $\mathcal{K}$ and $\mathcal{K}'$ be distinct arcs in C. The cross-correlation λ_c is the maximum number of points in the intersection of $\phi^i(\mathcal{K})$ and $\phi^j(\mathcal{K}')$. If $i \neq j$, then, as above, this number is at most k. However, if $i = j$, then $\phi^i(\mathcal{K})$ and $\phi^j(\mathcal{K}')$ are in a common hyperplane of Σ and can therefore share as many as t points. It follows that $\lambda_c = \max\{k,t\}$. □

Using the notation of the previous proof, a line of Σ intersects any member of $\mathcal{F}$ in at most 2 points. Hence, adding the $\mathcal{L}(k,q)$ codewords from Theorem 3.1 to C will not violate either correlation requirement. However, each line gives a codeword of weight $q+1$, whereas the weight of C is m. This poses no problem if $m \leq q+1$. Moreover, if $m \leq q+1$ the points of each of the $\mathcal{L}(k,q)$ lines may be arbitrarily subdivided into $\lfloor \frac{q+1}{m} \rfloor$ disjoint subsets (or, more generally, into subsets mutually intersecting in at most k points) of size m. Each of the resulting $\lfloor \frac{q+1}{m} \rfloor \cdot \mathcal{L}(k,q)$ subsets then correspond to a codeword of C. This gives the following corollary.

Corollary 4.4. *Let $\mathcal{F}$ be a t-family of m-arcs, $m \leq q+1$ in $\pi = PG(k,q)$. Let $\mu = \max\{k,t\}$. Then there exists a $(\frac{q^{k+2}-1}{q-1}, m, k, \mu)$-OOC consisting of $|\mathcal{F}| + \lfloor \frac{q+1}{m} \rfloor \cdot \mathcal{L}(k,q)$ codewords.*

4.1. An $(n, w, \lambda, \lambda+2)$ construction using normal rational curves.

The following theorem is a well known property of NRCs (see, e.g., [15]).

Theorem 4.5. *A $(d+3)$-arc in $PG(d,q)$ is contained in a unique normal rational curve.*

If $\mathcal{C}$ is an NRC in $PG(d,q)$, then the subgroup of $PGL(d+1,q)$ leaving $\mathcal{C}$ fixed is (isomorphic to) $PGL(2,q)$ (see [7, Theorem 27.5.3]). It follows that if $\nu(d,q)$ denotes the number of distinct normal rational curves in $PG(d,q)$, then

$$\nu(d,q) = \frac{|PGL(d+1,q)|}{|PGL(2,q)|} = \frac{(q^{d+1}-1)(q^{d+1}-q)\cdots(q^{d+1}-q^d)}{(q^2-1)(q^2-q)}. \tag{4.2}$$

Theorem 4.6. *For any prime power q and for each $k \geq 2$ there exists a $(\frac{q^{k+2}-1}{q-1}, q+1, k, k+2)$-OOC consisting of $\nu(k,q) + \mathcal{L}(k,q) \approx q^{k^2+2k-3}$ codewords.*

Proof. This follows immediately from Corollary 4.4 and Theorem 4.5. □

Taking $k = 2$ in Theorem 4.6 gives the following corollary.

Corollary 4.7. *For any prime power q there exists a $(q^3+q^2+q+1, q+1, 2, 4)$-OOC consisting of $q^5 - q^2 + q$ codewords.*

Remark 4.7.1. 1. Let $M(\frac{q^{k+2}-1}{q-1}, q+1, k, k+2)$ denote the size of the codes constructed in Theorem 4.6. We compare the size of our codes to other codes with similar correlation parameters in order to obtain some insight on the optimality of our codes. On the one hand (as one might expect), we have

$$M\left(\frac{q^{k+2}-1}{q-1}, q+1, k, k+2\right) < J\left(\frac{q^{k+2}-1}{q-1}, q+1, k+2\right) \approx q^{k^2+2k-1},$$

while on the other hand,

$$M\left(\frac{q^{k+2}-1}{q-1}, q+1, k, k+2\right) > J\left(\frac{q^{k+2}-1}{q-1}, q+1, k+1\right) \approx q^{k^2+k-1}.$$

Also, from the bound of Yang and Fuja (1.2), it follows that if $q^k > \frac{k+2}{q^4}$, then

$$M\left(\frac{q^{k+2}-1}{q-1}, q+1, k, k+2\right) > \Phi\left(\frac{q^{k+2}-1}{q-1}, q+1, k+2, k+1\right).$$

Thus, we have a strong indication that the codes constructed in Theorem 4.6 are quite robust. Moreover, we see that for the code parameters specific to the theorem and for q sufficiently large

$$\Phi\left(n, w, \lambda+1, \lambda\right) < \Phi\left(n, w, \lambda-1, \lambda+1\right).$$

2. Let C be a $(\frac{q^{k+2}-1}{q-1}, q+1, k, k+2)$-OOC constructed as in the theorem and let $\Sigma = PG(k+1, q)$. Let c_1, $c_2 \in C$ be two codewords. By the construction, it follows that there is at most one cyclic shift of c_2, say, c_2', for which the cross-correlation of c_1 and c_2' is greater than k (this will only occur if the NRCs $\mathcal{C}_1$ and $\mathcal{C}_2$ corresponding to c_1 and c_2', respectively, are contained in a common hyperplane of Σ, and intersect in more than k points).

4.2. An (n, w, λ) construction from m-arcs. In [10], Miyamoto, Mizuno, and Shinohara prove the existence of an asymptotically optimal family of $(n, w, 2)$-OOCs. Their proof utilizes a clever construction of a large 2-family of $(q+1)$-arcs in $PG(2, q)$. The construction relies heavily on the fact that the $(q+1)$-arcs concerned are conics (i.e., NRCs). In what follows we provide a construction for large families of arcs in $PG(k, q)$, $k \geq 2$. For $k = 2$ the corresponding OOCs form an asymptotically optimal family. Also, for $k = 2$ our code parameters match those of [10] for q even (see Corollary 4.11). Our construction is quite general in that it holds for arbitrary arcs; that is to say we do not rely on any correspondence between the arcs involved and algebraic curves. As such, our construction holds for arcs of size larger than $q+1$ in $PG(d, q)$ (which do not necessarily correspond to NRCs).

THEOREM 4.8. *Let $\pi = PG(k, q)$. If π contains an m-arc, then π contains a $(k+1)$-family $\mathcal{F}$ of m-arcs where $|F| = q^{k+1} - q^k$. Moreover, there exists a point P incident with each member of $\mathcal{F}$. Consequently, there exists a k-family consisting of $q^{k+1} - q^k$ distinct $(m-1)$-arcs.*

Proof. We work with the dual. Let $\mathcal{K} = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ be a dual m-arc in π. Consider π as embedded in $\Sigma = PG(k+1, q)$ and let $\Sigma^* = \Sigma \backslash \pi$ be the associated affine space. Let σ be any hyperplane of Σ on λ_m other than π. For each point $P \in \Sigma^* \setminus \sigma$ denote by ϕ_P the projection map taking π to σ through P. Each such ϕ_P fixes λ_m and carries $\mathcal{K}$ to a dual arc $\phi_P(\mathcal{K})$ in σ (containing λ_m). Let $S = \{\phi_P(\mathcal{K}) | P \in \Sigma^* \setminus \sigma\}$ be the set of $q^{k+1} - q^k$ dual m-arcs in σ obtained by projection. *We claim that apart from λ_m no two dual arcs S share as many as $k+1$ common members.* Let $\lambda \neq \lambda_m$ be a member of $\phi_P(\mathcal{K})$ and let $\psi = \lambda \cap \lambda_m$. Other than λ_m there is precisely one member of $\mathcal{K}$, say, λ', containing ψ (at most two members of $\mathcal{K}$ are incident with a given $(k-2)$-flat). So $\langle P, \lambda\rangle = \langle P, \lambda'\rangle = \langle \lambda, \lambda'\rangle$. It follows that if $\lambda \in \phi_Q(\mathcal{K})$ with $P \neq Q$, then the line $\langle P, Q\rangle$ intersects π in a point of λ'. Hence, if $\phi_P(\mathcal{K})$ and $\phi_Q(\mathcal{K})$ have $k+1$ common members other than λ_m, then the point at which the line $\langle P, Q\rangle$ intersects π will be incident with $k+1$ members of $\mathcal{K}$, which is a contradiction. Hence, S is a $(k+1)$-family of m-arcs where $|S| = q^{k+1} - q^k$. By removing λ_m from each member of S we obtain the k-family of dual $(m-1)$-arcs as required. □

Restricting to $k = 2$ we can give explicit coordinates for constructing the $q^3 - q^2$ conics of the projective plane described in Theorem 4.8. These coordinates are derived directly from the projection construction when q is odd. Let (x, y, z) represent the homogeneous coordinates for a projective point of π. Then, for fixed $a, b, c \in GF(q)$, $c \neq 0$, let $C_{a,b,c} = \{(1, a - cx^2, b - cx) : x \in GF(q)\} \cup \{(0, 1, 0)\}$. One can easily show that $C_{a,b,c}$ defines a conic of π. Varying a, b, and c gives a family of $q^3 - q^2$ conics that have the desired intersection property and which meet in the point $(0, 1, 0)$. As $c \neq 0$, we have exactly $q^3 - q^2$ conics of this form. These coordinates generate a similar set when q is even.

Lemma 4.9. *If there exists an m-arc in $PG(k, q)$, then there exists a $(\theta_{k+1,q}, m-1, k)$-OOC C where*

$$|C| = \begin{cases} q^{k+1} - q^k + \lfloor \frac{q+1}{m-1} \rfloor \cdot \mathcal{L}(k+1, q), & m \leq q+2, \\ q^{k+1} - q^k & \textit{otherwise}. \end{cases}$$

Proof. The proof follows immediately from Theorem 4.8 and Corollary 4.4. □

Corollary 4.10. *For any prime power q and $k \geq 2$ there exists a $(\theta_{k+1,q}, q, k)$-OOC consisting of $q^{k+1} - q^k + \mathcal{L}(k, q)$ codewords.*

Proof. Normal rational curves provide $(q+1)$-arcs in $PG(k, q)$, $k \geq 2$. The result follows from Lemma 4.9. □

Thus, for each $k \geq 2$ we have (via Corollary 4.10) an infinite family of OOCs. Moreover, for $k = 2$ it is easily verified that the family is asymptotically optimal. When $k = 2$ and q is even the fact that hyperovals exist in $PG(2, q)$ gives the following corollary yielding codes with parameters matching those of Miyamoto, Mizuno, and Shinohara [10].

Corollary 4.11. *For $q = 2^t$ there exists a $(\frac{q^4-1}{q-1}, q+1, 2)$-OOC consisting of $q^3 - q^2 + q$ codewords.*

4.3. An (n, w, λ) construction from $(k+1)$-arcs in $PG(k, q)$.

As observed above, $(k+2)$-arcs exist in $PG(k, q)$ for every k. Denote by $\mathcal{N}(k+1, q)$ the family of all $(k+1)$-arcs in $PG(k, q)$. As $\mathcal{N}(k+1, q)$ is a k-family of arcs in $PG(k, q)$, Theorem 4.3 gives the following.

Theorem 4.12. *For any prime power q and $k \geq 1$ there exists a $(\theta_{k+1,q}, k+1, k)$-OOC consisting of $|\mathcal{N}(k+1, q)|$ codewords.*

Corollary 4.13. *For any prime power q and $1 \leq k \leq q$ there exists a $(\theta_{k+1,q}, k+1, k)$-OOC consisting of $|\mathcal{N}(k+1, q)| + \lfloor \frac{q+1}{k+1} \rfloor \cdot \mathcal{L}(k, q)$ codewords.*

Observe the Johnson bound:

$$J(\theta_{k+1,q}, k+1, k) = \frac{(\theta_{k+1,q} - 1)(\theta_{k+1,q} - 2) \cdots (\theta_{k+1,q} - k)}{(k+1)!} \approx \frac{q^{(k+1)k}}{(k+1)!}. \tag{4.3}$$

By counting ordered $(k+2)$-tuples $(P_1, P_2, \ldots, P_{k+1}, \mathcal{K})$, where $\mathcal{K}$ is a $(k+1)$-arc in $PG(k,q)$ and $P_1, P_2, \ldots, P_{k+1}$ are the points in $\mathcal{K}$, we get

$$|\mathcal{N}(k+1,q)| = \frac{\theta_{k,q}\,(\theta_{k,q}-1)\,(\theta_{k,q}-\theta_{1,q})\,(\theta_{k,q}-\theta_{2,q})\cdots(\theta_{k,q}-\theta_{k-1,q})}{(k+1)!} \approx \frac{q^{k(k+1)}}{(k+1)!}. \tag{4.4}$$

It follows that the family of codes constructed as in Theorem 4.12 and Corollary 4.13 are asymptotically optimal.

5. An $(n, w, \lambda, \lambda+1)$ construction from arcs in $PG(k, q^2)$. Since $GF(q)$ is a subfield of $GF(q^n)$ for $n > 1$, the projective space $PG(k,q)$ is naturally embedded in $PG(k,q^n)$ once the coordinate system is fixed. In particular, any $PG(k,q)$ embedded in $PG(k,q^2)$ is called a *Baer subspace* (BSS) of $PG(k,q^2)$ (for an introduction to Baer subspaces, see [2] or [14]). A *frame* of a k-dimensional projective space is a set of $k+2$ points of which any $k+1$ points are a basis, that is, a $(k+2)$-arc. It is well known that a BSS of $PG(k,q^2)$ is uniquely determined by a frame. Denote by $\mathcal{B}(k,q^2)$ the number of BSSs of $PG(k,q^2)$. Then by counting ordered $(k+2)$-tuples or otherwise (see, e.g., [14]), we have

$$\mathcal{B}(k,q^2) = q^{\frac{k(k+1)}{2}} \prod_{i=2}^{k+1} (q^{\,i}+1) \approx q^{k^2+2k}. \tag{5.1}$$

THEOREM 5.1. *If $\Pi = PG(k,q)$ contains a $(k+1)$-family $\mathcal{F}$ of m-arcs, then there exists a $(k+1)$-family S of m-arcs in $\Sigma = PG(k,q^2)$, where $|S| = \mathcal{B}(k,q^2)\cdot|\mathcal{F}|$.*

Proof. Let Π_i, $1 \le i \le \mathcal{B}(k,q^2)$, denote the BSSs of Σ. By assumption, for each j, $1 \le j \le \mathcal{B}(k,q^2)$, there exists a $(k+1)$-family $\mathcal{F}_j$ of m-arcs in Π_j with $|\mathcal{F}_j| = |\mathcal{F}|$. Let

$$S = \bigcup_{j=1}^{\mathcal{B}(k,q^2)} \mathcal{F}_j.$$

As a BSS is uniquely determined by a frame, two distinct BSSs cannot share a $(k+2)$-arc. It follows that S is a $(k+1)$-family of m-arcs with $|S| = \mathcal{B}(k,q^2)\cdot|\mathcal{F}|$. □

THEOREM 5.2. *In $PG(k,q)$ there exists a $(k+1)$-family $\mathcal{F}$ of q-arcs where*

$$|F| = q^{k-1}\cdot\prod_{i=1}^{k-1}(q^{k+1}-q^i).$$

Proof. Denote by X_P the number of NRCs through an arbitrary fixed point $P \in \Sigma = PG(k,q)$. By counting ordered pairs $(\mathcal{C}, Q)$, where $\mathcal{C}$ is an NRC in Σ and Q is a point of $\mathcal{C}$, we get

$$\nu(k,q)(q+1) = \left(\frac{q^{k+1}-1}{q-1}\right)\cdot X_P,$$

which gives

$$X_P = q^{k-1}\cdot\prod_{i=1}^{k-1}(q^{k+1}-q^i).$$

Hence, removing P from each of the X_P NRCs through P yields a $(k+1)$-family $\mathcal{F}$ of q-arcs. □

Corollary 5.3. *For $k > 1$ and $q > k$ a prime power, there exists a $(\frac{q^{k+2}-1}{q-1}, q, k, k+1)$-OOC consisting of*

$$\begin{aligned}&\mathcal{B}(k, q^2) \cdot X_P + \mathcal{L}(k, q^2) \cdot \mathcal{B}(1, q^2) \\ &\quad = q^{\frac{k^2+3k-2}{2}} \left(q^{k+1} + 1\right) \prod_{i=1}^{k-1} \left[\left(q^{k+1} - q^i\right) \left(q^{i+1} + 1\right)\right] \\ &\quad + \left\lfloor \frac{q^{2(k+1)} - 1}{q^4 - 1} \right\rfloor (q^3 + q^2) \approx q^{2k^2+3k-2}\end{aligned}$$

codewords.

Proof. Fix $k > 1$ and $q > k$ and let $\Sigma = PG(k+1, q^2)$. From Theorems 5.1 and 4.3 there exists a $(\frac{q^{k+2}-1}{q-1}, q, k, k+1)$-OOC C consisting of $\mathcal{B}(k, q^2) \cdot X_P$ codewords. Let ℓ be one of the $\mathcal{L}(k+1, q^2)$ lines in Σ with full orbit. As a frame uniquely determines a BSS, it follows that any two Baer sublines of ℓ intersect in at most two points. Thus, as in Corollary 4.4 we may add $\mathcal{L}(k, q^2) \cdot \mathcal{B}(1, q^2)$ codewords to C. This gives a code of size $\mathcal{B}(k, q^2) \cdot X_P + \mathcal{L}(k, q^2) \cdot \mathcal{B}(1, q^2)$. □

Remark 5.3.1. Let $M(n, w, k, k+1)$ be the size of the codes constructed as in the theorem. Note that $J(n, w, k+1) \approx q^{2k^2+3k}$ and $J(n, w, k) \approx q^{2k^2+2k}$, so though the codes constructed in the corollary are not asymptotically optimal with respect to the Johnson bound, they appear to be of a competitive size. We also point out that from (1.2) it follows that $M(n, w, k, k+1) > \Phi(n, w, k+1, k)$ for $k > 2$.

Corollary 5.4. *For any prime power q, there exists a $(\frac{q^8-1}{q^2-1}, q+1, 2, 3)$-OOC consisting of $\mathcal{B}(2, q^2) \cdot (q^4 - q^2) + \mathcal{L}(3, q^2) \cdot \mathcal{B}(1, q^2) = q^{12} + q^9 - q^8 + q^4$ codewords. For $q = 2^t$ there exists a $(\frac{q^8-1}{q^2-1}, q+2, 2, 3)$-OOC consisting of $\mathcal{B}(2, q^2) \cdot (q^4 - q^2) = q^{12} + q^9 - q^8 - q^5$ codewords (in this case, we cannot include the lines).*

Proof. $(q+1)$-arcs (conics) exist in $PG(2, q)$ and if q is even, then $(q+2)$-arcs (hyperovals) exist. Appealing to Theorems 4.8 and 5.1 to construct the appropriate families of conics and hyperovals, the results then follow from Theorem 4.3 and Corollary 4.4. □

5.1. Codes from Baer subspaces.

One last consideration for constructing larger weight codes is to use BSSs of $PG(k, q^2)$ themselves to correspond to the codewords. The correlation numbers, in this case, are functions of q (for $k > 1$), which is probably not desirable. We provide the example nonetheless. Regarding the maximal intersection of two BSSs, we have the following result (see [8, Theorem 1.3]).

Theorem 5.5. *Let B_1 and B_2 be two BSSs of $PG(k, q^2)$. Then*

$$|B_1 \cap B_2| \leq \theta_{k-1,q} + 1.$$

This gives us the following theorem.

Theorem 5.6. *For any prime power q, there exists a $(\theta_{k+1,q^2}, \theta_{k,q}, \theta_{k-1,q}, \theta_{k-1,q}+1)$-OOC consisting of $\mathcal{B}(k, q^2)$ codewords.*

Proof. As in the previous sections, embed $\Pi = PG(k, q^2)$ into $\Sigma = PG(k+1, q^2)$ and consider the set of all BSSs in Π. Proceed with a construction as in Theorem 4.3 with BSSs in place of arcs. The auto-correlation, λ_a, is bounded above by the maximum intersection of two BSSs lying in different hyperplanes of Σ. As two such hyperplanes meet in a $(k-1)$-flat of Σ, this intersection is bounded by $\theta_{k-1,q}$. For the cross-correlation λ_c, we need to consider the intersection of two BSSs lying in the same hyperplane of Σ. By Theorem 5.5 we have $\lambda_c \leq \theta_{k-1,q} + 1$. □

6. Conclusion. We have exhibited several classes of OOCs generated by the same basic ideas in finite projective spaces. Our codes are derived from a nice geometric construction of sets of objects with small intersection sizes. Our hope was to find more examples of (asymptotically) optimal codes. Perhaps more research into the packing of various geometric objects into projective spaces, subject to a small intersection condition, may lead to further examples of optimal OOCs. We have exhibited constructions of code families wherein the auto-correlation is smaller than the cross-correlation. Perhaps these constructions will serve to motivate new investigations of upper bounds on $\Phi(n, w, \lambda_a, \lambda_c)$ with $\lambda_a < \lambda_c$.

REFERENCES

[1] C. M. Bird and A. D. Keedwell, *Design and applications of optical orthogonal codes—a survey*, Bull. Inst. Combin. Appl., 11 (1994), pp. 21–44.

[2] R. Casse, *Projective Geometry, An Introduction*, 1st ed., Oxford University Press, Oxford, UK, 2006.

[3] F. R. K. Chung, J. A. Salehi, and V. K. Wei, *Optical orthogonal codes: Design, analysis, and applications*, IEEE Trans. Inform. Theory, 35 (1989), pp. 595–604.

[4] H. Chung and P. V. Kumar, *Optical orthogonal codes—new bounds and an optimal construction*, IEEE Trans. Inform. Theory, 36 (1990), pp. 866–873.

[5] T. J. Healy, *Coding and decoding for code division multiple user communication systems*, IEEE Trans. Comm., 33 (1985), pp. 310–316.

[6] J. W. P. Hirschfeld, *The number of points on a curve, and applications. Arcs and curves: The legacy of Beniamino Segre*, Rend. Mat. Appl., 26 (2006), pp. 13–28.

[7] J. W. P. Hirschfeld and J. A. Thas, *General Galois geometries*, in Oxford Mathematical Monographs, Oxford Science Publications, The Clarendon Press, Oxford University Press, New York, 1991.

[8] I. Jagos, G. Kiss, and A. Pór, *On the intersection of Baer subgeometries of* PG(n, q^2), Acta Sci. Math. (Szeged), 69 (2003), pp. 419–429.

[9] S. V. Maric, O. Moreno, and C. Corrada, *Multimedia transmission in fiber-optic LANS using optical cdma*, J. Lightwave Technol., 14 (1996), pp. 2149–2153.

[10] N. Miyamoto, H. Mizuno, and S. Shinohara, *Optical orthogonal codes obtained from conics on finite projective planes*, Finite Fields Appl., 10 (2004), pp. 405–411.

[11] Q. A. Nguyen, L. Györfi, and J. L. Massey, *Constructions of binary constant-weight cyclic codes and cyclically permutable codes*, IEEE Trans. Inform. Theory, 38 (1992), pp. 940–949.

[12] R. Omrani, O. Moreno, and P. V. Kumar, *Optimum optical orthogonal codes with* $\lambda > 1$, in Proceedings of the International Symposium on Information Theory, Chicago, IL, 2004, p. 366.

[13] R. Omrani, O. Moreno, and P. V. Kumar, *Improved Johnson bounds for optical orthogonal codes with* $\lambda > 1$ *and some optimal constructions*, in Proceedings of the International Symposium on Information Theory, Adelaide, Australia, 2005, pp. 259–263.

[14] M. Sved, *Baer subspaces in the n-dimensional projective space*, in Combinatorial Mathematics, X (Adelaide, 1982), Lecture Notes in Math. 1036, Springer, Berlin, 1983, pp. 375–391.

[15] J. A. Thas, *Projective geometry over a finite field*, in Handbook of Incidence Geometry, North-Holland, Amsterdam, 1995, pp. 295–347.

[16] O. Veblen and J. W. Young, *Projective Geometry. Vol.* 1, Blaisdell Publishing Co., Ginn and Co., New York-Toronto-London, 1965.

[17] G.-C. Yang and T. E. Fuja, *Optical orthogonal codes with unequal auto- and cross-correlation constraints*, IEEE Trans. Inform. Theory, 41 (1995), pp. 96–106.

SIAM J. Discrete Math.
Vol. 21, No. 3, pp. 794–801

AVOIDING MONOCHROMATIC SEQUENCES WITH SPECIAL GAPS[*]

BRUCE M. LANDMAN[†] AND AARON ROBERTSON[‡]

Abstract. For $S \subseteq \mathbb{Z}^+$ and k and r fixed positive integers, denote by $f(S, k; r)$ the least positive integer n (if it exists) such that within every r-coloring of $\{1, 2, \ldots, n\}$ there must be a monochromatic sequence $\{x_1, x_2, \ldots, x_k\}$ with $x_i - x_{i-1} \in S$ for $2 \leq i \leq k$. We consider the existence of $f(S, k; r)$ for various choices of S, as well as upper and lower bounds on this function. In particular, we show that this function exists for all k if S is an odd translate of the set of primes and $r = 2$.

Key words. arithmetic progressions, Ramsey theory, primes in progression

AMS subject classifications. 05D10, 11B25, 11N13

DOI. 10.1137/S0895480103422196

1. Introduction. Van der Waerden's theorem on arithmetic progressions [10] states that for every partition of $\mathbb{Z}^+$ into r sets, at least one of the sets will contain arbitrarily long arithmetic progressions. An equivalent form of this theorem says that for all $k, r \in \mathbb{Z}^+$, there exists a least positive integer $n = w(k; r)$ such that within every r-coloring of $[1, n] = \{1, 2, \ldots, n\}$ there must be a monochromatic k-term arithmetic progression. By replacing the family of arithmetic progressions, AP, with another family $\mathcal{F}$ of sets, one may ask if the corresponding theorem holds, i.e., is it true that for all $k, r \in \mathbb{Z}^+$, there exists a positive integer $n = f(k; r)$ such that for every r-coloring of $[1, n]$, there is a monochromatic k-term member of $\mathcal{F}$? Rado's theorem involving monochromatic solutions to systems of linear homogeneous equations illustrates one way of choosing $\mathcal{F}$. Other examples may be found in [3, 4, 6, 7, 8, 9].

In [4], the authors considered replacing AP with the collection of those arithmetic progressions $\{x + id : 0 \leq i \leq k - 1\}$ whose common differences, d, belong to a prescribed set. Specifically, for $r \in \mathbb{Z}^+$ and $A \subseteq \mathbb{Z}^+$, call A an *r-large* set if for every r-coloring of $\mathbb{Z}^+$ there exist arbitrarily long monochromatic arithmetic progressions whose common differences belong to A. Define A to be *large* if it is r-large for every r. They gave several sufficient conditions and some necessary conditions for largeness and 2-largeness. They also conjectured that any set that is 2-large must be large.

In this paper we consider a property related to largeness. We consider sequences where the differences between consecutive terms belong to a prescribed set S; however, we do not insist that the sequence be an arithmetic progression.

We begin with the following notation and definitions. For any string u and any $t \in \mathbb{Z}^+$, we denote by u^t the string of t consecutive u's. For $t = 0$, we let u^t represent the empty string. For $S \subseteq \mathbb{Z}^+$, a sequence of positive integers $\{x_1, \ldots, x_k\}$ is called a k-term *S-diffsequence* if $x_i - x_{i-1} \in S$ for $2 \leq i \leq k$. For $r \in \mathbb{Z}^+$, S is called *r-accessible* if whenever $\mathbb{Z}^+$ is r-colored, there are arbitrarily long monochromatic S-diffsequences. The set S is called *accessible* if it is r-accessible for all $r \in \mathbb{Z}^+$. If S is not accessible,

[*] Received by the editors February 3, 2003; accepted for publication (in revised form) May 10, 2007; published electronically September 28, 2007.
http://www.siam.org/journals/sidma/21-3/42219.html

[†] Department of Mathematics, University of West Georgia, Carrollton, GA 30118 (landman@westga.edu).

[‡] Department of Mathematics, Colgate University, Hamilton, NY 13346 (aaron@math.colgate.edu).

the *degree of accessibility* of S, denoted $\mathrm{DA}(S)$, is the largest value of r such that S is r-accessible. Finally, we denote by $f(S,k;r)$ the least positive integer n (if it exists) such that for every r-coloring of $[1,n]$ there is a monochromatic k-term S-diffsequence. (Obviously, for S an r-accessible set, if $S \subseteq T$, then $f(S,k;r) \geq f(T,k;r)$.)

Denote the family of all accessible sets by $\mathcal{A}$ and the family of all r-accessible sets by $\mathcal{A}_r$. Likewise, denote the families of large sets and r-large sets by $\mathcal{L}$ and $\mathcal{L}_r$. Clearly, $\mathcal{L} \subseteq \mathcal{A}$ and $\mathcal{L}_r \subseteq \mathcal{A}_r$ for all r. As stated before, it is conjectured that $\mathcal{L} = \mathcal{L}_2$. As we shall see, $\mathcal{A} \neq \mathcal{A}_2$ and $\mathcal{A}_2 \neq \mathcal{L}_2$. Moreover, Jungić [5] has proved that $\mathcal{A} \neq \mathcal{L}$.

In section 2 we give some basic lemmas and consider a few elementary examples; in particular, we show that for any $d \in \mathbb{Z}^+$, there is a set with degree of accessibility d (this is in contrast to what is conjectured about large sets). In section 3 we prove that for each odd positive integer t there are arbitrarily long sequences of primes $p_1 < \cdots < p_k$ such that $p_i - p_{i-1} \in P + t$ for $2 \leq i \leq k$, where P is the set of primes. From this it follows that $P + t \in \mathcal{A}_2$. Section 4 contains some open questions and a table of computer-generated values of $f(S,k;2)$ for a few sets S and values k.

2. A few simple examples.

We begin with two useful lemmas.

Lemma 2.1. *Let $c \geq 0$ and $r \geq 2$, and let $S \subseteq \mathbb{Z}^+$. If every $(r-1)$-coloring of S yields arbitrarily long monochromatic $(S+c)$-diffsequences, then $S + c \in \mathcal{A}_r$.*

Proof. Let $S = \{s_i : i \in \mathbb{Z}^+\}$ and assume every $(r-1)$-coloring of S admits arbitrarily long monochromatic $(S+c)$-diffsequences. Let χ be an r-coloring of $\mathbb{Z}^+$. By induction on k, we show that, under χ, for all k there are k-term monochromatic $(S+c)$-diffsequences. Since there are trivially 1-term sequences, assume that under χ there is a monochromatic $(S+c)$-diffsequence $X = \{x_1, \ldots, x_k\}$. Say X is of color red. Consider $A = \{x_k + s_i + c : s_i \in S\}$. If some member of A is colored red, then we have a red $(k+1)$-term $(S+c)$-diffsequence. Otherwise, we have an $(r-1)$-coloring of A and hence, A must contain arbitrarily long monochromatic $(S+c)$-diffsequences. □

Remark. The converse of Lemma 2.1 is false. For example, let $S = \{2\} \cup (2\mathbb{Z}^+ - 1)$. Let χ be the 2-coloring of S defined by $\chi(x) = 1$ if $x \equiv 1 (\mathrm{mod}\ 4)$ or $x = 2$, and $\chi(x) = 0$ if $x \equiv 3 (\mathrm{mod}\ 4)$. Then χ does not yield arbitrarily long monochromatic S-diffsequences (there are none of length four). On the other hand, $S \in \mathcal{A}_3$ [8, Remark 5], and in fact $f(S,k;3) \leq 6k^2 - 13k + 6$ (see Theorem 2.6); more generally, from this same reference it follows that if m is even, and $j \in \mathbb{Z}^+$, then the set $\{jm\} \cup \{x : x \equiv \frac{m}{2} (\mathrm{mod}\ m)\}$ is 3-accessible.

We leave to the reader as an easy exercise the proof of the following result.

Lemma 2.2. *Let S be a set of positive integers, and let $k, r, j \in \mathbb{Z}^+$. If $f(S,k;r) = M$, then $f(jS,k;r) = j(M-1) + 1$.*

Using Lemma 2.1, with $c = 0$ and $r = 2$, it is clear that the set $\{2^i : i \geq 0\}$ is 2-accessible. The following result tells us more.

Theorem 2.3. *Let $a \in \mathbb{Z}^+ \setminus \{1,3\}$, and define $S = \{(a-1)a^j : j = 0,1,2,\ldots\} \cup \{(a-1)^2 a^j : j = 0,1,2,\ldots\}$. Then $2 \leq DA(S) \leq a$ and $f(S,k;2) \leq a^k - a + 1$.*

Proof. To show that $\mathrm{DA}(S) \leq a$, let χ be the $(a+1)$-coloring defined by $\chi(x) = x$ mod $(a+1)$. Assume that $\chi(y) = \chi(z)$ and that $z - y \in S$. By the definition of χ, $a+1$ divides $z-y$, and therefore either $(a+1)|(a-1)a^j$ or $(a+1)|(a-1)^2 a^j$ for some $j \geq 0$. Since $a \neq 1, 3$, neither of these is possible.

Now let $\alpha : [1, a^k - a + 1] \to \{0,1\}$. We will show, by induction on k, that under α there is a monochromatic k-term S-diffsequence. Obviously, it holds for $k = 1$. Now assume $k \geq 2$ and that it holds for $k-1$. Let $X = \{x_1, \ldots, x_{k-1}\}$ be a monochromatic S-diffsequence, say of color 0, that is contained in $[1, a^{k-1} - a + 1]$. Consider the set $A = \{x_{k-1} + (a-1)a^i : 0 \leq i \leq k-1\}$. Note that $A \subseteq [1, a^k - a + 1]$. If there exists

$y \in A$ of color 0, then $X \cup \{y\}$ is a monochromatic k-term S-diffsequence. If, on the other hand, no such y exists, then A is a monochromatic k-term S-diffsequence. □

Corollary 2.4. *If $S = \{2^i : i \geq 0\}$, then $DA(S) = 2$ and $8(k-3)+1 \leq f(S,k;2) \leq 2^k - 1$ for all $k \geq 3$.*

Proof. The fact that $\text{DA}(S) = 2$ and the upper bound follow from Theorem 2.3. For the lower bound, it is straightforward to show (by induction on k) that, for $k \geq 4$, the 2-coloring $\chi_k = (10010110)^{k-3}$ avoids monochromatic k-term S-diffsequences. □

The details of the proof of Corollary 2.4 are given in [9].

Remark. Note that $a = 2$ is the only value of a for which $\{a^i : i \geq 0\} \in \mathcal{A}_2$. This follows immediately from the observation that if $\gcd(i,m) = 1$ and $S = \{x \in \mathbb{Z}^+ : x \equiv i (\text{mod } m)\}$, then there exists a (fairly obvious) 2-coloring of $\mathbb{Z}^+$ that avoids monochromatic m-term S-diffsequences.

In [4] it was shown that if $A \notin \mathcal{L}_r$ and $B \notin \mathcal{L}_s$, then $A \cup B \notin \mathcal{L}_{rs}$ by using the canonical product coloring. The same simple argument can be used to prove the following lemma. We omit the details.

Lemma 2.5. *If $S \notin \mathcal{A}_r$ and $T \notin \mathcal{A}_s$ and either $S + T = \{s+t : s \in S, t \in T\} \subseteq S$ or $S + T \subseteq T$, then $S \cup T \notin \mathcal{A}_{rs}$.*

It is easy to see, using Lemma 2.1, that the set $S = \{2\} \cup (2\mathbb{Z}^+ - 1)$ is 2-accessible, since the set of odd numbers itself is an S-diffsequence. The next theorem tells us more about S. We omit the proof, which is given in [9].

Theorem 2.6. *If $S = \{2\} \cup (2\mathbb{Z}^+ - 1)$, then $DA(S) = 3$. Furthermore, $f(S,k;3) \leq 6k^2 - 13k + 6$, $f(S,k;2) = 3k - 3$ for k even, and $f(S,k;2) = 3k - 4$ for k odd.*

The proof of the following example is an easy exercise and is left to the reader.

Proposition 2.7. *Let $F = \{F_1, F_2, F_3, F_4 \ldots\} = \{1, 2, 3, 5 \ldots\}$ be the set of Fibonacci numbers. Then $f(F,k;2) \leq F_{k+2} - 2$ for $k \geq 1$.*

The following simple result provides us with examples of very sparse sets which are nonetheless accessible.

Proposition 2.8. *For $T \subseteq \mathbb{Z}^+$ infinite, $T - T = \{t - s : s < t \text{ and } s,t \in T\} \in \mathcal{A}$.*

Proof. Let $r \in \mathbb{Z}^+$ and let χ be an r-coloring of $\mathbb{Z}^+$. Let s be the minimum element in T and consider $T_s = \{t - s : t \in T\}$. Some color class must contain an infinite number of elements of T_s. This gives an infinitely long monochromatic $(T-T)$-diffsequence. Since this argument holds for all r, $T - T \in \mathcal{A}$. □

We now look at the accessibility of certain collections of congruence classes. In [4] it was proved that if $A \in \mathcal{L}_2$, then A must contain a multiple of every positive integer. We have seen that this is not true if we replace $\mathcal{L}_2$ with $\mathcal{A}_2$ (see, for example, Corollary 2.4 or Theorem 2.6). By the next proposition, we see that this condition is necessary in order for a set to be accessible. In addition to giving another example for which 2-accessible does not imply 2-large, it also shows that for all positive integers m, there exists a set having m as its degree of accessibility.

Proposition 2.9. *For $m \geq 2$, let $S_m = \{x \in \mathbb{Z}^+ : m \nmid x\}$. Then $DA(S_m) = m - 1$.*

Proof. That $\text{DA}(S_m) \leq m - 1$ is easily seen by considering the m-coloring χ of $\mathbb{Z}^+$ defined by $\chi(x) = x \pmod m$. To prove the reverse inequality, let χ be any $(m-2)$-coloring of S_m. Then some color must contain an infinite number of elements from each of at least two of the residue classes 1 (mod m), 2 (mod m), $\ldots, (m-1)$ (mod m). Thus, some color contains an infinite S_m-diffsequence and the proof is complete. □

For some results about the values of $f(S_m, k; 2)$ for specific choices of m, we refer the reader to [9].

3. Translations of the set of primes. In [4], the question was raised as to whether there exists a translation of P, the set of primes, that is large, or for that matter 2-large. Since a 2-large set must contain a multiple of every integer, $P+e \notin \mathcal{L}_2$ if e is even. Likewise, by Proposition 2.9, if e is even, then $P+e \notin \mathcal{A}$, and P itself is not 4-accessible. In fact, $P \notin \mathcal{A}_3$. To see this, color the multiples of 9 green, the remaining even numbers red, and the remaining odd numbers blue. It is easy to see that any sequences of nine reds, nine blues, or two greens must have numbers that differ by a nonprime. We do not know if P is 2-accessible or if some even translation of P is 2-accessible. On the other hand, as we shall see in this section, all odd translations of P are 2-accessible. In fact, we prove the following stronger result.

THEOREM 3.1. *Let $t \in \mathbb{Z}^+$ be odd. For any $k \geq 2$, there exist $p_1, p_2, \ldots, p_k \in P$ such that $p_i - p_{i-1} \in P + t$ for $2 \leq i \leq k$.*

Combining Theorem 3.1 with Lemma 2.1, we immediately get the following.

COROLLARY 3.2. *If t is odd, then $P + t \in \mathcal{A}_2$.*

We prove Theorem 3.1 as an application of a powerful number theoretic result due to Balog [1]. The application is stated as Theorem 3.3 below. Before stating the theorem, we mention some notation.

Let $\mathbf{b} = (b_1, b_2, \ldots, b_k) \in \mathbb{Z}^k$, $p \in P$, and $x \in \mathbb{R}^+$. We define the following:

(1) $\pi(x; \mathbf{b}) = |\{n : 1 < n + b_i \leq x \text{ is prime for every } 1 \leq i \leq k\}|$;

(2) $\rho(p) = \rho(p; \mathbf{b}) = |\{b_i \,(\mathrm{mod}\ p) : 1 \leq i \leq k\}|$;

(3) $\sigma(\mathbf{b}) = \prod_{p \in P}(1 - \frac{1}{p})^{-k}(1 - \frac{\rho(p)}{p})$;

(4) $T(x; \mathbf{b}) = \sum_{R(x)} \frac{1}{\log(n+b_1)\log(n+b_2)\ldots\log(n+b_k)}$, where $R(x) = \{n : 1 < n + b_i \leq x \text{ for all } i,\ 1 \leq i \leq k\}$.

Finally, we remind the reader of the following standard notation: For functions $f(x)$ and $g(x)$, we write $f(x) \gg g(x)$ if there exists a positive constant c such that $\liminf_{x\to\infty} \frac{f(x)}{g(x)} \geq c$. Furthermore, for a parameter k, we write $f(x) \gg_k g(x)$ if the constant c is dependent upon k.

THEOREM 3.3 (Balog). *Let $k \in \mathbb{Z}^+$, $x \in \mathbb{R}^+$, and $t \in \mathbb{Z}^+ \cup \{0\}$. Define*

$$B = \left\{ \left(0, q_1 + t, \ldots, \sum_{i=1}^{k-1}(q_i + t) \right) : q_i \in P, k < q_i \leq x/2k, 1 \leq i \leq k-1 \right\}.$$

Then

$$\sum_{\mathbf{b} \in B} |\pi(x; \mathbf{b}) - \sigma(\mathbf{b})T(x; \mathbf{b})| \ll_k \frac{x^k}{\log^{2k} x}.$$

Remark. This is a special case of Balog's theorem [1, p. 49] (where we use $A = 2k$, $c = 0$, $D = 1$, and $a_i = 1$ for $1 \leq i \leq k$).

We will need the following technical lemma.

LEMMA 3.4. *Let $k \geq 2$ and let $t \geq 1$ be odd. For $\mathbf{q} = (q_1, \ldots, q_{k-1}) \in P^{k-1}$, let*

$$\mathbf{b}(\mathbf{q}) = \mathbf{b}(\mathbf{q}, t) = \left(0, q_1 + t, \ldots, \sum_{i=1}^{k-1}(q_i + t) \right).$$

Let

$$M(x) = \left\{ \mathbf{q} \in P^{k-1} : \mathbf{q} \in \left(k, \frac{x}{2k}\right]^{k-1} \text{ and } \rho(p, \mathbf{b}(\mathbf{q})) < p \text{ for all } p \in P \right\}.$$

Then $|M(x)| \gg_k (\frac{x}{\log x})^{k-1}$.

Proof. Our goal is to show that "most" $\mathbf{q} \in P^{k-1}$ fail to define a complete set of residue classes modulo p for all $p \in P$. First of all, it is clear that $\rho(p, \mathbf{b}(\mathbf{q})) < p$ for any prime $p > k$ since $\rho(p, \mathbf{b}(\mathbf{q})) \leq k$. Hence, we need only to consider those primes $r_1, r_2, \ldots, r_d \leq k$; let $m = \prod_{i=1}^{d} r_i$. We will obtain a lower bound for the number of $\mathbf{q}$ such that $\rho(r_i, \mathbf{b}(\mathbf{q})) < r_i$ for all $1 \leq i \leq d$.

It suffices to show that for some integer h, all entries of $(h, h, \ldots, h) + \mathbf{b}(\mathbf{q})$ are not divisible by r_i for all i, $1 \leq i \leq d$. To this end, choose h such that $\gcd(h, m) = 1$ (note that h is odd). Obviously, this condition holds for the first entry.

In order to have that each r_i, $1 \leq i \leq d$, does not divide $h+q_1+t$ (the second entry in $(h, h, \ldots, h) + \mathbf{b}(\mathbf{q})$), it is sufficient that q_1 belong to some particular congruence class $c_1 \pmod{m}$, where $\gcd(c_1, m) = 1$. Since t and h are both odd, such a c_1 exists, and by Dirichlet's theorem for primes in arithmetic progressions, there are $\gg_k \frac{x}{\log x}$ choices for q_1.

Similarly, once $h, q_1, q_2, \ldots, q_{j-1}$ have been chosen, we need only consider those q_j so that for each r_i, q_j does not belong to any of the residue classes $-(h + q_1 + q_2 + \cdots q_{j-1} + jt) \pmod{r_i}$, $1 \leq i \leq d$. So it suffices to have q_j belong to one specific congruence class $c_j (\text{mod } m)$, with $\gcd(c_j, m) = 1$.

Using the above criteria, we have at least $\prod_{i=2}^{d}(r_i - 2)$ reduced residue classes modulo $m/2$ in which the entries of $(h, h, \ldots, h) + \mathbf{b}(\mathbf{q})$ may reside. To see this, note that for $2 \leq j \leq d$, for each prime r_j, we cannot have $c_j = 0 \pmod{r_j}$ or $-(h + q_1 + \cdots + q_{j-1} + jt) \pmod{r_j}$, giving $r_j - 2$ choices. Now, by Dirichlet's theorem we have $\gg_k \frac{x}{\log x}$ choices for each q_i, and thus $\gg_k (\frac{x}{\log x})^{k-1}$ valid choices for the $(k-1)$-tuple of primes $(q_1, q_2, \ldots, q_{k-1})$ that belong to $M(x)$. □

Using Theorem 3.3 and Lemma 3.4, we have the following result.

Lemma 3.5. *Let* $k \geq 2$, $t \geq 1$ *and odd, and* $x \in \mathbb{R}^+$. *If*

$$W(x) = \left\{(p, q_1, \ldots, q_{k-1}) : p, q_1, \ldots, q_{k-1} \textit{ are primes and } k < q_1, q_2, \ldots, q_{k-1} \leq \frac{x}{2k}\right\}$$

and

$$S(x) = \left\{ (p, q_1, \ldots, q_{k-1}) \in W : p + \sum_{j=1}^{i}(q_j + t) \leq x \textit{ is prime for all } i \, (0 \leq i \leq k) \right\},$$

then $|S(x)| \gg_k \frac{x^k}{\log^{2k-1} x}$.

Proof. We use the notation from Theorem 3.3 and Lemma 3.4. In order to apply Theorem 3.3, we first obtain effective bounds for ρ, σ, and T.

We first show that $0 < \sigma(\mathbf{b}(\mathbf{q})) < \infty$. The fact that $\sigma(\mathbf{b}(\mathbf{q})) < \infty$ is shown in [2]. We see that for all $\mathbf{q} \in M(x)$, by the definition of $M(x)$ we have $\rho(p; \mathbf{b}(\mathbf{q})) < p$ for all primes p. Since it is also true that $\rho(p; \mathbf{b}(\mathbf{q})) \leq k$ for any prime p, we have

$$(3.1) \qquad \sigma(\mathbf{b}(\mathbf{q})) \geq \prod_{p \leq k} \left(1 - \frac{1}{p}\right)^{-k} \left(1 - \frac{p-1}{p}\right) \prod_{p > k} \left(1 - \frac{1}{p}\right)^{-k} \left(1 - \frac{k}{p}\right) = \sigma_k,$$

a constant dependent upon only k. We next show that $\sigma_k > 0$.

Clearly, we have the finite product in (3.1) positive, so we must show that the infinite product in (3.1) converges to a positive constant. To this end, let $1 + a_p = (1 - 1/p)^{-k}(1 - k/p)$. By the binomial theorem, we have $a_p = \frac{-\sum_{i=2}^{k}(-1)^{k-i}\binom{k}{i}p^{-i}}{(1-1/p)^k}$.

Since $|a_p| \leq \frac{\sum_{i=2}^{k}\binom{k}{i}p^{-i}}{(1-1/p)^k} \leq \frac{\sum_{i=2}^{k}\binom{k}{i}p^{-2}}{(1-1/p)^k} \leq \frac{\sum_{i=2}^{k}\binom{k}{i}p^{-2}}{1/2^k} = 2^k(2^k-k-1)p^{-2}$, we see that $\sum_{p\in P} a_p$ converges absolutely. It follows that $\prod_{p\in P}(1+a_p)$ converges to a positive number. Thus, from (3.1),

$$\text{for all } \mathbf{q} \in M(x),\ \sigma(\mathbf{b}(\mathbf{q})) \geq \sigma_k > 0. \tag{3.2}$$

We next bound $T(x;\mathbf{b}(\mathbf{q}))$ by using

$$\begin{aligned} |\{n : 1 < n + b_i \leq x, 1 \leq i \leq k\}| &= (x - b_k) + O(1) \\ &= x - \sum_{i=1}^{k-1}(q_i + t) + O(1) \\ &> x - \sum_{i=1}^{k-1} q_i - kt \\ &\gg_{k,t} x - k\left(\tfrac{x}{2k}\right) \\ &= \tfrac{x}{2}. \end{aligned}$$

This gives us

$$T(x;\mathbf{b}(\mathbf{q})) \gg_{k,t} \frac{x}{2\log^k x}. \tag{3.3}$$

We now apply our above bounds to complete the proof. Noting that $\{\mathbf{b}(\mathbf{q}) : \mathbf{q} \in M(x)\} \subseteq B$ (where B is as defined in Theorem 3.3), Theorem 3.3 implies that

$$\sum_{\mathbf{q}\in M(x)} \Big|\pi(x;\mathbf{b}(\mathbf{q})) - \sigma(\mathbf{b}(\mathbf{q}))T(x;\mathbf{b}(\mathbf{q}))\Big| \ll_k \frac{x^k}{\log^{2k} x}. \tag{3.4}$$

Let $N(x) = P^{k-1} \cap (k, \frac{x}{2k}]^{k-1}$. Since we can write

$$W(x) = \bigcup_{p\in P} \bigcup_{\mathbf{q}\in N(x)} \{(p, q_1, q_2, \ldots, q_{k-1})\},$$

we have

$$|S(x)| \ \geq \sum_{\mathbf{q}\in M(x)} \pi(x;\mathbf{b}(\mathbf{q})).$$

Using (3.2) and (3.3) along with Lemma 3.4, inequality (3.4) yields

$$\begin{aligned} \sum_{\mathbf{q}\in M(x)} \pi(x;\mathbf{b}(\mathbf{q})) &\gg_k \sum_{\mathbf{q}\in M(x)} \sigma(\mathbf{b}(\mathbf{q}))T(x;\mathbf{b}(\mathbf{q})) - O\left(\tfrac{x^k}{\log^{2k} x}\right) \\ &\gg_{k,t} \sigma_k |M(x)| \left(\tfrac{x}{2} + O(1)\right)\left(\tfrac{1}{\log^k x}\right) - O\left(\tfrac{x^k}{\log^{2k} x}\right) \\ &\gg_{k,t} \sigma_k \left(\tfrac{x}{\log x}\right)^{k-1}\left(\tfrac{x}{\log^k x}\right) - O\left(\tfrac{x^k}{\log^{2k} x}\right) \\ &\gg_{k,t} \tfrac{x^k}{\log^{2k-1} x}. \qquad \square \end{aligned}$$

Having Lemma 3.5, we are now in a position to complete the proof of Theorem 3.1. We choose primes $p_1, q_1, \ldots, q_{k-1}$ so that $p_i = p_1 + \sum_{j=1}^{i-1}(q_j + t) \in P$ for $2 \leq i \leq k$. Since $p_i - p_{i-1} = q_{i-1} + t$ for $2 \leq i \leq k$, we are done.

4. Open questions and some exact values. There are many interesting questions left unanswered about accessibility. Here is a list of some.

1. Let $T = \{2^i : i \geq 0\}$. What is the exact value of $f(T, k; 2)$? In Table 1 we give the first few values of this function.

2. What is the exact value of $f(S, k; 3)$, where $S = \{2\} \cup (2\mathbb{Z}^+ - 1)$?

3. What is a formula for $f(S_m, k; 2)$? Calculations for the case $m = 6$ support our conjecture that for $k \geq 2$,

$$f(S_6, k; 2) = \begin{cases} (5k-4)/2 & \text{if } k \equiv 2 (\text{mod } 4), \\ (5k-5)/2 & \text{if } k \equiv 3 (\text{mod } 4), \\ (5k-6)/2 & \text{if } k \equiv 0 (\text{mod } 4), \\ (5k-7)/2 & \text{if } k \equiv 1 (\text{mod } 4). \end{cases}$$

Also note that (see [9]) if we let $am \leq k < (a+1)m$, a a nonnegative integer, we have $2k + 2a - 1 \leq f(S_m, k; 2)$, with equality when $a = 0$. We believe that this inequality is an equality for all $a \in \mathbb{Z}^+$.

4. If t is an odd positive integer, what is $\mathrm{DA}(P + t)$? Moreover, is it true that for every 2-coloring of P, there exist arbitrarily long monochromatic $(P+t)$-diffsequences? If the answer to the latter question is true, then by Lemma 2.1, $P + t \in \mathcal{A}_3$.

5. What is the order of magnitude of $f(P+t, k; 2)$ for a fixed odd positive integer t? Table 1 includes some specific values of this function.

6. As stated earlier, $P \notin \mathcal{A}_3$. Is $P \in \mathcal{A}_2$? If so, what is the magnitude of $f(P, k; 2)$? We have calculated the first several values of $f(P, k; 2)$ (see Table 1).

TABLE 1

$S \setminus k$	2	3	4	5	6	7	8
T	3	7	11	17	25	35	51
F	3	5	9	11	15	19	21
P	5	9	13	21	25	33	?
$P+1$	7	13	21	27	35	?	?
$P+2$	9	17	25	33	?	?	?
$P+3$	11	21	31	42	?	?	?
$P+4$	13	25	37	?	?	?	?
$P+5$	15	29	?	?	?	?	?
$P+6$	17	33	?	?	?	?	?
$P+7$	19	37	?	?	?	?	?
S_5	3	5	7	11	13	15	19
S_6	3	5	7	9	13	15	17

7. Let F be the set of Fibonacci numbers. What is $\mathrm{DA}(F)$? What is the order of magnitude of $f(F, k; 2)$?

8. For $k, m \geq 2$, what can we say about $\mathrm{DA}(S)$ and $f(S, k; 2)$, where S is the union of more than one congruence class modulo m?

Table 1 gives the exact value of $f(S, k; 2)$ for various S and k. The symbols T, F, and P denote $\{2^i : i \geq 0\}$, the set of Fibonacci numbers, and the set of primes, respectively.

Acknowledgments. We thank Andrew Granville for guiding us to Balog's theorem and for his invaluable assistance with the proofs of Lemmas 3.4 and 3.5. We also thank Scott Ahlgren for helping with some details of the proof of Lemma 3.5.

REFERENCES

[1] A. Balog, *The prime k-tuplets conjecture on average*, in Analytic Number Theory, Progr. Math. 85, Birkhäuser, Boston, 1990, pp. 47–75.

[2] P. T. Bateman and R. A. Horn, *A heuristic asymptotic formula concerning the distribution of prime numbers*, Math. Comp., 16 (1962), pp. 363–367.

[3] T. C. Brown, P. Erdős, and A. R. Freedman, *Quasi-progressions and descending waves*, J. Combin. Theory Ser. A, 53 (1990), pp. 81–95.

[4] T. C. Brown, R. L. Graham, and B. M. Landman, *On the set of common differences in van der Waerden's theorem on arithmetic progressions*, Canad. Math. Bull., 42 (1999), pp. 25–36.

[5] V. Jungić, *On a conjecture of Brown concerning accessible sets*, J. Combin. Theory Ser. A, 110 (2005), pp. 175–178.

[6] B. Landman, *Ramsey functions related to the van der Waerden numbers*, Discrete Math., 102 (1992), pp. 265–278.

[7] B. Landman, *On some generalizations of the van der Waerden number* $w(3)$, Discrete Math., 207 (1999), pp. 137–147.

[8] B. Landman, *Avoiding arithmetic progressions (mod m) and arithmetic progressions*, Util. Math., 52 (1997), pp. 173–182.

[9] B. Landman and A. Robertson, *Ramsey Theory on the Integers*, Stud. Math. Libr. 24, AMS, Providence, RI, 2004.

[10] B. L. van der Waerden, *Beweis einer Baudetschen Vermutung*, Nieuw Arch. Wisk., 15 (1928), pp. 212–216.

SIAM J. DISCRETE MATH.
Vol. 21, No. 3, pp. 802–804

ON ROTA'S BASIS CONJECTURE[*]

JIM GEELEN[†] AND KERRI WEBB[‡]

Abstract. Rota conjectured that if $(B_1,\dots,B_n)$ are disjoint bases in a rank-n matroid M, then there are n disjoint transversals of $(B_1,\dots,B_n)$ that are bases of M. We prove the weaker result that there are $O(\sqrt{n})$ disjoint transversals of $(B_1,\dots,B_n)$ that are bases. We also prove that if $(B_1,\dots,B_k)$ are disjoint bases of a rank-n matroid with $n > \binom{k+1}{2}$, then there are n disjoint independent transversals of $(B_1,\dots,B_k)$.

Key words. matroids, Rado's theorem, Rota's basis conjecture

AMS subject classification. 05B35

DOI. 10.1137/060666494

1. Introduction.

In 1989, Rota conjectured that, given n bases of a rank-n matroid, there is an n by n grid such that the rows contain the given bases and each column also contains a basis; see Huang and Rota [1]. By possibly adding parallel elements to the matroid, we can assume that the original n bases are disjoint, and thus we have the following equivalent conjecture.

CONJECTURE 1.1 (Rota's basis conjecture). *If $(B_1,\dots,B_n)$ are disjoint bases in a rank-n matroid M, then there are n disjoint transversals of $(B_1,\dots,B_n)$ that are bases of M.*

We prove the following related results.

THEOREM 1.2. *For $n \geq k^2-k+1$, if $(B_1,\dots,B_n)$ are disjoint bases in a rank-n matroid, then there are k disjoint transversals of $(B_1,\dots,B_n)$ that are bases.*

THEOREM 1.3. *If $(B_1,\dots,B_k)$ are disjoint bases in a rank-n matroid where $n \geq \binom{k+1}{2}+1$, then there are n disjoint independent transversals of $(B_1,\dots,B_k)$.*

We hope that the quadratic bounds in Theorems 1.2 and 1.3 will be improved to linear functions.

2. Disjoint independent transversals.

In this section we prove Theorem 1.2. For sets $(S_1,\dots,S_n)$ and $X \subseteq \{1,\dots,n\}$, we let $S(X)$ denote $\bigcup(S_i : i \in X)$. We use the following result; see Rado [3] or Oxley [2, p. 388].

THEOREM 2.1 (Rado's theorem). *Let $(S_1,\dots,S_n)$ be sets in a matroid. Then there is an independent transversal of $(S_1,\dots,S_n)$ if and only if $r(S(X)) \geq |X|$ for each $X \subseteq \{1,\dots,n\}$.*

As a corollary we have the following lemma.

LEMMA 2.2. *Let $t \leq n$, and let $(S_1,\dots,S_n)$ be independent sets of a matroid. If $|S_i| \geq \min(i, n-t)$ for each $i \in \{1,\dots,n\}$ and there are disjoint subsets $Y_1,\dots,Y_t$ of $\{1,\dots,n\}$ such that $S(Y_1),\dots,S(Y_t)$ each have rank at least n, then there is an independent transversal of $(S_1,\dots,S_n)$.*

[*] Received by the editors July 31, 2006; accepted for publication (in revised form) June 11, 2007; published electronically October 31, 2007. This research was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

http://www.siam.org/journals/sidma/21-3/66649.html

[†] Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 (jfgeelen@math.uwaterloo.ca).

[‡] Department of Mathematics and Computer Science, University of Lethbridge, Lethbridge, Alberta, Canada T1K 3M4 (kerri.webb@uleth.ca).

Proof. Let $X \subseteq \{1,\ldots,n\}$. If $Y_i \subseteq X$ for some $i \in \{1,\ldots,t\}$, then $r(S(X)) \geq r(S(Y_i)) \geq n \geq |X|$. If Y_i is not contained in X for each i, then $|X| \leq n-t$. Let k be the maximum index in X in this case. Then $|X| \leq k$, and since $r(S(X)) \geq r(S_k) = |S_k| \geq \min(k, n-t)$, it again follows that $r(S(X)) \geq |X|$. Hence, by Rado's theorem, there is an independent transversal of $(S_1,\ldots,S_n)$. □

Proof of Theorem 1.2. Let $l = \binom{k}{2}$; note that $n \geq 2l+1$. We can choose bases $X_1,\ldots,X_l$ such that $X_i \subseteq B_i \cup B_{n-i}$ and $|X_i \cap B_i| = n-i$. Let $S = (B_1 \cup \cdots \cup B_n) - (X_1 \cup \cdots \cup X_l)$. Now let $A_1 = \emptyset$ and, for each $i \in \{2,\ldots,k\}$, let $A_i = \left\{1,\ldots,\binom{i}{2}\right\}$. We claim that there are disjoint independent transversals $T_1,\ldots,T_k$ of $(B_1,\ldots,B_n)$ with $T_i \subseteq S \cup X(A_i)$ for each $i \in \{1,\ldots,k\}$. Certainly there exists an independent transversal $T_1 \subseteq S$ of $(B_1,\ldots,B_n)$. Assume that, for some $t \in \{1,\ldots,k-1\}$, we have found disjoint independent transversals $T_1,\ldots,T_t$ of $(B_1,\ldots,B_n)$ with $T_i \subseteq S \cup X(A_i)$ for each $i \in \{1,\ldots,t\}$. Let $T' = T_1 \cup \cdots \cup T_t$, let $S' = S \cup X(A_{t+1}) - T'$, and let $r = \binom{t+1}{2}$. Consider the independent sets $(S_1,\ldots,S_n)$, where

$$S_i = \begin{cases} B_{r+i} \cap S', & \text{if } 1 \leq i < n-2r; \\ B_{2r-n+1+i} \cap S' = B_{2r-n+1+i} - T', & \text{if } n-2r \leq i < n-r; \\ B_i \cap S' = B_i - T', & \text{if } n-r \leq i \leq n. \end{cases}$$

We claim that $|S_i| \geq \min(i, n-t)$ for each $i \in \{1,\ldots,n\}$.

First consider the case that $1 \leq i \leq l-r$. Then

$$S_i = B_{r+i} \cap S' = B_{r+i} - B_{r+i} \cap X_{r+i} - B_{r+i} \cap T'.$$

Since $r+i \notin A_t$, the sets X_{r+i} and T' are disjoint, and therefore $|S_i| = n-(n-(r+i))-t = i+r-t \geq i$. Similarly, if $n-l-r \leq i < n-2r$, then

$$S_i = B_{r+i} \cap S' = B_{r+i} - B_{r+i} \cap X_{n-(r+i)} - B_{r+i} \cap T',$$

and again $|S_i| = i+r-t \geq i$. It remains to consider the case that either $l-r < i < n-l-r$ or $n-2r \leq i \leq n$. Here $S_i = B_j - T'$ for some $j \in \{1,\ldots,r\} \cup \{l+1,\ldots,n-l-1\} \cup \{n-r,\ldots,n\}$, and thus $|S_i| = n-t$. Hence $|S_i| \geq \min(i, n-t)$ for each $i \in \{1,\ldots,n\}$.

Let $y_i = \binom{t}{2} + i$ for each $i \in \{1,\ldots,t\}$. Then $y_i \in A_{t+1} - A_t$, and thus the basis X_{y_i} is contained in $(B_{y_i} \cap S') \cup (B_{n-y_i} \cap S') = S_{n-2r-1+y_i} \cup S_{n-y_i}$. Hence if $Y_i = \{n-2r-1+y_i, n-y_i\}$, then $S(Y_i)$ has rank n for each $i \in \{1,\ldots,t\}$. By Lemma 2.2, there is an independent transversal $T_{t+1} \subseteq S'$ of $(B_1,\ldots,B_n)$, and we therefore inductively obtain the required transversals. □

3. Partitioning into independent transversals. In this section we prove Theorem 1.3 using the following lemma.

LEMMA 3.1. *Let $(S_1,\ldots,S_k)$ be disjoint k-element sets in a matroid, and let $(Y_1,Y_2,\ldots,Y_{k-1})$ be disjoint independent sets such that Y_i is an i-element transversal of $(S_1,\ldots,S_i)$ for each $i \in \{1,\ldots,k-1\}$. If $(S_1 \cup \cdots \cup S_k) - (Y_1 \cup \cdots \cup Y_{k-1})$ is independent, then there are k disjoint independent transversals of $(S_1,\ldots,S_k)$.*

Proof. Let $Z = (S_1 \cup \cdots \cup S_k) - (Y_1 \cup \cdots \cup Y_{k-1})$, and let $Y_0 = \emptyset$. For each $i \in \{0,\ldots,k-1\}$, there is a set $X_i \subseteq Z$ such that $|X_i| = |Y_i|$ and $(Z - X_i) \cup Y_i$ is independent. We can now find disjoint sets $(W_{k-1}, W_{k-2},\ldots,W_0)$, in that order, such that W_i is a transversal of $(S_1,\ldots,S_k)$ with $Y_i \subseteq W_i \subseteq (Z - X_i) \cup Y_i$, for each $i \in \{0,\ldots,k-1\}$. Note that each W_i is independent since it is contained in $(Z - X_i) \cup Y_i$. □

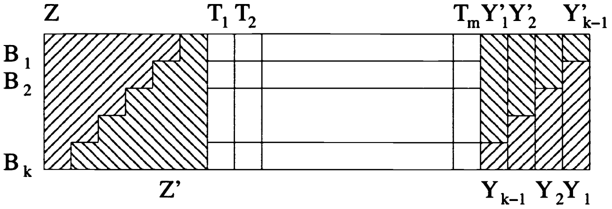


FIG. 1. *Proof of Theorem* 1.3.

Proof of Theorem 1.3. Since $n \geq \binom{k+1}{2}+1$, we can find disjoint independent sets Z and Z' such that $|Z \cap B_i| = k+1-i$ and $|Z' \cap B_i| = i$ for each $i \in \{1,\dots,k\}$. Let $m = (n-(k+1))-(k-1)$. By Rado's theorem, we can find disjoint independent transversals $(T_1,\dots,T_m)$ of $(B_1-(Z\cup Z'),\dots,B_k-(Z\cup Z'))$.

Let $S = (B_1\cup\cdots\cup B_k)-(Z\cup Z'\cup T_1\cup\cdots\cup T_m)$. We can find disjoint independent subsets $(Y_1',\dots,Y_{k-1}')$ of S such that Y_i' is a transversal of $(B_1,\dots,B_{k-i})$. Let $S' = S-(Y_1'\cup\cdots\cup Y_{k-1}')$. We can then find disjoint independent subsets $(Y_1,\dots,Y_{k-1})$ of S' such that Y_i is a transversal of $(B_{i+1},B_{i+2},\dots,B_k)$; see Figure 1. By Lemma 3.1, we can partition $Z\cup Y_1\cup\cdots\cup Y_{k-1}$ and $Z'\cup Y_1'\cup\cdots\cup Y_{k-1}'$ into independent transversals of $(B_1,\dots,B_k)$. ◻

Acknowledgment. We thank the referees for their constructive comments and suggestions.

REFERENCES

[1] R. HUANG AND G.-C. ROTA, *On the relations of various conjectures on Latin squares and straightening coefficients*, Discrete Math., 128 (1994), pp. 225–236.

[2] J. G. OXLEY, *Matroid Theory*, Oxford University Press, New York, 1992.

[3] R. RADO, *A theorem on independence relations*, Quart. J. Math., Oxford Ser., 13 (1942), pp. 83–89.

SIAM J. Discrete Math.
Vol. 21, No. 3, pp. 805–821

ON EXTREMAL k-GRAPHS WITHOUT REPEATED COPIES OF 2-INTERSECTING EDGES[*]

YEOW MENG CHEE[†] AND ALAN C. H. LING[‡]

Abstract. The problem of determining extremal hypergraphs containing at most r isomorphic copies of some element of a given hypergraph family was first studied by Boros et al. in 2001. There are not many hypergraph families for which exact results are known concerning the size of the corresponding extremal hypergraphs, except for those equivalent to the classical Turán numbers. In this paper, we determine the size of extremal k-uniform hypergraphs containing at most one pair of 2-intersecting edges for $k \in \{3, 4\}$. We give a complete solution when $k = 3$ and an almost complete solution (with eleven exceptions) when $k = 4$.

Key words. combinatorial design, hypergraph, packing

AMS subject classifications. 05B05, 05B07, 05B40, 05D05

DOI. 10.1137/060675915

1. Introduction.

A *set system* is a pair $G = (X, \mathcal{A})$, where X is a finite set and $\mathcal{A} \subseteq 2^X$. The members of X are called *vertices* or *points*, and the members of $\mathcal{A}$ are called *edges* or *blocks*. The *order* of G is the number of vertices $|X|$, and the *size* of G is the number of edges $|\mathcal{A}|$. The set K is called a *set of block sizes* for G if $|A| \in K$ for all $A \in \mathcal{A}$. G is called a *k-uniform hypergraph* (or *k-graph*) if $\{k\}$ is a set of block sizes for G. A 2-graph is also known simply as a *graph*.

A pair of edges is said to be *t-intersecting* if they intersect in at least t points. The k-graph of size two whose two edges intersect in exactly t points is denoted $\Lambda(k, t)$.

Let $\mathcal{F}$ be a family of k-graphs. Boros et al. [2] introduced the function $T(n, \mathcal{F}, r)$, which denotes the maximum number of edges in a k-graph of order n containing no r isomorphic copies of a member of $\mathcal{F}$. So $T(n, \mathcal{F}, 1)$ is just the classical Turán number $\mathrm{ex}(n, \mathcal{F})$ [1]. A family of k-graphs $\mathcal{F}$ is said to *grow polynomially* if there exist $c > 0$ and a nonnegative integer s such that, for every m, there are at most cm^s members in $\mathcal{F}$ having exactly m edges. The following theorem is established in [2].

Theorem 1.1 (Boros et al. [2]). *Let $\mathcal{F}$ be a family of k-graphs which grows polynomially with parameters c and s. Then, for n sufficiently large,*

$$T(n, \mathcal{F}, r) < \mathrm{ex}(n, \mathcal{F}) + (c \cdot (r-1) \cdot s! + 1)\mathrm{ex}(n, F)^{(s+1)/(s+2)} \\ + 2(c \cdot (r-1) \cdot s! + 1)^2 \mathrm{ex}(n, \mathcal{F})^{s/(s+2)}.$$

For $k \geq 3$, let $\mathcal{F}(k)$ be the family of k-graphs of two 2-intersecting edges; that is, $\mathcal{F}(k) = \{\Lambda(k, t) : 2 \leq t \leq k-1\}$. $T(n, \mathcal{F}(k), 1)$, which is the Turán number

[*] Received by the editors November 26, 2006; accepted for publication (in revised form) July 20, 2007; published electronically October 31, 2007.
http://www.siam.org/journals/sidma/21-3/67591.html

[†] Interactive Digital Media R&D Program Office, Media Development Authority, 140 Hill Street, 179369 Singapore, the Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, 637616 Singapore, and the Department of Computer Science, School of Computing, National University of Singapore, 117590 Singapore (ymchee@alumni.uwaterloo.ca). The research of this author was supported by the Singapore Ministry of Education Research Grant T206B2204.

[‡] Department of Computer Science, University of VT, Burlington, VT 05405 (aling@emba.uvm.edu).

$\mathrm{ex}(n, \mathcal{F}(k))$, is equal to the following well studied parameters in design theory and coding theory:

- $D(n, k, 2)$, the maximum number of blocks in a 2-$(n, k, 1)$ packing [11], and
- $A(n, 2(k-1), k)$, the maximum number of codewords in a binary code of length n, minimum distance $2(k-1)$, and constant weight k [10].

Despite much effort, the exact value of $T(n, \mathcal{F}(k), 1)$ is known for all n only when $k = 3$ [14, 15] and $k = 4$ [3]. Even for $k = 5$, there are an infinite number of n for which $T(n, \mathcal{F}(5), 1)$ is not yet determined. In this paper, we determine $T(n, \mathcal{F}(k), 2)$ for all n when $k = 3$ and for all but 11 values of n when $k = 4$.

2. Design-theoretic preliminaries.

Our determination of $T(n, \mathcal{F}(k), 2)$, $k \in \{3, 4\}$, makes extensive use of combinatorial designs. In this section, we review some design-theoretic constructs and review some prior results that are needed in our solution.

For positive integers $i \leq j$, the set $\{i, i+1, \ldots, j\}$ is denoted $[i, j]$. The set $[1, j]$ is further abbreviated as $[j]$. A k-graph $(X, \mathcal{A})$ of order n is a packing of pairs by k-tuples, or more commonly known as a 2-$(n, k, 1)$ *packing* if every 2-subset of X is contained in at most one block of $\mathcal{A}$. The *leave* of $(X, \mathcal{A})$ is the graph $L = (X, \mathcal{E})$, where $\mathcal{E}$ consists of all 2-subsets of X that are not contained in any blocks of $\mathcal{A}$. We also say that $(X, \mathcal{A})$ is a 2-$(n, k, 1)$ packing *leaving* L. Given a graph G, the maximum size of a 2-$(n, k, 1)$ packing whose leave contains G is denoted $m(n, k, G)$. Note that the maximum size of a 2-$(n, k, 1)$ packing, $D(n, k, 2)$, is the quantity $m(n, k, G)$ when G is the empty graph.

Theorem 2.1 (Schönheim [14], Spencer [15]). *For all $n \geq 0$, we have*

$$D(n,3,2) = \begin{cases} \left\lfloor \frac{n}{3} \left\lfloor \frac{n-1}{2} \right\rfloor \right\rfloor - 1 & \text{if } n \equiv 5 \pmod 6, \\ \left\lfloor \frac{n}{3} \left\lfloor \frac{n-1}{2} \right\rfloor \right\rfloor & \text{otherwise.} \end{cases}$$

Theorem 2.2 (Brouwer [3]). *For all $n \geq 0$, we have*

$$D(n,4,2) = \begin{cases} \left\lfloor \frac{n}{4} \left\lfloor \frac{n-1}{3} \right\rfloor \right\rfloor - 1 & \text{if } n \equiv 7 \text{ or } 10 \pmod{12} \text{ and } n \notin \{10, 19\}, \\ \left\lfloor \frac{n}{4} \left\lfloor \frac{n-1}{3} \right\rfloor \right\rfloor - 1 & \text{if } n \in \{9, 17\}, \\ \left\lfloor \frac{n}{4} \left\lfloor \frac{n-1}{3} \right\rfloor \right\rfloor - 2 & \text{if } n \in \{8, 10, 11\}, \\ \left\lfloor \frac{n}{4} \left\lfloor \frac{n-1}{3} \right\rfloor \right\rfloor - 3 & \text{if } n = 19, \\ \left\lfloor \frac{n}{4} \left\lfloor \frac{n-1}{3} \right\rfloor \right\rfloor & \text{otherwise.} \end{cases}$$

A *pairwise balanced design* (PBD) is a set system $(X, \mathcal{A})$ such that every 2-subset of X is contained in exactly one block of $\mathcal{A}$. If a PBD is of order n and has a set of block sizes K, we denote it by $\mathrm{PBD}(n, K)$. If a member $k \in K$ is superscripted with a "$\star$" (written "$k^\star$"), it means that the PBD has exactly one block of size k. We require the following result on the existence of PBDs.

Theorem 2.3 (Fort and Hedlund [5]). *There exists a* $\mathrm{PBD}(n, \{3, 5^\star\})$ *if and only if* $n \equiv 5 \pmod 6$.

Theorem 2.4 (Rees and Stinson [13]). *There exists a* $\mathrm{PBD}(n, \{4, f^\star\})$ *if and only if* $n \geq 3f + 1$, *and*

(i) $n \equiv 1$ *or* $4 \pmod{12}$ *and* $f \equiv 1$ *or* $4 \pmod{12}$ *or*

(ii) $n \equiv 7$ *or* $10 \pmod{12}$ *and* $f \equiv 7$ *or* $10 \pmod{12}$.

Let $(X, \mathcal{A})$ be a set system, and let $\mathcal{G} = \{G_1, \ldots, G_s\}$ be a partition of X into subsets, called *groups*. The triple $(X, \mathcal{G}, \mathcal{A})$ is a *group divisible design* (GDD) when every 2-subset of X not contained in a group appears in exactly one block, and

$|A \cap G| \leq 1$ for all $A \in \mathcal{A}$ and $G \in \mathcal{G}$. We denote a GDD $(X, \mathcal{G}, \mathcal{A})$ by K-GDD if K is a set of block sizes for $(X, \mathcal{A})$. The *type* of a GDD $(X, \mathcal{G}, \mathcal{A})$ is the multiset $[|G| : G \in \mathcal{G}]$. When more convenient, we use the exponentiation notation to describe the type of a GDD: A GDD of type $g_1^{t_1} \dots g_s^{t_s}$ is a GDD where there are exactly t_i groups of size g_i, $i \in [s]$. The following results on the existence of $\{4\}$-GDDs are useful.

Theorem 2.5 (Hanani [7]). *There exists a* $\{3\}$-GDD *of type* g^t *if and only if* $t \geq 3$, $g^2\binom{t}{2} \equiv 0 \pmod 3$, *and* $g(t-1) \equiv 0 \pmod 2$.

Theorem 2.6 (Brouwer, Schrijver, and Hanani [4]). *There exists a* $\{4\}$-GDD *of type* g^t *if and only if* $t \geq 4$ *and*

(i) $g \equiv 1$ *or* $5 \pmod 6$ *and* $t \equiv 1$ *or* $4 \pmod{12}$ *or*
(ii) $g \equiv 2$ *or* $4 \pmod 6$ *and* $t \equiv 1 \pmod 3$ *or*
(iii) $g \equiv 3 \pmod 6$ *and* $t \equiv 0$ *or* $1 \pmod 4$ *or*
(iv) $g \equiv 0 \pmod 6$,

with the two exceptions of types 2^4 *and* 6^4, *for which* $\{4\}$-GDD*s do not exist.*

Theorem 2.7 (Brouwer [3]). *A* $\{4\}$-GDD *of type* $2^u 5^1$ *exists if and only if* $u = 0$, *or* $u \equiv 0 \pmod 3$ *and* $u \geq 9$.

Theorem 2.8 (see [9]). *There exists a* $\{4\}$-GDD *of type* $3^t u^1$ *if and only if* $t = 0$, *or* $t \geq (2u+3)/3$ *and*

(i) $t \equiv 0$ *or* $1 \pmod 4$ *and* $u \equiv 0$ *or* $6 \pmod{12}$ *or*
(ii) $t \equiv 0$ *or* $3 \pmod 4$ *and* $u \equiv 3$ *or* $9 \pmod{12}$.

Theorem 2.9 (Ge and Ling [6]). *There exists a* $\{4\}$-GDD *of type* $2^t u^1$ *for* $t = 0$ *and for each* $t \geq 6$ *with* $t \equiv 0 \pmod 3$, $u \equiv 2 \pmod 3$, *and* $2 \leq u \leq t-1$, *except for* $(t, u) = (6, 5)$ *and except possibly for* $(t, u) \in \{(21, 17), (33, 23), (33, 29), (39, 35), (57, 44)\}$.

Theorem 2.10 (Ge and Ling [6]). *There exists a* $\{4\}$-GDD *of type* $12^t u^1$ *for* $t = 0$ *and for each* $t \geq 4$ *and* $u \equiv 0 \pmod 4$ *such that* $0 \leq u \leq 6(t-1)$.

An *incomplete transversal design of group size* n, *block size* k, and *hole size* h is a quadruple $(X, \mathcal{G}, H, \mathcal{A})$ such that

(i) $(X, \mathcal{A})$ is a k-graph of order nk;
(ii) $\mathcal{G}$ is a partition of X into k subsets (called *groups*), each of cardinality n;
(iii) $H \subseteq X$, with the property that, for each $G \in \mathcal{G}$, $|G \cap H| = h$; and
(iv) every 2-subset of X is
- contained in the *hole* H and not contained in any blocks or
- contained in a group and not contained in any blocks or
- contained in neither a hole nor a group and contained in exactly one block of $\mathcal{A}$.

Such an incomplete transversal design is denoted $\mathrm{TD}(k, n) - \mathrm{TD}(k, h)$.

Theorem 2.11 (Heinrich and Zhu [8]). *For* $n > h > 0$, *a* $\mathrm{TD}(4, n) - \mathrm{TD}(4, h)$ *exists if and only if* $n \geq 3h$ *and* $(n, h) \neq (6, 1)$.

3. Packings with leaves containing specified graphs. In this section, we relate the problem of determining $T(n, \mathcal{F}(k), 2)$ to that of determining $m(n, k, G)$ for G isomorphic to $K_4 - e$, $K_5 - e$, and $2 \circ K_4$ (edge-gluing of two K_4's) when $k \in \{3, 4\}$. These graphs are shown in Figures 3.1–3.3, respectively.

Lemma 3.1. *There exists a* 3-*graph of order* n *and size* m *containing exactly one copy of an element of* $\mathcal{F}(3)$ *if and only if there exists a* 2-$(n, 3, 1)$ *packing of size* $m-2$ *with a leave containing* $K_4 - e$ *as a subgraph.*

Proof. $\mathcal{F}(3)$ contains only a single 3-graph, $\Lambda(3, 2)$. Let $(X, \mathcal{A})$ be a 3-graph of order n and size m containing exactly one copy of $\Lambda(3, 2)$. Then there exist exactly two blocks $A, B \in \mathcal{A}$, with $|A \cap B| = 2$. Let $P = (X, \mathcal{A} \setminus \{A, B\})$. Then P is a

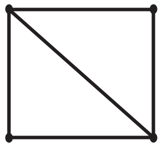


FIG. 3.1. $K_4 - e$.

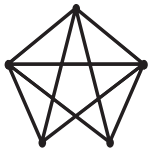


FIG. 3.2. $K_5 - e$.

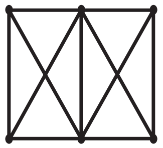


FIG. 3.3. $2 \circ K_4$.

2-$(n, 3, 1)$ packing of size $m - 2$ with a leave containing the 2-subsets in X that occurs in A and B, which together form a $K_4 - e$. This construction is reversible. □

COROLLARY 3.2. *The following holds:*

$$T(n, \mathcal{F}(3), 2) = \max\{T(n, \mathcal{F}(3), 1), m(n, 3, K_4 - e) + 2\}.$$

Proof. If a 3-graph contains no two isomorphic copies of $\Lambda(3, 2)$, then either it contains no copies, in which case its maximum size is given by $T(n, \mathcal{F}(3), 1)$, or else it contains exactly one copy, in which case its maximum size is given by $m(n, 3, K_4 - e) + 2$. □

The proofs for the following two lemmas are similar to that for Lemma 3.1 and are thus omitted.

LEMMA 3.3. *There exists a* 4*-graph of order* n *and size* m *containing exactly one copy of* $\Lambda(4, 2)$ *if and only if there exists a* 2-$(n, 4, 1)$ *packing of size* $m - 2$ *with a leave containing* $2 \circ K_4$ *as a subgraph.*

LEMMA 3.4. *There exists a* 4*-graph of order* n *and size* m *containing exactly one copy of* $\Lambda(4, 3)$ *if and only if there exists a* 2-$(n, 4, 1)$ *packing of size* $m - 2$ *with a leave containing* $K_5 - e$ *as a subgraph.*

COROLLARY 3.5. *The following holds:*

$$T(n, \mathcal{F}(4), 2) = \max\{T(n, \mathcal{F}(4), 1), m(n, 4, 2 \circ K_4) + 2, m(n, 4, K_5 - e) + 2\}.$$

Proof. $\mathcal{F}(4)$ contains the graphs $\Lambda(4, 2)$ and $\Lambda(4, 3)$. So if a 4-graph contains no two isomorphic copies of an element of $\mathcal{F}(4)$, then either it contains none of them, in which case its maximum size is given by $T(n, \mathcal{F}(4), 1)$, or else it contains exactly one of $\Lambda(4, 2)$ or $\Lambda(4, 3)$. In the former case, its maximum size is $m(n, 4, 2 \circ K_4) + 2$ by Lemma 3.3, and, in the latter case, its maximum size is $m(n, 4, K_5 - e)$ by Lemma 3.4. □

4. Determining $T(n,\mathcal{F}(3),2)$. When $n \equiv 1$ or $3 \pmod 6$, a 2-$(n,3,1)$ packing of size $T(n,\mathcal{F}(3),1)$ has the property that every pair of distinct points is contained in exactly one block. Such a 2-$(n,3,1)$ packing is called a *Steiner triple system* of order n and is denoted $\mathrm{STS}(n)$.

Let $P=(X,\mathcal{A})$ be a 2-$(n,3,1)$ packing. When $n \equiv 1$ or $3 \pmod 6$, the leave $L=(X,\mathcal{E})$ of P must satisfy:

(i) $|\mathcal{E}| \equiv 0 \pmod 3$, and

(ii) the degree of every vertex in L is even.

Any L containing $K_4 - e$ as a subgraph and satisfying conditions (i) and (ii) above has at least nine edges. Hence, the maximum size of a 2-$(n,3,1)$ packing with a leave containing K_4-e is at most $\frac{1}{3}(\binom{n}{2}-9)$. We show below that there indeed exists such a 2-$(n,3,1)$ packing of size $\frac{1}{3}(\binom{n}{2}-9)$.

Lemma 4.1. *There exists a* 2-$(n,3,1)$ *packing of size* $\frac{1}{3}(\binom{n}{2}-9)$*, with a leave containing* K_4-e*, for every* $n \equiv 1$ *or* $3 \pmod 6$.

Proof. Let $(X,\mathcal{A})$ be an $\mathrm{STS}(n)$. Suppose there exist three blocks in $\mathcal{A}$ of the form $\{1,2,3\}$, $\{1,4,5\}$, and $\{3,4,a\}$. Then deleting these three blocks gives a 2-$(n,3,1)$ packing of size $\frac{1}{3}(\binom{n}{2}-9)$ with a leave containing K_4-e. Hence, it suffices to show that we can always find such a 3-block configuration in any $\mathrm{STS}(n)$. To see that this is true, pick any two intersecting blocks in an $\mathrm{STS}(n)$, say, $\{1,2,3\}$ and $\{1,4,5\}$. As the third block, take the unique block containing the 2-subset $\{3,4\}$. □

Next, we consider $n \equiv 5 \pmod 6$. In this case, $\binom{n}{2} \equiv 1 \pmod 3$. So if the leave of a 2-$(n,3,1)$ packing contains K_4-e, then it must contain at least seven edges. Therefore, such a packing can have at most $\frac{1}{3}(\binom{n}{2}-7)$ blocks. We show below that this upper bound can be met using pairwise balanced designs.

Lemma 4.2. *There exists a* 2-$(n,3,1)$ *packing of size* $\frac{1}{3}(\binom{n}{2}-7)$*, with a leave containing* K_4-e*, for every* $n \equiv 5 \pmod 6$.

Proof. Let $(X,\mathcal{A})$ be a $\mathrm{PBD}(n,\{3,5^\star\})$ with $[5]$ as the block of size five. The existence of such a PBD is provided by Theorem 2.3. Deleting the block of size five from this PBD and adding the block $\{1,2,3\}$ yield the desired 2-$(n,3,1)$ packing. □

For $n \equiv 0$, 2, or $4 \pmod 6$, every vertex in the leave L of a 2-$(n,3,1)$ packing is of odd degree. If L contains K_4-e, then L must have at least four vertices of degree at least three. The minimum possible number of edges in L, if L contains K_4-e, is therefore $n/2+4$. It follows that the number of blocks in a 2-$(n,3,1)$ packing with a leave containing K_4-e is at most $\left\lfloor \frac{1}{3}(\binom{n}{2}-\frac{n}{2}-4)\right\rfloor$.

Lemma 4.3. *There exists a* 2-$(n,3,1)$ *packing of size* $\frac{1}{3}(\binom{n}{2}-\frac{n}{2}-4)$*, with a leave containing* K_4-e*, for every* $n \equiv 4 \pmod 6$.

Proof. Let $(X,\mathcal{A})$ be a $\mathrm{PBD}(n+1,\{3,5^\star\})$ which exists by Theorem 2.3. Let x be a point contained in the block of size five. Then $(X\setminus\{x\},\mathcal{B})$, where

$$\mathcal{B}=\{A\in\mathcal{A} : x\notin A \text{ and } |A|=3\}$$

is the desired 2-$(n,3,1)$ packing. □

Lemma 4.4. *There exists a* 2-$(n,3,1)$ *packing of size* $\frac{1}{3}(\binom{n}{2}-\frac{n}{2}-6)$*, with a leave containing* K_4-e*, for every* $n \equiv 0$ *or* $2 \pmod 6$.

Proof. Consider a $\{3\}$-GDD of type $2^{n/2}$, which exists whenever $n \equiv 0$ or $2 \pmod 6$ by Theorem 2.5. Without loss of generality, we may assume $\{1,2\}$ is a group and $\{1,3,4\}$ is a block in this GDD. There is a unique block of the form $\{2,3,a\}$. Deleting the blocks $\{1,3,4\}$ and $\{2,3,a\}$ from this GDD gives a 2-$(n,3,1)$ packing of size $\frac{1}{3}(\binom{n}{2}-\frac{n}{2}-6)$, with a leave containing K_4-e. □

This completes our determination of $m(n,3,K_4-e)$. We summarize our results above as follows.

THEOREM 4.5. *For all $n \geq 0$, we have $m(n,3,K_4-e) = \frac{1}{3}(\binom{n}{2} - f(n))$, where*

$$f(n) = \begin{cases} n/2+6 & \text{if } n \equiv 0 \text{ or } 2 \pmod 6, \\ 9 & \text{if } n \equiv 1 \text{ or } 3 \pmod 6, \\ n/2+4 & \text{if } n \equiv 4 \pmod 6, \\ 7 & \text{if } n \equiv 5 \pmod 6. \end{cases}$$

5. Determining $T(n,\mathcal{F}(4),2)$.

We now determine $T(n,\mathcal{F}(4),2)$.

5.1. The case $n \equiv 1$ or $4 \pmod{12}$.

The leave $L = (X,\mathcal{E})$ of a 2-$(n,4,1)$ packing must satisfy:

(i) $|\mathcal{E}| \equiv 0 \pmod 6$, and

(ii) every vertex in L has degree $\equiv 0 \pmod 3$.

Any leave of P containing K_5-e or $2 \circ K_4$ as a subgraph and satisfying conditions (i) and (ii) above has at least 18 edges. So $m(n,4,G) \leq \frac{1}{6}(\binom{n}{2} - 18)$ for $G \in \{K_5-e, 2\circ K_4\}$. We show below that this bound can be met with a finite number of possible exceptions.

The *cocktail party graph* CP(n) is the unique $(2n-2)$-regular graph on $2n$ vertices. We begin with an observation on CP(4) (shown in Figure 5.1).

LEMMA 5.1. CP(4) *contains an edge-disjoint union of a K_5-e and a K_4.*

Proof. Without loss of generality, we may take the vertex set and edge set of the CP(4) as $[8]$ and $\{A \subset [8] : |A| = 2\} \setminus \{\{i,i+4\} : i \in [4]\}$, respectively. Consider the subsets of edges $\mathcal{E}_1 = \{A \subset \{1,2,3,5,8\} : |A| = 2\} \setminus \{\{1,5\}\}$ and $\mathcal{E}_2 = \{A \subset \{2,4,6,7\} : |A| = 2\}$. $\mathcal{E}_1$ is the edge set of a K_5-e, $\mathcal{E}_2$ is the edge set of a K_4, and they are disjoint. □

LEMMA 5.2. CP(4) *contains an edge-disjoint union of a $2\circ K_4$ and a K_4.*

Proof. Without loss of generality, we may take the vertex set and edge set of the CP(4) as $[8]$ and $\{A \subset [8] : |A| = 2\} \setminus \{\{i,i+4\} : i \in [4]\}$, respectively. Consider the subsets of edges $\mathcal{E}_1 = \{A \subset [4] : |A| = 2\} \cup (\{A \subset [3,6] : |A| = 2\} \setminus \{\{3,4\}\})$ and $\mathcal{E}_2 = \{A \subset \{1,6,7,8\} : |A| = 2\}$. $\mathcal{E}_1$ is the edge set of a $2\circ K_4$, $\mathcal{E}_2$ is the edge set of a K_4, and they are disjoint. □

LEMMA 5.3. *Let $G \in \{K_5-e, 2\circ K_4\}$ and $n \equiv 1$ or $4 \pmod{12}$. If there exists a 2-$(n,4,1)$ packing leaving* CP(4)*, then there exists a 2-$(n,4,1)$ packing of size $\frac{1}{6}(\binom{n}{2}-18)$ with a leave containing G.*

Proof. A 2-$(n,4,1)$ packing whose leave is CP(4) has size $\frac{1}{6}(\binom{n}{2}-24)$. We have seen from Lemmas 5.1 and 5.2 that we can add one more block of size four to this packing to give a 2-$(n,4,1)$ packing with a leave containing G. □

In view of the above lemma, we now focus on constructing 2-$(n,4,1)$ packings leaving CP(4).

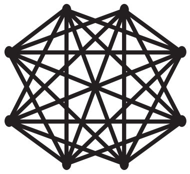


FIG. 5.1. CP(4).

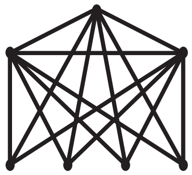


FIG. 5.2. $K_{3,4}+3e$.

LEMMA 5.4. *Let $n \geq 6$. If there exists a* $\mathrm{PBD}(n+f,\{4,f^\star\})$*, then there exists a* 2-$(4n+f,4,1)$ *packing leaving* CP(4).

Proof. Take a $\mathrm{TD}(4,n){-}\mathrm{TD}(4,2)$ $(X,\mathcal{G},H,\mathcal{A})$, which exists by Theorem 2.11, and for each $G\in\mathcal{G}$, let $(G\cup F,\mathcal{A}_G)$ be a $\mathrm{PBD}(n+f,\{4,f^\star\})$, where F is the block of size f in the PBD. Consider the set system $(Y,\mathcal{B})$, where $Y=X\cup F$, and $\mathcal{B}=\mathcal{A}\cup(\cup_{G\in\mathcal{G}}\mathcal{A}_G)$ (note that the block of size F is included only once). $(Y,\mathcal{B})$ is a 4-graph of order $4n+f$ having the property that every 2-subset of $X\cup F$ is contained in exactly one block of $\mathcal{B}$, except for those 2-subsets $\{a,b\}$, with $a\in G\cap H$ and $b\in G'\cap H$ for distinct $G,G'\in\mathcal{G}$, which are not contained in any blocks of $\mathcal{B}$. $(Y,\mathcal{B})$ therefore gives the required 2-$(4n+f,4,1)$ packing leaving CP(4). □

LEMMA 5.5. *Let $n\equiv 1$ or 4 (mod 12) such that $n\geq 40$ and $n\notin\{73,76,85\}$. Then there exists a* 2-$(n,4,1)$ *packing leaving* CP(4).

Proof. Taking a $\mathrm{PBD}(n+f,\{4,f^\star\})$, with $(n,f)\in\{(9,4), (12,1), (13,0), (15,1), (16,0), (21,4), (24,1), (25,0), (27,1), (28,0)\}$, whose existence is provided by Theorem 2.4, and applying Lemma 5.4 give 2-$(n,4,1)$ packings leaving CP(4) for $n\in\{40, 49, 52, 61, 64, 88, 97, 100, 109, 112\}$. By Theorem 2.4, there exists a $\mathrm{PBD}(n,\{4,40^\star\})$ for all $n\equiv 1$ or 4 (mod 12) and $n\geq 121$. Break up the block of size 40 in this PBD with the blocks of a 2-$(40,4,1)$ packing leaving CP(4) to obtain a 2-$(n,4,1)$ packing leaving CP(4). □

COROLLARY 5.6. *Let $n\equiv 1$ or 4 (mod 12) such that $n\geq 40$ and $n\notin\{73,76,85\}$. Then $m(n,4,G)=\frac{1}{6}(\binom{n}{2}-18)$ for $G\in\{K_5-e,2\circ K_4\}$.*

5.2. The case $n\equiv 7$ or 10 (mod 12). The leave $L=(X,\mathcal{E})$ must satisfy:

(i) $|\mathcal{E}|\equiv 3$ (mod 6), and

(ii) every vertex in L has degree $\equiv 0$ (mod 3).

We first consider the case when L contains K_5-e. Any such L satisfying the conditions (i) and (ii) above must have at least 15 edges. So $m(n,4,K_5-e)\leq\frac{1}{6}(\binom{n}{2}-15)$.

When L contains $2\circ K_4$, L must also have at least 15 edges. Suppose L contains $2\circ K_4$ and has 15 edges. Then L must have at least two vertices, each of degree at least six. Let a be the number of degree three vertices, and let b be the number of vertices with degree greater than three in L. Then we have $3a+6b\leq 30$ (counting the edges), $b\geq 2$ (considering the two vertices of degree five in $2\circ K_4$), and $a+b\geq 7$ (considering the presence of vertices with degree at least six). These inequalities imply that $2\leq b\leq 3$ and $a+b\leq 8$. So the possible degree sequences for L are $\mathcal{D}_1=(6,6,6,3,3,3,3)$ and $\mathcal{D}_2=(6,6,3,3,3,3,3,3)$. Note that we suppress including vertices of degree zero in the degree sequence of L. There is a unique graph with degree sequence $\mathcal{D}_1$, namely, the graph in Figure 5.2, obtained by adding to $K_{3,4}$ three edges connecting the vertices in the part of the bipartition with three vertices. This graph does not contain $2\circ K_4$. Hence, L cannot have degree sequence $\mathcal{D}_1$. If L contains $2\circ K_4$ and has degree sequence $\mathcal{D}_2$, then since $2\circ K_4$ has degree sequence

$(5,5,3,3,3,3)$, the two vertices of nonzero degree not in $2\circ K_4$ cannot both be adjacent to the two vertices of degree five in $2\circ K_4$. But this prevents these two vertices having degree three, a contradiction. Hence L cannot have degree sequence $\mathcal{D}_2$. It follows that the leave of any 2-$(n,4,1)$ packing containing $2\circ K_4$ must have at least 21 edges, and we have $m(n,4,2\circ K_4)\leq\frac{1}{6}(\binom{n}{2}-21)$.

The following shows that these bounds can be met.

Lemma 5.7. *K_7 contains an edge-disjoint union of a K_5-e and a K_4.*

Proof. Take the vertex set of the K_7 as $[7]$. Consider the subsets of edges $\mathcal{E}_1=\{A\subset[5]:|A|=2\}\setminus\{\{4,5\}\}$ and $\mathcal{E}_2=\{A\subset[4,7]:|A|=2\}$. Then $\mathcal{E}_1$ is the edge set of a K_5-e, $\mathcal{E}_2$ is the edge set of a K_4, and they are disjoint. □

Lemma 5.8. *Let $n\equiv 7$ or 10 (mod 12) such that $n\geq 7$ and $n\notin\{10,19\}$. Then $m(n,4,K_5-e)=\frac{1}{6}(\binom{n}{2}-15)$.*

Proof. Let $(X,\mathcal{A})$ be a PBD$(n,\{4,7^\star\})$ with F as the block of size seven, whose existence is provided by Theorem 2.4, and let B be any 4-subset of F. Then $(X,(\mathcal{A}\cup\{B\})\setminus\{F\})$ is a 2-$(n,4,1)$ packing of size $\frac{1}{6}(\binom{n}{2}-15)$ leaving K_7-K_4, which contains K_5-e by Lemma 5.7. □

Lemma 5.9. *Let $n\equiv 7$ or 10 (mod 12) such that $n\geq 7$ and $n\notin\{10,19\}$. Then $m(n,4,2\circ K_4)=\frac{1}{6}(\binom{n}{2}-21)$.*

Proof. Observe that any 2-$(n,4,1)$ packing leaving K_7 has size $\frac{1}{6}(\binom{n}{2}-21)$. The theorem now follows for $n=7$ trivially and for $n\geq 22$ from the existence of a PBD$(n,\{4,7^\star\})$ provided by Theorem 2.4. □

5.3. The case $n\equiv 2$, 5, 8, or 11 (mod 12).

The leave $L=(X,\mathcal{E})$ must have vertices all of degree 1 (mod 3). Furthermore, $|\mathcal{E}|\equiv 1$ (mod 6) when $n\equiv 2$ or 11 (mod 12), and $|\mathcal{E}|\equiv 4$ (mod 6) when $n\equiv 5$ or 8 (mod 12).

If L contains K_5-e, then L must have at least five vertices, each of degree at least four and the remaining vertices each of degree at least one. Hence, L must have at least $\frac{1}{2}(n+15)$ edges when $n\equiv 5$ or 11 (mod 12) and at least $\frac{1}{2}(n+24)$ edges when $n\equiv 2$ or 8 (mod 12). Consequently,

$$m(n,4,K_5-e)\leq\begin{cases}\frac{1}{6}(\binom{n}{2}-\frac{n+15}{2}) & \text{if } n\equiv 5 \text{ or } 11 \pmod{12},\\ \frac{1}{6}(\binom{n}{2}-\frac{n+24}{2}) & \text{if } n\equiv 2 \text{ or } 8 \pmod{12}.\end{cases}$$

If L contains $2\circ K_4$, then L must have at least two vertices, each of degree at least seven, at least four vertices each of degree at least four, and the rest of the vertices each of degree one. Hence, L must have at least $\frac{1}{2}(n+24)$ edges when $n\equiv 2$ or 8 (mod 12) and at least $\frac{1}{2}(n+27)$ edges when $n\equiv 5$ or 11 (mod 12). Consequently,

$$m(n,4,2\circ K_4)\leq\begin{cases}\frac{1}{6}(\binom{n}{2}-\frac{n+24}{2}) & \text{if } n\equiv 2 \text{ or } 8 \pmod{12},\\ \frac{1}{6}(\binom{n}{2}-\frac{n+27}{2}) & \text{if } n\equiv 5 \text{ or } 11 \pmod{12}.\end{cases}$$

These bounds can be met with the following constructions.

5.3.1. The value of $m(n,4,K_5-e)$.

Lemma 5.10. *Let $n\equiv 5$ or 11 (mod 12) such that $n=5$ or $n\geq 23$. Then we have $m(n,4,K_5-e)=\frac{1}{6}(\binom{n}{2}-\frac{1}{2}(n+15))$.*

Proof. Let $(X,\mathcal{G},\mathcal{A})$ be a $\{4\}$-GDD of type $2^{(n-5)/2}5^1$, which exists by Theorem 2.7. Then $(X,\mathcal{A})$ is a 2-$(n,4,1)$ packing of size $\frac{1}{6}(\binom{n}{2}-\frac{1}{2}(n+15))$ with a leave containing K_5, and hence K_5-e. □

Lemma 5.11. *There exists a 2-$(14,4,1)$ packing of size 12 having a leave containing K_5-e.*

Proof. Let $(X,\mathcal{A})$ be a maximum 2-$(13,4,1)$ packing, which has size 13 by Theorem 2.2. Let $\infty \notin X$ and $A \in \mathcal{A}$. Then $(X\cup\{\infty\},\mathcal{A}\setminus\{A\})$ is a 2-$(14,4,1)$ packing of size 12 with a leave containing K_5 (whose edges are the 2-subsets of $A\cup\{\infty\}$). □

LEMMA 5.12. *Let $n \equiv 2$ or $8 \pmod{12}$ such that $n = 14$ or $n \geq 44$. Then we have $m(n,4,K_5-e) = \frac{1}{6}(\binom{n}{2} - \frac{1}{2}(n+24))$.*

Proof. Let $(X,\mathcal{G},\mathcal{A})$ be a $\{4\}$-GDD of type $2^{(n-14)/2}14^1$, which exists by Theorem 2.9. Let $G \in \mathcal{G}$ be the group of cardinality 14, and let $(G,\mathcal{B})$ be a 2-$(14,4,1)$ packing of size 12 having a leave containing K_5-e, whose existence is provided by Theorem 5.11. Then $(X,\mathcal{A}\cup\mathcal{B})$ is a 2-$(n,4,1)$ packing having a leave containing K_5-e. The size of this packing is $\frac{1}{6}(\binom{n}{2} - \frac{1}{2}(n-14) - \binom{14}{2}) + 12 = \frac{1}{6}(\binom{n}{2} - \frac{1}{2}(n+24))$. □

5.3.2. The value of $m(n,4,2\circ K_4)$.

LEMMA 5.13. *If there exists a $\{4\}$-GDD of type $[g_1,\dots,g_s]$ with $s \geq 3$ and a $\{4\}$-GDD of type $2^{g_i/2+1}$ for each $i \in [s]$, then there exists a 2-$(n,4,1)$ packing of size $\frac{1}{6}(\binom{n}{2} - \frac{1}{2}(n+24))$ with a leave contaning $2\circ K_4$, where $n = 2+\sum_{i=1}^{s} g_i$.*

Proof. Suppose that $(X,\mathcal{G},\mathcal{A})$ is a $\{4\}$-GDD of type $[g_1,\dots,g_s]$, where $\mathcal{G} = \{G_1,\dots,G_s\}$ and $|G_i| = g_i$ for $i \in [s]$. Let $Y = \{\infty_1,\infty_2\}$, where $\infty_1,\infty_2 \notin X$, and let $(G_i\cup Y,\mathcal{H}_{G_i},\mathcal{A}_{G_i})$ be a $\{4\}$-GDD of type $2^{g_i/2+1}$ such that

$$\begin{cases} Y \in \mathcal{H}_{G_i} & \text{if } i \in [s-2],\\ Y \text{ is contained in a block } A_{G_i} \in \mathcal{A}_{G_i} & \text{if } i \in \{s-1,s\}.\end{cases}$$

Construct a 4-graph $(X\cup Y,\mathcal{B})$ of order $2+\sum_{i=1}^{s} g_i$, where

$$\mathcal{B} = \mathcal{A} \cup \left(\bigcup_{i=1}^{s} \mathcal{A}_{G_i}\right) \setminus \{A_{G_{s-1}}, A_{G_s}\}.$$

It is easy to see that $(X\cup Y,\mathcal{B})$ is a 2-$(2+\sum_{i=1}^{s} g_i,4,1)$ packing. Also, the 2-subsets of $A_{G_{s-1}}$ and A_{G_s} are not contained in any blocks of $\mathcal{B}$. So the leave of $(X\cup Y,\mathcal{B})$ contains $2\circ K_4$ as a subgraph. It remains to compute the size of $(X\cup Y,\mathcal{B})$. The 2-subsets of $X\cup Y$ that are not contained in any blocks of $\mathcal{B}$ are precisely the elements of $\mathcal{H}_{G_i}$ for $i \in [s]$ and the 2-subsets of $A_{G_{s-1}}$ and A_{G_s}. Since Y appears precisely s times among these 2-subsets, the total number of distinct 2-subsets of $X\cup Y$ that are not contained in any blocks of $\mathcal{B}$ is $\sum_{i=1}^{s}(g_i/2+1) + 12 - (s-1) = n/2+12$, where $n = 2+\sum_{i=1}^{s} g_i$. Hence $|\mathcal{B}| = \frac{1}{6}(\binom{n}{2} - \frac{1}{2}(n+24))$, as required. □

LEMMA 5.14. *If there exists a $\{4\}$-GDD of type $[g_1,\dots,g_s]$ with $s \geq 3$, a $\{4\}$-GDD of type $2^{g_i/2+1}$ for each $i \in [s-1]$, and a $\{4\}$-GDD of type $2^{(g_s-3)/2}5^1$, then there exists a 2-$(n,4,1)$ packing of size $\frac{1}{6}(\binom{n}{2} - \frac{1}{2}(n+27))$ with a leave containing $2\circ K_4$, where $n = 2+\sum_{i=1}^{s} g_i$.*

Proof. Suppose that $(X,\mathcal{G},\mathcal{A})$ is a $\{4\}$-GDD of type $[g_1,\dots,g_s]$, where $\mathcal{G} = \{G_1,\dots,G_s\}$ and $|G_i| = g_i$ for $i \in [s]$. Let $Y = \{\infty_1,\infty_2\}$, where $\infty_1,\infty_2 \notin X$, and let $(G_i\cup Y,\mathcal{H}_{G_i},\mathcal{A}_{G_i})$ be a $\{4\}$-GDD of type $2^{g_i/2+1}$ such that

$$\begin{cases} Y \in \mathcal{H}_{G_i} & \text{if } i \in [s-3],\\ Y \text{ is contained in a block } A_{G_i} \in \mathcal{A}_{G_i} & \text{if } i \in \{s-2,s-1\}.\end{cases}$$

Further, let $(G_s\cup Y,\mathcal{H}_{G_s},\mathcal{A}_{G_s})$ be a $\{4\}$-GDD of type $2^{(g_s-3)/2}5^1$ such that Y is contained in the group $H \in \mathcal{H}_{G_s}$ of cardinality five. Now form the 4-graph $(X\cup Y,\mathcal{B})$ of order $2+\sum_{i=1}^{s} g_i$, where

$$\mathcal{B} = \mathcal{A} \cup \left(\bigcup_{i=1}^{s} \mathcal{A}_{G_i}\right) \cup \{H\setminus\{\infty_1\}\} \setminus \{A_{G_{s-2}}, A_{G_{s-1}}\}.$$

It is easy to see that $(X \cup Y, \mathcal{B})$ is a 2-$(2 + \sum_{i=1}^{s} g_i, 4, 1)$ packing. Also, the 2-subsets of $A_{G_{s-1}}$ and A_{G_s} are not contained in any blocks of $\mathcal{B}$. So the leave of $(X \cup Y, \mathcal{B})$ contains $2 \circ K_4$ as a subgraph. It remains to compute the size of $(X \cup Y, \mathcal{B})$. The 2-subsets of $X \cup Y$ that are not contained in any blocks of $\mathcal{B}$ are precisely the 2-subsets of $A_{G_{s-2}}$ and $A_{G_{s-1}}$ and the 2-subsets of elements of $\mathcal{H}_{G_i}$ for $i \in [s]$, except for the 2-subsets of $H \setminus \{\infty_1\}$. Since Y appears precisely s times among these 2-subsets, the total number of distinct 2-subsets of $X \cup Y$ that are not contained in any blocks of $\mathcal{B}$ is $\sum_{i=1}^{s-1}(g_i/2 + 1) + (g_s - 3)/2 + (10 - 6) - 1 + 12 - (s - 1) = \frac{1}{2}(n + 27)$, where $n = 2 + \sum_{i=1}^{s} g_i$. Hence $|\mathcal{B}| = \frac{1}{6}(\binom{n}{2} - \frac{1}{2}(n + 27))$, as required. □

Corollary 5.15. *For all $n \equiv 2 \pmod{12}$, $n \geq 50$, there exists a 2-$(n,4,1)$ packing of size $\frac{1}{6}(\binom{n}{2} - \frac{1}{2}(n + 24))$ with a leave containing $2 \circ K_4$.*

Proof. Apply Lemma 5.13 with $\{4\}$-GDDs of type $12^{(n-2)/12}$ and type 2^7, which exist by Theorem 2.6. □

Corollary 5.16. *For $n = 29$ and for all $n \equiv 5 \pmod{12}$, $n \geq 101$, there exists a 2-$(n,4,1)$ packing of size $\frac{1}{6}(\binom{n}{2} - \frac{1}{2}(n + 27))$ with a leave containing $2 \circ K_4$.*

Proof. Apply Lemma 5.14 with $\{4\}$-GDDs of type $12^{(n-29)/12}27^1$, which exists by Theorem 2.10, $\{4\}$-GDDs of type 2^7, which exists by Theorem 2.6, and $\{4\}$-GDDs of type $2^{12}5^1$, which exists by Theorem 2.7. □

Corollary 5.17. *For $n = 20$ and for all $n \equiv 8 \pmod{12}$, $n \geq 68$, there exists a 2-$(n,4,1)$ packing of size $\frac{1}{6}(\binom{n}{2} - \frac{1}{2}(n + 24))$ with a leave containing $2 \circ K_4$.*

Proof. Apply Lemma 5.13 with $\{4\}$-GDDs of type $12^{(n-20)/12}18^1$, which exists by Theorem 2.10, and $\{4\}$-GDDs of types 2^7 and 2^{10}, which exists by Theorem 2.6. □

Corollary 5.18. *For $n = 23$ and for all $n \equiv 11 \pmod{12}$, $n \geq 83$, there exists a 2-$(n,4,1)$ packing of size $\frac{1}{6}(\binom{n}{2} - \frac{1}{2}(n + 27))$ with a leave containing $2 \circ K_4$.*

Proof. Apply Lemma 5.14 with $\{4\}$-GDDs of type $12^{(n-23)/12}21^1$, which exists by Theorem 2.10, $\{4\}$-GDDs of type 2^7, which exists by Theorem 2.6, and $\{4\}$-GDDs of type $2^9 5^1$, which exists by Theorem 2.7. □

5.4. The case $n \equiv 0$, 3, 6, or 9 (mod 12). The leave $L = (X, \mathcal{E})$ must have vertices all of degree 2 (mod 3). Furthermore, $|\mathcal{E}| \equiv 0 \pmod 6$ when $n \equiv 0$ or 9 (mod 12), and $|\mathcal{E}| \equiv 3 \pmod 6$ when $n \equiv 3$ or 6 (mod 12).

If L contains $K_5 - e$ or $2 \circ K_4$, then L must have at least six vertices each of degree at least five and the remaining vertices each of degree at least two. Hence, L must have at least $n + 9$ edges when $n \equiv 6$ or 9 (mod 12) and at least $n + 12$ edges when $n \equiv 0$ or 3 (mod 12). Consequently, for $G \in \{K_5 - e, 2 \circ K_4\}$, we have

$$m(n,4,G) \leq \begin{cases} \frac{1}{6}(\binom{n}{2} - (n+9)) & \text{if } n \equiv 6 \text{ or } 9 \pmod{12}, \\ \frac{1}{6}(\binom{n}{2} - (n+12)) & \text{if } n \equiv 0 \text{ or } 3 \pmod{12}. \end{cases}$$

These bounds can again be met with the following constructions.

Lemma 5.19. *For $n = 6$ and for all $n \equiv 6$ or 9 (mod 12), $n \geq 21$ there exists a 2-$(n,4,1)$ packing of size $\frac{1}{6}(\binom{n}{2} - (n + 9))$ with a leave containing G, where $G \in \{K_5 - e, 2 \circ K_4\}$.*

Proof. Let $(X, \mathcal{G}, \mathcal{A})$ be a $\{4\}$-GDD of type $3^{(n-6)/3}6^1$, which exists by Theorem 2.8. Then $(X, \mathcal{A})$ is a 2-$(n,4,1)$ packing with a leave containing K_6, and hence $K_5 - e$ and $2 \circ K_4$. The size of $(X, \mathcal{A})$ is easily verified: $|\mathcal{A}| = \frac{1}{6}(\binom{n}{2} - \frac{n-6}{3}\binom{3}{2} - \binom{6}{2}) = \frac{1}{6}(\binom{n}{2} - (n + 9))$. □

Lemma 5.20. *There exists a 2-$(15,4,1)$ packing of size 13 with a leave containing G, where $G \in \{K_5 - e, 2 \circ K_4\}$.*

Proof. The 13 blocks of a 2-$(15, 4, 1)$ packing with a leave containing $K_5 - e$ are

{2,6,13,14}, {3,6,9,10}, {4,7,9,13}, {4,5,6,12}, {1,6,11,15},
{3,7,11,14}, {2,7,8,15}, {1,8,9,14}, {3,12,13,15}, {2,9,11,12},
{1,7,10,12}, {5,10,14,15}, {5,8,11,13}.

The 13 blocks of a 2-$(15, 4, 1)$ packing with a leave containing $2 \circ K_4$ are

{1,8,12,13}, {6,8,11,14}, {4,6,9,15}, {3,7,8,9}, {2,8,10,15},
{2,9,13,14}, {4,5,7,14}, {1,6,7,10}, {1,5,11,15}, {2,7,11,12},
{4,10,11,13}, {3,12,14,15}, {5,9,10,12}. $\square$

LEMMA 5.21. *For all $n \equiv 0$ or $3 \pmod{12}$, $n \geq 48$, there exists a* 2-$(n, 4, 1)$ *packing of size $\frac{1}{6}(\binom{n}{2} - (n+12))$ with a leave containing G, where $G \in \{K_5 - e, 2 \circ K_4\}$.*

Proof. Let $(X, \mathcal{G}, \mathcal{A})$ be a $\{4\}$-GDD of type $3^{(n-15)/3}15^1$, which exists by Theorem 2.8. Let Y be the group of cardinality 15 in $\mathcal{G}$ and $(Y, \mathcal{B})$ be a 2-$(15, 4, 1)$ packing of size 13 with a leave containing G, which exists by Lemma 5.20. Then $(X, \mathcal{A} \cup \mathcal{B})$ is a 2-$(n, 4, 1)$ packing with a leave containing G. The size of $(X, \mathcal{A} \cup \mathcal{B})$ is easily verified: $|\mathcal{A} \cup \mathcal{B}| = \frac{1}{6}(\binom{n}{2} - \frac{n-12}{3}\binom{3}{2} - 2\binom{6}{2}) + 13 = \frac{1}{6}(\binom{n}{2} - (n+12))$. $\square$

5.5. Remaining small orders. The values of n for which $m(n, 4, K_5 - e)$ and $m(n, 4, 2 \circ K_4)$ remain undetermined are as follows:

	Unsettled n												
$m(n, 4, K_5 - e)$	8	9	10	11	12	13	16	17	18	19	20	24	25
	26	27	28	32	36	37	38	39	73	76	85		
$m(n, 4, 2 \circ K_4)$	8	9	10	11	12	13	14	16	17	18	19	24	25
	26	27	28	32	35	36	37	38	39	41	44	47	53
	56	59	65	71	73	76	77	85	89				

For $n = 19$, we have the following tighter upper bound.

LEMMA 5.22. *For $G \in \{K_5 - e, 2 \circ K_4\}$, we have $m(19, 4, G) \leq 24$.*

Proof. Suppose we have a 2-$(19, 4, 1)$ packing of size 25 with a leave containing G, and then we can add a K_4 in G to this packing, giving a 2-$(19, 4, 1)$ packing of size 26. This is a contradiction, since $D(19, 4, 2) = 25$. $\square$

For values of $n < 16$, it is possible to determine $m(n, 4, G)$, $G \in \{K_5 - e, 2 \circ K_4\}$, via exhaustive search. Let H be a specific subgraph of K_n isomorphic to G. We form a graph Γ_n whose vertex set is the set of all K_4's of $K_n - H$, and two vertices in Γ_n are adjacent if and only if the corresponding K_4's are edge-disjoint. Then $m(n, 4, G)$ is equal to the size of a maximum clique in Γ_n. We used Cliquer, an implementation of Östergård's exact algorithm for maximum cliques [12], to determine the size of maximum cliques in Γ_n, for $n \leq 15$.

When $n \geq 16$, it is infeasible to use Cliquer, so we resort to a stochastic local search heuristic to construct packings of the required size directly. The results of our computation are summarized in Table 5.1, while the blocks of the actual packings are listed in Appendices A and B.

5.6. Piecing things together. The results in previous subsections can be summarized as follows.

Table 5.1

Values of $m(n,4,K_5-e)$ and $m(n,4,2\circ K_4)$ for some small values of n. A blank entry indicates an unknown value.

	n												
n	8	9	10	11	12	13	16	17	18	19	20	24	25
$m(n,4,K_5-e)$	1	2	3	4	6	9			21	24	28	40	
n	26	27	28	32	36	37	38	39	73	76	85		
$m(n,4,K_5-e)$	50	52			97								
n	8	9	10	11	12	13	14	16	17	18	19	24	25
$m(n,4,2\circ K_4)$	1	2	3	4	6	9	11			21	24	40	
n	26	27	28	32	35	36	37	38	39	41	44	47	53
$m(n,4,2\circ K_4)$		52											
n	56	59	65	71	73	76	77	85	89				
$m(n,4,2\circ K_4)$													

Theorem 5.23. *For all $n \geq 5$, we have $m(n,4,K_5-e) = \frac{1}{6}(\binom{n}{2} - f(n))$, where*

$$f(n) = \begin{cases} 18 & \text{if } n \equiv 1 \text{ or } 4 \pmod{12},\ n \neq 13, \\ 15 & \text{if } n \equiv 7 \text{ or } 10 \pmod{12},\ n \notin \{10,19\}, \\ (n+24)/2 & \text{if } n \equiv 2 \text{ or } 8 \pmod{12},\ n \neq 8, \\ (n+15)/2 & \text{if } n \equiv 5 \text{ or } 11 \pmod{12},\ n \neq 11, \\ n+9 & \text{if } n \equiv 6 \text{ or } 9 \pmod{12},\ n \neq 9, \\ n+12 & \text{if } n \equiv 0 \text{ or } 3 \pmod{12},\ n \neq 12, \\ 22 & \text{if } n = 8, \\ 24 & \text{if } n = 9, \\ 27 & \text{if } n = 10, \\ 31 & \text{if } n = 11, \\ 30 & \text{if } n = 12, \\ 24 & \text{if } n = 13, \\ 27 & \text{if } n = 19, \end{cases}$$

except possibly for $n \in \{16,17,25,28,32,37,38,39,73,76,85\}$.

Theorem 5.24. *For all $n \geq 6$, we have $m(n,4,2\circ K_4) = \frac{1}{6}(\binom{n}{2} - f(n))$, where*

$$f(n) = \begin{cases} 18 & \text{if } n \equiv 1 \text{ or } 4 \pmod{12},\ n \neq 13, \\ 21 & \text{if } n \equiv 7 \text{ or } 10 \pmod{12},\ n \notin \{10,19\}, \\ (n+24)/2 & \text{if } n \equiv 2 \text{ or } 8 \pmod{12},\ n \notin \{8,14\}, \\ (n+27)/2 & \text{if } n \equiv 5 \text{ or } 11 \pmod{12},\ n \neq 11, \\ n+9 & \text{if } n \equiv 6 \text{ or } 9 \pmod{12},\ n \neq 9, \\ n+12 & \text{if } n \equiv 0 \text{ or } 3 \pmod{12},\ n \neq 12, \\ 22 & \text{if } n = 8, \\ 24 & \text{if } n = 9, \\ 27 & \text{if } n = 10, \\ 31 & \text{if } n = 11, \\ 30 & \text{if } n = 12, \\ 24 & \text{if } n = 13, \\ 25 & \text{if } n = 14, \\ 27 & \text{if } n = 19, \end{cases}$$

except possibly for $n \in \{16, 17, 25, 26, 28, 32, 35, 36, 37, 38, 39, 41, 44, 47, 53, 56, 59, 65, 71, 73, 76, 77, 85, 89\}$.

6. Conclusion. Theorems 4.5, 5.23, and 5.24 can be expressed more succinctly in terms of $D(n,3,2)$ and $D(n,4,2)$ as follows.

THEOREM 6.1. *For all* $n \geq 4$,

$$m(n,3,K_4-e)+2 = \begin{cases} D(n,3,2) & \textit{if } n \equiv 0,\ 2,\ \textit{or } 5 \pmod 6, \\ D(n,3,2)-1 & \textit{if } n \equiv 1,\ 3,\ \textit{or } 4 \pmod 6. \end{cases}$$

THEOREM 6.2. *For all* $n \geq 5$,

$$m(n,4,K_5-e)+2 = \begin{cases} D(n,4,2)+1 & \textit{if } n \equiv 5,\ 6,\ 7,\ 9,\ 10,\ \textit{or } 11 \pmod{12}, \\ & n \notin \{9,10,11\}, \\ D(n,4,2) & \textit{if } n \equiv 0,\ 2,\ 3,\ \textit{or } 8 \pmod{12},\ n \notin \{8,12\}, \\ D(n,4,2)-1 & \textit{if } n \equiv 1 \textit{ or } 4 \pmod{12},\ n \neq 13, \\ n-5 & \textit{if } n \in \{8,9,10,11\}, \\ 8 & \textit{if } n = 12, \\ 11 & \textit{if } n = 13, \end{cases}$$

except possibly for $n \in \{16,17,25,28,32,37,38,39,73,76,85\}$.

THEOREM 6.3. *For all* $n \geq 6$,

$$m(n,4,2\circ K_4)+2 = \begin{cases} D(n,4,2)+1 & \textit{if } n \equiv 6 \textit{ or } 9 \pmod{12},\ n \neq 9, \\ D(n,4,2) & \textit{if } n \equiv 0,\ 2,\ 3,\ 5,\ 7,\ 8,\ 10,\ \textit{or } 11 \pmod{12}, \\ & n \notin \{8,10,11,12,14\}, \\ D(n,4,2)-1 & \textit{if } n \equiv 1 \textit{ or } 4 \pmod{12},\ n \neq 13, \\ n-5 & \textit{if } n \in \{8,9,10,11\}, \\ 8 & \textit{if } n = 12, \\ 11 & \textit{if } n = 13, \\ 13 & \textit{if } n = 14, \end{cases}$$

except possibly for $n \in \{16, 17, 25, 26, 28, 32, 35, 36, 37, 38, 39, 41, 44, 47, 53, 56, 59, 65, 71, 73, 76, 77, 85, 89\}$.

These have the following consequences.

COROLLARY 6.4. *For all* $n \geq 4$, $T(n,\mathcal{F}(3),2) = D(n,3,2)$.

COROLLARY 6.5. *For all* $n \geq 6$,

$$T(n,\mathcal{F}(4),2) = \begin{cases} D(n,4,2)+1 & \textit{if } n \equiv 5,\ 6,\ 7,\ 9,\ 10,\ \textit{or } 11 \pmod{12}, \\ & n \notin \{9,10,11\}, \\ D(n,4,2) & \textit{if } n \equiv 0,\ 1,\ 2,\ 3,\ 4,\ \textit{or } 8 \pmod{12}, \\ & n \notin \{8,12,13\}, \\ n-5 & \textit{if } n \in \{8,9,10,11\}, \\ 8 & \textit{if } n = 12, \\ 11 & \textit{if } n = 13, \end{cases}$$

except possibly for $n \in \{16,17,25,28,32,37,38,39,73,76,85\}$.

Appendix A. Some maximum 2-$(n,4,1)$ packings with a leave containing $K_5 - e$.
In each case, the edges of the $K_5 - e$ in the leave are $\binom{[5]}{2} \setminus \{\{4,5\}\}$.

A.1. The blocks of a maximum 2-$(10,4,1)$ packing with a leave containing $K_5 - e$. $\{4,5,6,7\}$, $\{3,7,8,9\}$, $\{1,6,8,10\}$.

A.2. The blocks of a maximum 2-$(18,4,1)$ packing with a leave containing $K_5 - e$.

{4,8,12,16}, {3,6,7,8}, {3,11,13,16}, {2,9,15,16}, {10,11,12,14},
{2,7,11,17}, {4,9,13,14}, {1,6,9,17}, {5,13,17,18}, {3,14,15,17},
{2,8,14,18}, {4,7,10,15}, {2,6,10,13}, {1,8,11,15}, {4,6,11,18},
{5,8,9,10}, {1,10,16,18}, {5,7,14,16}, {3,9,12,18}, {1,7,12,13},
{5,6,12,15}.

A.3. The blocks of a maximum 2-$(19,4,1)$ packing with a leave containing $K_5 - e$.

{8,14,17,18}, {2,9,13,14}, {3,7,12,14}, {1,10,14,19}, {4,5,10,18},
{4,6,14,16}, {6,11,18,19}, {4,11,13,17}, {3,8,15,19}, {5,12,13,19},
{1,9,12,18}, {3,13,16,18}, {2,7,15,18}, {3,9,10,17}, {4,7,9,19},
{2,16,17,19}, {5,6,7,17}, {2,6,8,12}, {10,12,15,16}, {7,8,10,13},
{5,8,9,16}, {5,11,14,15}, {1,7,11,16}, {1,6,13,15}.

A.4. The blocks of a maximum 2-$(20,4,1)$ packing with a leave containing $K_5 - e$.

{4,6,16,18}, {3,12,16,20}, {1,10,11,15}, {9,12,14,19}, {2,7,10,12},
{6,7,15,19}, {9,10,17,18}, {4,9,11,13}, {4,12,15,17}, {4,5,10,19},
{1,8,12,18}, {3,13,18,19}, {5,8,14,17}, {1,16,17,19}, {1,7,13,14},
{2,6,13,17}, {11,14,18,20}, {2,8,11,19}, {5,6,11,12}, {5,13,15,20},
{3,8,9,15}, {8,10,13,16}, {3,7,11,17}, {2,14,15,16}, {3,6,10,14},
{5,7,9,16}, {4,7,8,20}, {1,6,9,20}.

A.5. The blocks of a maximum 2-$(24,4,1)$ packing with a leave containing $K_5 - e$.

{12,14,15,18}, {3,6,16,18}, {6,9,10,13}, {5,9,15,22}, {3,9,11,21},
{4,8,15,19}, {1,18,21,22}, {12,16,17,19}, {11,12,22,23}, {4,9,23,24},
{4,5,6,12}, {3,10,12,24}, {5,8,21,24}, {6,14,17,21}, {1,8,12,13},
{6,19,22,24}, {4,16,20,21}, {2,18,19,23}, {1,7,17,23}, {3,17,20,22},
{1,11,16,24}, {2,13,14,16}, {2,7,10,21}, {5,7,14,20}, {8,10,17,18},
{13,18,20,24}, {2,9,12,20}, {7,8,16,22}, {3,7,13,19}, {2,15,17,24},
{5,11,13,17}, {13,15,21,23}, {10,11,19,20}, {1,9,14,19}, {4,7,11,18},
{1,6,15,20}, {3,8,14,23}, {2,6,8,11}, {5,10,16,23}, {4,10,14,22}.

A.6. The blocks of a maximum 2-$(26,4,1)$ packing with a leave containing $K_5 - e$.

{4,17,22,24}, {3,11,17,20}, {5,7,18,22}, {4,16,18,23}, {1,7,19,25},
{14,21,22,23}, {1,10,18,26}, {2,11,21,26}, {3,6,7,23}, {11,14,16,19},
{12,20,24,26}, {4,7,14,26}, {3,9,16,22}, {6,10,15,16}, {3,10,12,19},
{7,8,15,17}, {4,9,13,19}, {5,12,13,21}, {15,19,22,26}, {5,19,23,24},
{4,12,15,25}, {3,15,18,21}, {8,9,21,25}, {6,12,17,18}, {5,8,16,26},
{2,7,9,12}, {9,17,23,26}, {1,8,20,22}, {5,9,11,15}, {7,10,21,24},
{1,13,14,15}, {6,19,20,21}, {7,13,16,20}, {10,11,22,25}, {2,6,13,22},
{2,16,24,25}, {9,14,18,20}, {2,8,18,19}, {1,6,9,24}, {4,6,8,11},
{5,6,14,25}, {8,10,13,23}, {11,13,18,24}, {2,10,14,17}, {3,13,25,26},
{3,8,14,24}, {2,15,20,23}, {1,11,12,23}, {4,5,10,20}, {1,16,17,21}.

A.7. The blocks of a maximum 2-$(27,4,1)$ packing with a leave containing $K_5 - e$.

{2,7,16,21}, {7,20,26,27}, {5,17,25,27}, {5,15,21,23}, {5,6,11,22},
{13,21,22,27}, {3,8,11,26}, {6,15,17,24}, {4,5,7,19}, {1,6,18,27},
{3,18,21,24}, {2,11,12,13}, {9,13,16,23}, {10,11,14,15}, {3,14,23,27},
{4,8,15,18}, {14,19,22,24}, {1,10,19,23}, {3,12,16,20}, {2,8,23,24},
{5,8,9,20}, {4,12,14,21}, {4,9,11,27}, {3,6,10,25}, {8,14,16,17},
{2,15,19,27}, {9,12,15,22}, {3,7,13,15}, {1,8,12,25}, {3,9,17,19},
{19,20,21,25}, {2,6,14,20}, {6,8,13,19}, {7,11,24,25}, {1,11,17,21},
{4,10,17,20}, {9,10,21,26}, {10,16,24,27}, {4,16,22,25}, {7,12,17,18},
{1,7,9,14}, {2,17,22,26}, {11,16,18,19}, {5,12,24,26}, {1,15,16,26},
{5,10,13,18}, {1,13,20,24}, {18,20,22,23}, {2,9,18,25}, {4,6,23,26},
{13,14,25,26}, {7,8,10,22}.

A.8. The blocks of a maximum 2-$(36,4,1)$ packing with a leave containing $K_5 - e$.

{7,10,17,35}, {11,15,26,36}, {6,16,25,29}, {1,12,24,28}, {3,13,34,35},
{30,31,35,36}, {21,23,28,34}, {1,14,19,35}, {8,9,28,32}, {15,18,21,25},
{3,18,26,27}, {1,8,25,30}, {3,10,23,29}, {6,28,30,33}, {15,19,23,24},
{4,14,17,34}, {7,13,26,28}, {10,19,22,36}, {6,11,12,34}, {1,7,11,29},
{5,13,17,25}, {14,24,26,31}, {13,19,27,29}, {1,20,23,26}, {2,22,31,34},
{14,23,25,36}, {5,16,24,33}, {4,18,29,33}, {4,21,26,32}, {8,22,26,29},
{9,11,22,25}, {12,18,20,32}, {2,11,20,21}, {11,13,31,32}, {10,14,30,32},
{3,9,33,36}, {3,11,24,30}, {24,29,32,36}, {7,18,24,34}, {7,19,21,31},
{3,7,12,25}, {2,8,13,24}, {2,7,14,16}, {5,7,8,20}, {10,11,16,28},
{5,6,18,31}, {8,11,14,18}, {3,17,19,32}, {10,20,25,31}, {4,5,11,19},
{16,18,19,30}, {16,20,34,36}, {3,6,15,20}, {4,8,10,12}, {6,9,13,14},
{9,17,20,24}, {13,20,22,33}, {4,6,7,36}, {1,13,18,36}, {5,26,30,34},
{1,6,22,32}, {16,21,27,35}, {12,13,21,30}, {2,9,18,35}, {12,17,29,31},
{8,17,21,36}, {7,9,23,30}, {20,28,29,35}, {2,15,29,30}, {4,20,27,30},
{1,15,16,17}, {1,9,27,31}, {4,15,28,31}, {12,14,15,33}, {9,12,19,26},
{25,26,33,35}, {6,10,24,27}, {3,14,21,22}, {2,23,32,33}, {5,14,27,28},
{5,12,22,23}, {25,27,32,34}, {3,8,16,31}, {4,13,16,23}, {8,19,33,34},
{2,6,17,26}, {5,15,32,35}, {2,19,25,28}, {9,10,15,34}, {6,8,23,35},
{1,10,21,33}, {7,15,22,27}, {4,22,24,35}, {11,17,27,33}, {2,12,27,36},
{5,9,21,29}, {17,18,22,28}.

Appendix B. Some maximum 2-$(n, 4, 1)$ packings with a leave containing $2 \circ K_4$.

B.1. The blocks of a maximum 2-$(18, 4, 1)$ packing with a leave containing $2 \circ K_4$.

{3,9,10,12}, {1,7,9,16}, {4,7,17,18}, {1,5,13,17}, {5,9,11,14},
{1,6,14,18}, {4,6,8,9}, {2,7,10,14}, {3,14,16,17}, {5,8,10,18},
{1,8,12,15}, {6,10,15,17}, {3,7,8,13}, {3,11,15,18}, {6,7,11,12},
{2,12,16,18}, {10,11,13,16}, {2,8,11,17}, {2,9,13,15}, {4,5,15,16},
{4,12,13,14}.

B.2. The blocks of a maximum 2-$(19, 4, 1)$ packing with a leave containing $2 \circ K_4$.

{3,8,9,16}, {5,12,16,18}, {11,13,15,16}, {1,12,13,19}, {6,8,10,11},
{1,10,14,16}, {6,12,15,17}, {6,7,9,13}, {4,13,14,18}, {1,9,15,18},
{5,9,14,19}, {4,6,16,19}, {4,7,8,12}, {5,8,13,17}, {2,8,14,15},
{3,10,17,18}, {4,5,10,15}, {2,9,10,12}, {1,5,7,11}, {2,7,16,17},
{4,9,11,17}, {3,7,15,19}, {3,11,12,14}, {2,11,18,19}.

B.3. The blocks of a maximum 2-$(24, 4, 1)$ packing with a leave containing $2 \circ K_4$.

{4,13,14,21}, {3,12,16,20}, {3,15,21,22}, {3,17,18,23}, {7,13,15,23},
{4,12,19,22}, {1,9,15,19}, {4,7,8,18}, {6,10,15,18}, {9,11,17,21},
{3,10,11,13}, {5,9,14,22}, {1,11,14,18}, {1,6,8,21}, {6,7,14,20},
{2,13,19,24}, {10,19,20,21}, {5,11,12,15}, {8,11,16,24}, {2,8,10,17},
{2,16,21,23}, {5,8,13,20}, {4,6,9,16}, {1,7,10,16}, {2,7,11,22},
{2,9,18,20}, {14,15,16,17}, {8,9,12,23}, {10,12,14,24}, {3,7,9,24},
{13,16,18,22}, {6,17,22,24}, {1,12,13,17}, {5,7,17,19}, {3,8,14,19},
{4,15,20,24}, {1,20,22,23}, {4,5,10,23}, {6,11,19,23}, {5,18,21,24}.

B.4. The blocks of a maximum 2-$(27, 4, 1)$ packing with a leave containing $2 \circ K_4$.

{6,12,17,21}, {1,5,10,19}, {4,8,10,27}, {4,14,21,22}, {19,21,24,25},
{2,11,23,26}, {3,10,20,24}, {3,9,15,21}, {12,20,26,27}, {8,11,12,25},
{10,15,18,26}, {3,8,19,26}, {4,11,13,20}, {9,22,24,26}, {3,7,22,27},
{1,15,22,23}, {5,14,24,27}, {3,12,14,23}, {9,12,13,19}, {2,9,17,27},
{3,13,18,25}, {4,7,12,15}, {6,14,16,20}, {5,7,9,11}, {6,15,25,27},
{10,17,22,25}, {5,20,23,25}, {2,10,12,16}, {4,6,18,19}, {5,12,18,22},
{3,11,16,17}, {6,8,13,22}, {1,13,16,27}, {2,13,15,24}, {5,8,15,17},
{5,16,21,26}, {13,14,17,26}, {7,10,13,21}, {2,19,20,22}, {1,6,11,24},
{7,17,18,20}, {1,8,20,21}, {1,7,25,26}, {11,14,15,19}, {4,9,16,25},
{7,16,19,23}, {8,16,18,24}, {11,18,21,27}, {4,17,23,24}, {1,9,14,18},
{2,7,8,14}, {6,9,10,23}.

REFERENCES

[1] B. Bollobás, *Extremal Graph Theory*, London Math. Soc. Monogr. 11, Academic Press, Harcourt Brace Jovanovich, London, 1978.

[2] E. Boros, Y. Caro, Z. Füredi, and R. Yuster, *Covering non-uniform hypergraphs*, J. Combin. Theory Ser. B, 82 (2001), pp. 270–284.

[3] A. E. Brouwer, *Optimal packings of K_4's into a K_n*, J. Combin. Theory Ser. A, 26 (1979), pp. 278–297.

[4] A. E. Brouwer, A. Schrijver, and H. Hanani, *Group divisible designs with block-size four*, Discrete Math., 20 (1977), pp. 1–10.

[5] M. K. Fort, Jr., and G. A. Hedlund, *Minimal coverings of pairs by triples*, Pacific J. Math., 8 (1958), pp. 709–719.

[6] G. Ge and A. C. H. Ling, *Group divisible designs with block size four and group type $g^u m^1$ for small g*, Discrete Math., 285 (2004), pp. 97–120.

[7] H. Hanani, *Balanced incomplete block designs and related designs*, Discrete Math., 11 (1975), pp. 255–369.

[8] K. Heinrich and L. Zhu, *Existence of orthogonal Latin squares with aligned subsquares*, Discrete Math., 59 (1986), pp. 69–78.

[9] D. L. Kreher and D. R. Stinson, *Small group-divisible designs with block size four*, J. Statist. Plann. Inference, 58 (1997), pp. 111–118.

[10] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.

[11] W. H. Mills and R. C. Mullin, *Coverings and packings*, in Contemporary Design Theory, J. H. Dinitz and D. R. Stinson, eds., Wiley-Intersci. Ser. Discrete Math. Optim., Wiley, New York, 1992, pp. 371–399.

[12] P. R. J. Östergård, *A fast algorithm for the maximum clique problem*, Discrete Appl. Math., 120 (2002), pp. 197–207.

[13] R. Rees and D. R. Stinson, *On the existence of incomplete designs of block size four having one hole*, Util. Math., 35 (1989), pp. 119–152.

[14] J. Schönheim, *On maximal systems of k-tuples*, Studia Sci. Math. Hungar., 1 (1966), pp. 363–368.

[15] J. Spencer, *Maximal consistent families of triples*, J. Combin. Theory, 5 (1968), pp. 1–8.

SIAM J. DISCRETE MATH.
Vol. 21, No. 4, pp. 823–843

NEIGHBORHOOD BROADCASTING IN HYPERCUBES[*]

JEAN-CLAUDE BERMOND[†], AFONSO FERREIRA[†], STÉPHANE PÉRENNES[†], AND JOSEPH G. PETERS[‡]

Abstract. In the broadcasting problem, one node needs to broadcast a message to all other nodes in a network. If nodes can only communicate with one neighbor at a time, broadcasting takes at least $\lceil \log_2 N \rceil$ rounds in a network of N nodes. In the *neighborhood broadcasting* problem, the node that is broadcasting needs to inform only its neighbors. In a binary hypercube with N nodes, each node has $\log_2 N$ neighbors, so neighborhood broadcasting takes at least $\lceil \log_2 \log_2(N+1) \rceil$ rounds. In this paper, we present asymptotically optimal neighborhood broadcast protocols for binary hypercubes.

Key words. broadcasting, hypercubes, neighborhood communication

AMS subject classifications. 94A05, 68M10, 68M12

DOI. 10.1137/040617716

1. Introduction.

In the broadcasting problem, a single *originator* is required to disseminate a piece of information to all other nodes of a network (modelled as a graph) as quickly as possible. In the *unit-cost single-port* communication model, each message transmission requires one time unit or *round*, and each node can communicate with at most one adjacent node (*neighbor*) at any given time. It is well known that broadcasting in an n-dimensional binary hypercube, or *n-cube*, under this model requires $n = \log_2 N$ rounds of communication to inform all $N = 2^n$ nodes and that this is optimal. In this paper, we address a variant of this problem called *neighborhood broadcasting* in which the originator needs to inform only its n neighbors in a hypercube. We show that this can be accomplished exponentially faster than normal (complete) broadcasting. A lower bound on the number of rounds for a neighborhood broadcast is $\lceil \log_2(n+1) \rceil = \lceil \log_2 \log_2(N+1) \rceil$. We present two neighborhood broadcast protocols and prove that the second protocol achieves the lower bound asymptotically. More precisely, we prove that a neighborhood broadcast can be completed in at most $log_2 n + \lceil \sqrt{2 log_2 n}\ \rceil$ rounds (so the ratio of the upper bound for the second protocol and the lower bound tends to 1 as n tends to infinity). The exact analyses of our protocols are difficult, so, for each protocol, we introduce a sequence of truncated protocols and prove that their performances approach the lower bound.

The neighborhood broadcasting problem was introduced by Cosnard and Ferreira [3] who outlined a simple $\mathrm{O}(\log_2 n)$ protocol. They proved that the number of neigh-

[*] Received by the editors October 27, 2004; accepted for publication (in revised form) March 13, 2007; published electronically November 7, 2007. This research was done while the first author was visiting Simon Fraser University and while the fourth author was visiting Project MASCOTTE - Sophia-Antipolis.

http://www.siam.org/journals/sidma/21-4/61771.html

[†] Project MASCOTTE, CNRS - UNSA - INRIA, 2004 Route des Lucioles, BP 93, F-06902 Sophia-Antipolis Cedex, France (bermond@sophia.inria.fr, ferreira@sophia.inria.fr, speren@sophia.inria.fr). The research of these authors was supported by the European Project FET IST AEOLUS. The second author is currently on leave from CNRS. Currently Science Officer for Telecommunications & IST at COST, Brussels, Belgium.

[‡] School of Computing Science, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada (peters@cs.sfu.ca). This author's research was supported by NSERC of Canada, Université de Nice - Sophia-Antipolis, CNRS, and INRIA.

bors of the originator informed by their protocol after t rounds satisfies a Fibonacci recurrence and is proportional to $(1.618)^t$. Thus, the number of rounds to complete a neighborhood broadcast using their protocol is proportional to $1.4404 \log_2 n$. In section 2, we generalize the protocol from [3] to obtain the first of our new protocols called Protocol **A**. We were unable to find a closed form expression for the performance of Protocol **A**, but we can give generalized Fibonacci recurrence relations for truncated versions of Protocol **A**. The truncated protocol $\mathbf{A}_k$, $k \geq 2$ is obtained from Protocol **A** by discarding all communications that involve a node at a distance greater than k from the originator. Protocol $\mathbf{A}_2$ is the protocol from [3]. For Protocol $\mathbf{A}_3$, the number of neighbors of the originator informed after t rounds is proportional to $(1.839)^t$, for Protocol $\mathbf{A}_4$ it is proportional to $(1.913)^t$, and for Protocol $\mathbf{A}_{12}$ it is $(1.991)^t$.

In section 3, we describe and analyze a more sophisticated, and more efficient, protocol called Protocol **B**. We show that for any fixed $\epsilon > 0$ and sufficiently large t, the number of neighbors of the originator informed after t rounds of Protocol **B** is at least $(2 - \epsilon)^t$. We also derive recurrence relations for the truncated protocols $\mathbf{B}_k$, $k \geq 2$. For example, the number of neighbors of the originator informed after t rounds of Protocol $\mathbf{B}_5$ is proportional to $(1.999)^t$. We think that Protocol **B** is not just asymptotically optimal, but that it is optimal or near-optimal in the sense that no protocol can inform the neighbors of the originator faster. Unfortunately, our attempts to significantly improve the lower bound have not succeeded, so improving the lower bound and determining the optimal performance exactly remain as open problems.

The protocol in section 2 was first presented at a workshop in 1991 [2], including the closed form solution for a truncated version of the protocol and empirical evidence that the (untruncated) protocol is asymptotically optimal. An incomplete manuscript [1] of the present paper, including the protocols in sections 2 and 3 and parts of the analysis, has been in circulation since 1998. The workshop presentation and the manuscript have stimulated considerable interest in neighborhood communication problems [5, 6, 10, 11, 12, 13, 16, 19, 20].

Hypercubes are Cayley graphs and many of the ideas in this paper can be modified or extended to other classes of Cayley graphs such as *star graphs*, which are Cayley graphs on permutation groups. The first bounds for broadcasting in star graphs appeared in [10]. The bounds were improved in [19], and an alternative protocol (with a weaker bound) was presented in [20]. The best current upper bounds for neighborhood broadcasting in star graphs are $1.3125 \log_2 n + \mathrm{O}(\log_2 \log_2 \mathrm{n})$ [12] and $\log_2 n + \mathrm{O}(\sqrt{\log_2 \mathrm{n}})$ [11]. A larger class of Cayley graphs on permutation groups is studied in [16].

Neighborhood gossiping in hypercubes was studied in [13]. In the neighborhood gossiping problem, each node starts with a unique piece of information and must learn the information of all of its neighbors. Normal (complete) gossiping in an n-cube takes at least $1.44n + \mathrm{O}(1)$ rounds [4, 18] and at most $1.88n + \mathrm{O}(1)$ rounds [17] using *half-duplex* links, and exactly n rounds using *full-duplex* (i.e., bidirectional) links (see [14]). The bounds in [13] on the numbers of rounds, $h_1(n)$ and $h_2(n)$, for half-duplex and full-duplex neighborhood gossiping in an n-cube, respectively, are $2.88 \log_2 n + \mathrm{O}(1) \leq \mathrm{h}_1(\mathrm{n}) \leq 3.76 \log_2 \mathrm{n} + \mathrm{O}(1)$ and $h_2(n) = 2 \log_2 n + \mathrm{O}(1)$. The ideas in [13] were extended to star graphs in [10]. Note that while the distinction between the half-duplex and full-duplex links is important for gossiping problems, it can be ignored for broadcasting problems because the (single) message in a broadcast protocol never needs to traverse any link in both directions.

In *k-neighborhood communication* problems, nodes that are at distance at most k are required to communicate. The neighborhood broadcasting and gossiping problems are examples of 1-neighborhood communication. Bounds for k-neighborhood broadcasting and gossiping in paths, trees, 2-dimensional grids, 2-dimensional tori, and cycles were derived in [5, 6]. The results are optimal in most cases and within an additive constant of optimal in the other cases.

There are many papers describing protocols that minimize the time for a normal (complete) broadcast on various interconnection networks such as hypercubes and meshes. See [15] for a discussion of models and results for broadcasting and gossiping with unit-cost models and [9, 14] for comprehensive surveys.

2. A simple protocol. In Cosnard and Ferreira's neighborhood broadcast protocol [3], the originator in a hypercube sends its message to a new neighbor during each round. Each informed neighbor of the originator broadcasts to its neighbors (except the originator, of course). These neighbors of the neighbors do not need to know the message, but each of them can inform one new neighbor of the originator. It is not difficult to show directly that this protocol takes $\mathrm{O}(\log_2 n)$ rounds to inform all neighbors of the originator in an n-cube, but we will take the opportunity to introduce some notation that we will use to analyze our new protocols.

We will identify each vertex in an n-cube by a binary string of length n. Without loss of generality, the originator is labelled with a string of n 0s: $00\cdots00$. Each neighbor of the originator has exactly one 1 in its label. Each neighbor of a neighbor of the originator (except the originator) has two 1s in its label. In general, a node at (Hamming) distance k from the originator has exactly k 1s in its label. We will say that nodes at distance k from the originator are at *level k*. In the neighborhood broadcasting problem, all level 1 nodes must be informed, and we want to do this as quickly as possible.

It will often be convenient to have a compact way to write node labels. When we write $\delta_1\delta_2\delta_3\delta_4$, $\delta_1 < \delta_2 < \delta_3 < \delta_4$, we mean that the label contains 1s in the indicated positions and 0s in all other positions, so this is a level 4 node. The label $\delta_1\bar{\delta}_2\delta_3$ has 1s in positions δ_1 and δ_3, a 0 in position δ_2, and 0s elsewhere, so this is a level 2 node. We will sometimes insert commas into labels to avoid ambiguity. For example, 1,4,21 is the level 3 node shown in Figure 1 with 1s in positions 1, 4, and 21.

In our figures, we will draw the originator on the left and Hamming distance from the originator will increase from left to right. When we say that a node is informed *from the left* or *from the right*, we are referring to this left to right arrangement of increasing levels.

To analyze our protocols, we use the following notation:

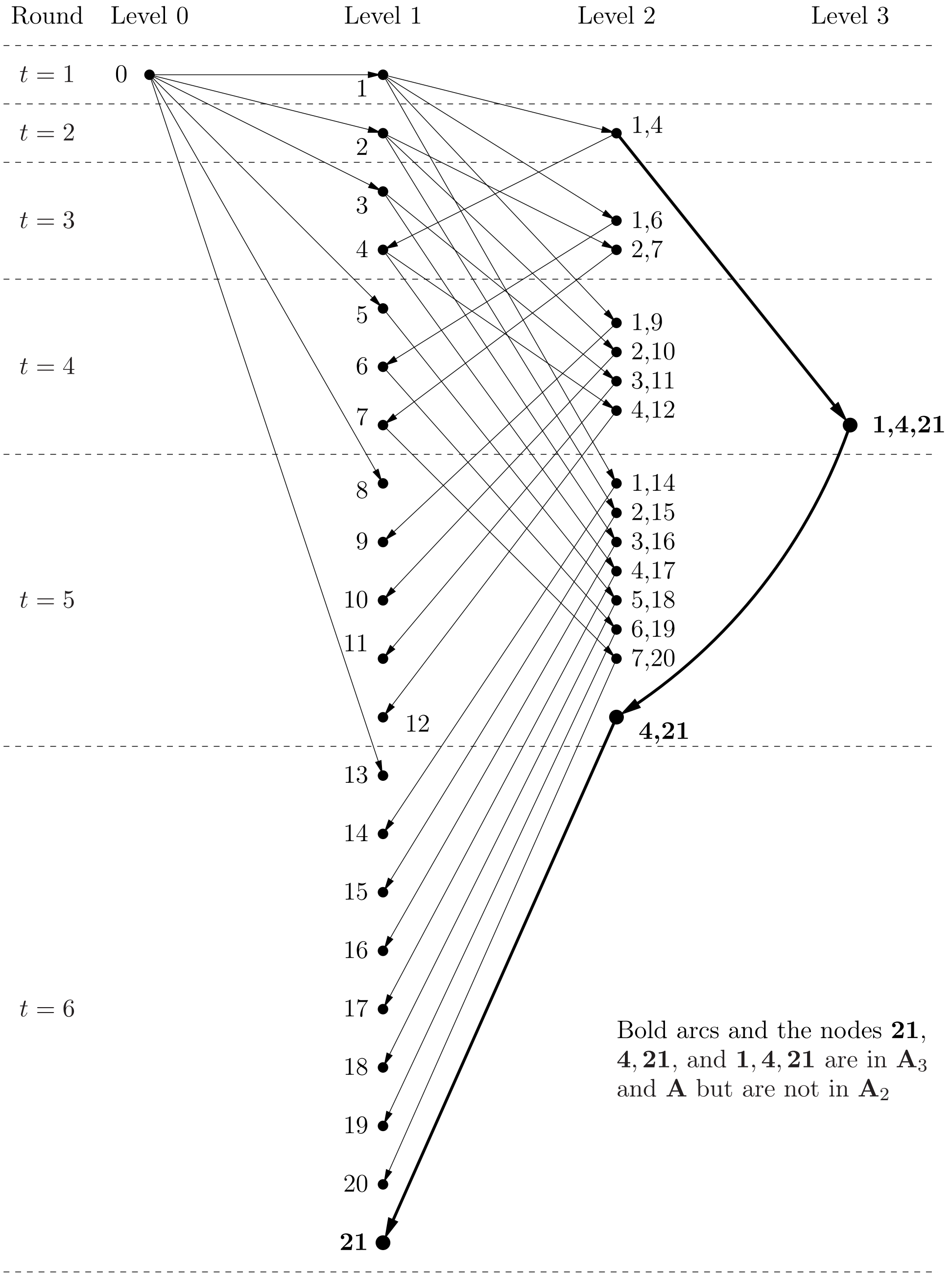
$L_k^t(\mathbf{P})$: maximum number of level k nodes informed by level $k-1$ nodes (i.e., from the left) during round t of Protocol $\mathbf{P}$

$R_k^t(\mathbf{P})$: maximum number of level k nodes informed by level $k+1$ nodes (i.e., from the right) during round t of Protocol $\mathbf{P}$

$N_k^t(\mathbf{P}) = L_k^t(\mathbf{P}) + R_k^t(\mathbf{P})$: maximum total number of level k nodes informed during round t of Protocol $\mathbf{P}$

$T_k^t(\mathbf{P}) = \sum_{i=1}^{t} N_k^i(\mathbf{P})$: maximum total number of level k nodes informed during the first t rounds of Protocol $\mathbf{P}$

We will often omit the name of the protocol to simplify the notation when the protocol $\mathbf{P}$ is clear from the context.



Bold arcs and the nodes **21**, $\mathbf{4}, \mathbf{21}$, and $\mathbf{1}, \mathbf{4}, \mathbf{21}$ are in $\mathbf{A}_3$ and $\mathbf{A}$ but are not in $\mathbf{A}_2$

FIG. 1. *Node labels during the first six rounds of Protocols* $\mathbf{A}_2$*,* $\mathbf{A}_3$*, and* $\mathbf{A}$.

In the analyses of our protocols, we will show several things. For each protocol $\mathbf{P}$, we will develop recurrence relations for $T_k^t(\mathbf{P})$. The value of $T_k^t(\mathbf{P})$ is an upper bound on the number of informed level k nodes after t rounds of Protocol $\mathbf{P}$. To prove that Protocol $\mathbf{P}$ achieves these bounds, we need to show that it informs *exactly* $T_k^t(\mathbf{P})$ level k nodes during the first t rounds. We do this by showing that all newly informed nodes are distinct and that all level 1 nodes are eventually informed. We will then determine the rate at which Protocol $\mathbf{P}$ informs level 1 nodes as a function of t. We do this by determining the value of the largest root a_k of the associated polynomial of the recurrence relation $T_1^t(\mathbf{P})$. The number of level 1 nodes informed by Protocol $\mathbf{P}$ is proportional to a_k^t.

We will begin by considering the protocol from [3]. We will call this Protocol $\mathbf{A}_2$ because it is a truncated version of Protocol $\mathbf{A}$, the first of our new protocols which we will introduce later in this section. If x is a node that is informed during round t of Protocol $\mathbf{A}_2$, then x informs uninformed nodes as follows:

Protocol $\mathbf{A}_2$ [3]

(i) If x is the originator, inform type L_1 nodes during rounds $t+1, t+2, \ldots$.

(ii) If x is a level 1 node, inform type L_2 nodes during rounds $t+1, t+2, \ldots$.

(iii) If x is a level 2 node, inform a type R_1 node during round $t+1$.

The next theorem and corollary from [3] are restated using our notation.

THEOREM 1 (see [3]). $T_1^t(\mathbf{A}_2) = T_1^{t-1}(\mathbf{A}_2) + T_1^{t-2}(\mathbf{A}_2) + 1$.

Proof. First, we get $L_1^t = 1$, $t \geq 1$, because the originator informs one neighbor during each round. We also have $L_2^t = T_1^{t-1}$, $t \geq 2$, because each level 1 node that was informed during the first $t-1$ rounds can potentially inform a new level 2 node during round t. Finally, $R_1^t = L_2^{t-1}$, $t \geq 3$, because each informed level 2 node can potentially inform one new neighbor of the originator immediately after it receives the message. Thus, $R_1^t = T_1^{t-2}$, and for $t \geq 3$, we can write $T_1^t = T_1^{t-1} + N_1^t = T_1^{t-1} + L_1^t + R_1^t = T_1^{t-1} + T_1^{t-2} + 1$. □

COROLLARY 1 (see [3]). $T_1^t(\mathbf{A}_2) \sim 1.618^t$.

Proof. Since $T_1^1 = 1$ and $T_1^2 = 2$, we get $T_1^t = F_{t+2} - 1$, where F_i is the ith Fibonacci number (with starting values $F_1 = F_2 = 1$). The associated polynomial of T_1^t is $x^2 - x - 1 = 0$ and its largest root is $a_2 = \frac{1+\sqrt{5}}{2}$. It follows that the potential number of informed neighbors of the originator after t rounds is proportional to $(\frac{1+\sqrt{5}}{2})^t \sim 1.618^t$. □

To show that the bound of Theorem 1 can be attained, we need to show that every level 1 node is informed and that no nodes are informed more than once. To do this, we have to specify which nodes are informed during each round. We use the following method: During round t, the originator (which we will refer to as node 0) will inform node $T_1^{t-1} + 1$ at level 1 (i.e., the node whose label has a 1 in position $T_1^{t-1} + 1$), and any level 1 node δ, $1 \leq \delta \leq T_1^{t-1}$, that was informed during the first $t-1$ rounds will inform node $\delta, \delta + T_1^t + 1$ at level 2 if $\delta + T_1^t + 1 \leq n$. If $\delta + T_1^t + 1 > n$, then we can assume that node δ is idle because communications to the right will not result in any more informed level 1 nodes before the end of the protocol. Then, during round $t+1$, each level 2 node $\delta, \delta + T_1^t + 1$ that was informed during round t will inform node $\delta + T_1^t + 1$ at level 1. Figure 1 shows how this can be done for $n \leq T_1^6 = 20$. (In Figure 1, the three bold arcs and the nodes with 21 in their labels are not part of Protocol $\mathbf{A}_2$ and should be ignored at this point.) The following lemma establishes the correctness of this pattern.

LEMMA 1. *All level* 1 *nodes* δ *with* $1 \leq \delta \leq \min(n, T_1^t(\mathbf{A}_2))$ *are informed in* t *rounds.*

Proof. The proof is by induction. The claim is true for $t = 1$ and $t = 2$. Now, suppose that the claim is true after round t. If $n \leq T_1^t$, we are done. If $n > T_1^t$, then the new level 1 nodes informed during round $t+1$ are node $T_1^t + 1$, which is informed by node 0, and all nodes δ with $T_1^t + 2 \leq \delta \leq \min(n, T_1^t + T_1^{t-1} + 1)$, which are informed by the level 2 nodes $\delta, \delta + T_1^t + 1$ with $1 \leq \delta \leq T_1^{t-1}$. By Theorem 1, $T_1^{t+1} = T_1^t + T_1^{t-1} + 1$, so the new level 1 nodes informed during round $t+1$ are all nodes δ with $T_1^t + 1 \leq \delta \leq \min(n, T_1^{t+1})$. □

The first of our new protocols is a natural generalization of the protocol from [3]. Each node x that is informed during round t informs the following uninformed nodes:

Protocol **A**

(i) If x is the originator, inform type L_1 nodes during rounds $t+1, t+2, \ldots$.

(ii) If x is a level 1 node, inform type L_2 nodes during rounds $t+1, t+2, \ldots$.

(iii) If x is a level $k \geq 2$ node, inform a type R_{k-1} node during round $t+1$ and type L_{k+1} nodes during rounds $t+2, t+3, \ldots$.

In Protocol **A**, each newly informed node at level $k \geq 2$ immediately informs one level $k-1$ node and then informs level $k+1$ nodes until the protocol terminates. The intuition is that each communication to the right can introduce a new dimension, which can eventually result in a new level 1 node being informed. So, in Protocol **A**, a node that has been informed from the left immediately initiates a path of communications going back to the level 1 node with the new dimension. Newly informed nodes that have received the message from the right continue to forward the message left towards the level 1 node. Additional communications to the left will not lead directly to more informed nodes at level 1 because no new dimensions are being introduced. (We will see later in Protocol **B** how more new dimensions can be introduced indirectly.)

Protocol **A** informs level 1 nodes faster than Protocol $\mathbf{A}_2$. Figure 1 shows that Protocol **A** can inform 21 level 1 nodes during the first six rounds while Protocol $\mathbf{A}_2$ can inform at most 20. The third protocol, $\mathbf{A}_3$, shown in Figure 1 will be described later. Protocol $\mathbf{A}_3$ can inform the same number of level 1 nodes as Protocol **A** during the first six rounds, but eventually (when the number of rounds is nine or greater) Protocol **A** informs level 1 nodes faster than Protocol $\mathbf{A}_3$.

The recurrence equations for Protocol **A** are as follows:

$$L_1^t(\mathbf{A}) = 1 \qquad t \geq 1$$

$$L_2^1(\mathbf{A}) = 0$$

$$(1) \qquad L_2^t(\mathbf{A}) = \sum_{i=1}^{t-1}(L_1^i(\mathbf{A}) + R_1^i(\mathbf{A})) = T_1^{t-1}(\mathbf{A}) \qquad t \geq 2$$

$$L_k^t(\mathbf{A}) = 0 \qquad t \leq 2k-3,\ k \geq 2$$

$$(2) \qquad L_k^t(\mathbf{A}) = \sum_{i=1}^{t-2}(L_{k-1}^i(\mathbf{A}) + R_{k-1}^i(\mathbf{A})) \qquad t \geq 2k-2,\ k \geq 3$$

$$R_k^t(\mathbf{A}) = 0 \qquad t \leq 2k,\ k \geq 1$$

$$(3) \qquad R_k^t(\mathbf{A}) = L_{k+1}^{t-1}(\mathbf{A}) + R_{k+1}^{t-1}(\mathbf{A}) \qquad t \geq 2k+1,\ k \geq 1$$

$$(4) \qquad N_k^t(\mathbf{A}) = L_k^t(\mathbf{A}) + R_k^t(\mathbf{A}) \qquad t \geq 1,\ k \geq 1$$

$$T_k^t(\mathbf{A}) = \sum_{i=1}^{t} N_k^i(\mathbf{A}) = \sum_{i=1}^{t}(L_k^i(\mathbf{A}) + R_k^i(\mathbf{A})) \qquad t \geq 1,\ k \geq 1.$$

We begin our analysis of Protocol $\mathbf{A}$ by simplifying the expression for $T_1^t(\mathbf{A})$. We can express $N_1^t(\mathbf{A})$ as a function of the $L_k^t(\mathbf{A})$ by using (3) repeatedly:

$$(5)\quad N_1^t = L_1^t + R_1^t = 1 + L_2^{t-1} + R_2^{t-1} = 1 + L_2^{t-1} + L_3^{t-2} + L_4^{t-3} + \cdots + L_k^{t-k+1} + \cdots .$$

Then we use $T_1^t = T_1^{t-1} + N_1^t$, (5), and $L_2^{t-1} = T_1^{t-2}$ (from (1)) to obtain

$$(6)\qquad T_1^t = T_1^{t-1} + T_1^{t-2} + 1 + \sum_{i \geq 3} L_i^{t-i+1}.$$

To show that this bound for $T_1^t(\mathbf{A})$ is attained by Protocol $\mathbf{A}$, we have to specify which nodes are informed during each round. We also have to show that no nodes are informed more than once and that every neighbor of the originator is informed.

During round t, node 0 (the originator) will inform node $T_1^{t-1} + 1$ at level 1. Each level 1 node δ, $1 \leq \delta \leq T_1^{t-1}$, that was informed during the first $t-1$ rounds will inform node $\delta, \delta + T_1^t + 1$ if $\delta + T_1^t + 1 \leq n$ and will be idle if $\delta + T_1^t + 1 > n$. Once a node becomes idle, it remains idle until the end of the protocol.

To describe the behavior of the level 2 nodes during round t, let us rank the nodes $\delta_1\delta_2$, $\delta_1 < \delta_2$, that are informed during the first $t-2$ rounds in increasing order according to the value of δ_2. (We will prove below that there are exactly $T_2^{t-2} = L_3^t$ such nodes and that they all have different values of δ_2.) If $\delta_1\delta_2$ is the jth node in this ranking, it will inform the level 3 node $\delta_1\delta_2\delta_3$, where $\delta_3 = T_1^{t+1} + 1 + L_2^{t+1} + j$, if $\delta_3 \leq n$ and will be idle otherwise.

To describe the pattern by which level $k-1$ nodes inform level k nodes during round t (and the way that new dimensions are introduced), let us rank the level $k-1$ nodes $\delta_1\delta_2\cdots\delta_{k-1}$, $\delta_1 < \delta_2 < \cdots < \delta_{k-1}$, that are informed during the first $t-2$ rounds in increasing order according to the value of δ_{k-1}. (We will prove below that there are exactly $T_{k-1}^{t-2} = L_k^t$ such nodes and that they all have different values of δ_{k-1}.) Then, if $\delta_1\delta_2\cdots\delta_{k-1}$ is the jth node in this ranking, it will inform the level k node $\delta_1\delta_2\cdots\delta_{k-1}\delta_k$, where $\delta_k = T_1^{t+k-2} + 1 + L_2^{t+k-2} + L_3^{t+k-3} + \cdots + L_{k-1}^{t+1} + j$, if $\delta_k \leq n$ and will be idle otherwise.

Finally, each level $k \geq 2$ node $\delta_1\delta_2\cdots\delta_k$, $\delta_1 < \delta_2 < \cdots < \delta_k$, that is informed during round $t-1$ will inform the level $k-1$ node $\rho_1\rho_2\cdots\rho_{k-1} = \delta_2\delta_3\cdots\delta_k$ during round t (i.e., we always delete the leftmost index from the label of the level k node to obtain the label of the level $k-1$ node).

Claim 1. *During round t, the nodes informed by Protocol $\mathbf{A}$ are as follows:*

(i) *all level 1 nodes δ_1 such that $\delta_1 = T_1^{t-1}(\mathbf{A}) + j$, where $1 \leq j \leq N_1^t(\mathbf{A})$;*

(ii) *all level 2 nodes $\delta_1\delta_2$, $\delta_1 < \delta_2$ such that $\delta_2 = T_1^t(\mathbf{A}) + 1 + j$, where $1 \leq j \leq N_2^t(\mathbf{A})$;*

(iii) *all level k nodes $\delta_1\delta_2\cdots\delta_k$, $\delta_1 < \delta_2 < \cdots < \delta_k$ such that $\delta_k = T_1^{t+k-2}(\mathbf{A}) + 1 + L_2^{t+k-2}(\mathbf{A}) + L_3^{t+k-3}(\mathbf{A}) + \cdots + L_{k-1}^{t+1}(\mathbf{A}) + j$, where $1 \leq j \leq N_k^t(\mathbf{A})$.*

Proof. First, let us prove that if the claim is true, then the level k nodes informed during round t have a different rightmost index than the nodes informed during the first $t-1$ rounds, so $T_k^t = \sum N_k^t$. For level 1, it is clear that $\delta_1 > T_1^{t-1}$. The level 2 nodes informed before round t have $\delta_2 \leq T_1^{t-1} + 1 + N_2^{t-1}$ and the nodes informed during round t have $\delta_2 \geq T_1^t + 2 = T_1^{t-1} + N_1^t + 2 = T_1^{t-1} + R_1^t + 3 = T_1^{t-1} + N_2^{t-1} + 3$. The level k nodes informed before round t have $\delta_k \leq T_1^{t+k-3} + 1 + L_2^{t+k-3} + \cdots + L_{k-1}^t + N_k^{t-1} \leq T_1^{t+k-2}$ and the nodes informed during round t have $\delta_k \geq T_1^{t+k-2} + 2 + L_2^{t+k-2} + \cdots + L_{k-1}^{t+1} > T_1^{t+k-2}$.

Now suppose that the claim is true until round $t-1$. We prove that the claim is true for round t by induction on t. We showed above that $T_k^{t-1} = \sum N_k^{t-1}$ if the claim is true for round $t-1$. The level 1 nodes that are informed during round t are node $T_1^{t-1}+1$, which is informed by the originator, and each node $\rho_1 = \bar{\delta}_1\delta_2$ such that $\delta_1\delta_2$ is a level 2 node that was informed during round $t-1$. By the induction hypothesis, these nodes informed by level 2 nodes are of the form $\rho_1 = T_1^{t-1}+1+j$, where $1 \le j \le N_2^{t-1}$. So, altogether, the level 1 nodes informed during round t are the nodes $T_1^{t-1}+j$, where $1 \le j \le 1+N_2^{t-1} = N_1^t$. This last equation is true because $L_1^t = 1$ and $N_2^{t-1} = R_1^t$ by (3) and (4).

The level 2 nodes that are informed during round t are as follows:

(i) every node $\delta_1\delta_2$ informed by a level 1 node δ_1 such that $\delta_2 = T_1^t+1+j$, where $1 \le j \le T_1^{t-1} = L_2^t$ (by (1));

(ii) every node $\rho_1\rho_2$ informed by a level 3 node $\delta_1\delta_2\delta_3$ which was informed during round $t-1$ such that $\rho_2 = \delta_3$, where $\rho_2 = T_1^t+1+L_2^t+j$, $1 \le j \le N_3^{t-1}$ by the induction hypothesis (at level 3).

Altogether, the level 2 nodes informed during round t are the nodes with rightmost index T_1^t+1+j, where $1 \le j \le L_2^t+N_3^{t-1} = N_2^t$. This last equation is true because $N_3^{t-1} = R_2^t$ and $L_2^t+R_2^t = N_2^t$ by (3) and (4).

The level k nodes that are informed during round t are as follows:

(i) every node $\delta_1\delta_2\cdots\delta_k$ informed by a level $k-1$ node which was informed during round $t-1$ such that $\delta_k = T_1^{t+k-2}+1+L_2^j$, where $1 \le j \le T_1^{t-1} = L_2^{t+k-2} + L_3^{t+k-3}+\cdots+L_{k-1}^{t+1}+j$, $1 \le j \le T_{k-1}^{t-2} = L_k^t$ (by (1));

(ii) every node $\rho_1\rho_2\cdots\rho_k$ informed by a level $k+1$ node $\delta_1\delta_2\cdots\delta_{k+1}$ which was informed during round $t-1$ such that the rightmost index $\rho_k = \delta_{k+1}$ satisfies $\rho_k = T_1^{t+k-2}+1+L_2^{t+k-2}+L_3^{t+k-3}+\cdots+L_{k-1}^{t+1}+j$, $1 \le j \le N_{k+1}^{t-1}$ by the induction hypothesis.

Altogether, the level k nodes informed during round t are the nodes with rightmost index $T_1^{t+k-2}+1+L_2^{t+k-2}+L_3^{t+k-3}+\cdots+L_{k-1}^{t+1}+j$, where $1 \le j \le L_k^t+N_{k+1}^{t-1} = N_k^t$. This last equation is true because $N_{k+1}^{t-1} = R_k^t$ and $L_k^t+R_k^t = N_k^t$ by (3) and (4). □

If we *truncate* Protocol **A** at some level k, that is, we discard all parts of the protocol involving levels greater than k, then we get a Protocol $\mathbf{A}_k$ that approximates Protocol **A**. In fact, Protocol $\mathbf{A}_2$ is exactly the protocol from [3]. Figure 1 shows the first six rounds of Protocol $\mathbf{A}_3$. Notice that Protocol $\mathbf{A}_3$ can inform one more level 1 node than Protocol $\mathbf{A}_2$ in six rounds (using the bold arcs). The sequence $\mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4, \ldots$ is a sequence of increasingly accurate approximations of Protocol **A**. We will solve the recurrence equations for Protocol $\mathbf{A}_k$, but, unfortunately, we have not been able to solve the recurrence equations for Protocol **A** without truncation.

Now, let us show how to find an expression for $T_1^t(\mathbf{A}_k)$ for the truncated protocol $\mathbf{A}_k$. First, note that for Protocol $\mathbf{A}_k$ we have to truncate (6) at level k. This is done by deleting the terms $L_i^{t-i+1}(\mathbf{A})$ for all $i > k$. Our aim will therefore be to express $T_1^t(\mathbf{A})$ for any k as the sum of two functions, the first depending on the $L_i^{t-i+1}(\mathbf{A})$ for $i \le k$, and the second depending on the $L_i^{t-i+1}(\mathbf{A})$ for $i > k$. Furthermore, we will show how to express the first function as a polynomial in the $T_1^j(\mathbf{A})$ for $j \le t-1$.

In summary, we want to obtain $T_1^t(\mathbf{A}) = P_k^t + g_k^t$, where P_k^t is a polynomial in the $T_1^j(\mathbf{A})$ with $j \le t-1$ and g_k^t is a function of the $L_i^{t-i+1}(\mathbf{A})$ with $i > k$. Therefore, for $\mathbf{A}_k$ we will obtain $T_1^t(\mathbf{A}_k) = Q_k^t$, where Q_k^t is the polynomial obtained from P_k^t by replacing the $T_1^j(\mathbf{A})$ by the $T_1^j(\mathbf{A}_k)$, $j \le t-1$. $T_1^t(\mathbf{A}_k)$ satisfies a generalized Fibonacci type of recurrence relation for which the asymptotic behavior is determined by the largest root of the associated polynomial.

For $k=2$, (6) gives $P_2^t = T_1^{t-1}+T_1^{t-2}+1$ and $g_2^t = \sum_{i\ge 3} L_i^{t-i+1}$, so we obtain

$T_1^t(\mathbf{A}_2) = T_1^{t-1}(\mathbf{A}_2) + T_1^{t-2}(\mathbf{A}_2) + 1$, which is Theorem 1.

For $k \geq 3$, we have to compute the L_i^{t-i+1} as functions of the T_1^j. This cannot be done directly, but it can be done using differences. For this purpose, we introduce a difference operator D such that for any function $f(t)$, $D[f(t)] = f(t) - f(t-1)$.

Using $T_1^t = T_1^{t-1} + D[T_1^t]$, (6) becomes

$$T_1^t = T_1^{t-1} + D[P_2^t] + \sum_{i\geq 3} D[L_i^{t-i+1}]. \tag{7}$$

Using $D[P_2^t] = T_1^{t-1} - T_1^{t-3}$, we get $T_1^t = P_3^t + g_3^t$, where

$$P_3^t = 2T_1^{t-1} - T_1^{t-3} + D[L_3^{t-2}] \text{ and } g_3^t = \sum_{i\geq 4} D[L_i^{t-i+1}]. \tag{8}$$

By (2) and (3), $D[L_3^{t-2}] = L_3^{t-2} - L_3^{t-3} = L_2^{t-4} + R_2^{t-4} = R_1^{t-3}$. By (4),

$$D[L_3^{t-2}] = R_1^{t-3} = N_1^{t-3} - L_1^{t-3} = T_1^{t-3} - T_1^{t-4} - 1. \tag{9}$$

Using (9) in (8), we get $P_3^t = 2T_1^{t-1} - T_1^{t-4} - 1$. This gives the following result.

THEOREM 2. $T_1^t(\mathbf{A}_3) = 2T_1^{t-1}(\mathbf{A}_3) - T_1^{t-4}(\mathbf{A}_3) - 1$.

This is a generalized Fibonacci sequence. The largest root of the associated polynomial $x^4 - 2x^3 + 1 = 0$ is $a_3 \approx 1.839$. Thus we have the following corollary.

COROLLARY 2. $T_1^t(\mathbf{A}_3) \sim 1.839^t$.

We will compute the polynomials for $k \geq 4$ using the following theorem.

THEOREM 3. $P_k^t = T_1^{t-1}(\mathbf{A}) + P_{k-1}^t - P_{k-1}^{t-1} + T_1^{t-3}(\mathbf{A}) - T_1^{t-4}(\mathbf{A}) - P_{k-2}^{t-3} + P_{k-2}^{t-4}$, $k \geq 4$.

Proof. First, we prove by induction that

$$P_k^t = T_1^{t-1} + D[P_{k-1}^t] + D^{k-2}[L_k^{t-k+1}] \text{ and } g_k^t = \sum_{i\geq k+1} D^{k-2}[L_i^{t-i+1}]. \tag{10}$$

This is true for $k = 2$ by (1) and (6) and for $k = 3$ by (8). Suppose that it is true for k. Then using $T_1^t = T_1^{t-1} + D[T_1^t]$, we obtain

$$T_1^t = T_1^{t-1} + D[P_k^t] + D^{k-1}[L_{k+1}^{t-k}] + \sum_{i\geq k+2} D^{k-1}[L_i^{t-i+1}],$$

so

$$P_{k+1}^t = T_1^{t-1} + D[P_k^t] + D^{k-1}[L_{k+1}^{t-k}] \text{ and } g_{k+1}^t = \sum_{i\geq k+2} D^{k-1}[L_i^{t-i+1}].$$

Note that the formula of the theorem can be rewritten as

$$P_k^t = T_1^{t-1} + D[P_{k-1}^t] + D[T_1^{t-3} - P_{k-2}^{t-3}]. \tag{11}$$

So, using (10), the theorem can be proved by proving that

$$D^{k-2}[L_k^{t-k+1}] = D[T_1^{t-3} - P_{k-2}^{t-3}]. \tag{12}$$

For $k \geq 3$, we can use (2) and (3) to obtain $D[L_k^t] = L_k^t - L_k^{t-1} = L_{k-1}^{t-2} + R_{k-1}^{t-2} = R_{k-2}^{t-1}$. So, for $k \geq 4$ we can use $D[L_{k-1}^{t+1}] = L_{k-2}^{t-1} + R_{k-2}^{t-1}$ to obtain

$$D[L_k^t] = D[L_{k-1}^{t+1}] - L_{k-2}^{t-1}. \tag{13}$$

By (13),

$$\begin{aligned} D^{k-2}[L_k^{t-k+1}] &= D^{k-2}[L_{k-1}^{t-k+2}] - D^{k-3}[L_{k-2}^{t-k}] \\ &= D[D^{k-3}[L_{k-1}^{t-k+2}] - D^{k-4}[L_{k-2}^{t-k}]]. \end{aligned} \tag{14}$$

By induction, (12) with $k-1$ substituted for k gives

$$D^{k-3}[L_{k-1}^{t-k+2}] = D[T_1^{t-3} - P_{k-3}^{t-3}], \tag{15}$$

and (12) with $k-2$ substituted for k and $t-3$ substituted for t gives

$$D^{k-4}[L_{k-2}^{t-k}] = D[T_1^{t-6} - P_{k-4}^{t-6}]. \tag{16}$$

Equation (11) with $k-2$ substituted for k and $t-3$ substituted for t gives

$$P_{k-2}^{t-3} = T_1^{t-4} + D[P_{k-3}^{t-3}] + D[T_1^{t-6} - P_{k-4}^{t-6}]. \tag{17}$$

Combining (15), (16), and (17), we obtain

$$\begin{aligned} &D^{k-3}[L_{k-1}^{t-k+2}] - D^{k-4}[L_{k-2}^{t-k}] \\ &\qquad = D[T_1^{t-3} - P_{k-3}^{t-3}] - P_{k-2}^{t-3} + T_1^{t-4} + D[P_{k-3}^{t-3}] = T_1^{t-3} - P_{k-2}^{t-3}. \qquad \square \end{aligned}$$

Using Theorem 3, we are able to compute all of the polynomials P_k^t for $k \geq 4$ and therefore the recurrence relations for $T_1^k(\mathbf{A}_k)$. For example, we obtain the following theorems.

THEOREM 4. $T_1^t(\mathbf{A}_4) = 3T_1^{t-1}(\mathbf{A}_4) - 2T_1^{t-2}(\mathbf{A}_4) + T_1^{t-3}(\mathbf{A}_4) - 3T_1^{t-4}(\mathbf{A}_4) + T_1^{t-5}(\mathbf{A}_4) + T_1^{t-6}(\mathbf{A}_4)$.

THEOREM 5. $T_1^t(\mathbf{A}_5) = 4T_1^{t-1}(\mathbf{A}_5) - 5T_1^{t-2}(\mathbf{A}_5) + 4T_1^{t-3}(\mathbf{A}_5) - 7T_1^{t-4}(\mathbf{A}_5) + 6T_1^{t-5}(\mathbf{A}_5) - T_1^{t-8}(\mathbf{A}_5)$.

The following table, Table 1, shows the value of the largest root a_k of the associated polynomial of $T_1^t(\mathbf{A}_k)$ for $k \leq 13$. The number of level 1 nodes informed by Protocol $\mathbf{A}_k$ is proportional to a_k^t.

TABLE 1
Asymptotic values for Protocol $\mathbf{A}_k$.

Protocol	Largest root	Protocol	Largest root	Protocol	Largest root
$\mathbf{A}_2$	1.61803	$\mathbf{A}_6$	1.96277	$\mathbf{A}_{10}$	1.98703
$\mathbf{A}_3$	1.83929	$\mathbf{A}_7$	1.97297	$\mathbf{A}_{11}$	1.98933
$\mathbf{A}_4$	1.91286	$\mathbf{A}_8$	1.97948	$\mathbf{A}_{12}$	1.99107
$\mathbf{A}_5$	1.94552	$\mathbf{A}_9$	1.98390	$\mathbf{A}_{13}$	1.99241

3. A sophisticated protocol. In Protocol **A**, each newly informed node at level $k \geq 3$ informs only one level $k-1$ node before broadcasting to the right. This leaves some nodes at levels 2 through $k-1$ uninformed. The idea of our second protocol, Protocol **B**, is to inform as many nodes as possible at the lower levels, because these nodes can introduce new dimensions by communicating to the right and this will lead to new level 1 nodes. A new dimension introduced by a level k node in a communication during round t can result in a newly informed node at level 1 as early as round $t + k$.

To describe Protocol **B** more precisely, we need to extend the notation used for Protocol **A**. We will distinguish nodes informed from the right by a node x according to the number of communications to the left that have been made by x. If x is a node at level k, then the first node that it informs at level $k-1$ is a type $R_{k-1,1}$ node, the second node that it informs at level $k-1$ is a type $R_{k-1,2}$ node, and so on. This gives the following notation:

$L_k^t(\mathbf{P})$: maximum number of level k nodes informed by level $k-1$ nodes during round t of Protocol $\mathbf{P}$

$R_{k,1}^t(\mathbf{P})$: maximum number of level k nodes informed during round t of Protocol $\mathbf{P}$ by level $k+1$ nodes which have not communicated to the left before round t

$R_{k,j}^t(\mathbf{P})$: maximum number of level k nodes informed during round t of Protocol $\mathbf{P}$ by level $k+1$ nodes which have informed exactly $j-1$ level k nodes before round t

$R_k^t(\mathbf{P}) = \sum_{j=1}^{k} R_{k,j}^t(\mathbf{P})$: maximum total number of level k nodes informed by level $k+1$ nodes during round t of Protocol $\mathbf{P}$

$N_k^t(\mathbf{P}) = L_k^t(\mathbf{P}) + R_k^t(\mathbf{P})$: maximum total number of level k nodes informed during round t of Protocol $\mathbf{P}$

$T_k^t(\mathbf{P}) = \sum_{i=1}^{t} N_k^i(\mathbf{P})$: maximum total number of level k nodes informed during the first t rounds of Protocol $\mathbf{P}$

Now we can describe Protocol **B** precisely. If x is a node that is informed during round t, then x informs the following uninformed nodes:

Protocol **B**

(i) If x is the originator, inform type L_1 nodes during rounds $t+1, t+2, \ldots$.

(ii) If x is a level 1 node, inform type L_2 nodes during rounds $t+1, t+2, \ldots$.

(iii) If x is a type L_k node, $k \geq 2$, inform a type $R_{k-1,1}$ node during round $t+1$, a type $R_{k-1,2}$ node during round $t+2, \ldots$, a type $R_{k-1,k-1}$ node during round $t+k-1$, and type L_{k+1} nodes during rounds $t+k, t+k+1, \ldots$.

(iv) If x is a type $R_{k,j}$ node, $k \geq 2$, $1 \leq j \leq k$, inform a type $R_{k-1,j}$ node during round $t+1$, a type $R_{k-1,j+1}$ node during round $t+2, \ldots$, a type $R_{k-1,k-1}$ node during round $t+k-j$, and type L_{k+1} nodes during rounds $t+k-j+1, t+k-j+2, \ldots$.

Before we analyze Protocol **B**, it will be helpful to look at an example of part of the protocol. Figure 2 shows a path from the originator to a level 5 node labelled $\delta_1\delta_2\delta_3\delta_4\alpha$. The tree of all communications to the left from node $\delta_1\delta_2\delta_3\delta_4\alpha$ is also shown, but all other communications have been omitted to keep the figure simple. In our example, dimension α is introduced in the communication right from node $\delta_1\delta_2\delta_3\delta_4$ to node $\delta_1\delta_2\delta_3\delta_4\alpha$ during round t. The rounds during which other nodes are informed and the types of the nodes are indicated in the figure.

Figure 2 illustrates some properties that we will use in our analysis. First, consider the path of type L_k nodes from the originator to node $\delta_1\delta_2\delta_3\delta_4\alpha$ along the top of the diagram. Each of the communications to the right shown in the figure introduces a new dimension, but the rounds during which these communications occur are not consecutive because communications to the left by the type L_k nodes are done before communications to the right. The type L_2 node labelled $\delta_1\delta_2$ makes one communication to the left (to a type $R_{1,1}$ node) before informing the type L_3 node $\delta_1\delta_2\delta_3$,

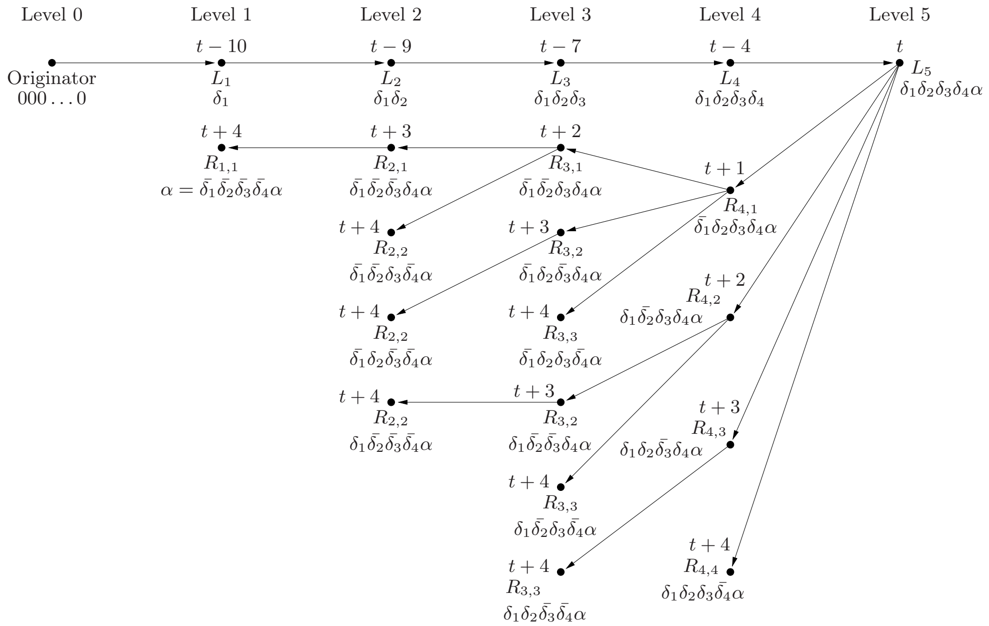

Level 0
Level 1
Level 2
Level 3
Level 4
Level 5
$t-10$
$t-9$
$t-7$
$t-4$
t
Originator
$000\ldots0$
L_1
δ_1
L_2
$\delta_1\delta_2$
L_3
$\delta_1\delta_2\delta_3$
L_4
$\delta_1\delta_2\delta_3\delta_4$
L_5
$\delta_1\delta_2\delta_3\delta_4\alpha$
$t+4$
$R_{1,1}$
$\alpha = \bar{\delta}_1\bar{\delta}_2\bar{\delta}_3\bar{\delta}_4\alpha$
$t+3$
$R_{2,1}$
$\bar{\delta}_1\bar{\delta}_2\bar{\delta}_3\delta_4\alpha$
$t+2$
$R_{3,1}$
$\bar{\delta}_1\bar{\delta}_2\delta_3\delta_4\alpha$
$t+1$
$R_{4,1}$
$\bar{\delta}_1\delta_2\delta_3\delta_4\alpha$
$t+4$
$R_{2,2}$
$\bar{\delta}_1\bar{\delta}_2\delta_3\bar{\delta}_4\alpha$
$t+3$
$R_{3,2}$
$\bar{\delta}_1\delta_2\bar{\delta}_3\delta_4\alpha$
$t+2$
$R_{4,2}$
$\delta_1\bar{\delta}_2\delta_3\delta_4\alpha$
$t+4$
$R_{2,2}$
$\bar{\delta}_1\delta_2\bar{\delta}_3\bar{\delta}_4\alpha$
$t+4$
$R_{3,3}$
$\bar{\delta}_1\delta_2\delta_3\bar{\delta}_4\alpha$
$t+4$
$R_{2,2}$
$\delta_1\bar{\delta}_2\bar{\delta}_3\bar{\delta}_4\alpha$
$t+3$
$R_{3,2}$
$\delta_1\bar{\delta}_2\bar{\delta}_3\delta_4\alpha$
$t+3$
$R_{4,3}$
$\delta_1\delta_2\bar{\delta}_3\delta_4\alpha$
$t+4$
$R_{3,3}$
$\delta_1\bar{\delta}_2\delta_3\bar{\delta}_4\alpha$
$t+4$
$R_{3,3}$
$\delta_1\delta_2\bar{\delta}_3\bar{\delta}_4\alpha$
$t+4$
$R_{4,4}$
$\delta_1\delta_2\delta_3\bar{\delta}_4\alpha$

FIG. 2. *The broadcast tree of* $T_{\delta_1\delta_2\delta_3\delta_4\alpha}$.

the type L_3 node makes two communications to the left, and, in general, a type L_k node, $k \geq 2$, will make $k-1$ communications to the left before communicating to the right. So, a type L_k node will receive the message $\sum_{i=1}^{k-1} i = \frac{k(k-1)}{2}$ rounds after the originator initiates the path to the right. Next, we can consider node $\delta_1\delta_2\delta_3\delta_4\alpha$ to be the root of a broadcast tree, which we denote by $T_{\delta_1\delta_2\delta_3\delta_4\alpha}$, going left and starting in round $t+1$. The tree $T_{\delta_1\delta_2\delta_3\delta_4\alpha}$ is a complete binomial tree of depth 4 and contains all nodes at levels 1 through 5 with a 1 in position α. Notice that the number of level $i+1$ nodes in $T_{\delta_1\delta_2\delta_3\delta_4\alpha}$ is $\binom{4}{i}$, $1 \leq i+1 \leq 5$. In particular, $T_{\delta_1\delta_2\delta_3\delta_4\alpha}$ contains one new level 1 node. Another useful property of $T_{\delta_1\delta_2\delta_3\delta_4\alpha}$ is that all $\sum_{i=0}^{4}\binom{4}{i} = 2^4$ nodes, including $\delta_1\delta_2\delta_3\delta_4\alpha$, finish their communications to the left during the same round $t+4$, so they can all start communicating to the right simultaneously in round $t+5$. Each of these communications to the right introduces a new dimension, and each node that receives the message from the left during round $t+5$ is the root of a broadcast tree going left that contains a new level 1 node. In general, the broadcast tree of a type L_k node that is informed during round t contains 2^{k-1} nodes, including one level 1 node that is informed during round $t+k-1$, and all nodes of this tree communicate to the right during round $t+k$ introducing 2^{k-1} new dimensions.

With this intuition, we can write the recurrence equations for Protocol **B**:

$$
\begin{array}{rlll}
& L_1^t(\mathbf{B}) &= 1 & t \ge 1 \\[2ex]
& L_k^t(\mathbf{B}) &= 0 & t \le \dfrac{k(k-1)}{2},\ k \ge 2 \\[2ex]
(18) & L_k^t(\mathbf{B}) &= \displaystyle\sum_{i=1}^{t-k+1} L_{k-1}^i(\mathbf{B}) + \sum_{j=1}^{k-1} \sum_{i=1}^{t-k+j} R_{k-1,j}^i(\mathbf{B}) & t \ge \dfrac{k(k-1)}{2} + 1,\ k \ge 2 \\[2ex]
& R_{k,j}^t(\mathbf{B}) &= 0 & 1 \le k < j \\[2ex]
& R_{k,j}^t(\mathbf{B}) &= 0 & t \le \dfrac{k(k+1)}{2} + j,\ 1 \le j \le k \\[2ex]
(19) & R_{k,j}^t(\mathbf{B}) &= L_{k+1}^{t-j}(\mathbf{B}) + \displaystyle\sum_{i=1}^{j} R_{k+1,i}^{t-j+i-1}(\mathbf{B}) & t \ge \dfrac{k(k+1)}{2} + j + 1, \\
& & & 1 \le j \le k \\[2ex]
& R_k^t(\mathbf{B}) &= 0 & t \le \dfrac{k(k+1)}{2} + 1,\ k \ge 1 \\[2ex]
& R_k^t(\mathbf{B}) &= \displaystyle\sum_{j=1}^{k} R_{k,j}^t(\mathbf{B}) & t \ge \dfrac{k(k+1)}{2} + 2,\ k \ge 1 \\[2ex]
& N_k^t(\mathbf{B}) &= L_k^t(\mathbf{B}) + R_k^t(\mathbf{B}) & t \ge 1,\ k \ge 1 \\[2ex]
& T_k^t(\mathbf{B}) &= \displaystyle\sum_{i=1}^{t} N_k^i(\mathbf{B}) = \sum_{i=1}^{t} (L_k^i(\mathbf{B}) + R_k^i(\mathbf{B})) & t \ge 1,\ k \ge 1.
\end{array}
$$

THEOREM 6. *Protocol* **B** *informs* 2^t *level* 1 *nodes no later than round* $t + \lceil \frac{\sqrt{8t+1}-1}{2} \rceil$.

Proof. During each round of Protocol **B**, each informed node informs an uninformed neighbor, so the total number of informed nodes after t rounds is 2^t. By (18), the most distant informed node from the originator after t rounds is at level at most k_t, where $t \le \frac{k_t(k_t+1)}{2}$. So, $k_t = \lceil \frac{\sqrt{8t+1}-1}{2} \rceil$.

From the discussion above, a type L_k node x that is informed during round t is the root of a broadcast tree T_x with 2^k nodes that are all informed during round $t+k-1$. In particular, T_x includes a level 1 node which we will call $f_1(x)$.

Now we will show that at time $t + k_t$ there are at least 2^t informed level 1 nodes. To prove this, we will associate with each of the 2^t nodes informed during the first t rounds, a level 1 node that is informed no later than round $t + k_t$.

If node x is of type $L_k^{t'}$, the associated level 1 node is $f_1(x)$ of the broadcast tree T_x, and $f_1(x)$ is informed no later than round $t' + k - 1 \le t + k_t - 1$.

If node x is of type R_m, it belongs to the broadcast tree of a type L_k^{t-h} node $r(x)$ with $k \le k_t$; therefore, $m \le k_t - 1$.

Case 1. If $h \geq k-1$, then all of the nodes of the broadcast tree of $r(x)$ are informed during round t. During round $t+1$, x will inform a type L_{m+1}^{t+1} node y, which in turn informs the level 1 node $f_1(y)$ m rounds later, that is, during round $t+m+1 \leq t+k_t$.

Case 2. If $h < k-1$, then only $2^h - 1$ nodes of the broadcast tree of $r(x)$ are informed during the first t rounds. We will show that we can associate at least 2^h-1 informed level 1 nodes with this broadcast tree. Indeed, all the nodes of the broadcast tree of $r(x)$ are informed during round $t-h+k-1$. During round $t-h+k$, any type R_p node of the broadcast tree will inform a type L_{p+1} node, which in turn informs a level 1 node during round $t-h+k+p$. So, no later than round $t+k_t$ we have at least as many informed level 1 nodes as the number of R_p nodes with $p \leq h$. The number of such R_p nodes is $1+\binom{k-1}{1}+\cdots+\binom{k-1}{h-1} > 1+\binom{h}{1}+\cdots+\binom{h}{h-1} = 2^h-1$. □

Corollary 3. *In the hypercube with $N = 2^n$ nodes, neighborhood broadcasting can be done in at most $log_2 n + \lceil\sqrt{2log_2 n}\,\rceil$ rounds.*

Corollary 4. *For any fixed $\epsilon > 0$ and sufficiently large t, the number of level* 1 *nodes informed by Protocol* **B** *in t rounds is at least $(2-\epsilon)^t$.*

Proof. After $t = u + \sqrt{2u+1}$ rounds, we have 2^u level 1 nodes informed. Solving for u we get $u = t+1-\sqrt{2t+1}$. So, at time t there are at least $2^{t+1-\sqrt{2t+1}}$ informed level 1 nodes. For any fixed ϵ and sufficiently large t, $2^{t+1-\sqrt{2t+1}} \geq (2-\epsilon)^t$. □

We can *truncate* Protocol **B** at some level $k \geq 3$, in the same way that we truncated Protocol **A**, to get a sequence $\mathbf{B}_3, \mathbf{B}_4, \mathbf{B}_5, \ldots$, of increasingly accurate approximations of Protocol **B**. Protocol $\mathbf{B}_2$ is exactly the same as Protocol $\mathbf{A}_2$. We begin our analysis in the same way as we did for Protocol **A** (cf. (6)) by simplifying the expression for $T_1^t(\mathbf{B})$:

$$T_1^t = T_1^{t-1} + N_1^t = T_1^{t-1} + 1 + L_2^{t-1} + L_3^{t-2} + \cdots + L_k^{t-k+1} + \cdots .$$

Noting that $L_2^{t-1} = T_1^{t-2}$, we get

$$T_1^t = T_1^{t-1} + T_1^{t-2} + 1 + \sum_{i\geq 3} L_i^{t-i+1}. \tag{20}$$

Using the difference operator with (18), we get

$$D[L_k^t] = L_{k-1}^{t-k+1} + \sum_{j=1}^{k-1} R_{k-1,j}^{t-k+j}.$$

By repeated use of (19), we get

$$D[L_3^t] = L_2^{t-2} + 2L_3^{t-3} + 3L_4^{t-4} + \cdots + (i-1)L_i^{t-i} + \cdots , \tag{21}$$

$$D[L_4^t] = L_3^{t-3} + 3L_4^{t-4} + 6L_5^{t-5} + \cdots + \binom{i-1}{2} L_i^{t-i} + \cdots , \tag{22}$$

and more generally

$$D[L_k^t] = \sum_{i\geq k-1} \binom{i-1}{k-2} L_i^{t-i}. \tag{23}$$

Theorem 7. $T_1^t(\mathbf{B}_3) = 2T_1^{t-1}(\mathbf{B}_3) + T_1^{t-3}(\mathbf{B}_3) - 2T_1^{t-4}(\mathbf{B}_3) - T_1^{t-5}(\mathbf{B}_3) - 2.$

Proof. Truncating (20) at level 3 gives

$$T_1^t = T_1^{t-1} + T_1^{t-2} + 1 + L_3^{t-2}. \tag{24}$$

Applying the difference operator to (24) we get

$$T_1^t = T_1^{t-1} + D[T_1^t] = 2T_1^{t-1} - T_1^{t-3} + D[L_3^{t-2}].$$

By (21), $D[L_3^{t-2}] = L_2^{t-4} + 2L_3^{t-5} = T_1^{t-5} + 2L_3^{t-5}$, so we get

$$T_1^t = 2T_1^{t-1} - T_1^{t-3} + T_1^{t-5} + 2L_3^{t-5}. \tag{25}$$

Substituting $t-3$ for t in (24) gives $L_3^{t-5} = T_1^{t-3} - (T_1^{t-4} + T_1^{t-5} + 1)$ and so (25) becomes $T_1^t = 2T_1^{t-1} + T_1^{t-3} - 2T_1^{t-4} - T_1^{t-5} - 2$. □

Corollary 5. *$T_1^t(\mathbf{B}_3) \sim 1.913^t$.*

It is interesting to note that $T_1^t(\mathbf{B}_3) = T_1^t(\mathbf{A}_4)$ (compare Theorems 7 and 4) even though the protocols are different. The originator and nodes of types L_1 and L_2 behave the same in the two protocols. In Protocol $\mathbf{A}_4$, each level 3 node informs a type $R_{2,1}$ node and then informs level 4 nodes until the end of the protocol. Each level 4 node informs one level 3 node and then becomes idle. In Protocol $\mathbf{B}_3$, each level 3 node informs a type $R_{2,1}$ node and a type $R_{2,2}$ node and then becomes idle. The type $R_{2,1}$ nodes behave the same in the two protocols. To see that the two protocols inform the same level 1 nodes during each round, we will compare the parts of the protocols that are different. Figure 3 shows parts of the broadcast trees rooted at a level 3 node $\delta_1\delta_2\delta_3$. In both protocols, node $\delta_1\delta_2\delta_3$ informs the type $R_{2,1}$ node $\delta_2\delta_3$ during round $t+1$. Node $\delta_2\delta_3$ behaves the same in both protocols, so it is not shown. In Figure 3, communications that are in Protocol $\mathbf{A}_4$ are shown in normal typeface and communications that are in Protocol $\mathbf{B}_3$ are shown in bold typeface. Notice that the two protocols inform different level 3 nodes, but the same level 2 nodes are informed. In both protocols, node $\delta_3\alpha_1$ will inform the new level 1 node α_1 during round $t+5$, node $\delta_3\alpha_2$ will inform the new level 1 node α_2 during round $t+6$, and so on.

The proofs of the next two theorems appear in the appendix.

Theorem 8.

$$\begin{aligned} T_1^t(\mathbf{B}_4) = 3T_1^{t-1}(\mathbf{B}_4) - 2T_1^{t-2}(\mathbf{B}_4) + T_1^{t-3}(\mathbf{B}_4) - 5T_1^{t-5}(\mathbf{B}_4) \\ + T_1^{t-6}(\mathbf{B}_4) + 3T_1^{t-8}(\mathbf{B}_4) + T_1^{t-9}(\mathbf{B}_4) + 3. \end{aligned}$$

Corollary 6. *$T_1^t(\mathbf{B}_4) \sim 1.9867^t$.*

Theorem 9.

$$\begin{aligned} T_1^t(\mathbf{B}_5) = 4T_1^{t-1}(\mathbf{B}_5) &- 5T_1^{t-2}(\mathbf{B}_5) + 3T_1^{t-3}(\mathbf{B}_5) - T_1^{t-4}(\mathbf{B}_5) \\ &- T_1^{t-5}(\mathbf{B}_5) - 6T_1^{t-6}(\mathbf{B}_5) + 7T_1^{t-7}(\mathbf{B}_5) - T_1^{t-8}(\mathbf{B}_5) \\ &+ 4T_1^{t-9}(\mathbf{B}_5) + 7T_1^{t-10}(\mathbf{B}_5) - 4T_1^{t-11}(\mathbf{B}_5) - 2T_1^{t-12}(\mathbf{B}_5) \\ &- 4T_1^{t-13}(\mathbf{B}_5) - T_1^{t-14}(\mathbf{B}_5) - 4. \end{aligned}$$

Corollary 7. *$T_1^t(\mathbf{B}_5) \sim 1.9989^t$.*

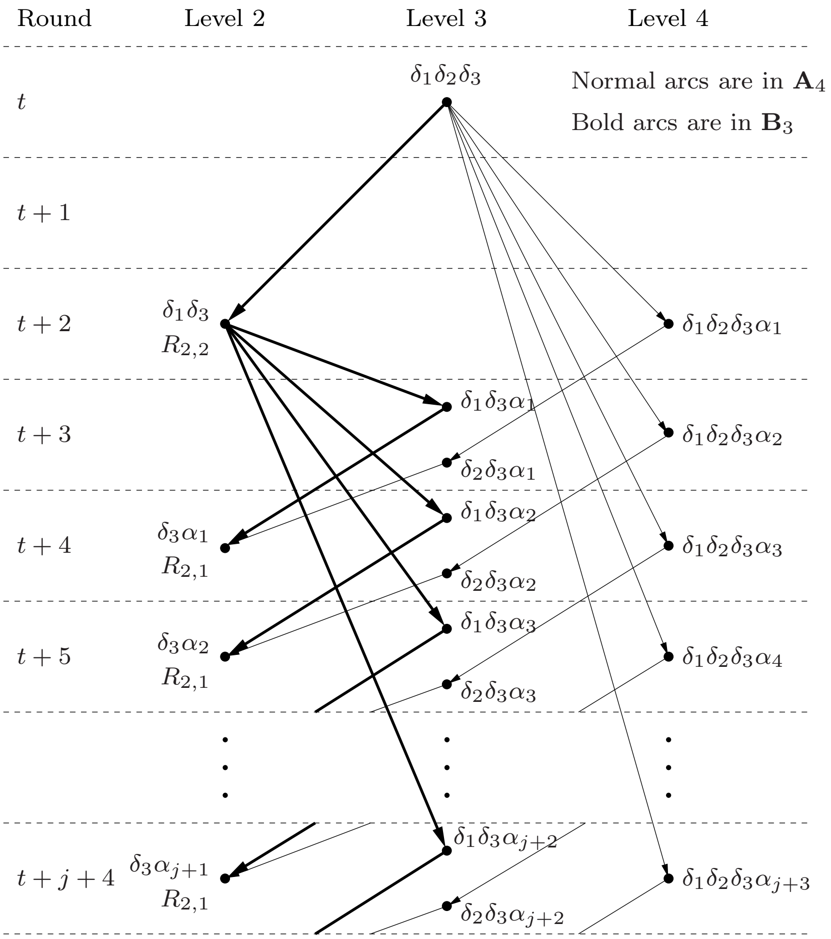

Round
Level 2
Level 3
Level 4
$\delta_1\delta_2\delta_3$
Normal arcs are in $\mathbf{A}_4$
Bold arcs are in $\mathbf{B}_3$
t
$t+1$
$t+2$
$\delta_1\delta_3$
$R_{2,2}$
$\delta_1\delta_2\delta_3\alpha_1$
$t+3$
$\delta_1\delta_3\alpha_1$
$\delta_1\delta_2\delta_3\alpha_2$
$\delta_2\delta_3\alpha_1$
$t+4$
$\delta_3\alpha_1$
$R_{2,1}$
$\delta_1\delta_3\alpha_2$
$\delta_1\delta_2\delta_3\alpha_3$
$\delta_2\delta_3\alpha_2$
$t+5$
$\delta_3\alpha_2$
$R_{2,1}$
$\delta_1\delta_3\alpha_3$
$\delta_1\delta_2\delta_3\alpha_4$
$\delta_2\delta_3\alpha_3$
$t+j+4$
$\delta_3\alpha_{j+1}$
$R_{2,1}$
$\delta_1\delta_3\alpha_{j+2}$
$\delta_1\delta_2\delta_3\alpha_{j+3}$
$\delta_2\delta_3\alpha_{j+2}$

FIG. 3. *Differences between Protocols* $\mathbf{A}_4$ *and* $\mathbf{B}_3$.

4. Conclusions. Table 2 shows the numbers of informed level 1 nodes for several protocols. These numbers were obtained using programs based on the recurrence relations in this paper. The numbers for the truncated protocols can also be obtained using the theorems in this paper. The protocols in Table 2 are ordered left to right according to an increasing number of informed level 1 nodes. An entry shown in bold font indicates the first round during which a protocol is better than the protocol on its left.

It is interesting to examine the last row of Table 2 which shows the numbers of informed nodes after 30 rounds. Protocol $\mathbf{A}_3$ nearly doubles the number of informed nodes compared to Protocol $\mathbf{A}_2$, and Protocol $\mathbf{A}_4$ more than doubles it again. Protocol **B** is so much better than Protocol **A** that even the truncated Protocol $\mathbf{B}_4$ outperforms the untruncated Protocol **A**. We know from Corollary 4 that Protocol **B** is asymptotically optimal. The last two columns suggest that Protocol $\mathbf{B}_4$ is almost as good as the untruncated Protocol **B**. To examine this further, we used programs based on the recurrence relations to determine lower bounds on the rates that the truncated protocols inform level 1 nodes. More precisely, the number of level 1 nodes informed by each truncated Protocol $\mathbf{A}_k$ is proportional to a_k^t, where a_k is the largest

TABLE 2
Level 1 *nodes informed.*

Round	$\mathbf{A}_2 = \mathbf{B}_2$	$\mathbf{A}_3$	$\mathbf{A}_4 = \mathbf{B}_3$	$\mathbf{A}$	$\mathbf{B}_4$	$\mathbf{B}$
1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	4	4	4	4	4	4
4	7	7	7	7	7	7
5	12	12	12	12	12	12
6	20	**21**	21	21	21	21
7	33	37	37	37	37	37
8	54	66	66	66	66	66
9	88	119	**120**	120	120	120
10	143	216	221	221	**222**	222
11	232	394	411	411	416	416
12	376	721	771	**772**	788	788
13	609	1322	1455	1461	1507	1507
14	986	2427	2757	2780	2905	2905
15	1596	4459	5240	5316	5634	**5635**
20	17710	93723	132662	142644	163510	164203
25	196417	1972659	3392169	4013545	4958328	5039922
30	2178308	41523767	86856182	115996781	152476127	158120581

root of the associated polynomial of $T_1^t(\mathbf{A}_k)$. Similarly, the performance of Protocol $\mathbf{B}_k$ is proportional to b_k^t, where b_k is the largest root of the associated polynomial of $T_1^t(\mathbf{B}_k)$. The results are shown in Figure 4. The lower curve shows the sequence $\{a_k\}$, $k = 3, 4, 5, \ldots$, and the upper curve shows the sequence $\{b_k\}$, $k = 3, 4, 5, \ldots$. (We have omitted the value $a_2 = b_2 = \frac{1+\sqrt{5}}{2} \approx 1.618$ for Protocol $\mathbf{A}_2$ = Protocol $\mathbf{B}_2$ to reduce the range of the vertical scale of the graph.) The graph shows that the sequence $\{b_k\}$ converges very quickly with increasing k towards the optimal value 2 (shown as a horizontal line at the top of the graph). The sequence $\{a_k\}$ converges more slowly, but it is clear that it is also approaching the optimal value.

An alternative approach to solving the recurrence relations in this paper is to use the matrix approach described in [7, 8]. We have applied this approach to the protocols in this paper and obtained the same polynomials for the truncated protocols.

We note that the recurrence relations that we have presented in this paper apply to k-neighborhood broadcasting for any $k \geq 1$. It is possible to extend our analysis to determine expressions for the truncated protocols for $k > 1$, but the derivations might be quite long.

Finally, we reiterate that improvement of the lower bound for neighborhood broadcasting or a proof that no protocol can inform the neighbors of the originator faster than Protocol $\mathbf{B}$ are open problems.

Appendix: Proofs of Theorems 8 and 9.

THEOREM 8.

$$T_1^t(\mathbf{B}_4) = 3T_1^{t-1}(\mathbf{B}_4) - 2T_1^{t-2}(\mathbf{B}_4) + T_1^{t-3}(\mathbf{B}_4) - 5T_1^{t-5}(\mathbf{B}_4)$$
$$+T_1^{t-6}(\mathbf{B}_4) + 3T_1^{t-8}(\mathbf{B}_4) + T_1^{t-9}(\mathbf{B}_4) + 3.$$

Proof. In this case, (20) becomes

$$T_1^t = T_1^{t-1} + T_1^{t-2} + 1 + L_3^{t-2} + L_4^{t-3} \tag{26}$$

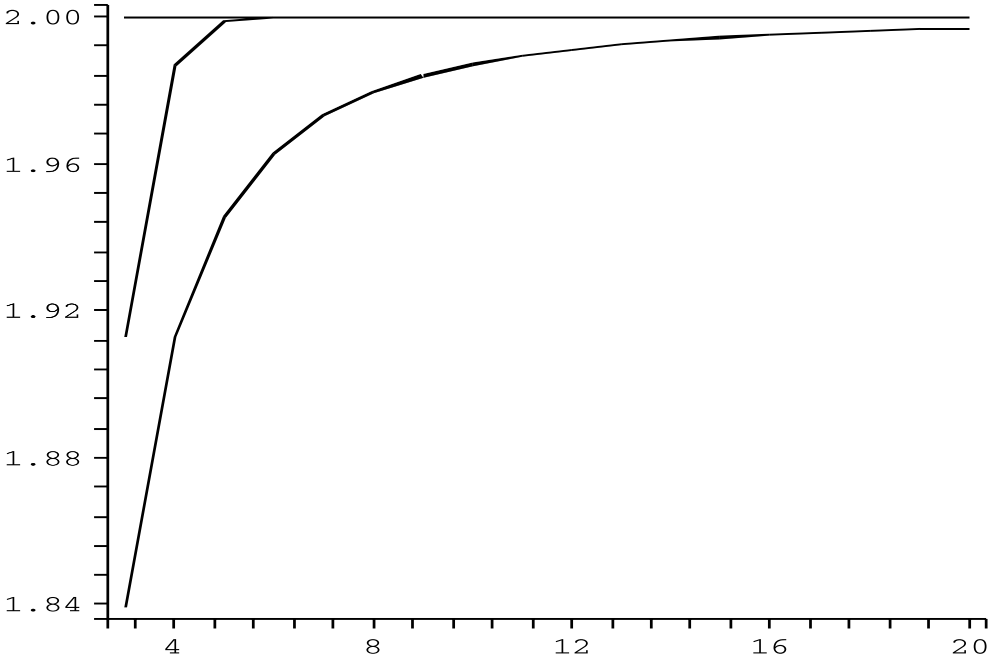


FIG. 4. *Asymptotic convergence of largest roots a_k and b_k for $k \geq 3$.*

and by difference

$$T_1^t = T_1^{t-1} + D[T_1^t] = 2T_1^{t-1} - T_1^{t-3} + D[L_3^{t-2}] + D[L_4^{t-3}]. \tag{27}$$

Truncating (21) and (22) at level 4 and using the value $L_3^{t-3} + L_4^{t-4} = T_1^{t-1} - T_1^{t-2} - T_1^{t-3} - 1$ deduced from (26) (with $t-1$ substituted for t) gives

$$D[L_3^t] = T_1^{t-3} + 2L_3^{t-3} + 3L_4^{t-4} = 2T_1^{t-1} - 2T_1^{t-2} - T_1^{t-3} - 2 + L_4^{t-4}, \tag{28}$$

$$D[L_4^t] = L_3^{t-3} + 3L_4^{t-4} = T_1^{t-1} - T_1^{t-2} - T_1^{t-3} - 1 + 2L_4^{t-4}. \tag{29}$$

Using (27), (28) with $t-2$ substituted for t, and (29) with $t-3$ substituted for t we get

$$T_1^t = 2T_1^{t-1} + T_1^{t-3} - T_1^{t-4} - 2T_1^{t-5} - T_1^{t-6} - 3 + L_4^{t-6} + 2L_4^{t-7}. \tag{30}$$

We can write (30) as $T_1^t = P^t + F^t(L_4)$, where

$$P^t = 2T_1^{t-1} + T_1^{t-3} - T_1^{t-4} - 2T_1^{t-5} - T_1^{t-6} - 3 \tag{31}$$

and

$$F^t(L_4) = L_4^{t-6} + 2L_4^{t-7} = T_1^t - P^t. \tag{32}$$

Using the difference operator we get

$$T_1^t = T_1^{t-1} + D[T_1^t] = T_1^{t-1} + D[P^t] + D[L_4^{t-6}] + 2D[L_4^{t-7}]. \tag{33}$$

By (29),

$$\begin{aligned} D[L_4^{t-6}]+2D[L_4^{t-7}] &= (T_1^{t-7}+T_1^{t-8}-3T_1^{t-9}-2T_1^{t-10}-3)+2L_4^{t-10}+4L_4^{t-11} \\ &= (T_1^{t-7}+T_1^{t-8}-3T_1^{t-9}-2T_1^{t-10}-3)+2F^{t-4}(L_4). \qquad (34) \end{aligned}$$

Using (33), (34), (32) with $t-4$ substituted for t, and (31), we get

$$\begin{aligned} T_1^t &= T_1^{t-1}+(P^t-P^{t-1})+(T_1^{t-7}+T_1^{t-8}-3T_1^{t-9}-2T_1^{t-10}-3)+2(T_1^{t-4}-P^{t-4}) \\ &= 3T_1^{t-1}-2T_1^{t-2}+T_1^{t-3}-5T_1^{t-5}+T_1^{t-6}+3T_1^{t-8}+T_1^{t-9}+3. \qquad \Box \end{aligned}$$

THEOREM 9.

$$\begin{aligned} T_1^t(\mathbf{B}_5) &= 4T_1^{t-1}(\mathbf{B}_5)-5T_1^{t-2}(\mathbf{B}_5)+3T_1^{t-3}(\mathbf{B}_5)-T_1^{t-4}(\mathbf{B}_5) \\ &\quad -T_1^{t-5}(\mathbf{B}_5)-6T_1^{t-6}(\mathbf{B}_5)+7T_1^{t-7}(\mathbf{B}_5)-T_1^{t-8}(\mathbf{B}_5) \\ &\quad +4T_1^{t-9}(\mathbf{B}_5)+7T_1^{t-10}(\mathbf{B}_5)-4T_1^{t-11}(\mathbf{B}_5)-2T_1^{t-12}(\mathbf{B}_5) \\ &\quad -4T_1^{t-13}(\mathbf{B}_5)-T_1^{t-14}(\mathbf{B}_5)-4. \end{aligned}$$

Proof. Truncating (20) at level 5 gives

$$T_1^t = T_1^{t-1}+T_1^{t-2}+1+L_3^{t-2}+L_4^{t-3}+L_5^{t-4}. \tag{35}$$

Using the value of $L_3^{t-3}+L_4^{t-4}+L_5^{t-5}$ deduced from (35) (with $t-1$ substituted for t) in (21), (22), and (23) gives

$$D[L_3^t] = L_2^{t-2}+2L_3^{t-3}+3L_4^{t-4}+4L_5^{t-5} = 2T_1^{t-1}-2T_1^{t-2}-T_1^{t-3}-2+L_4^{t-4}+2L_5^{t-5}, \tag{36}$$

$$D[L_4^t] = L_3^{t-3}+3L_4^{t-4}+6L_5^{t-5} = T_1^{t-1}-T_1^{t-2}-T_1^{t-3}-1+2L_4^{t-4}+5L_5^{t-5}, \tag{37}$$

$$D[L_5^t] = L_4^{t-4}+4L_5^{t-5}. \tag{38}$$

By difference we get

$$T_1^t = T_1^{t-1}+D[T_1^t] = 2T_1^{t-1}-T_1^{t-3}+D[L_3^{t-2}]+D[L_4^{t-3}]+D[L_5^{t-4}].$$

Using (36), (37), and (38) with $t-2$, $t-3$, and $t-4$ substituted for t, respectively, gives $T_1^t = Q^t + F^t(L_4, L_5)$, where

$$Q^t = 2T_1^{t-1}+T_1^{t-3}-T_1^{t-4}-2T_1^{t-5}-T_1^{t-6}-3 \text{ and} \tag{39}$$

$$F^t(L_4,L_5) = L_4^{t-6}+2L_4^{t-7}+L_4^{t-8}+2L_5^{t-7}+5L_5^{t-8}+4L_5^{t-9} = T_1^t - Q^t. \tag{40}$$

By difference using $t-6$, $t-7$, $t-8$ substituted for t in (37) and $t-7$, $t-8$, $t-9$ substituted for t in (38), we get

$$\begin{aligned} T_1^t &= T_1^{t-1}+(Q^t-Q^{t-1})+(F^t(L_4,L_5)-F^{t-1}(L_4,L_5)) \qquad (41) \\ &= T_1^{t-1}+(Q^t-Q^{t-1})+(T_1^{t-7}-T_1^{t-8}-T_1^{t-9}-1) \\ &\quad +2(T_1^{t-8}-T_1^{t-9}-T_1^{t-10}-1)+(T_1^{t-9}-T_1^{t-10}-T_1^{t-11}-1) \\ &\quad +2L_4^{t-10}+6L_4^{t-11}+7L_4^{t-12}+4L_4^{t-13} \\ &\quad +5L_5^{t-11}+18L_5^{t-12}+25L_5^{t-13}+16L_5^{t-14}. \end{aligned}$$

The last two lines of (41) involving terms in L_4^j and L_5^j can be written as

$$(42)\qquad 2F^{t-4}(L_4, L_5) + 4F^{t-5}(L_4, L_5) - 2L_4^{t-11} - 3L_4^{t-12} + L_5^{t-11} - 3L_5^{t-13}.$$

Using (40) to deduce the values of $F^{t-4}(L_4, L_5)$ and $F^{t-5}(L_4, L_5)$ in (42) (by substituting $t-4$ and $t-5$ for t, respectively), (41) becomes $T_1^t = S^t + G(L_4, L_5)$, where

$$S^t = 3T_1^{t-1} - 2T_1^{t-2} + T_1^{t-3} - T_1^{t-5} - 7T_1^{t-6} - T_1^{t-8} + 6T_1^{t-9} + 7T_1^{t-10} + 3T_1^{t-11} + 14$$

and $G(L_4, L_5) = -2L_4^{t-11} - 3L_4^{t-12} + L_5^{t-11} - 3L_5^{t-13}$.

By difference again, using $t-11$, $t-12$ substituted for t in (37), and $t-11$, $t-13$ substituted for t in (38), we get

$$\begin{aligned}(43)\qquad T_1^t &= T_1^{t-1} + (S^t - S^{t-1})\\ &\quad -2(T_1^{t-12} - T_1^{t-13} - T_1^{t-14} - 1) - 3(T_1^{t-13} - T_1^{t-14} - T_1^{t-15} - 1)\\ &\quad -3L_4^{t-15} - 6L_4^{t-16} - 3L_4^{t-17} - 6L_5^{t-16} - 15L_5^{t-17} - 12L_5^{t-18}.\end{aligned}$$

The third line of (43) involving terms in L_4^j and L_5^j is exactly $-3F^{t-9}(L_4, L_5)$. Using (40) with $t-9$ substituted for t to get an expression for $-3F^{t-9}(L_4, L_5)$, (43) becomes

$$\begin{aligned}T_1^t &= 4T_1^{t-1} - 5T_1^{t-2} + 3T_1^{t-3} - T_1^{t-4} - T_1^{t-5} - 6T_1^{t-6} + 7T_1^{t-7} - T_1^{t-8}\\ &\quad +4T_1^{t-9} + 7T_1^{t-10} - 4T_1^{t-11} - 2T_1^{t-12} - 4T_1^{t-13} - T_1^{t-14} - 4. \qquad \square\end{aligned}$$

Acknowledgments. We thank the referees for their suggestions for improving the presentation of this paper.

REFERENCES

[1] J.-C. Bermond, A. Ferreira, S. Pérennes, and J. G. Peters, *Neighbourhood Broadcasting in Hypercubes*, manuscript, 1998.

[2] J.-C. Bermond, A. Ferreira, and J. G. Peters, *Partial broadcasting in hypercubes*, in the International Workshop on Interconnection Networks (IWIN), Luminy, France, 1991.

[3] M. Cosnard and A. Ferreira, *On the real power of loosely coupled parallel architectures*, Parallel Process. Lett., 1 (1991), pp. 103–111.

[4] S. Even and B. Monien, *On the number of rounds necessary to disseminate information*, in Proceedings of the 1st ACM Symposium on Parallel Algorithms and Architectures, Santa Fe, New Mexico, 1989, pp. 318–327.

[5] G. Fertin and A. Raspaud, *k-neighborhood broadcasting*, in Proceedings of the International Colloquium on Structural Information and Communication Complexity (SIROOCO 8), Vall de Núria, Spain, 2001, Proceedings in Informatics, Vol. 11, Carleton Scientific Ontario, 2001, pp. 133–146.

[6] G. Fertin and A. Raspaud, *Neighborhood communications in networks*, in Proceedings of the Euroconference on Combinatorics, Graph Theory and Applications (COMB01), Barcelona, Spain, 2001, Electron. Notes Discrete Math., Vol. 10, Elsevier Amsterdam, 2001.

[7] M. Flammini and S. Pérennes, *On the optimality of general lower bounds for broadcasting and gossiping*, SIAM J. Discrete Math., 14 (2001), pp. 267–282.

[8] M. Flammini and S. Pérennes, *Lower bounds on the broadcasting and gossiping time for restricted protocols*, SIAM J. Discrete Math., 17 (2004), pp. 521–540.

[9] P. Fraigniaud and E. Lazard, *Methods and problems of communication in usual networks*, Discrete Applied Math., 53 (1994), pp. 79–133.

[10] S. Fujita, *Neighborhood information dissemination in the star graph*, IEEE Trans. Comput., 49 (2000), pp. 1366–1370.

[11] S. Fujita, *Optimal neighborhood broadcast in star graphs*, J. Interconnection Networks, 4 (2003), pp. 419–428.

[12] S. Fujita, *Time-efficient multicast to local vertices in star interconnection networks under the single-port model*, IEICE Trans. Information Systems, E87-D (2004), pp. 315–321.

[13] S. Fujita, S. Pérennes, and J. G. Peters, *Neighbourhood gossiping in hypercubes*, Parallel Process. Lett., 8 (1998), pp. 189–195.

[14] S. M. Hedetniemi, S. T. Hedetniemi, and A. L. Liestman, *A survey of gossiping and broadcasting in communication networks*, Networks, 18 (1986), pp. 319–349.

[15] J. Hromovič, R. Klasing, A. Pelc, P. Ružička, and W. Unger, *Dissemination of information in communication networks: Broadcasting, gossiping, leader election, and fault-tolerance*, Texts in Theoretical Computer Science, Springer-Verlag, Berlin, 2005.

[16] D. D. Kouvatsos and I. M. Mkwawa, *Neighbourhood broadcasting schemes for Cayley graphs with background traffic*, in Proceedings of the 4th EPSRC/BCS PG Symposium on the Convergence of Telecommunications, Networking and Broadcasting (PG Net 2003), M. Merabti, ed., Liverpool, 2003, pp. 143–148.

[17] D. W. Krumme, *Fast gossiping for the hypercube*, SIAM J. Comput., 21 (1992), pp. 365–380.

[18] D. W. Krumme, G. Cybenko, and K. N. Venkataraman, *Gossiping in minimal time*, SIAM J. Comput., 21 (1992), pp. 111–139.

[19] I. M. Mkwawa and D. D. Kouvatsos, *An optimal neighbourhood broadcasting scheme for star interconnection networks*, J. Interconnection Networks, 4 (2003), pp. 103–112.

[20] K. Qiu and S. K. Das, *A novel neighbourhood broadcasting algorithm on star graphs*, in Proceedings of the 9th International Conference on Parallel and Distributed Systems (ICPADS'02), Taiwan, 2002, pp. 37–41.

SIAM J. Discrete Math.
Vol. 21, No. 4, pp. 844–850

MOD $(2p + 1)$-ORIENTATIONS AND $K_{1,2p+1}$-DECOMPOSITIONS*

HONG-JIAN LAI[†]

Abstract. In this paper, we establish an equivalence between the contractible graphs with respect to the mod $(2p + 1)$-orientability and the graphs with $K_{1,2p+1}$-decompositions. This is applied to disprove a conjecture proposed by Barat and Thomassen that every 4-edge-connected simple planar graph G with $|E(G)| \equiv 0 \pmod 3$ has a claw decomposition.

Key words. nowhere zero flows, circular flows, mod $(2p+1)$-orientations, $K_{1,2p+1}$-decompositions

AMS subject classification. 05C99

DOI. 10.1137/060676945

1. Introduction.

Graphs in this paper are finite and loopless and may have multiple edges. See [2] for undefined notations and terminologies. In particular, $\kappa'(G)$ denotes the edge connectivity of a graph G, and if X is an edge subset or a vertex subset of a graph G, then $G[X]$ denotes the subgraph of G induced by X. A connected loopless graph with 3 edges and a vertex of degree 3 is called *a generalized claw.* When restricted to simple graphs, a generalized claw must be isomorphic to a $K_{1,3}$. A graph G with $|E(G)| \equiv 0 \pmod 3$ has a *claw decomposition* if $E(G)$ can be partitioned into disjoint unions $E(G) = X_1 \cup X_2 \cup \cdots \cup X_k$ such that, for each i with $1 \leq i \leq k$, $G[X_i]$ is a generalized claw. Barat and Thomassen [1] showed that the claw-decomposition problem is closely related to the nowhere zero 3-flow problem. In particular, the following conjecture is proposed.

Conjecture 1.1 (Barat and Thomassen [1]). *Every 4-edge-connected simple planar graph G with $|E(G)| \equiv 0 \pmod 3$ has a claw decomposition.*

The purpose of this note is to disprove this conjecture. In section 2, we shall introduce contractible graphs with respect to the mod $(2p+1)$-orientability and discuss their properties and their relationship to the graphs with $K_{1,2p+1}$-decompositions. In section 3, we disprove the conjecture above.

2. M^o_{2p+1} and $K_{1,2p+1}$-decompositions.

Throughout this section, $p > 0$ denotes an integer. We shall extend the definition of claw decomposition to $K_{1,2p+1}$-decomposition as follows. A connected loopless graph with $2p + 1$ edges and a vertex of degree $2p + 1$ is called *a generalized* $K_{1,2p+1}$. A graph G with $|E(G)| \equiv 0$ $(\text{mod } 2p + 1)$ has *a $K_{1,2p+1}$-decomposition* if $E(G)$ can be partitioned into disjoint unions $E(G) = X_1 \cup X_2 \cup \cdots \cup X_k$ such that, for each i with $1 \leq i \leq k$, $G[X_i]$ is a generalized $K_{1,2p+1}$. In this case, we say that G has a $K_{1,2p+1}$-decomposition $\mathcal{X} = \{X_1, X_2, \ldots, X_k\}$.

Let $D = D(G)$ be an orientation of an undirected graph G. If an edge $e \in E(G)$ is directed from a vertex u to a vertex v, then let $\text{tail}(e) = u$ and $\text{head}(e) = v$. For a vertex $v \in V(G)$, let

$$E^+_D(v) = \{e \in E(D) \ : \ v = \text{ tail}(e)\} \text{ and } E^-_D(v) = \{e \in E(D) \ : \ v = \text{ head}(e)\}.$$

*Received by the editors December 6, 2006; accepted for publication (in revised form) August 3, 2007; published electronically November 7, 2007.

http://www.siam.org/journals/sidma/21-4/67694.html

[†]Department of Mathematics, West Virginia University, Morgantown, WV 26506 (hjlai@math.wvu.edu).

We shall denote $d_D^+(v) = |E_D^+(v)|$ (the *out degree* of v) and $d_D^-(v) = |E_D^-(v)|$ (the *in degree* of v). The subscript D may be omitted when $D(G)$ is understood from the context. Let A be an (additive) Abelian group. If $f : E(G) \mapsto A$ is a function, then the *boundary* of f is a map $\partial f : V(G) \mapsto A$ such that

$$\partial f(v) = \sum_{e \in E_D^+(v)} f(e) - \sum_{e \in E_D^-(v)} f(e), \ \forall v \in V(G).$$

Let $k > 0$ be an integer, and assume that G has a fixed orientation D. A *mod k-orientation* of G is a function $f : E(G) \mapsto \{1, -1\}$ such that for all $v \in V(G)$, $\partial f(v) \equiv 0 \pmod{k}$. The collection of all graphs admitting a mod k-orientation is denoted by M_k. Note that, by definition, $K_1 \in M_k$. Jaeger has conjectured [7] that every $4k$-edge-connected graph is in M_{2k+1}. This conjecture is still open.

Throughout this note, $\mathbf{Z}$ denotes the set of all integers. For integers $a_1, a_2, \dots a_k$ such that not all of them are zero, let $gcd(a_1, a_2, \dots, a_k)$ denote the greatest common divisor of $a_1, a_2, \dots a_k$. For an $m \in \mathbf{Z}$, $\mathbf{Z}_m$ denotes the set of integers modulo m, as well as the additive cyclic group on m elements. For a graph G, a function $b : V(G) \mapsto \mathbf{Z}_m$ is a *zero sum function* in $\mathbf{Z}_m$ if $\sum_{v \in V(G)} b(v) \equiv 0 \pmod{m}$. The set of all zero sum functions in $\mathbf{Z}_m$ of G is denoted by $Z(G, \mathbf{Z}_m)$. When $k = 2p+1 > 0$ is an odd number, we define M_{2p+1}^o to be the collection of graphs such that $G \in M_{2p+1}^o$ if and only if for all $b \in Z(G, \mathbf{Z}_{2p+1})$, $\exists f : E(G) \mapsto \{1, -1\}$ such that for all $v \in V(G)$, $\partial f(v) \equiv b(v)$ $\pmod{2p+1}$.

Note that if a function $f : E(G) \mapsto \{1, -1\}$ is given, then one can reverse the orientation of e for each $e \in E(G)$ with $f(e) = -1$ to obtain an orientation D' of G such that for all $v \in V(G)$, $d_{D'}^+(v) - d_{D'}^-(v) = \partial f(v)$. Thus we have the following proposition.

Proposition 2.1. *$G \in M_{2p+1}^o$ if and only if for all $b \in Z(G, \mathbf{Z}_{2p+1})$, G has an orientation D with the property that for all $v \in V(G)$, $d_D^+(v) - d_D^-(v) \equiv b(v)$* (mod$2p$+1).

For a subgraph H of G, define the set of *vertices of attachments* of H in G to be $A_G(H) = \{v \in V(H) : v \text{ is adjacent to a vertex in } G - V(H)\}$.

Proposition 2.2. *For any integer $p \geq 1$, M_{2p+1}^o is a family of connected graphs such that each of the following holds.*

(C1) *$K_1 \in M_{2p+1}^o$.*

(C2) *If $e \in E(G)$ and if $G - e \in M_{2p+1}^o$, then $G \in M_{2p+1}^o$.*

(C3) *If H is a subgraph of G, and if $H, G/H \in M_{2p+1}^o$, then $G \in M_{2p+1}^o$.*

Proof. (C1) and (C2) are straightforward, and so we verify only (C3).

Suppose that G has a fixed orientation, H is a subgraph of G, and both $H \in M_{2p+1}^o$ and $G/H \in M_{2p+1}^o$. Thus the edges in both H and G/H are oriented by the orientation of G. By (C2), we may assume that H is an induced subgraph of G, and so $E(G)$ is the disjoint union of $E(H)$ and $E(G/H)$. Note that H is connected, and so H will be contracted to a vertex v_H (say) in G/H. Let $b : V(G) \mapsto \mathbf{Z}_{2p+1}$ such that $\sum_{v \in V(G)} b(v) \equiv 0 \pmod{2p+1}$, and let $a_0 = \sum_{v \in V(H)} b(v)$. Define $b_1 : V(G/H) \rightarrow A$ by setting $b_1(z) = b(z)$ if $z \neq v_H$, and $b_1(v_H) = a_0$. Then $\sum_{z \in V(G/H)} b_1(z) = \sum_{z \in V(G)} b(z) \equiv 0 \pmod{2p+1}$. Since $G/H \in M_{2p+1}^o$, there exists $f_1 : E(G/H) \mapsto \{1, -1\}$ such that $\partial f_1 = b_1$. For each $z \in V(H)$, define

$$b_2(z) = \begin{cases} b(z) + \sum_{e \in E_{G/H}^-(v_H) \cap E_G^-(z)} f_1(e) - \sum_{e \in E_{G/H}^+(v_H) \cap E_G^+(z)} f_1(e) & \text{if } z \in A_G(H), \\ b(z) & \text{otherwise.} \end{cases}$$

Then $\sum_{z \in V(H)} b_2(z) \equiv 0 \pmod{2p+1}$. Since $H \in M_{2p+1}^o$, there exists $f_2 : E(G/H) \mapsto$

$\{1,-1\}$ such that $\partial f_2 = b_2$. Now for each $e \in E(G)$, define $f(e) = f_1(e) + f_2(e)$. As $E(G)$ is a disjoint union of $E(H)$ and $E(G/H)$, it is routine to verify that $\partial f(z) \equiv b(z) \pmod{2p+1}$, and so $G \in M^o_{2p+1}$. □

Catlin [3] (see also [4], [5]) called families of connected graphs satisfying (C1), (C2), and (C3) complete families. Complete families seem to be useful in applying certain reduction methods ([3], [4], [5]).

For a subgraph H of a graph G, define

$$\partial(H) = \{uv \in E(G) : u \in V(H), v \in V(G) - V(H)\}.$$

Let D be an orientation of G. Let $d^+_D(H)$ denote the number of edges in $\partial(H)$ that are oriented in D from H to $G - V(H)$, and $d^-_D(H) = |\partial(H)| - d^+_D(H)$.

To demonstrate the relationship between M^o_{2p+1} and all of the graphs with $K_{1,2p+1}$-decompositions, we make the following definitions.

(i) $k_{c,2p+1}$ denotes the smallest integer $k > 0$ such that every k-edge-connected graph G is in M^o_{2p+1}.

(ii) $k^{c,2p+1}$ denotes the smallest integer $k > 0$ such that every k-edge-connected graph G with $|E(G)| \equiv 0 \pmod{2p+1}$ has a $K_{1,2p+1}$-decomposition.

The main result of this section is the following relationship.

THEOREM 2.3. *For any positive integer $p > 0$, if one of $k_{c,2p+1}$ and $k^{c,2p+1}$ exists as a finite number, then $k_{c,2p+1} = k^{c,2p+1}$.*

To prove this theorem, we need to establish some lemmas. In each of the following lemmas, G is a graph and H is a subgraph of G. Suppose that G has a $K_{1,2p+1}$-decomposition $\mathcal{X} = \{X_1, X_2, \ldots, X_k\}$, where each $G[X_i]$ is a generalized $K_{1,2p+1}$ for all i. For each $G[X_i]$, we orient the edges from the vertex v_i of degree $2p+1$ in $G[X_i]$ to all other vertices of $G[X_i]$. This yields an orientation $D = D(\mathcal{X})$ induced by the decomposition $\mathcal{X}$. For each i, the vertex v_i is called the *center* of the oriented X_i.

LEMMA 2.4. *Suppose that G has a $K_{1,2p+1}$-decomposition $\mathcal{X} = \{X_1, X_2, \ldots, X_k\}$, and let $D = D(\mathcal{X})$. Then for any subgraph H of G,*

$$|E(H)| + d^+_D(H) \equiv 0 \pmod{2p+1}.$$

Proof. Let $[H, G - V(H)]$ denote the set of edges in $\partial(H)$ that are oriented in $D(\mathcal{X})$ from H to $G - V(H)$. Then $|[H, G - V(H)]| = d^+_D(H)$.

By the definition of $D(\mathcal{X})$, the edge subset $E(H) \cup [H, G - V(H)]$ is the disjoint union of the oriented X_i's whose centers are in $V(H)$. It follows that $|E(H)| + d^+_D(H) = |E(H) \cup [H, G - V(H)]| \equiv 0 \pmod{2p+1}$. □

LEMMA 2.5. *Let $b \in \mathbf{Z}$ be a number, and let $d = |\partial(H)|$. Suppose that G has a $K_{1,2p+1}$-decomposition $\mathcal{X}$ and that H is a subgraph of G. If $2|E(H)| \equiv -d - b \pmod{2p+1}$, then, in the orientation $D = D(\mathcal{X})$,*

$$d^+_D(H) - d^-_D(H) \equiv b \pmod{2p+1}.$$

Proof. Let $d^+ = d^+_D(H)$ and $d^- = d^-_D(H)$. Then $d = d^+ + d^-$. By Lemma 2.4, $|E(H)| \equiv -d^+ \pmod{2p+1}$, and so $b \equiv -d - 2|E(H)| \equiv -d + 2d^+ \equiv (-d + d^+) + d^+ \equiv d^+ - d^- \pmod{2p+1}$. □

The following below is well-known in number theory. For a reference, see Theorem 1.5 of [12].

LEMMA 2.6. *Let $a_1, a_2, \ldots, a_k$ be integers, not all zero. Then $gcd(a_1, a_2, \ldots, a_k) = 1$ if and only if there exist integers $x_1, x_2, \ldots, x_k$ such that $a_1x_1 + a_2x_2 + \cdots + a_kx_k = 1$.*

LEMMA 2.7. *Let $k, l, p \in \mathbf{Z}$ such that $k > 0$, $p > 0$, and $0 \le l \le 2p$. Each of the following holds.*

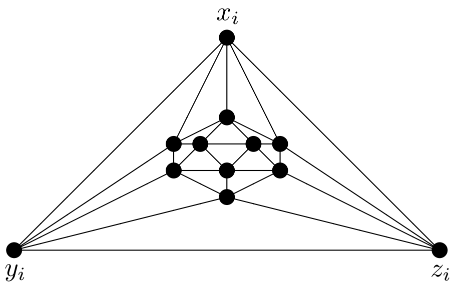


FIG. 1. *$I_{12}(i)$ (isomorphic to icosahedron) with specified x_i, y_i, z_i.*

(i) *There exists a planar graph H with $\kappa'(H) \geq k$ and $2|E(H)| \equiv l \pmod{2p+1}$.*

(ii) *There exists a simple graph H with $\kappa'(H) \geq k$ and $2|E(H)| \equiv l \pmod{2p+1}$.*

(iii) *If $2 \leq k \leq 5$, then there exists a simple planar graph H with $\kappa'(H) \geq k$ and $2|E(H)| \equiv l \pmod{2p+1}$.*

Proof. (i) For any integer $n > 0$, let nK_2 denote the connected loopless graph with two vertices and n multiple edges. Let $s > 0$ be an integer such that $s(2p+1) \geq k$. Define the desired H as follows:

$$H = \begin{cases} (2ps+s+t)K_2 & \text{if } l = 2t \text{ is even,} \\ ((2p+1)(s+1)-(p-t))K_2 & \text{if } l = 2t+1 \text{ is odd.} \end{cases}$$

(ii) Take an integer $m \geq 4p+2+k$, and let $H_v = K_m - W$ for some edge set $W \subset E(K_m)$ such that $|W| \leq 4p+1$ and $2(|E(K_m)| - |W|) \equiv l \pmod{2p+1}$.

(iii) Since $gcd(10, 18, 2p+1) = 1$, by Lemma 2.6, there are integers a_0, b_0, c_0 such that $10a_0 + 18b_0 + (2p+1)c_0 = 1$. Choose $x_0 = la_0 + l(|a_0|+1)(2p+1)$ and $y_0 = lb_0 + l(|b_0|+1)(2p+1)$. Then x_0, y_0 are positive integers such that

$$10x_0 + 18y_0 \equiv l \mod (2p+1)$$

holds. Let $t = (2p+1)(x_0+y_0+1)$, and let $I_{12}(i)$, $1 \leq i \leq t-1$, be a graph isomorphic to icosahedron defined below (see Figure 1). Define H to be the graph obtained from $I_{12}(1), I_{12}(2), \ldots, I_{12}(t)$ by identifying z_i and y_{i+1}, $1 \leq i \leq t-1$, and by adding $x_0 + 2y_0$ new vertices $u_1, u_2, \ldots, u_{x_0}, v_1, v_2, \ldots, v_{y_0}, w_1, w_2, \ldots, w_{y_0}$ with $N(u_k) = \{x_{5k-4}, x_{5k-3}, x_{5k-2}, x_{5k-1}, x_{5k}\}$, $N(v_{k'}) = \{x_{5x_0+8k'-7}, x_{5x_0+8k'-6}, x_{5x_0+8k'-5}, x_{5x_0+8k'-4}, w_{k'}\}$, and $N(w_{k'}) = \{x_{5x_0+8k'-3}, x_{5x_0+8k'-2}, x_{5x_0+8k'-1}, x_{5x_0+8k'}, v_{k'}\}$, where $1 \leq k \leq x_0$ and $1 \leq k' \leq y_0$. So H is a simple planar graph with $\kappa(H) \geq k$ and $2|E(H)| = 60t + 10x_0 + 18y_0 = 60(2p+1)(x_0+y_0+1) + 10x_0 + 18y_0 \equiv l \mod (2p+1)$. □

LEMMA 2.8. (i) *Let $k > 0$ be an integer. If every k-edge-connected (simple) graph G with $|E(G)| \equiv 0 \pmod{2p+1}$ has a $K_{1,2p+1}$-decomposition, then every k-edge-connected (simple) graph $L \in M^o_{2p+1}$.*

(ii) *Let $k > 0$ be an integer. If every k-edge-connected planar graph G with $|E(G)| \equiv 0 \pmod{2p+1}$ has a $K_{1,2p+1}$-decomposition, then every k-edge-connected planar graph $L \in M^o_{2p+1}$.*

(iii) *Let $2 \leq k \leq 5$. If every k-edge-connected simple planar graph G with $|E(G)| \equiv 0 \pmod{2p+1}$ has a $K_{1,2p+1}$-decomposition, then every k-edge-connected simple planar graph $L \in M^o_{2p+1}$.*

Proof. We shall prove (i) and assume first that every k-edge-connected (simple) graph G with $|E(G)| \equiv 0 \pmod{2p+1}$ has a $K_{1,2p+1}$-decomposition. By contradiction, we assume that there exists a k-edge-connected (simple) graph L such that $L \notin M^o_{2p+1}$.

Therefore, $\exists b \in Z(L, \mathbf{Z}_{2p+1})$ such that L does not have an orientation D satisfying $d^+_D(v) - d^-_D(v) \equiv b(v) \pmod{2p+1}$ for all $v \in V(L)$.

Let $l_v \in \mathbf{Z}$, with $0 \leq l_v \leq 2p$ such that $l_v \equiv -b(v) - d_L(v) \pmod{2p+1}$ for all $v \in V(L)$. By Lemma 2.7 (ii), there exists a simple graph H_v with $2|E(H_v)| \equiv l_v \equiv -b(v) - d_L(v) \pmod{2p+1}$ such that H_v is also k-edge-connected. For each $v \in V(L)$, replace v by H_v in such a way that the resulting graph G is also a k-edge-connected (simple) graph.

Since $b \in Z(L, \mathbf{Z}_{2p+1})$, $2|E(G)| = \sum_{v\in V(L)} 2|E(H_v)| + 2|E(L)| = -\sum_{v\in V(L)} b(v) - \sum_{v\in V(L)} d_L(v) + 2|E(L)| \equiv -\sum_{v\in V(L)} b(v) \equiv 0 \pmod{2p+1}$. By the fact that 2 and $2p+1$ are relatively prime, $|E(G)| \equiv 0 \pmod{2p+1}$. By the assumption of this lemma, G has a $K_{1,2p+1}$-decomposition $\mathcal{X}$. By the construction of G, $|\partial(H_v)| = d_L(v)$. Since $2|E(H_v)| \equiv l_v \equiv -b(v) - d_L(v) \pmod{2p+1}$, it follows by Lemma 2.5 that, in the orientation $D = D(\mathcal{X})$ for all $v \in V(L)$, $d^+_D(H_v) - d^-_D(H_v) \equiv b(v) \pmod{2p+1}$, contrary to the assumption that L is a counterexample.

The proofs for (ii) and (iii) are similar except that we shall use Lemma 2.7 (i) and (iii) instead of Lemma 2.7 (ii). Thus we omit the detailed proofs. □

Lemma 2.9. *If G has an orientation D such that for all $v \in V(G)$, $d^+_D(v) \equiv 0 \pmod{2p+1}$, then G is $K_{1,2p+1}$-decomposable.*

Proof. Note that if D is an orientation of G, then $E(G) = \cup_{v\in V(G)} E^+_D(v)$ is a disjoint union. As for all $v \in V(G)$, $d^+_D(v) \equiv 0 \pmod{2p+1}$, each $E^+_D(v)$ is a disjoint union of generalized $K_{1,2p+1}$'s, and so G is $K_{1,2p+1}$-decomposable. □

Lemma 2.10. *Suppose that $G \in M^o_{2p+1}$. If $|E(G)| \equiv 0 \pmod{2p+1}$, then G has a $K_{1,2p+1}$-decomposition.*

Proof. For all $v \in V(G)$, pick an $x(v) \in \{0, 1, \ldots, 2p\}$ such that $d(v) \equiv x(v) \pmod{2p+1}$. Define $b(v) = d(v) - 2x(v)$. First, we shall show that $b \in Z(G, \mathbf{Z}_{2p+1})$. Since $x(v) \equiv d(v) \pmod{2p+1}$, we have $d(v) - 2x(v) \equiv -x(v) \equiv -d(v) \pmod{2p+1}$. Note also that $\sum_{v\in V(G)} d(v) = 2|E(G)| \equiv 0 \pmod{2p+1}$. Thus

$$\sum_{v\in V(G)} b(v) = \sum_{v\in V(G)} (d(v) - 2x(v)) = -\sum_{v\in V(G)} d(v) \equiv 0 \pmod{2p+1}.$$

Hence $b \in Z(G, \mathbf{Z}_{2p+1})$.

Since $G \in M^o_{2p+1}$, there exists an orientation D of G such that, under this orientation, at each vertex $v \in V(G)$, $d^+(v) - d^-(v) = b(v) = d(v) - 2x(v)$. Since $d^+(v) + d^-(v) = d(v)$, we have $2d^+(v) = 2d(v) - 2x(v) = 2(d(v) - x(v))$. Since 2 and $2p+1$ are relatively prime, $d^+(v) \equiv d(v) - x(v) \equiv 0 \pmod{2p+1}$. Therefore, by Lemma 2.9, G has a $K_{1,2p+1}$-decomposition. □

Now we can easily prove Theorem 2.3. By Lemma 2.8, $k_{c,2p+1} \leq k^{c,2p+1}$ and by Lemma 2.10, $k_{c,2p+1} \geq k^{c,2p+1}$. Thus Theorem 2.3 follows.

By (ii) and (iii) of Lemma 2.8 and by Lemma 2.10, and noting that the edge connectivity of a simple planar graph cannot exceed 5 (Corollary 9.5.3 of [2]), we also have the following corollary.

Corollary 2.11. (i) *Let k' denote the smallest positive integer such that every k'-edge-connected planar graph G with $|E(G)| \equiv 0 \pmod{2p+1}$ has a $K_{1,2p+1}$-decomposition, and let k'' denote the smallest positive integer such that every k''-edge-connected planar graph is in M^o_{2p+1}. Then $k' = k''$.*

(ii) *Let l' denote the smallest positive integer such that every l'-edge-connected simple planar graph G with $|E(G)| \equiv 0 \pmod{2p+1}$ has a $K_{1,2p+1}$-decomposition, and let l'' denote the smallest positive integer such that every l''-edge-connected simple planar graph is in M^o_{2p+1}. Then $l' = l''$.*

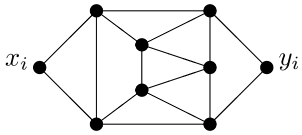


FIG. 2. *The building block H_i.*

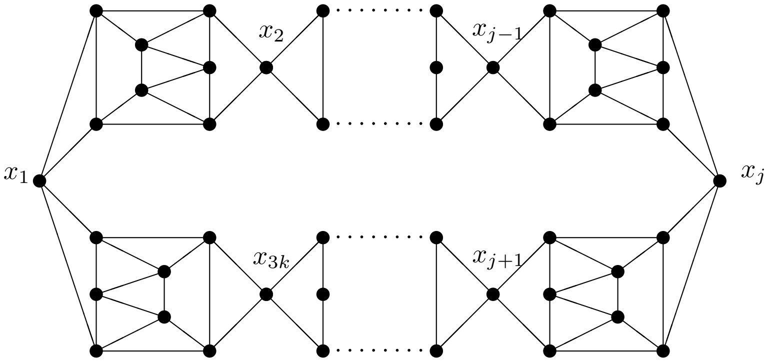


FIG. 3. *The graph $G = G(k)$.*

3. Planar graphs. When $p = 1$, graphs in M_3^o are also called $\mathbf{Z}_3$-connected graphs [8], [10], [11]. The following has been recently proved.

THEOREM 3.1 (Theorem 3 of [9]). *There exists a family of* 4*-edge-connected simple planar graphs that are not in M_3^o.*

In fact, the dual version of Theorem 3.1 is proved in [9]. The equivalence between Theorem 3 of [9] and Theorem 3.1 here was pointed out without a proof in [8], and a formal proof of this equivalence can be found in [6].

COROLLARY 3.2. *There exists a* 4*-edge-connected simple planar graph G with* $|E(G)| \equiv 0 \pmod 3$ *which does not have a claw decomposition.*

Proof. Suppose, to the contrary, that Conjecture 1.1 holds. Then by (ii) of Corollary 2.11, every 4-edge-connected simple planar graph must be in M_3^o, which contradicts Theorem 3.1. □

Corollary 3.2 disproves Conjecture 1.1. In fact, we can also directly construct an infinite family of 4-edge-connected simple planar graphs G with $|E(G)| \equiv 0 \pmod 3$ which does not have a claw decomposition. We present the construction as follows.

Let $k > 0$ be an integer. For each i with $1 \le i \le 3k$, define H_i to be the graph depicted below. See Figure 2.

A graph $G = G(k)$ can be constructed from the disjoint H_i's by identifying y_i and x_{i+1}, where $x_{3k+1} = x_1$ and where $i = 1, 2, \dots 3k$.

EXAMPLE 3.3. *For each $k > 0$, $G = G(k)$ defined in Figure* 3 *is a* 4*-regular and* 4*-edge-connected simple planar graph with* $|E(G)| \equiv 0 \pmod 3$*, and G has no claw decomposition.*

Proof. Suppose G has a claw decomposition $\mathcal{X} = \{X_1, X_2, \dots, X_m\}$, and let $D = D(\mathcal{X})$. Since G is 4-regular, for all $v \in V(G)$, $|E_D^+(v)| \in \{0, 3\}$. Note that $|V(G)| = 24k$ and $|E(G)| = 48k$. Thus G has $m = 48k/3 = 16k$ edge-disjoint claws. Let W denote the set of vertices v with $|E_D^+(v)| = 0$. Then $|W| = |V(G)| - m = 24k - 16k = 8k$. Note that no two vertices in W are adjacent in G, and so, for each $i = 1, 2, \ldots, 3k$ (mod $3k$), $|W \cap V(H_i \cup H_{i+1} - \{y_{i+1}\})| \leq 5$. It follows that $16k = 2|W| = \sum_{i=1}^{3k} |V(H_i \cup H_{i+1} - \{y_{i+1}\}) \cap W| \leq 5 \times 3k = 15k$, a contradiction. □

It is an open problem whether $k_{c,2p+1}$, or, equivalently, $k^{c,2p+1}$, exists as a finite number. We conjecture that it does. In view of Corollary 3.2 and Example 3.3, we further conjecture that $k_{c,2p+1} = 4p + 1$.

REFERENCES

[1] J. Barat and C. Thomassen, *Claw-decompositions and Tutte-orientations*, J. Graph Theory, 52 (2006), pp. 135–146.

[2] J. A. Bondy and U. S. R. Murty, *Graph Theory with Applications*, Elsevier, New York, 1976.

[3] P. A. Catlin, *Double cycle covers and the Petersen graph*, J. Graph Theory, 13 (1989), pp. 95–116.

[4] P. A. Catlin, *The reduction of graph families under contraction*, Discrete Math., 160 (1996), pp. 67–80.

[5] P. A. Catlin, A. M. Hobbs, and H.-J. Lai, *Graph families operations*, Discrete Math., 230 (2001), pp. 71–97.

[6] Z. H. Chen, H.-J. Lai, X. Lei, and X. Zhang, *Group coloring and group connectivity of graphs*, Congr. Numer., 134 (1998), pp. 123–130.

[7] F. Jaeger, *Nowhere-Zero Flow Problems*, in Selected Topics in Graph Theory 3., L. Beineke et al., eds., Academic Press, London, New York, 1988, pp. 91–95.

[8] F. Jaeger, N. Linial, C. Payan, and M. Tarsi, *Group connectivity of graphs - A nonhomogeneous analogue of nowhere-zero flow properties*, J. Combin. Theory Ser. B, 56 (1992), pp. 165–182.

[9] D. Kral, O. Pangrac, and H.-J. Voss, *A note on group colorings*, J. Graph Theory, 50 (2005), pp. 123–129.

[10] H.-J. Lai, *Group connectivity of* 3*-edge-connectied chordal graphs*, Graphs Combin., 16 (2000), pp. 165–176.

[11] H.-J. Lai, *Nowhere-zero* 3*-flows in locally connected graphs*, J. Graph Theory, 42 (2003), pp. 211–219.

[12] M. B. Nathanson, *Elementary Methods in Number Theory*, Springer, New York, 1999.

SIAM J. Discrete Math.
Vol. 21, No. 4, pp. 851–864

ON THE EXISTENCE OF $(K_5 \setminus e)$-DESIGNS WITH APPLICATION TO OPTICAL NETWORKS[*]

GENNIAN GE[†] AND ALAN C. H. LING[‡]

Abstract. Motivated by the connection between graph decompositions and traffic grooming in optical networks, we continue the investigation of the existence problem for $(K_5 \setminus e)$-designs of order n. It is proved that the necessary conditions for the existence of such designs are also sufficient with 3 definite exceptions ($n = 9, 10, 18$) and 12 possible exceptions with $n = 234$ being the largest. This gives a near solution for the long standing problem posed by Bermond et al. in [*Ars Combin.*, 10 (1980), pp. 211–254]. As a consequence, we also give an optimal grooming on n nodes with $C = 9$ when such a $(K_5 \setminus e)$-design of order n exists.

Key words. $(K_5 \setminus e)$-design, optical networks, traffic grooming

AMS subject classifications. 05B05, 68M10, 68R05

DOI. 10.1137/060660084

1. Introduction and definitions. Let $\mathcal{G} = \{G_1, G_2, \ldots, G_t\}$ and H be (finite, simple, undirected) graphs. A $\mathcal{G}$-decomposition of H is a partition of the edges of H into classes so that the edges within each class form a graph isomorphic to G_i for some i. When H is a complete graph of *order* n (denoted by K_n), the graphs in a $\mathcal{G}$-decomposition of H form a $\mathcal{G}$-design of order n (and *index* one, since each edge of H appears in exactly one of the graphs chosen). When $\mathcal{G} = \{G\}$, we simply denote it as G-design. A K_k-design of order n is just a *Steiner system* $S(2, k, n)$.

We shall be interested not only in taking H to be complete, but also taking H to be "nearly" complete. To this end, define a complete multipartite graph to be of *type* $g_1^{u_1} g_2^{u_2} \cdots g_s^{u_s}$ if it has exactly $\sum_{i=1}^{s} u_i$ classes in the multipartition, and there are u_i parts of size g_i for $i = 1, 2, \ldots, s$. A $\mathcal{G}$-decomposition of the complete multipartite graph of type $g_1^{u_1} g_2^{u_2} \cdots g_s^{u_s}$ is termed as a *$\mathcal{G}$-group divisible design* of *type* $g_1^{u_1} g_2^{u_2} \cdots g_s^{u_s}$, and is often called a $\mathcal{G}$-GDD for short. Again, if $\mathcal{G} = \{G\}$, it is called a G-GDD. The special case when a G-GDD has type $1^m h^1$ is an *incomplete* G-design of order $m + h$ with a *hole* of size h; in graph-theoretic vernacular this is a partition of the edges of $K_{m+h} \setminus K_h$ into copies of G. Another special case is when $\mathcal{G} = \{K_{k_1}, K_{k_2}, \ldots, K_{k_r}\}$, then a $\mathcal{G}$-design is simply denoted by $\{k_1, k_2, \ldots, k_r\}$-GDD. Furthermore, if $r = 1$, then it is simply denoted by k_1-GDD.

Let $K_5 \setminus e$ be a graph with 9 edges on 5 vertices. The problem of determining the existence of $(K_n, K_5 \setminus e)$-designs has been studied for a long time. The first result on the problem was stated in [18], but the result is not in any of the references cited there.

THEOREM 1.1 (see [18]). *If* $n \equiv 1 \pmod{18}$ *and* $n \neq 37, 55, 73, 109, 397, 415, 469, 487, 505, 541, 613, 685$*, then there exists a* $(K_n, K_5 \setminus e)$*-design.*

[*] Received by the editors May 16, 2006; accepted for publication (in revised form) May 11, 2007; published electronically November 16, 2007.
http://www.siam.org/journals/sidma/21-4/66008.html

[†] Department of Mathematics, Zhejiang University, Hangzhou 310027, Zhejiang, People's Republic of China (gnge@zju.edu.cn). This author's research was supported by National Natural Science Foundation of China under grant 10471127 and Zhejiang Provincial Natural Science Foundation of China under grant R604001.

[‡] Department of Computer Science, University of Vermont, Burlington, VT 05405 (aling@emba.uvm.edu). This author's research was supported by ARO grant 19-01-1-0406 and a DOE grant.

Recently, the problem has been considered in [21], where all possible exceptions stated in the above theorem have been eliminated, leading to the following theorem.

THEOREM 1.2 (see [21]). *If $n \equiv 1 \pmod{18}$, then there exists a $(K_5 \setminus e)$-design of order n.*

However, a simple computation shows that a necessary condition for the existence of a $(K_5 \setminus e)$-design of order n is $n \equiv 0, 1 \pmod 9$. Therefore, the above theorem leaves the cases $n \equiv 0, 9, 10 \pmod{18}$ wide open.

Another one of our motivations for this problem is the connection between graph decompositions and grooming in optical networks which we briefly describe next. *Traffic grooming* is the process of packing low-rate signals into higher-rate streams which share a wavelength. In optical networks, particularly in SONET ring networks, grooming has received much attention; surveys are given in [9, 11, 22, 24]. The setting is a wavelength-division multiplexed (WDM) network; each wavelength is an optical communication medium which connects all nodes in a circle and may be unidirectional or bidirectional. An *add-drop multiplexer* (ADM) is required on each wavelength at each node at which traffic is added or dropped. In general, there are two main goals given a set of traffic requirements between nodes. The first is to minimize the number of wavelengths employed, while the second is to minimize the total number of ADMs (the *drop cost*).

A case of substantial interest (see [3, 4, 5, 6] and references therein) arises with symmetric uniform traffic requirements on a unidirectional ring. In this scenario, for every source node i and every target node j, the traffic requirement is for the fixed fraction $1/C$ of a wavelength. The quantity C is the *grooming ratio*, because we can "groom" C circles onto the same wavelength. Bermond and Coudert [5] establish that minimizing drop cost (that is the total number of ADMs used, denoted $A(C, n)$) of a grooming on n nodes with grooming ratio C can be expressed as an optimization problem of graphs, as follows. Partition the edges of K_n into subgraphs $G_1, G_2, \ldots, G_t$ so that each G_i contains at most C edges, and the sum of the numbers of vertices of nonzero degree in the $\{G_i\}$ is minimized. Such a partition of K_n is a *C-grooming.*

Optimal constructions for given grooming ratio C have been obtained using tools of graph and design theory [10]. In particular, results are available for grooming ratio $C = 3$ [2], $C = 4$ [19, 6], $C = 5$ [4], $C = 6$ [3], and $C \leq n(n-1)/6$ [6]. The problem is also solved for large values of C [6]. Related problems have been studied both in the context of variable traffic requirements [9, 12, 17, 25, 27] and the case of fixed traffic requirements [2, 4, 5, 6, 11, 15, 16, 19, 20, 22, 26, 28].

Let $\rho(G_\ell)$ denote the ratio for the subgraph G_ℓ, $\rho(G_\ell) = \frac{|E(G_\ell)|}{|V(G_\ell)|}$, and $\rho(m)$ the maximum ratio of a subgraph with m edges. Let $\rho_{\max}(C)$ denote the maximum ratio of subgraphs with $m \leq C$ edges. We have $\rho_{\max}(C) = \max\{\rho(G_\ell) \mid |E(G_\ell)| \leq C\} = \max_{m \leq C} \rho(m)$. For the sake of illustration, Table 1 gives the values of $\rho_{\max}(C)$ for small values of C. For example for $C = 9$, $\rho_{\max}(9) = \frac{9}{5}$, the bound being attained for $K_5 \setminus e$.

The grooming problem is closely connected to problems in combinatorial design theory. Indeed an $(n, k, 1)$-design is exactly a partition of the edges of K_n into subgraphs isomorphic to K_k. That corresponds to requiring in our partitioning problem that all the subgraphs G_ℓ be isomorphic to K_k. Our interest in the existence of a G-design is shown by the following proposition.

PROPOSITION 1.3. *If there exists a G-design of order n, where G is a graph with at most C edges and ratio $\rho_{\max}(C)$, then $A(C, n) = \frac{n(n-1)}{2\rho_{\max}(C)}$.*

TABLE 1
Values of $\rho_{\max}(C)$ for small C.

C	1	2	3	4	5	6	7	8	9	10
$\rho_{\max}(C)$	$\frac{1}{2}$	$\frac{2}{3}$	1	1	$\frac{5}{4}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{8}{5}$	$\frac{9}{5}$	2
C	11	12	13	14	15	16	24	32	48	64
$\rho_{\max}(C)$	2	2	$\frac{13}{6}$	$\frac{14}{6}$	$\frac{5}{2}$	$\frac{5}{2}$	3	$\frac{32}{9}$	$\frac{9}{2}$	$\frac{64}{11}$

NECESSARY CONDITIONS 1.4 (existence of a G-design). *If there exists a G-design, then*

(i) $\frac{n(n-1)}{2}$ *is a multiple of* $E(G)$,

(ii) $n-1$ *is a multiple of the greatest common divisor of the degrees of the vertices of* G.

The purpose of this paper is to obtain a strong existence result for $(K_5 \setminus e)$-designs of order n with $n \equiv 0, 9, 10 \pmod{18}$, thus settling a problem that was first considered many years ago. The result in this paper also provides an optimal grooming on n nodes with $C = 9$ when such a $(K_5 \setminus e)$-design of order n exists.

2. Direct constructions. In this section, we present various direct constructions that are useful in obtaining the main result. We use the representation $[a, b, c, d, e]$ to denote a $K_5 \setminus e$ on the vertices $\{a, b, c, d, e\}$ with the edge $\{d, e\}$ removed. Similarly, we use $[a, b, c, d]$ to denote a $K_4 \setminus e$ on the vertices $\{a, b, c, d\}$ with the edge $\{c, d\}$ removed. In a $\mathcal{G}$-design, a *parallel class* is a set of graphs in the $\mathcal{G}$-design such that each point appears exactly once in the "interested part" of the graph. Here, we normally try to add a new point to each parallel class to form copies of $K_5 \setminus e$. Take a parallel class composed of copies of K_4 as an example. We are only interested in adding three edges to link the new point and three of the four vertices in the K_4, and the "interested part" of the K_4 are the three distinguished vertices, not all four vertices in the K_4.

LEMMA 2.1. *There exists a* $(K_5 \setminus e)$-*design of order* n *for each* $n \in \{37, 55\}$.

Proof. For a given n, the desired design is constructed cyclically on $V = Z_n$ and developed from the following base blocks:

$n = 37$: $[1, 2, 4, 9, 16]$, $[1, 5, 11, 22, 29]$,

$n = 55$: $[1, 6, 17, 35, 44]$, $[1, 2, 8, 21, 4]$, $[1, 11, 33, 41, 42]$. □

LEMMA 2.2. *There exists a* $(K_5 \setminus e)$-*design of order* 28.

Proof. Let the point set be Z_{28}. Developing the following base blocks by $+4$ mod 28 gives the desired design:

$$[1, 5, 14, 28, 16], \quad [1, 4, 11, 13, 6], \quad [4, 7, 8, 24, 14],$$
$$[3, 9, 11, 27, 16], \quad [1, 7, 21, 22, 18], \quad [2, 3, 6, 12, 22]. \quad □$$

LEMMA 2.3. *There exists a* $(K_5 \setminus e)$-*design of order* n *for each* $n \in \{46, 82\}$.

Proof. For a given n, the desired design is constructed on Z_n with the subgroup of order $\frac{n}{2}$ generated by the element 2 acting on the points. The base blocks are as follows:

$n = 46$: $[2, 3, 15, 44, 13]$, $[1, 8, 10, 16, 17]$, $[1, 5, 25, 26, 19]$, $[2, 17, 25, 36, 30]$, $[1, 18, 28, 44, 4]$.

$n = 82$: $[1, 34, 44, 71, 52]$, $[1, 16, 70, 77, 38]$, $[1, 12, 42, 63, 18]$, $[2, 36, 61, 80, 38]$, $[2, 31, 44, 55, 70]$, $[1, 15, 57, 64, 75]$, $[1, 66, 81, 82, 4]$, $[1, 33, 37, 67, 28]$, $[1, 36, 39, 48, 11]$. □

LEMMA 2.4. *There exists a* $(K_5 \setminus e)$*-design of order* 45.

Proof. The design is constructed on $Z_{44} \cup \{\infty\}$ with the subgroup of order 11 generated by the element 4 acting on the point set, where ∞ is a fixed point under the group action. The base blocks are as follows:

$[1,17,27,30,13]$, $[2,34,37,38,32]$, $[4,5,16,35,29]$, $[1,7,26,35,9]$,
$[2,3,15,26,7]$, $[4,11,26,33,32]$, $[3,8,12,23,29]$, $[2,16,19,25,36]$,
$[2,4,30,41,12]$, $[\infty,3,4,10,1]$. □

LEMMA 2.5. *There exists a* $(K_5 \setminus e)$*-GDD of type* $1^{44}19^1$.

Proof. Let $V = (Z_{11} \times \{0,1,2,3,x\}) \cup \{\infty_0, \infty_1, \dots, \infty_7\}$. The hole of size 19 is $(Z_{11} \times \{x\}) \cup \{\infty_0, \dots, \infty_7\}$. The desired graphs are as follows:

$[(i,0),(1+i,1),(i,2),(i,x),(i-1,x)]$, $[(2+i,0),(4+i,1),(3+i,2),(i,x),(i-2,x)]$,
$[(7+i,0),(2+i,2),(9+i,3),(i,x),(i-1,x)]$,
$[(3+i,1),(9+i,2),(4+i,3),(8+i,0),(i,x)]$,
$[(10+i,1),(8+i,2),(6+i,3),(0+i,1),(i,x)]$,
$[(i,x),(7+i,1),(4+i,2),(10+i,0),(2+i,3)]$,
$[(i,x),(3+i,0),(5+i,0),(6+i,0),(9+i,0)]$,
$[(i,x),(i,1),(9+i,1),(5+i,1),(8+i,1)]$,
$[(i,x),(5+i,2),(10+i,2),(6+i,2),(7+i,2)]$,
$[(i,x),(i,3),(8+i,3),(2+i,3),(7+i,3)]$,
$[\infty_0,(i,1),(i,3),(i,2),(i,0)]$, $[\infty_1,(2+i,2),(3+i,3),(i,0),(1+i,1)]$,
$[\infty_2,(4+i,2),(6+i,3),(i,0),(2+i,1)]$, $[\infty_3,(3+i,1),(9+i,3),(i,0),(6+i,2)]$,
$[\infty_4,(4+i,1),(8+i,2),(i,0),(1+i,3)]$, $[\infty_5,(i,0),(10+i,2),(5+i,1),(4+i,3)]$,
$[\infty_6,(i,0),(7+i,1),(3+i,2),(10+i,3)]$, $[\infty_7,(i,0),(8+i,3),(10+i,1),(9+i,2)]$,
for $i \in Z_{11}$. □

LEMMA 2.6. *There exists a* $(K_5 \setminus e)$*-GDD of type* $g^n h^1$ *for each* $(g,n,h) \in \{(1,80,19),\ (1,80,28),\ (1,80,37),\ (16,5,40),\ (1,116,37),\ (1,116,55),\ (1,143,37),\ (1,143,46),\ (1,152,46),\ (1,161,46),\ (28,8,37),\ (1,260,46)\}$.

Proof. We apply the difference method on Z_{gn} with the groups generated by the cosets of the subgroup of order g. The base blocks for each design are listed in Table 2. The table consists of three columns: the first column corresponds to the parameters of the design, the second column is an integer u that divides gn, and the elements of the corresponding base blocks in the third column are distinct modulo u so that the base blocks, after applying the subgroup generated by u, form a "parallel class" (or equivalently u points in the hole). When u is undefined, it corresponds to a graph $K_5 \setminus e$ which will be simply developed over Z_{gn}. Note that we have a total of three types of graphs besides $K_5 \setminus e$. The first type is a $K_4 \setminus e$, where for the new point in the hole and each vertex of the graph $K_4 \setminus e$ we add an edge between them so as to form a $K_5 \setminus e$. The second type is a K_4, where we need to adjoin the new point to only three of the four vertices (called the "interested part" of the K_4 at the beginning of this section) to form a $K_5 \setminus e$. We label the three points by underlying the three vertices in a K_4. The final type is a K_3, where we need to add two vertices, both of them are in the hole, to form a $K_5 \setminus e$. In our representation, there will be two copies of the same K_3, but with different translates over Z_{gn}. We label the two occurrences of the same K_3 by the same superscript. The two occurrences of the block of K_3 indicate which two new points we are going to add to the K_3. Take the following type $1^{80}37^1$ as an example. The two base blocks $\{37,44,72\}^1$ and $\{38,45,73\}^1$ are given the same superscript 1, and we can translate the first block to the second one by adding 1 modulo 80. This means we are going to add ∞_0 and

∞_1 to the block $\{38, 45, 73\}^1$, and ∞_0 and $\infty_{-1} \equiv \infty_{15}$ to the block $\{37, 44, 72\}^1$. If $gn \equiv 0 \pmod 2$, then $[0, \frac{gn}{2}, a, a + \frac{gn}{2}]$ generates only $\frac{gn}{2}$ blocks (short orbit) of $K_4 \setminus e$ when $4a \not\equiv 0 \pmod{gn}$. Furthermore, when 4 divides gn and a is odd, the short orbit $[0, \frac{gn}{2}, a, a + \frac{gn}{2}]$ generates two parallel classes on Z_{gn}. We superscript the short orbit with s to distinguish the blocks. This observation is very useful in constructing the designs.

The required base blocks for the designs are presented in Table 2. □

LEMMA 2.7. *There exists a $(K_5 \setminus e)$-GDD of type $18^4 27^1$.*

Proof. The design is constructed on Z_{72}. Developing the following base blocks by $+1$ mod 72 gives the desired design

$$[5, 39, 70, 72, 48],\ \{\underline{1}, \underline{2}, \underline{12}, 59\},\ \{35, 52, 65\},\ \{3, 9, 30\},\ \{12, 58, 61\},\ \{7, 26, 44\}.$$

Here, the base block of size 4 (K_4) would allow us to add three infinite points as the three underlined points are distinct modulo 3. The points in all the base blocks of size three are distinct modulo 12. These base blocks of size three can be used to generate twelve parallel classes, each of which allows us to add two infinite points. Hence, we can attach twenty-four infinite points to these twelve parallel classes and in total twenty-seven infinite points to the design generated by the above base blocks. □

LEMMA 2.8. *There exists a $(K_5 \setminus e)$-GDD of type $18^n 45^1$ for each $n \in \{4, 5, 6, 7, 8, 9\}$.*

Proof. The designs are constructed on Z_{18n}. Every block of size 4 (K_4) would allow us to add three infinite points as the three underlined points are distinct modulo 3. Each parallel class composed of blocks of size three allows us to add two infinite points. The designs are given in Table 3. The table has three columns: the first column corresponds to the order of the design, the second column corresponds to u, a divisor of $18n$ such that the set of blocks in the third column are distinct modulo u. If the blocks in the third column are blocks of size three, we can attach $2u$ infinite points to the u parallel classes generated by the blocks. If the block is a K_4, the u is going to be 3 in the second column, and hence, only three points can be added. If the block is $K_5 \setminus e$ in the third column, the corresponding u is labeled by $-$, which means we will not add any infinite point to the block. □

LEMMA 2.9. *There exists a $(K_5 \setminus e)$-GDD of type 4^9.*

Proof. The construction is on $Z_{32} \cup \{\infty_0, \infty_1, \infty_2, \infty_3\}$. The four infinite points are attached to the four parallel classes generated by the following base block $K_4 \setminus e$. The required base blocks are

$$[1, 2, 4, 8, 13], [1, 6, 16, 19]. \qquad □$$

LEMMA 2.10. *There exists a $(K_5 \setminus e)$-GDD of type 9^n for each $n \in \{6, 10\}$.*

Proof. For each given n, the desired design is constructed on $V = Z_{9n}$. The groups are generated by $\{0, n, 2n, \ldots, 8n\}$ and the design is generated from the following base blocks with the subgroup of order $\frac{n}{2}$ generated by the element 2 as the automorphism group:

$n = 6$: $[2, 3, 19, 34, 6]$, $[1, 5, 15, 26, 28]$, $[1, 10, 36, 47, 2]$, $[1, 30, 35, 40, 33]$, $[2, 9, 16, 37, 54]$.

$n = 10$: $[1, 33, 42, 77, 27]$, $[2, 34, 47, 59, 9]$, $[2, 4, 86, 89, 20]$, $[1, 28, 72, 89, 87]$, $[1, 29, 37, 90, 48]$, $[1, 2, 38, 43, 25]$, $[1, 23, 44, 82, 40]$, $[2, 14, 36, 69, 78]$,$[1, 8, 19, 35, 70]$. □

LEMMA 2.11. *There exists a $(K_5 \setminus e)$-GDD of type 9^n for each $n \in \{5, 7, 9, 11, 13, 15, 17\}$.*

Table 2

(g,n,h)	u	Base blocks
$(1,80,19)$	5	$[0,40,1,41]^s$, $\{\underline{4},\underline{58},\underline{67},64\}$,
	4	$[11,56,69,66]$,
	10	$\{\underline{4},\underline{37},\underline{53},41\}$, $\{\underline{2},\underline{9},\underline{68},20\}$, $[26,45,50,11]$,
	–	$[31,33,61,69,4]$.
$(1,80,28)$	2	$[0,40,1,41]^s$,
	10	$[4,25,71,57]$, $\{\underline{9},\underline{53},\underline{78},70\}$, $\{\underline{2},\underline{40},\underline{66},35\}$,
	16	$[16,18,46,61]$, $[5,56,76,79]$, $[35,38,42,20]$, $[1,11,23,25]$.
$(1,80,37)$	5	$[0,40,1,41]^s$, $\{\underline{2},\underline{58},\underline{74},60\}$,
	16	$\{37,44,72\}^1$, $\{38,45,73\}^1$, $\{17,42,63\}^2$, $\{18,43,64\}^2$,
		$[3,20,23,30]$,
	16	$[12,54,63,24]$, $\{\underline{34},\underline{67},\underline{71},2\}$, $\{\underline{4},\underline{9},\underline{53},27\}$,
		$\{45,58,64\}^3$, $\{46,59,65\}^3$.
$(16,5,40)$	40	$[24,5,31,3]$, $[11,42,14,25]$, $[36,74,23,80]$, $[16,4,20,62]$,
		$\{9,32,33\}^1$, $\{49,72,73\}^1$, $\{7,15,48\}^2$, $\{47,55,8\}^2$,
		$\{70,1,38\}^3$, $\{30,41,78\}^3$, $\{59,6,77\}^4$, $\{19,46,37\}^4$.
$(1,116,37)$	29	$\{\underline{37},\underline{41},\underline{109},57\}$, $\{\underline{60},\underline{84},\underline{97},42\}$, $\{\underline{52},\underline{93},\underline{98},95\}$, $\{\underline{5},\underline{56},\underline{112},22\}$,
		$\{\underline{24},\underline{100},\underline{108},78\}$, $\{\underline{36},\underline{72},\underline{86},107\}$, $\{\underline{9},\underline{78},\underline{106},15\}$,
		$\{46,61,73\}^1$, $\{47,62,74\}^1$,
	4	$[16,26,109,103]$,
	4	$[5,83,114,16]$.
$(1,116,55)$	2	$[0,58,1,59]^s$,
	29	$\{50,55,103\}^1$, $\{52,57,105\}^1$, $\{7,15,59\}^2$, $\{9,17,61\}^2$,
		$\{4,95,107\}^3$, $\{6,97,109\}^3$, $\{\underline{25},\underline{53},\underline{71},103\}$,
		$[40,60,116,56]$, $[5,14,41,106]$,
	4	$[61,112,115,26]$,
	4	$[4,75,109,14]$,
	4	$[27,29,50,68]$,
	4	$[17,90,96,43]$,
	4	$[42,49,116,71]$,
	4	$[61,75,92,42]$.
$(1,143,37)$	11	$\{\underline{23},\underline{117},\underline{130},99\}$, $[15,21,112,137]$, $[3,99,107,116]$,
	13	$\{\underline{29},\underline{53},\underline{93},112\}$, $\{\underline{36},\underline{39},\underline{97},77\}$, $\{\underline{34},\underline{89},\underline{126},103\}$, $[4,72,77,70]$,
	13	$\{\underline{82},\underline{94},\underline{104},138\}$, $\{\underline{5},\underline{119},\underline{123},91\}$, $\{\underline{1},\underline{64},\underline{99},49\}$, $[24,113,114,124]$,
	–	$[36,45,52,78,117]$.
$(1,143,46)$	11	$\{\underline{1},\underline{38},\underline{79},117\}$, $[29,30,105,92]$, $[33,119,131,80]$,
	11	$\{\underline{1},\underline{108},\underline{125},26\}$, $[6,10,79,58]$, $[29,121,129,126]$,
	11	$\{\underline{82},\underline{91},\underline{98},131\}$, $[34,57,88,4]$, $[28,31,41,7]$,
	13	$\{\underline{32},\underline{43},\underline{120},61\}$, $\{\underline{5},\underline{11},\underline{126},25\}$, $\{\underline{7},\underline{65},\underline{67},33\}$, $[8,23,79,116]$.
$(1,152,46)$	38	$\{\underline{15},\underline{19},\underline{48},45\}$, $\{\underline{52},\underline{80},\underline{131},137\}$, $\{\underline{34},\underline{54},\underline{104},15\}$, $\{\underline{6},\underline{13},\underline{60},65\}$,
		$\{\underline{11},\underline{59},\underline{147},105\}$, $\{\underline{20},\underline{56},\underline{100},31\}$, $\{\underline{5},\underline{123},\underline{145},67\}$, $\{\underline{74},\underline{101},\underline{111},14\}$,
		$\{\underline{8},\underline{23},\underline{32},152\}$, $\{\underline{12},\underline{83},\underline{151},52\}$, $[0,76,1,77]^s$,
		$\{2,105,140\}^1$, $\{3,106,141\}^1$,
	4	$[26,69,87,28]$,
	4	$[2,23,109,40]$.
$(1,161,46)$	23	$\{\underline{33},\underline{40},\underline{120},79\}$, $\{\underline{85},\underline{105},\underline{113},117\}$, $\{\underline{49},\underline{65},\underline{71},12\}$, $\{\underline{12},\underline{30},\underline{116},126\}$,
		$\{\underline{4},\underline{9},\underline{77},60\}$, $[14,130,149,68]$, $[43,52,161,110]$,
	23	$\{\underline{104},\underline{106},\underline{137},107\}$, $\{\underline{27},\underline{48},\underline{159},96\}$, $\{\underline{80},\underline{144},\underline{158},10\}$,
		$\{\underline{39},\underline{100},\underline{115},75\}$, $\{\underline{3},\underline{41},\underline{153},130\}$, $[9,33,93,128]$, $[7,42,97,86]$.
$(28,8,37)$	7	$[50,65,103,220]$, $\{\underline{112},\underline{174},\underline{193},60\}$,
	14	$[118,141,217,223]$, $[67,206,208,184]$, $\{\underline{61},\underline{134},\underline{185},95\}$, $\{\underline{32},\underline{65},\underline{140},18\}$,
	8	$[23,27,64,30]$, $[2,129,213,28]$,
	8	$[24,51,118,7]$, $[1,50,165,196]$,
	–	$[30,51,143,188,101]$, $[22,77,183,203,16]$, $[26,96,105,191,31]$,
		$[184,209,219,220,116]$, $[73,85,131,162,103]$.
$(1,260,46)$	26	$[0,130,1,131]^s$, $\{\underline{32},\underline{125},\underline{147},197\}$, $\{\underline{25},\underline{60},\underline{256},220\}$, $\{\underline{111},\underline{122},\underline{245},155\}$,
		$\{\underline{171},\underline{205},\underline{220},139\}$, $\{\underline{106},\underline{218},\underline{258},234\}$, $\{\underline{29},\underline{72},\underline{134},12\}$, $\{\underline{92},\underline{146},\underline{149},33\}$,
		$\{\underline{113},\underline{117},\underline{187},26\}$,
	20	$[89,99,191,178]$, $[83,104,201,167]$, $[6,115,142,193]$, $[136,174,188,192]$,
		$[90,137,145,60]$,
	–	$[26,122,127,210,175]$, $[111,134,205,251,136]$, $[45,87,106,191,113]$,
		$[13,64,95,170,58]$, $[72,84,225,234,205]$, $[100,139,167,225,180]$.

Table 3

Order	u	Base blocks
$18^4 45^1$	3	$\{\underline{1}, \underline{2}, \underline{12}, 51\}$,
	9	$\{6, 11, 57\}$, $\{34, 37, 71\}$, $\{4, 18, 59\}$,
	12	$\{45, 58, 64\}$, $\{14, 29, 59\}$, $\{6, 8, 15\}$, $\{13, 31, 60\}$.
$18^5 45^1$	3	$\{\underline{37}, \underline{68}, \underline{75}, 34\}$,
	9	$\{4, 20, 52\}$, $\{12, 78, 89\}$, $\{54, 73, 77\}$,
	9	$\{2, 14, 16\}$, $\{37, 58, 80\}$, $\{18, 51, 57\}$,
	3	$\{1, 2, 30\}$,
	–	$[3, 40, 49, 57, 57]$.
$18^6 45^1$	3	$\{\underline{5}, \underline{18}, \underline{22}, 68\}$,
	12	$\{2, 9, 28\}$, $\{47, 49, 108\}$, $\{67, 70, 90\}$, $\{3, 17, 56\}$,
	9	$\{2, 39, 72\}$, $\{34, 55, 98\}$, $\{5, 6, 85\}$,
	–	$[3, 38, 43, 54, 11]$, $[9, 34, 86, 108, 19]$.
$18^7 45^1$	3	$\{\underline{21}, \underline{106}, \underline{122}, 54\}$,
	9	$\{33, 64, 111\}$, $\{2, 61, 126\}$, $\{44, 50, 94\}$,
	9	$\{40, 93, 104\}$, $\{8, 74, 96\}$, $\{19, 70, 99\}$,
	3	$\{1, 9, 35\}$,
	–	$[14, 15, 51, 54, 69]$, $[38, 50, 65, 95, 33]$, $[46, 65, 69, 89, 56]$.
$18^8 45^1$	3	$\{\underline{26}, \underline{72}, \underline{109}, 126\}$,
	9	$\{21, 64, 76\}$, $\{5, 71, 90\}$, $\{2, 52, 105\}$,
	12	$\{41, 48, 90\}$, $\{4, 25, 135\}$, $\{9, 115, 143\}$, $\{20, 22, 26\}$,
	–	$[19, 37, 105, 136, 106]$, $[66, 96, 118, 129, 143]$, $[75, 80, 89, 140, 60]$, $[70, 73, 109, 144, 47]$.
$18^9 45^1$	3	$\{\underline{61}, \underline{156}, \underline{161}, 140\}$,
	9	$\{63, 78, 118\}$, $\{48, 49, 83\}$, $\{44, 115, 158\}$,
	9	$\{117, 119, 147\}$, $\{59, 106, 132\}$, $\{31, 80, 91\}$,
	3	$\{1, 51, 104\}$,
	–	$[11, 55, 156, 159, 17]$, $[21, 45, 52, 127, 137]$, $[47, 111, 131, 144, 123]$, $[8, 45, 74, 113, 49]$, $[8, 50, 82, 101, 60]$.

Proof. Let $V = Z_{9n}$. The groups are generated by $\{0, n, 2n, \ldots, 8n\}$. The desired designs are obtained by developing the following base blocks cyclically:

$n = 5$: $[1, 2, 20, 34, 43]$, $[1, 9, 18, 25, 7]$.

$n = 7$: $[1, 6, 17, 47, 2]$, $[1, 11, 35, 38, 55]$, $[1, 7, 39, 52, 62]$.

$n = 9$: $[1, 15, 41, 71, 48]$, $[1, 30, 43, 58, 20]$, $[1, 60, 65, 81, 62]$, $[1, 13, 44, 50, 9]$.

$n = 11$: $[1, 58, 68, 70, 8]$, $[1, 6, 26, 49, 96]$, $[1, 38, 53, 72, 32]$, $[1, 4, 28, 86, 87]$, $[1, 19, 54, 55, 92]$.

$n = 13$: $[1, 10, 22, 110, 77]$, $[1, 43, 71, 74, 3]$, $[1, 47, 83, 108, 113]$, $[1, 5, 24, 38, 98]$, $[1, 12, 28, 46, 60]$, $[1, 8, 65, 103, 2]$.

$n = 15$: $[1, 69, 75, 77, 89]$, $[1, 79, 92, 132, 80]$, $[1, 37, 70, 86, 88]$, $[1, 11, 33, 38, 8]$, $[1, 24, 63, 94, 82]$, $[1, 10, 81, 107, 102]$, $[1, 12, 29, 53, 101]$.

$n = 17$: $[1, 12, 14, 22, 85]$, $[1, 26, 53, 57, 79]$, $[1, 4, 51, 87, 142]$, $[1, 36, 41, 117, 134]$, $[1, 29, 58, 88, 90]$, $[1, 47, 96, 110, 80]$, $[1, 7, 46, 55, 8]$, $[1, 25, 113, 135, 136]$. □

Lemma 2.12. *There exists a $(K_5 \backslash e)$-GDD of type 9^n for each* $n \in \{8, 12, 14, 16, 18, 20\}$.

Proof. The construction is on $Z_{9(n-1)} \cup \{\infty_0, \infty_1, \ldots, \infty_8\}$. The nine infinite points are attached to either nine classes generated by three base blocks of K_4, or three classes generated by one base block of K_3 together with three classes generated by one base block of K_4. Here, the nine underlined elements coming from the three base blocks of size four are distinct modulo 9, and besides, the three elements in the K_3 and the three underlined elements in the unique K_4 are both distinct modulo 3.

The required base blocks are as follows:

$n = 8$: $[6, 38, 56, 60, 4]$,
$\{\underline{25}, \underline{41}, \underline{42}, 15\}$, $\{\underline{2}, \underline{8}, \underline{46}, 41\}$, $\{\underline{3}, \underline{18}, \underline{58}, 15\}$.
$n = 12$: $[2, 31, 94, 96, 70]$, $[45, 48, 60, 73, 44]$, $[1, 9, 27, 44, 46]$, $[7, 16, 37, 64, 83]$,
$\{16, 26, 66\}$,
$\{\underline{1}, \underline{15}, \underline{53}, 94\}$.
$n = 14$: $[59, 74, 112, 117, 13]$, $[6, 14, 74, 86, 95]$, $[1, 2, 4, 8, 36]$, $[11, 86, 95, 117, 66]$,
$\{\underline{17}, \underline{57}, \underline{104}, 3\}$, $\{\underline{45}, \underline{55}, \underline{96}, 72\}$, $\{\underline{2}, \underline{25}, \underline{94}, 75\}$.
$n = 16$: $[34, 39, 47, 83, 66]$, $[39, 93, 109, 121, 5]$, $[1, 2, 4, 8, 41]$, $[12, 60, 69, 133, 89]$,
$[37, 79, 96, 117, 61]$,
$\{\underline{58}, \underline{108}, \underline{119}, 36\}$, $\{\underline{53}, \underline{86}, \underline{132}, 63\}$, $\{\underline{1}, \underline{52}, \underline{93}, 26\}$.
$n = 18$: $[30, 50, 71, 80, 86]$, $[5, 38, 101, 113, 93]$, $[19, 85, 101, 128, 48]$, $[1, 2, 4, 8, 50]$,
$[39, 44, 57, 130, 116]$, $[19, 41, 80, 120, 140]$,
$\{\underline{72}, \underline{119}, \underline{142}, 100\}$, $\{\underline{32}, \underline{58}, \underline{96}, 127\}$, $\{\underline{1}, \underline{12}, \underline{26}, 36\}$.
$n = 20$: $[9, 23, 119, 132, 107]$, $[34, 43, 69, 101, 26]$, $[42, 112, 122, 163, 73]$, $[35, 69, 141, 159, 169]$,
$[1, 2, 4, 8, 57]$, $[2, 23, 46, 158, 128]$, $[15, 39, 79, 156, 108]$,
$\{\underline{3}, \underline{14}, \underline{169}, 77\}$, $\{\underline{45}, \underline{105}, \underline{130}, 78\}$, $\{\underline{2}, \underline{44}, \underline{127}, 24\}$. □

LEMMA 2.13. *There exists a* $(K_5 \setminus e)$*-GDD of type* 9^{19}*.*

Proof. The design is constructed on $Z_{19} \times Z_9$. The groups are $\{i\} \times Z_9$ for $i \in Z_{19}$. The blocks are $[(4^i j, k), (4^i j, k+1), (4^i(j+2), k+3), (4^i(j+4), k+1), (4^i(8+j), 5+k)]$ for $j \in Z_{19}$, $k \in Z_9$ and $i = 0, 1, 2, \ldots, 8$. □

LEMMA 2.14. *There exists a* $(K_5 \setminus e)$*-GDD of type* 18^n *for each* $n \geq 4$*.*

Proof. By [7], we have a $(K_5 \setminus e)$-GDD of type 6^4. We also have a $(K_5 \setminus e)$-GDD of type 6^7 which is constructed on $V = Z_{42}$ with the groups being generated by the subgroup $\{0, 7, 14, \ldots, 35\}$. The blocks are obtained by developing the following base blocks cyclically:

$$[1, 2, 4, 17, 35], [1, 7, 19, 24, 39].$$

By [13, Theorem 1.3], there exists a $\{4, 7\}$-GDD of type 3^n for each $n \geq 4$ and $n \neq 6$. Give weight 6 to each point of this GDD using $(K_5 \setminus e)$-GDDs of types 6^4 and 6^7 as input designs to obtain a $(K_5 \setminus e)$-GDD of type 18^n for each $n \geq 4$ and $n \neq 6$. This leaves $n = 6$ to be considered, which is constructed on $V = Z_{108}$ with the groups being generated by the subgroup $\{0, 6, 12, \ldots, 102\}$. The blocks are obtained by developing the following base blocks cyclically:

$$[1, 41, 87, 92, 2], [1, 5, 20, 30, 36], [1, 14, 34, 77, 81], [1, 72, 98, 106, 45], [1, 8, 57, 95, 107]. \quad □$$

LEMMA 2.15. *There exists a* $(K_5 \setminus e)$*-GDD of type* $9^n 18^1$ *for each* $n \in \{4, 5, 6\}$*.*

Proof. When $n = 4$, the GDD is constructed on Z_{36} with eighteen infinite points x_i for $i = 0, 1, \ldots, 17$. The groups of the GDD are $\{i, i+4, \ldots, i+32\}$ for $i = 0, 1, 2, 3$ and $\{x_0, x_1, \ldots, x_{17}\}$. The design is generated by the action of Z_{36} on the blocks. Two classes of $K_4 \setminus e$ are generated by $[0, 18, 1, 19]$. Four additional classes of $K_4 \setminus e$ are generated by $[1, 14, 23, 8]$ by noting that the four points are distinct modulo 4. Six parallel classes of K_3 are generated by the base blocks $\{5, 8, 10\}$ and $\{3, 13, 24\}$ by noting that these points are distinct modulo 6. We attach one infinite point to each class of $K_4 \setminus e$ and two infinite points to each class of K_3.

When $n = 5$, the GDD is constructed on Z_{45} with eighteen infinite points x_i for $i = 0, 1, \ldots, 17$. A base block of $K_5 \setminus e$ is $[2, 11, 30, 44, 3]$, and nine parallel classes of

K_3 are generated by $\{19, 26, 42\}$, $\{12, 23, 36\}$, $\{2, 4, 43\}$. The nine points are distinct modulo 9, and we can attach eighteen infinite points to these nine parallel classes of blocks of size three.

When $n = 6$, the GDD is constructed on Z_{54} with eighteen infinite points x_i for $i = 0, 1, \ldots, 17$. The groups of the GDD are $\{i, i+6, \ldots, i+48\}$ for $i = 0, 1, \ldots, 5$ and $\{x_0, x_1, \ldots, x_{17}\}$. The design is generated from the base blocks as follows by adding 2 modulo 54: Three base blocks of $K_5 \setminus e$ are $[17, 22, 45, 49, 12]$, $[33, 35, 43, 44, 26]$, $[16, 37, 50, 54, 18]$; the first three classes of K_3 are generated by $\{6, 17, 20\}$, $\{10, 25, 45\}$, where the six points are distinct modulo 6; the second three classes of K_3 are generated by $\{8, 9, 16\}$, $\{18, 31, 47\}$, where the six points are distinct modulo 6; and the last three classes of K_3 are generated by $\{1, 16, 41\}$, $\{20, 48, 51\}$, where the six points are distinct modulo 6. □

3. Recursive constructions. In this section, we obtain the main result of this paper. Before we prove the main theorem of the paper, we need to introduce some tools.

A *transversal design* TD(k, n) is a k-GDD of type n^k. The following theorems about TDs can be found in [1].

THEOREM 3.1. *There exists a TD$(p+1, p)$ for any prime power p.*

THEOREM 3.2. *There exists a TD$(6, n)$ for any odd $n \geq 5$.*

THEOREM 3.3. *There exists a TD$(6, n)$ for any $n \equiv 0 \pmod 4$ and $n \geq 8$.*

Next, we need to state the main recursive constructions of the paper.

CONSTRUCTION 3.4 (Wilson's fundamental construction [8]). *Let $(X, \mathcal{G}, \mathcal{B})$ be a GDD, and let $w : X \to Z^+ \cup \{0\}$ be a weight function on X. Suppose that for each block $B \in \mathcal{B}$, there exists a $(K_5 \setminus e)$-GDD of type $\{w(x) : x \in B\}$. Then there is a $(K_5 \setminus e)$-GDD of type $\{\sum_{x \in G} w(x) : G \in \mathcal{G}\}$.*

CONSTRUCTION 3.5 (inflation [8]). *If there exists a $(K_5 \setminus e)$-design of type T, then there exists a $(K_5 \setminus e)$-design of type mT if a TD$(4, m)$ exists.*

LEMMA 3.6. *There exists a $(K_5 \setminus e)$-GDD of type $9^5 36^1$.*

Proof. There exists a 3-RGDD of type 9^5 [23]. It has eighteen parallel classes, and we add two infinite points for every class and the thirty-six new infinite points form a new group. □

We first give a simple proof of the result stated in [18, 21].

LEMMA 3.7. *There exists a $(K_5 \setminus e)$-design of order n for any $n \equiv 1 \pmod{18}$, $n \geq 19$.*

Proof. When $n = 19$, the design is constructed in [7]. When $n = 37, 55$, the designs are constructed in Lemma 2.1. By Lemma 2.14, we have a $(K_5 \setminus e)$-GDD of type 18^t for each $t \geq 4$. Adjoining one infinite point and filling in the holes with a $(K_5 \setminus e)$-design of order 19, we obtain a $(K_5 \setminus e)$-design of order $18t + 1$. This settles all $n \geq 73$. □

Before we proceed, we need to give a solution for $(K_5 \setminus e)$-GDDs of type 9^n.

LEMMA 3.8. *For every $n \geq 5$, there exists a $(K_5 \setminus e)$-GDD of type 9^n.*

Proof. When $5 \leq n \leq 20$, the required designs are constructed in Lemmas 2.10–2.13. Take a TD$(5, t)$ for $t = 4, 5$, give weight nine to points in the first four groups, and weight nine or eighteen to points in the last group as both a $(K_5 \setminus e)$-GDD of type 9^5 and a $(K_5 \setminus e)$-GDD of type $9^4 18^1$ (Lemma 2.15) exist. We obtain a $(K_5 \setminus e)$-GDD of type $36^4(36+9k)^1$ for each $k = 0, 1, 2, 3, 4$ and a $(K_5 \setminus e)$-GDD of type $45^4(45 + 9m)^1$ for each $m = 0, 1, 2, 3, 4, 5$. Take the former GDD and add nine points and fill in the groups with either a $(K_5 \setminus e)$-GDD of type 9^5 or a $(K_5 \setminus e)$-GDD of type 9^{5+k}. We obtain a $(K_5 \setminus e)$-GDD of type 9^{21+k}. This settles $21 \leq n \leq 25$. Take the latter GDD, add nine infinite points, and fill in the groups with $(K_5 \setminus e)$-GDDs of types 9^6 and 9^{6+m} to obtain a $(K_5 \setminus e)$-GDD of type 9^{26+m}. This settles $26 \le n \le 31$.

For the remaining values of $n \ge 32$, take a $\mathrm{TD}(6,t)$ and remove a point to redefine a $\{6,t\}$-GDD of type $5^t(t-1)^1$. Give weight nine to points in the first t groups of size five, weight $0, 9, 18, 36$ to points in the last group (the required input $(K_5 \setminus e)$-GDDs of types $9^5 18^1$ and $9^5 36^1$ exist by Lemmas 2.15 and 3.6). If there also exists a $(K_5 \setminus e)$-GDD of type 9^t, then we obtain a $(K_5 \setminus e)$-GDD of type $45^t(9x)^1$ for each $x = 5, \ldots, 4(t-1) - 2$. We obtain a $(K_5 \setminus e)$-GDD of type 9^{5t+x} whenever $(K_5 \setminus e)$-GDDs of types 9^t and 9^x exist. Take all odd $t \ge 5$ and use simple induction to establish the result. □

LEMMA 3.9. *For every* $n \ne 0, 3$, *there exists a* $(K_5 \setminus e)$*-design of order* $18n + 10$.

Proof. When $n = 1, 2, 4$, the designs are constructed in Lemmas 2.2 and 2.3. When $n = 5$, we have a $(K_5 \setminus e)$-GDD of type $18^4 27^1$ by Lemma 2.7. Add one infinite point and fill in the holes with a $(K_5 \setminus e)$-design of order 19 and a $(K_5 \setminus e)$-design of order 28 to obtain a $(K_5 \setminus e)$-design of order 100. Furthermore, there exists a $(K_5 \setminus e)$-GDD of type $18^t 45^1$ for each $t = 4, 5, 6, 7, 8, 9$ by Lemma 2.8. Add one infinite point and fill in the holes with a $(K_5 \setminus e)$-design of order 19 and a $(K_5 \setminus e)$-design of order 46 to obtain a $(K_5 \setminus e)$-design of order $18t + 46$. This settles $n = 6, 7, 8, 9, 10, 11$. Finally, take a $\mathrm{TD}(6,t)$ and remove one point to redefine a $\{6,t\}$-GDD of type $5^t(t-1)^1$. Give weight nine to points in the first t groups of size five, weight $0, 9, 18, 36$ to points in the last group (the required input $(K_5 \setminus e)$-GDDs exist by Lemmas 2.15, 3.6 and 3.8). We obtain a $(K_5 \setminus e)$-GDD of type $45^t(18x)^1$ for each $x = 1, 2, 3, \ldots, 2(t-1)$. Add one point and fill in the holes with a $(K_5 \setminus e)$-design of order 46 and a $(K_5 \setminus e)$-design of order $18x + 1$ to obtain a $(K_5 \setminus e)$-design of order $45t + 18x + 1$. Take odd $t \ge 5$ to prove the result. □

Next, we consider the case when the order $n \equiv 0 \pmod 9$. First, we need the following result.

LEMMA 3.10. *There exists a* $(K_5 \setminus e)$*-GDD of type* $1^n(\frac{2(n-1)}{5})^1$ *for each* $n \equiv 16$ $(\mathrm{mod}\ 20)$ *and* $n \ge 16$ *except possibly for* $n = 116, 296$.

Proof. Take a resolvable $(K_4 - e)$-design of order n from [14] and complete all the parallel classes to obtain the designs as desired. □

LEMMA 3.11. *There exists a* $(K_5 \setminus e)$*-design of order* $9n$ *for each* $n \in \{5, 7, 11, 12, 13, 14, 17, 19, 20, 21, 22, 23, 34\}$.

Proof. When $n = 5$, the design is constructed in Lemma 2.4. When $n = 14$, take a $(K_5 \setminus e)$-GDD of type $16^5 40^1$ (Lemma 2.6), add six infinite points, and fill in a $(K_5 \setminus e)$-GDD of type $1^{16} 6^1$ (Lemma 3.10) and a $(K_5 \setminus e)$-GDD of type 1^{46} to obtain a $(K_5 \setminus e)$-design of order 9×14. For the remaining values of n, the desired designs are obtained by taking the $(K_5 \setminus e)$-GDDs of type $1^m h^1$ constructed in Lemmas 2.5–2.6 and filling in the holes with a $(K_5 \setminus e)$-design of order h (such input $(K_5 \setminus e)$-designs exist by Lemmas 3.7 and 3.9). Here, the corresponding parameters are $(m, h) \in \{(44, 19), (80, 19), (80, 28), (80, 37), (116, 37), (116, 55), (143, 37), (143, 46), (152, 46), (161, 46),$ and $(260, 46)\}$. □

LEMMA 3.12. *There exists a* $(K_5 \setminus e)$*-design of order* $9n$ *for each* $n \in \{27, 28, 29, 31, 33, 41\}$.

Proof. For $n = 27$, we take a $\mathrm{TD}(5,5)$. Give weight nine to points in the first four groups. We give weight eighteen to two points in the last group and weight nine to the remaining points. Since $(K_5 \setminus e)$-GDDs of types 9^5 and $9^4 18^1$ exist (Lemmas 3.8 and 2.15), we obtain a $(K_5 \setminus e)$-GDD of type $45^4 63^1$. Fill in the holes with a $(K_5 \setminus e)$-design of order $9t$ for $t = 5, 7$ to obtain the design as desired.

For $n = 28$, inflate a $(K_5 \setminus e)$-GDD of type 4^9 coming from Lemma 2.9 by seven to obtain a $(K_5 \setminus e)$-GDD of type 28^9. Filling in the holes with a $(K_5 \setminus e)$-design of order 28 gives the desired design. For $n = 29$, take a $(K_5 \setminus e)$-GDD of type $28^8 37^1$ coming from Lemma 2.6 and fill in the holes with a $(K_5 \setminus e)$-design of order 28 and a $(K_5 \setminus e)$-design of order 37 to obtain the design as desired.

For $n = 31$, we first consider a design on Z_{60} with base blocks

$$\{27, 29, 56\}, \{42, 48, 59\}, \{7, 14, 33\}, \{1, 4, 22\}, \{\underline{1}, \underline{2}, \underline{15}, 24\}.$$

These five base blocks cover all differences except for the multiples of four and five. So, it basically has two sets of holes that intersect on three points. Furthermore, the twelve points in the four blocks of K_3 are distinct modulo 12, and the three special points in the K_4 are distinct modulo 3. So, ignoring the holes, we can add twenty-seven points to the design to obtain a $(K_5 \setminus e)$-design with two sets of spanning holes. This is essentially a $(K_5 \setminus e)$-HGDD of type $(3^5)^4 27^1$, where we treat it as a $(K_5 \setminus e)$-GDD of type $15^4 27^1$ but with a set of five spanning holes, each is a 3^4 that cuts across each group in three points. Now, we give weight three, the hole of size 3^4 will be inflated to become a hole of size 9^4, and we plug in a $(K_5 \setminus e)$-GDD of type $9^4 18^1$ (Lemma 2.15) to obtain a $(K_5 \setminus e)$-GDD of type $45^4 99^1$. Fill in the holes with a $(K_5 \setminus e)$-design of order $9k$ for $k = 5, 11$ to obtain a $(K_5 \setminus e)$-design of order 9×31.

For $n = 33$, we first consider a design on Z_{84} with four base blocks

$$[5, 14, 16, 31, 71], [1, 2, 52, 55, 47], \{28, 41, 51\}, \{\underline{1}, \underline{20}, \underline{42}, 79\}.$$

These four base blocks cover all differences except for the multiples of four and seven. So, it basically has two sets of holes that intersect on three points. Furthermore, the three points in the unique block of K_3 are distinct modulo 3, and the three special points in the K_4 are distinct modulo 3. So, ignoring the holes, we can add nine points to the design to obtain a $(K_5 \setminus e)$-design with two sets of spanning holes. This is essentially a $(K_5 \setminus e)$-HGDD of type $(3^7)^4 9^1$. Now, we give weight three and the hole of size 3^4 will be inflated to become a hole of size 9^4 and we plug in a $(K_5 \setminus e)$-GDD of type $9^4 18^1$ (Lemma 2.15) to obtain a $(K_5 \setminus e)$-GDD of type $63^4 45^1$. Fill in the holes with a $(K_5 \setminus e)$-design of order $9k$ for $k = 5, 7$ to obtain a $(K_5 \setminus e)$-design of order 9×33.

For $n = 41$, take a TD$(7, 7)$ and pick up two blocks that intersect in one point. Remove all points in the two blocks except for the point of intersection. It becomes a $\{5, 6, 7\}$-GDD of type $5^6 7^1$. Give weight nine to points in the six groups of size five, and weight eighteen to four points in the last group and weight nine to the remaining points. Since a $(K_5 \setminus e)$-GDD of type 9^t exists for any integer $t \geq 5$ and a $(K_5 \setminus e)$-GDD of type $9^i 18^1$ exists for $i = 4, 5, 6$ (Lemma 2.15), we have a $(K_5 \setminus e)$-GDD of type $45^6 99^1$. Fill in the holes to obtain the result. □

LEMMA 3.13. *There exists a $(K_5 \setminus e)$-design of order $9n$ for each $n \in \{37, 43, 44\}$.*

Proof. If there exists a $\{5, 6, 7, 8\}$-GDD of type $7^i 5^j$, we can give weight nine with a $(K_5 \setminus e)$-GDD of type 9^k for $k = 5, 6, 7, 8$ and fill in the holes with $(K_5 \setminus e)$-designs of orders 45 or 63 to obtain a $(K_5 \setminus e)$-design of order $9 \times (7i + 5j)$. Therefore, the conclusion follows if we construct $\{5, 6, 7, 8\}$-GDDs of types $7^1 5^6$, $7^4 5^3$, and $7^2 5^6$.

To construct a GDD of type $7^1 5^6$, take a TD$(7, 7)$ containing two disjoint blocks. Remove the points in the two blocks except for the two points in the same group. Every block has at least five points. This settles $n = 37$.

To construct a GDD of type $7^4 5^3$, we again take a TD$(7,7)$. We first remove two points from each of the two selected groups. This defines only four blocks of size five. Then, we can pick two points in the third group so that they are disjoint from the four blocks of size five and remove those two points. It gives a GDD of type $7^4 5^3$.

To construct a GDD of type $7^2 5^6$, we look at the construction of a TD$(8,7)$ on $Z_7 \times (Z_7 \cup \{\infty\})$. The groups are $Z_7 \times \{i\}$ for $i \in Z_7 \cup \{\infty\}$. The blocks are $\{(i,\infty), (j,0), (i+j,1), (i+2j,2), (i+3j,3), (i+4j,4), (i+5j,5), (i+6j,6)\}$ for $i,j \in Z_7$. Consider the four blocks when $(i,j) = (0,0), (0,1), (4,1), (4,0)$ and restrict our attention to groups $0,1,2,3,4,6$. It is easy to check that if we remove the points $(0,0)$, $(0,1)$, $(0,2)$ from the first block, $(1,0)$, $(1,1)$, $(1,2)$ from the second block, $(0,3)$, $(1,4)$, $(3,6)$ from the third block, and $(4,3)$, $(4,4)$, $(4,6)$ from the last block, then no block in the design would intersect these twelve points in four points. Hence, by removing these twelve points, we obtain the desired GDD of type $7^2 5^6$. □

Lemma 3.14. *If* $n \equiv 0 \pmod 5$ *and* $n \neq 10, 15$, *then there exists a* $(K_5 \setminus e)$*-design of order* $9n$.

Proof. Take a $(K_5 \setminus e)$-GDD of type 9^t for $t \geq 5$ by Lemma 3.8 and inflate it with a TD$(4,5)$ to obtain a $(K_5 \setminus e)$-GDD of type 45^t. Fill in the holes with a $(K_5 \setminus e)$-design of order 45. This settles $n \geq 25$. When $n = 20$, the result follows from Lemma 3.11. □

Lemma 3.15. *If* $n \not\equiv 0 \pmod 5$ *and* $n \notin \{1,6,16,26,2,3,8,18,4,9,24\}$, *then there exists a* $(K_5 \setminus e)$*-design of order* $9n$.

Proof. Take a TD$(6,k)$ and remove one point to obtain a $\{6,k\}$-GDD of type $5^k(k-1)^1$. We note that the blocks of size k are disjoint from the group of size $k-1$. Give weight nine to points in the k groups of size five, and weight $0,9,18,36$ to points in the last group (the input $(K_5 \setminus e)$-GDDs exist by Lemmas 3.8, 2.15, and 3.6). This gives a $(K_5 \setminus e)$-GDD of type $45^k(9m)^1$ for all $7 \leq m \leq 4k-6$. If there exists a $(K_5 \setminus e)$-design of order $9m$, then we can obtain a $(K_5 \setminus e)$-design of order $9(5k+m)$.

When $n \equiv 1 \pmod 5$: When $n = 11, 21, 31, 41$, the required designs are solved in Lemmas 3.11 and 3.12. We take k to be odd, $k \geq 5$ and $m = 11$ to obtain all $n \equiv 6 \pmod{10}$, and $n \geq 36$. On the other hand, we take $k \equiv 0 \pmod 4$, $k \geq 8$ and $m = 11, 21$ to solve $n \equiv 1 \pmod{10}$ and $n \geq 51$.

When $n \equiv 2 \pmod 5$: When $n = 7, 12, 17, 22, 27, 37$, the required designs are solved in Lemmas 3.11, 3.12, and 3.13. When $n \equiv 2 \pmod{10}$, and $n \geq 32$, take odd $k \geq 5$ and $m = 7$. When $n \equiv 7 \pmod{10}$ and $n \geq 47$, take $k \equiv 0 \pmod 4$, $k \geq 8$ and $m = 7, 17$.

When $n \equiv 3 \pmod 5$: When $n = 13, 23, 28, 33, 43$, the required designs are constructed in Lemmas 3.11, 3.12, and 3.13. When $n \equiv 8 \pmod{10}$ and $n \geq 38$, take odd $k \geq 5$ and $m = 13$. When $n \equiv 3 \pmod{10}$ and $n \geq 53$, take $k \equiv 0 \pmod 4$, $k \geq 8$ and $m = 13, 23$.

When $n \equiv 4 \pmod 5$: When $n = 14, 19, 29, 34, 44$, the designs are obtained in Lemmas 3.11, 3.12, and 3.13. When $n \equiv 9 \pmod{10}$ and $n \geq 39$, take odd $k \geq 5$ with $m = 14$ to obtain the result. When $n \equiv 4 \pmod{10}$ and $n \geq 54$, take odd $k \geq 7$ with $m = 19$ to obtain the result. □

Theorem 3.16. *Let* $A_1 = \{64\}$, $A_2 = \{27,36,54,72,81,90,135,144,162,216,234\}$, *and* $E = \{9,10,18\}$. *If* $n \equiv 0,1 \pmod 9$ *and* $n \notin A_1 \cup A_2 \cup E$, *then there exists a* $(K_5 \setminus e)$*-design of order* n. *Furthermore, if* $n \in E$, *then no such design exists.*

Proof. For the nonexistence of a $(K_5 \setminus e)$-design of order n with $n \in \{9,10,18\}$, see [18]. The conclusion then follows by combining Lemmas 3.7, 3.9, 3.14, and 3.15. □

Acknowledgment. The authors cordially thank the anonymous referees for their valuable comments which led to the improvement of this paper.

REFERENCES

[1] R. J. R. Abel, A. E. Brouwer, C. J. Colbourn, and J. H. Dinitz, *Mutually orthogonal Latin squares*, in CRC Handbook of Combinatorial Designs, C. J. Colbourn and J. H. Dinitz, eds., CRC Press, Boca Raton, FL, 1996, pp. 111–142.

[2] J.-C. Bermond and S. Ceroi, *Minimizing SONET ADMs in unidirectional WDM rings with grooming ratio* 3, Networks, 41 (2003), pp. 83–86.

[3] J.-C. Bermond, C. J. Colbourn, D. Coudert, G. Ge, A. C. H. Ling, and X. Muñoz, *Traffic grooming in unidirectional wavelength-division multiplexed rings with grooming ratio* $C = 6$, SIAM J. Discrete Math., 19 (2005), pp. 523–542.

[4] J.-C. Bermond, C. J. Colbourn, A. C. H. Ling, and M. L. Yu, *Grooming in unidirectional rings:* $K_4 - e$ *designs*, Discrete Math., 284 (2004), pp. 57–62.

[5] J.-C. Bermond and D. Coudert, *Traffic grooming in unidirectional WDM ring networks using design theory*, in Proceedings of ICC 2003, Anchorage, AK, 2003, pp. 11–15.

[6] J.-C. Bermond, D. Coudert, and X. Muñoz, *Traffic grooming in unidirectional WDM ring networks: The all-to-all unitary case*, in Proceedings of the ONDM03, Seventh IFIP Workshop Optical Network Design and Modelling, Budapest, Hungary, 2003, pp. 1135–1153.

[7] J.-C. Bermond, C. Huang, A. Rosa, and D. Sotteau, *Decomposition of complete graphs into isomorphic subgraphs with five vertices*, Ars Combin., 10 (1980), pp. 211–254.

[8] T. Beth, D. Jungnickel, and H. Lenz, *Design Theory*, 2nd ed., Cambridge University Press, Cambridge, UK, 1999.

[9] A. L. Chiu and E. H. Modiano, *Traffic grooming algorithms for reducing electronic multiplexing costs in WDM ring networks*, in IEEE/OSA J. Lightwave Technol., 18 (2000), pp. 2–12.

[10] C. J. Colbourn and J. H. Dinitz, eds., *The CRC Handbook of Combinatorial Designs*, CRC Press, Boca Raton, FL, 1996.

[11] R. Dutta and N. Rouskas, *Traffic grooming in WDM networks: Past and future*, IEEE Network, 16 (2002), pp. 46–56.

[12] R. Dutta and N. Rouskas, *On optimal traffic grooming in WDM rings*, IEEE Journal of Selected Areas in Communications, 20 (2002), pp. 1–12.

[13] G. Ge and A. C. H. Ling, *Constructions of quad-rooted double covers*, Graphs Combin., 21 (2005), pp. 231–238.

[14] G. Ge and A. C. H. Ling, *On the existence of resolvable* $K_4 - e$ *designs*, J. Combin. Designs, 15 (2007), pp. 502–510.

[15] O. Gerstel, P. Lin, and G. Sasaki, *Wavelength assignment in a WDM ring to minimize cost of embedded SONET rings*, in IEEE Infocom, San Francisco, CA, 1998, pp. 94–101.

[16] O. Gerstel, R. Ramaswani, and G. Sasaki, *Cost-effective traffic grooming in WDM rings*, IEEE/ACM Transactions on Networking, 8 (2000), pp. 618–630.

[17] O. Goldschmidt, D. Hochbaum, A. Levin, and E. Olinick, *The SONET edge-partition problem*, Networks, 41 (2003), pp. 13–23.

[18] K. Heinrich, *Graph decompositions and designs*, in CRC Handbook of Combinatorial Designs, C. J. Colbourn and J. H. Dinitz, eds., CRC Press, Boca Raton, FL, 1996, pp. 361–366.

[19] J. Q. Hu, *Optimal traffic grooming for wavelength-division-multiplexing rings with all-to-all uniform traffic*, OSA Journal of Optical Networks, 1 (2002), pp. 32–42.

[20] J. Q. Hu, *Traffic grooming in WDM ring networks: A linear programming solution*, OSA Journal of Optical Networks, 1 (2002), pp. 397–408.

[21] Q. Li and Y. Chang, *A few more* $(K_v, K_5 \setminus e)$*-designs*, Bull. Inst. Combin. Appl., 45 (2005), pp. 11–16.

[22] E. Modiano and P. Lin, *Traffic grooming in WDM networks*, IEEE Commun. Mag., 39 (2001), pp. 124–129.

[23] R. S. Rees, *Two new direct product type constructions for resolvable group-divisible designs*, J. Combin. Des., 1 (1993), pp. 15–26.

[24] A. Somani, *Survivable traffic grooming in WDM networks*, in Proceedings of the International Conference on Broadband Optical Fiber Communications Technology, Jalgon, India, 2001, pp. 17–45.

[25] P. J. Wan, G. Calinescu, L. Liu, and O. Frieder, *Grooming of arbitrary traffic in SONET/WDM BLSRs*, IEEE Journal of Selected Areas in Communications, 18 (2000), pp. 1995–2003.

[26] J. Wang, W. Cho, V. Vemuri, and B. Mukherjee, *Improved approaches for cost-effective traffic grooming in WDM ring networks: ILP formulations and single-hop and multihop connections*, IEEE/OSA J. Lightwave Technol., 19 (2001), pp. 1645–1653.

[27] X. Yuan and A. Fulay, *Wavelength assignment to minimize the number of SONET ADMs in WDM rings*, in IEEE ICC, New York, 2002, pp. 2917–2921.

[28] X. Zhang and C. Qiao, *An effective and comprehensive approach for traffic grooming and wavelength assignment in SONET/WDM rings*, IEEE/ACM Transactions on Networking, 8 (2000), pp. 608–617.

SIAM J. Discrete Math.
Vol. 21, No. 4, pp. 865–899

IMPROVED ASYMPTOTIC BOUNDS FOR CODES USING DISTINGUISHED DIVISORS OF GLOBAL FUNCTION FIELDS[*]

HARALD NIEDERREITER[†] and FERRUH ÖZBUDAK[‡]

Abstract. For a prime power q, let α_q be the standard function in the asymptotic theory of codes, that is, $\alpha_q(\delta)$ is the largest asymptotic information rate that can be achieved for a given asymptotic relative minimum distance δ of q-ary codes. In recent years the Tsfasman–Vlăduţ–Zink lower bound on $\alpha_q(\delta)$ was improved by Elkies, Xing, Niederreiter and Özbudak, and Maharaj. In this paper we show further improvements on these bounds by using distinguished divisors of global function fields. We also show improved lower bounds on the corresponding function α_q^{lin} for linear codes.

Key words. asymptotic theory of codes, Gilbert–Varshamov bound, global function fields, Tsfasman–Vlăduţ–Zink bound, Xing bound

AMS subject classifications. Primary, 11T71, 94B27, 94B65; Secondary, 11R58, 14G50

DOI. 10.1137/060674478

1. Introduction.

Let $\mathbb{F}_q$ be the finite field of order q, where q is an arbitrary prime power. For a code C over $\mathbb{F}_q$ (or in other words a q-ary code), we denote by $n(C)$ its length and by $d(C)$ its minimum distance. We write $|M|$ for the cardinality of a finite set M.

For any prime power q, let α_q and α_q^{lin} denote the important functions in the asymptotic theory of codes which are defined by

$$\alpha_q(\delta) = \sup\{R \in [0,1] : (\delta, R) \in U_q\} \qquad \text{for } 0 \le \delta \le 1 \tag{1.1}$$

and

$$\alpha_q^{\mathrm{lin}}(\delta) = \sup\{R \in [0,1] : (\delta, R) \in U_q^{\mathrm{lin}}\} \qquad \text{for } 0 \le \delta \le 1. \tag{1.2}$$

Here U_q (resp., U_q^{lin}) is the set of all ordered pairs $(\delta, R) \in [0,1]^2$ for which there exists a sequence $\{C_i\}_{i=1}^{\infty}$ of not necessarily linear (resp., linear) codes over $\mathbb{F}_q$ such that $n(C_i) \to \infty$ as $i \to \infty$ and

$$\delta = \lim_{i\to\infty} \frac{d(C_i)}{n(C_i)}, \quad R = \lim_{i\to\infty} \frac{\log_q |C_i|}{n(C_i)},$$

where $\log_q$ is the logarithm to the base q. We refer to [10, section 1.3.1] for some basic properties of the functions α_q and α_q^{lin}. In particular, both functions are nonincreasing on the interval $[0,1]$. Furthermore, we have the known values $\alpha_q(0) = \alpha_q^{\mathrm{lin}}(0) = 1$

[*] Received by the editors November 9, 2006; accepted for publication (in revised form) May 17, 2007; published electronically November 16, 2007. This research was supported by the DSTA grant R-394-000-025-422 with Temasek Laboratories in Singapore. This paper was written while the second author was visiting Temasek Laboratories and Department of Mathematics at the National University of Singapore.

http://www.siam.org/journals/sidma/21-4/67447.html

[†] Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543, Republic of Singapore (nied@math.nus.edu.sg).

[‡] Temasek Laboratories, National University of Singapore, 5 Sports Drive 2, 117508 Singapore, Republic of Singapore and Department of Mathematics, Middle East Technical University, İnönü Bulvarı, 06531, Ankara, Turkey (ozbudak@metu.edu.tr).

and $\alpha_q(\delta) = \alpha_q^{\mathrm{lin}}(\delta) = 0$ for $(q-1)/q \le \delta \le 1$. It is trivial that $\alpha_q(\delta) \ge \alpha_q^{\mathrm{lin}}(\delta)$ for $0 \le \delta \le 1$.

A central problem in the asymptotic theory of codes is to find lower bounds on $\alpha_q(\delta)$ for $0 < \delta < (q-1)/q$. A classical lower bound is the asymptotic Gilbert–Varshamov bound, which says that

$$\alpha_q^{\mathrm{lin}}(\delta) \ge R_{\mathrm{GV}}(\delta) := 1 - \delta \log_q(q-1) + \delta \log_q \delta + (1-\delta)\log_q(1-\delta) \tag{1.3}$$

for $0 < \delta < (q-1)/q$. It is well known (see [6, section 6.2]) that for sufficiently large composite q and for certain ranges of the parameter δ, one can beat the asymptotic Gilbert–Varshamov bound by the Tsfasman–Vlăduţ–Zink bound [11]

$$\alpha_q^{\mathrm{lin}}(\delta) \ge 1 - \delta - \frac{1}{A(q)} \qquad \text{for } 0 \le \delta \le 1. \tag{1.4}$$

Here

$$A(q) := \limsup_{g\to\infty} \frac{N_q(g)}{g},$$

where $N_q(g)$ denotes the maximum number of rational places that a global function field of genus g with full constant field $\mathbb{F}_q$ can have. We recall from [6, Chapter 5] that $A(q) > 0$ for all q and that $A(q) = \sqrt{q} - 1$ if q is a square. For nonsquares q the exact value of $A(q)$ is not known, but we have lower and upper bounds on $A(q)$ (see again [6, Chapter 5]). We note, in particular, the recent bound in [1] which says that for any cube q we have

$$A(q) \ge \frac{2(q^{2/3}-1)}{q^{1/3}+2}. \tag{1.5}$$

The bound (1.4) for $\alpha_q^{\mathrm{lin}}(\delta)$ was improved, although not uniformly in δ, by Vlăduţ [12] (see also [10, Chapter 3.4]) and Xing [13]. Elkies [2] and Xing [14] considered not necessarily linear codes and Xing [14] improved the bound (1.4) for $\alpha_q(\delta)$ uniformly in δ. Shortly thereafter, Niederreiter and Özbudak [4, Corollary 5.4] improved the bound in Xing [14] by showing that

$$\alpha_q(\delta) \ge 1 - \delta - \frac{1}{A(q)} + \log_q\left(1 + \frac{1}{q^3}\right) \qquad \text{for } 0 \le \delta \le 1. \tag{1.6}$$

Later, Stichtenoth and Xing [8] gave a simpler proof of (1.6) and Maharaj [3] refined their approach.

Recently, Niederreiter and Özbudak [5] improved the bound (1.6) for certain values of q and δ. In this paper we extensively refine and complement the methods of [5] and we obtain further improvements on lower bounds for $\alpha_q(\delta)$ and $\alpha_q^{\mathrm{lin}}(\delta)$ for certain values of q and δ (see Theorem 6.3 and Corollary 6.4). In section 2 we present our basic code construction. We obtain the cardinality of an important auxiliary set in this construction in section 3. Asymptotic upper bounds on the cardinality of this set are given in sections 4 and 5. We present our main results in section 6. Section 7 is devoted to some examples demonstrating the improvements obtained by the main results. We conclude in section 8.

2. The basic code construction. In this section we present our basic construction of q-ary codes (see Theorem 2.9 and Corollary 2.10). First we give a short overview of the contents of this section. In Theorem 2.9, we construct a q-ary code as the image of a certain subset $N_{\mathbf{c}}$ (see (2.7)) of a suitable Riemann-Roch space of a global function field under an $\mathbb{F}_q$-linear map ψ (see (2.3)). In Corollary 2.10, we show that such a q-ary code is linear over $\mathbb{F}_q$ in a special case. The existence of a suitable Riemann-Roch space for Theorem 2.9 is guaranteed if the condition in (2.5) holds. An appropriate existence result for distinguished divisors of global function fields is proved in Proposition 2.4. The set in (2.5) is important and is introduced in a more general form in Definition 2.3. We also introduce further notation in Definitions 2.1, 2.2, and 2.5, and we prove related results in Lemmas 2.6, 2.7, and 2.8, which are used in the proof of Theorem 2.9.

Next we recall some definitions and explain the terminology that we use throughout the paper. A *global function field* F over $\mathbb{F}_q$ is an extension field of $\mathbb{F}_q$ such that there exists an element $z \in F$ that is transcendental over $\mathbb{F}_q$ and for which F is a finite extension of the rational function field $\mathbb{F}_q(z)$. Moreover, we call $\mathbb{F}_q$ the *full constant field* of F if $\mathbb{F}_q$ is algebraically closed in F. A *place* of F is the maximal ideal of some valuation ring of F. Let $\mathbb{Z}$ denote the set of integers. A *normalized discrete valuation* of F is a surjective map $\nu : F \to \mathbb{Z} \cup \{\infty\}$ satisfying the following:

(i) $\nu(x) = \infty \iff x = 0$;
(ii) $\nu(xy) = \nu(x) + \nu(y)$ for all $x, y \in F$;
(iii) $\nu(x + y) \geq \min(\nu(x), \nu(y))$ for all $x, y \in F$;
(iv) $\nu(a) = 0$ for all $a \in \mathbb{F}_q \setminus \{0\}$.

There is a bijective correspondence between the places of F and the normalized discrete valuations of F. Let v_P be the normalized discrete valuation of F corresponding to the place P of F. The valuation ring of P is

$$\mathcal{O}_P = \{x \in F : v_P(x) \geq 0\}$$

and the maximal ideal of $\mathcal{O}_P$ is

$$M_P = \{x \in \mathcal{O}_P : v_P(x) > 0\}.$$

If $\mathbb{F}_q$ is the full constant field of F, then the residue class field $\mathcal{O}_P/M_P$ can be identified with a finite extension of $\mathbb{F}_q$. The degree of this extension is called the *degree* of the place P. A place of degree 1 is called *rational*. For detailed background on global function fields we refer to the book of Stichtenoth [7].

From now on we assume that F is a global function field with full constant field $\mathbb{F}_q$ and with at least one rational place. Let $n \geq 1$ be the number of rational places of F and let $P_1, \ldots, P_n$ be all rational places of F. Let h be the class number of F (see, for example, [7, section V.1]). Let $\mathbb{P}_F$ be the set of all places of F. For $f \in F \setminus \{0\}$,

$$(f) = \sum_{P \in \mathbb{P}_F} v_P(f) P$$

denotes the principal divisor of f and

$$(f)_0 = \sum_{\substack{P \in \mathbb{P}_F \\ v_P(f) \geq 1}} v_P(f) P$$

denotes the zero divisor of f. For an arbitrary divisor

$$G = \sum_{P \in \mathbb{P}_F} m_P P$$

of F, we write $v_P(G)$ for the coefficient m_P of P. We use the standard notation

$$\mathcal{L}(G) = \{f \in F : v_P(f) \geq -v_P(G) \text{ for all } P \in \mathbb{P}_F\}$$

for the Riemann–Roch space of G. In this section and in section 3, all places and divisors are from the given global function field F. We fix an integer $m \geq 1$.

Definition 2.1. *For a positive divisor D, let $\overline{D}$ be the divisor*

$$\overline{D} = a_1 P_1 + \cdots + a_n P_n,$$

where $a_i = \min(m+1, v_{P_i}(D))$ for $1 \leq i \leq n$.

Definition 2.2. *For a positive divisor D, let*

$$\begin{aligned} \textsc{j}_0(D) &= \left|\{i \in \{1,\dots,n\} : v_{P_i}(D) = m\}\right|, \\ \textsc{j}_1(D) &= \left|\{i \in \{1,\dots,n\} : v_{P_i}(D) = m-1\}\right|, \\ &\vdots \\ \textsc{j}_m(D) &= \left|\{i \in \{1,\dots,n\} : v_{P_i}(D) = 0\}\right|. \end{aligned}$$

Moreover, we define

$$J_m(D) = 2\textsc{j}_1(D) + 3\textsc{j}_2(D) + \cdots + (m+1)\textsc{j}_m(D). \tag{2.1}$$

Definition 2.3. *For integers $r \geq s \geq 0$ and nonnegative integers $X_1, X_2, \dots, X_m$, let $\mathcal{V}_m(r, s; X_1, X_2, \dots, X_m)$ be the set consisting of the positive divisors D of the global function field F satisfying all of the following conditions:*

- *C1: $\deg(D) = r$ and $\deg\left(\overline{D}\right) \geq s$.*
- *C2:*

$$\begin{aligned} \textsc{j}_m(D) &\leq 2X_m, \\ \textsc{j}_{m-1}(D) &\leq 2X_{m-1} + X_m, \\ \textsc{j}_{m-2}(D) &\leq 2X_{m-2} + (X_{m-1} + X_m), \\ &\vdots \\ \textsc{j}_1(D) &\leq 2X_1 + (X_2 + X_3 + \cdots + X_m). \end{aligned}$$

- *C3: $J_m(D) \leq 2\left(2X_1 + 3X_2 + \cdots + (m+1)X_m\right)$.*

Proposition 2.4. *For integers $r \geq s \geq 0$ and nonnegative integers $X_1, \dots, X_m$, if*

$$|\mathcal{V}_m(r, s; X_1, \dots, X_m)| < h,$$

then there exists a divisor G of degree r such that $\operatorname{supp}(G) \cap \{P_1, \dots, P_n\} = \emptyset$ and for each $f \in \mathcal{L}(G) \setminus \{0\}$, if $E = (f)_0$ satisfies conditions C2 and C3 of Definition 2.3 with the given $X_1, \dots, X_m$, then $\deg\left(\overline{E}\right) \leq s-1$.

Proof. As $|\mathcal{V}_m(r, s; X_1, \dots, X_m)| < h$, there exists a degree r divisor G such that $G \not\sim V$ for any $V \in \mathcal{V}_m(r, s; X_1, \dots, X_m)$. Using the weak approximation theorem [7, Theorem I.3.1], we can assume that $\operatorname{supp}(G) \cap \{P_1, \dots, P_n\} = \emptyset$ without loss of generality (compare with [5, Proof of Corollary 2.2]). Let $f \in \mathcal{L}(G) \setminus \{0\}$, $D = G + (f)$, and $E = (f)_0$. Since $\operatorname{supp}(G) \cap \{P_1, \dots, P_n\} = \emptyset$ and D is positive, we have $\overline{D} = \overline{E}$.

Assume that conditions C2 and C3 of Definition 2.3 are satisfied by E. If $\deg\left(\overline{E}\right) \geq s$, then $D \in \mathcal{V}_m(r, s; X_1, \ldots, X_m)$ and hence $D \not\sim G$, which is a contradiction. Thus, we must have $\deg\left(\overline{E}\right) \leq s-1$. $\square$

Now give another definition related to our construction.

Definition 2.5. *For* $\boldsymbol{\alpha} = (\alpha_1^{(1)}, \ldots, \alpha_m^{(1)}, \alpha_1^{(2)}, \ldots, \alpha_m^{(2)}, \ldots\ldots, \alpha_1^{(n)}, \ldots, \alpha_m^{(n)}) \in \mathbb{F}_q^{mn}$, *let* $I_m(\boldsymbol{\alpha}), I_{m-1}(\boldsymbol{\alpha}), \ldots, I_1(\boldsymbol{\alpha})$ *be the subsets of* $\{1, \ldots, n\}$ *defined by*

$$\begin{aligned} I_m(\boldsymbol{\alpha}) &= \left\{ i \in \{1, \ldots, n\} : \alpha_m^{(i)} \neq 0 \right\}, \\ I_{m-1}(\boldsymbol{\alpha}) &= \left\{ i \in \{1, \ldots, n\} : \alpha_m^{(i)} = 0,\ \alpha_{m-1}^{(i)} \neq 0 \right\}, \\ &\vdots \\ I_1(\boldsymbol{\alpha}) &= \left\{ i \in \{1, \ldots, n\} : \alpha_m^{(i)} = \cdots = \alpha_2^{(i)} = 0,\ \alpha_1^{(i)} \neq 0 \right\}. \end{aligned}$$

The following two lemmas are related to Definition 2.5 and important for our construction.

Lemma 2.6. *For* $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{F}_q^{mn}$, *we have*

$$\begin{aligned} &2\,|I_1(\boldsymbol{\alpha}-\boldsymbol{\beta})| + 3\,|I_2(\boldsymbol{\alpha}-\boldsymbol{\beta})| + \cdots + (m+1)\,|I_m(\boldsymbol{\alpha}-\boldsymbol{\beta})| \\ &\quad \leq 2\,|I_1(\boldsymbol{\alpha})| + 3\,|I_2(\boldsymbol{\alpha})| + \cdots + (m+1)\,|I_m(\boldsymbol{\alpha})| \\ &\qquad + 2\,|I_1(\boldsymbol{\beta})| + 3\,|I_2(\boldsymbol{\beta})| + \cdots + (m+1)\,|I_m(\boldsymbol{\beta})|. \end{aligned}$$

Proof. Let $\boldsymbol{\alpha} = \left(\alpha_1^{(1)}, \ldots, \alpha_m^{(1)}, \ldots\ldots, \alpha_1^{(n)}, \ldots, \alpha_m^{(n)}\right)$ and $\boldsymbol{\beta} = \left(\beta_1^{(1)}, \ldots, \beta_m^{(1)}, \ldots \ldots, \beta_1^{(n)}, \ldots, \beta_m^{(n)}\right)$. Let $A \subseteq \{1, \ldots, n\}$ be the set consisting of the $i \in \{1, \ldots, n\}$ such that $(\alpha_1^{(i)}, \ldots, \alpha_m^{(i)}) \neq \mathbf{0}$ or $(\beta_1^{(i)}, \ldots, \beta_m^{(i)}) \neq \mathbf{0}$. If $A = \emptyset$, then $\boldsymbol{\alpha} = \boldsymbol{\beta} = \boldsymbol{\alpha} - \boldsymbol{\beta} = \mathbf{0}$ and the result follows immediately. If $A \neq \emptyset$, then for each $i \in A$, let $1 \leq \ell_i \leq m$ be the largest integer such that $\alpha_{\ell_i}^{(i)} \neq 0$ or $\beta_{\ell_i}^{(i)} \neq 0$. For each $i \in A$, we have

$$i \notin \bigcup_{\ell_i < j \leq m} I_j(\boldsymbol{\alpha} - \boldsymbol{\beta}),$$

and also $i \in I_{\ell_i}(\boldsymbol{\alpha})$ or $i \in I_{\ell_i}(\boldsymbol{\beta})$. Hence for each $i \in A$ we obtain

$$\begin{aligned} &2\,|\{i\} \cap I_1(\boldsymbol{\alpha}-\boldsymbol{\beta})| + 3\,|\{i\} \cap I_2(\boldsymbol{\alpha}-\boldsymbol{\beta})| + \cdots + (m+1)\,|\{i\} \cap I_m(\boldsymbol{\alpha}-\boldsymbol{\beta})| \\ &\quad \leq 2\,|\{i\} \cap I_1(\boldsymbol{\alpha})| + 3\,|\{i\} \cap I_2(\boldsymbol{\alpha})| + \cdots + (m+1)\,|\{i\} \cap I_m(\boldsymbol{\alpha})| \\ &\qquad + 2\,|\{i\} \cap I_1(\boldsymbol{\beta})| + 3\,|\{i\} \cap I_2(\boldsymbol{\beta})| + \cdots + (m+1)\,|\{i\} \cap I_m(\boldsymbol{\beta})|. \end{aligned}$$

We complete the proof by summing over all $i \in A$. $\square$

Lemma 2.7. *For* $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{F}_q^{mn}$, *we have the following containment relations:*

$$\begin{aligned} I_m(\boldsymbol{\alpha}-\boldsymbol{\beta}) &\subseteq I_m(\boldsymbol{\alpha}) \cup I_m(\boldsymbol{\beta}), \\ I_{m-1}(\boldsymbol{\alpha}-\boldsymbol{\beta}) &\subseteq I_{m-1}(\boldsymbol{\alpha}) \cup I_{m-1}(\boldsymbol{\beta}) \cup \{I_m(\boldsymbol{\alpha}) \cap I_m(\boldsymbol{\beta})\}, \\ I_{m-2}(\boldsymbol{\alpha}-\boldsymbol{\beta}) &\subseteq I_{m-2}(\boldsymbol{\alpha}) \cup I_{m-2}(\boldsymbol{\beta}) \cup \{I_{m-1}(\boldsymbol{\alpha}) \cap I_{m-1}(\boldsymbol{\beta})\} \cup \{I_m(\boldsymbol{\alpha}) \cap I_m(\boldsymbol{\beta})\}, \\ &\vdots \\ I_1(\boldsymbol{\alpha}-\boldsymbol{\beta}) &\subseteq I_1(\boldsymbol{\alpha}) \cup I_1(\boldsymbol{\beta}) \cup \bigcup_{2 \leq \nu \leq m} \{I_\nu(\boldsymbol{\alpha}) \cap I_\nu(\boldsymbol{\beta})\}. \end{aligned}$$

Proof. First we consider the case of the subscript m and assume that $i \in I_m(\boldsymbol{\alpha} - \boldsymbol{\beta})$. Then $\alpha_m^{(i)} \neq \beta_m^{(i)}$ and at least one of $\alpha_m^{(i)}$ and $\beta_m^{(i)}$ is nonzero. Hence $i \in I_m(\boldsymbol{\alpha}) \cup I_m(\boldsymbol{\beta})$.

Next we consider the case of the subscript $m-1$ and assume that $i \in I_{m-1}(\boldsymbol{\alpha}-\boldsymbol{\beta})$. We have $\alpha_m^{(i)} = \beta_m^{(i)}$ and $\alpha_{m-1}^{(i)} \neq \beta_{m-1}^{(i)}$. If $\alpha_m^{(i)} = \beta_m^{(i)} \neq 0$, then $i \in I_m(\boldsymbol{\alpha}) \cap I_m(\boldsymbol{\beta})$. If $\alpha_m^{(i)} = \beta_m^{(i)} = 0$, then since at least one of $\alpha_{m-1}^{(i)}$ and $\beta_{m-1}^{(i)}$ is nonzero, we get $i \in I_{m-1}(\boldsymbol{\alpha}) \cup I_{m-1}(\boldsymbol{\beta})$.

Now we consider the case of the subscript $m-2$. Assume that $i \in I_{m-2}(\boldsymbol{\alpha}-\boldsymbol{\beta})$. Then $\alpha_m^{(i)} = \beta_m^{(i)}$, $\alpha_{m-1}^{(i)} = \beta_{m-1}^{(i)}$, and $\alpha_{m-2}^{(i)} \neq \beta_{m-2}^{(i)}$. If $\alpha_m^{(i)} = \beta_m^{(i)} \neq 0$, then $i \in I_m(\boldsymbol{\alpha}) \cap I_m(\boldsymbol{\beta})$. If $\alpha_m^{(i)} = \beta_m^{(i)} = 0$ and $\alpha_{m-1}^{(i)} = \beta_{m-1}^{(i)} \neq 0$, then $i \in I_{m-1}(\boldsymbol{\alpha}) \cap I_{m-1}(\boldsymbol{\beta})$. Finally, if $\alpha_m^{(i)} = \beta_m^{(i)} = 0$ and $\alpha_{m-1}^{(i)} = \beta_{m-1}^{(i)} = 0$, then since $\alpha_{m-2}^{(i)}$ and $\beta_{m-2}^{(i)}$ are distinct, we get $i \in I_{m-2}(\boldsymbol{\alpha})$ or $i \in I_{m-2}(\boldsymbol{\beta})$. We complete the proof similarly for each subscript $1 \leq \nu \leq m$. □

For each $i = 1, \dots, n$, let t_i be a local parameter of F at P_i. Assume that G is a divisor with $\operatorname{supp}(G) \cap \{P_1, \dots, P_n\} = \emptyset$ and $\dim(\mathcal{L}(G)) \geq 1$. For f in the Riemann–Roch space $\mathcal{L}(G)$, the local expansion of f at P_i has the form

$$f = \sum_{l=0}^{\infty} f^{(l)}(P_i) t_i^l$$

with $f^{(l)}(P_i) \in \mathbb{F}_q$ for $1 \leq i \leq n$ and $l \geq 0$. For each $i = 1, \dots, n$, let

$$\begin{aligned} \phi_i : \mathcal{L}(G) &\to \mathbb{F}_q^m \\ f &\mapsto \left(f^{(m-1)}(P_i), \dots, f^{(1)}(P_i), f^{(0)}(P_i) \right). \end{aligned}$$

Let $\boldsymbol{\Phi}$ be the $\mathbb{F}_q$-linear map defined by

$$\begin{aligned} \Phi : \mathcal{L}(G) &\to \mathbb{F}_q^{mn} \\ f &\mapsto (\phi_1(f), \dots, \phi_n(f)). \end{aligned} \tag{2.2}$$

Moreover, let ψ be the $\mathbb{F}_q$-linear map

$$\begin{aligned} \psi : \mathcal{L}(G) &\to \mathbb{F}_q^n \\ f &\mapsto \left(f^{(m)}(P_1), \dots, f^{(m)}(P_n) \right). \end{aligned} \tag{2.3}$$

LEMMA 2.8. *For a divisor G with* $\operatorname{supp}(G) \cap \{P_1, \dots, P_n\} = \emptyset$ *and* $\dim(\mathcal{L}(G)) \geq 1$, *let $f \in \mathcal{L}(G) \setminus \{0\}$. Moreover, let $E = (f)_0$ be the zero divisor of f and $\boldsymbol{\alpha} := \boldsymbol{\Phi}(f) \in \mathbb{F}_q^{mn}$. Then*

$$\mathrm{J}_1(E) = |I_1(\boldsymbol{\alpha})|,\ \mathrm{J}_2(E) = |I_2(\boldsymbol{\alpha})|, \dots, \mathrm{J}_m(E) = |I_m(\boldsymbol{\alpha})|,$$

and

$$J_m(E) = 2\,|I_1(\boldsymbol{\alpha})| + 3\,|I_2(\boldsymbol{\alpha})| + \cdots + (m+1)\,|I_m(\boldsymbol{\alpha})|.$$

Proof. For each $1 \leq i \leq n$ and $1 \leq \ell \leq m$, using Definition 2.5 we obtain that $i \in I_\ell(\boldsymbol{\alpha}) \iff v_{P_i}(E) = m - \ell$. Hence by Definition 2.2 we have

$$\mathrm{J}_m(E) = |I_m(\boldsymbol{\alpha})|,\ \mathrm{J}_{m-1}(E) = |I_{m-1}(\boldsymbol{\alpha})|, \dots, \mathrm{J}_1(E) = |I_1(\boldsymbol{\alpha})|.$$

Using (2.1) we complete the proof. □

For $\boldsymbol{c} \in \mathbb{F}_q^{mn}$ and nonnegative real numbers $x_1, \dots, x_m$ with $x_1 + \cdots + x_m \leq 1$, let $M(x_1, \dots, x_m; \boldsymbol{c})$ be the subset of $\mathbb{F}_q^{mn}$ defined by

$$M(x_1, \dots, x_m; \boldsymbol{c}) = \left\{ \boldsymbol{\alpha} \in \mathbb{F}_q^{mn} : |I_1(\boldsymbol{\alpha} - \boldsymbol{c})| \leq \lfloor x_1 n \rfloor, \dots, |I_m(\boldsymbol{\alpha} - \boldsymbol{c})| \leq \lfloor x_m n \rfloor \right\}.$$

We have

$$\begin{aligned}
&|M(x_1, \dots, x_m; \boldsymbol{c})| = |M(x_1, \dots, x_m; \mathbf{0})| \\
&\geq |\{\boldsymbol{\alpha} \in \mathbb{F}_{q^{mn}} : |I_1(\boldsymbol{\alpha})| = \lfloor x_1 n \rfloor, \dots, |I_m(\boldsymbol{\alpha})| = \lfloor x_m n \rfloor\}| \\
&= \binom{n}{\lfloor x_m n \rfloor} (q-1)^{\lfloor x_m n \rfloor} q^{(m-1)\lfloor x_m n \rfloor} \\
&\quad \times \binom{n - \lfloor x_m n \rfloor}{\lfloor x_{m-1} n \rfloor} (q-1)^{\lfloor x_{m-1} n \rfloor} q^{(m-2)\lfloor x_{m-1} n \rfloor} \\
&\quad \times \cdots \times \binom{n - (\lfloor x_m n \rfloor + \lfloor x_{m-1} n \rfloor + \cdots + \lfloor x_2 n \rfloor)}{\lfloor x_1 n \rfloor} (q-1)^{\lfloor x_1 n \rfloor}.
\end{aligned} \tag{2.4}$$

Now we are ready to give our basic code construction. Assume that $r \geq s \geq 0$ are integers and $x_1, \dots, x_m \geq 0$ are real numbers such that

$$|\mathcal{V}_m(r, s; \lfloor x_1 n \rfloor, \lfloor x_2 n \rfloor, \dots, \lfloor x_m n \rfloor)| < h. \tag{2.5}$$

Let G be a divisor of degree r obtained using (2.5) and Proposition 2.4. Recall the linear maps $\mathbf{\Phi}$ and ψ defined in (2.2) and (2.3), respectively, using the chosen divisor G. The map $\mathbf{\Phi}$ is not necessarily surjective. If

$$|\mathcal{L}(G)| \cdot |M(x_1, \dots, x_m; \mathbf{0})| > q^{mn}, \tag{2.6}$$

then there exists $\boldsymbol{c} \in \mathbb{F}_q^{mn}$ such that for the set

$$N_{\boldsymbol{c}} := \{ f \in \mathcal{L}(G) : \mathbf{\Phi}(f) \in M(x_1, \dots, x_m; \boldsymbol{c}) \} \tag{2.7}$$

we have

$$|N_{\boldsymbol{c}}| \geq \frac{|\mathcal{L}(G)| \cdot |M(x_1, \dots, x_m; \mathbf{0})|}{q^{mn}} > 1. \tag{2.8}$$

THEOREM 2.9. *Assume that $r \geq s \geq 0$ are integers and that $x_1, \dots, x_m$ are nonnegative real numbers with $x_1 + \cdots + x_m \leq 1$ satisfying* (2.5)*. Let G be a divisor of degree r obtained using* (2.5) *and Proposition* 2.4*. Assume also that* (2.6) *holds and that*

$$(m+1)n \geq s + 2 \sum_{l=1}^{m} (l+1) \lfloor x_l n \rfloor. \tag{2.9}$$

Using the chosen divisor G and (2.6)*, let $\boldsymbol{c} \in \mathbb{F}_q^{mn}$ be such that the set $N_{\boldsymbol{c}}$ satisfies* (2.8)*. Let C be the q-ary code of length n given by $C = \psi(N_{\boldsymbol{c}})$. Then for the cardinality $|C|$ of C we have*

$$|C| \geq \left\lceil \frac{\mathcal{L}(G) \cdot |M(x_1, \dots, x_m; \mathbf{0})|}{q^{mn}} \right\rceil$$

and for the minimum distance $d(C)$ of C we have

$$d(C) \geq (m+1)n + 1 - s - 2\sum_{l=1}^{m}(l+1)\lfloor x_l n \rfloor.$$

Proof. Let $f_1, f_2 \in N_{\boldsymbol{c}}$ be such that $f_1 \neq f_2$ and put $f = f_1 - f_2 \in \mathcal{L}(G)$. Let E be the zero divisor of f and

$$\overline{E} = a_1 P_1 + \cdots + a_n P_n$$

be the divisor defined in Definition 2.1. Let $\boldsymbol{\Phi}(f_1) = \boldsymbol{\alpha}$ and $\boldsymbol{\Phi}(f_2) = \boldsymbol{\beta}$. We have

$$\boldsymbol{\Phi}(f) = \boldsymbol{\alpha} - \boldsymbol{\beta}. \tag{2.10}$$

As $\boldsymbol{\alpha}, \boldsymbol{\beta} \in M(x_1, \ldots, x_m; \boldsymbol{c})$, we also have

$$|I_i(\boldsymbol{\alpha} - \boldsymbol{c})| \leq \lfloor x_i n \rfloor \quad \text{and} \quad |I_i(\boldsymbol{\beta} - \boldsymbol{c})| \leq \lfloor x_i n \rfloor \quad \text{for } 1 \leq i \leq n. \tag{2.11}$$

Using (2.10), (2.11), and Lemmas 2.8 and 2.6, we obtain that

$$\begin{aligned}
J_m(E) &= 2\,|I_1(\boldsymbol{\alpha} - \boldsymbol{\beta})| + 3\,|I_2(\boldsymbol{\alpha} - \boldsymbol{\beta})| + \cdots + (m+1)\,|I_m(\boldsymbol{\alpha} - \boldsymbol{\beta})| \\
&\leq 2\,|I_1(\boldsymbol{\alpha} - \boldsymbol{c})| + 3\,|I_2(\boldsymbol{\alpha} - \boldsymbol{c})| + \cdots + (m+1)\,|I_m(\boldsymbol{\alpha} - \boldsymbol{c})| \\
&\quad + 2\,|I_1(\boldsymbol{\beta} - \boldsymbol{c})| + 3\,|I_2(\boldsymbol{\beta} - \boldsymbol{c})| + \cdots + (m+1)\,|I_m(\boldsymbol{\beta} - \boldsymbol{c})| \\
&\leq 2\left(2\lfloor x_1 n \rfloor + 3\lfloor x_2 n \rfloor + \cdots + (m+1)\lfloor x_m n \rfloor\right).
\end{aligned}$$

Moreover, using (2.10), (2.11), and Lemmas 2.8 and 2.7, we further obtain that

$$\begin{aligned}
\mathrm{J}_m(E) &= |I_m((\boldsymbol{\alpha} - \boldsymbol{c}) - (\boldsymbol{\beta} - \boldsymbol{c}))| \leq |I_m(\boldsymbol{\alpha} - \boldsymbol{c})| + |I_m(\boldsymbol{\beta} - \boldsymbol{c})| \leq 2\lfloor x_m n \rfloor, \\
\mathrm{J}_{m-1}(E) &= |I_{m-1}((\boldsymbol{\alpha} - \boldsymbol{c}) - (\boldsymbol{\beta} - \boldsymbol{c}))| \\
&\leq |I_{m-1}(\boldsymbol{\alpha} - \boldsymbol{c})| + |I_{m-1}(\boldsymbol{\beta} - \boldsymbol{c})| + |I_m(\boldsymbol{\alpha} - \boldsymbol{c}) \cap I_m(\boldsymbol{\beta} - \boldsymbol{c})| \\
&\leq 2\lfloor x_{m-1} n \rfloor + \lfloor x_m n \rfloor, \\
&\vdots \\
\mathrm{J}_1(E) &= |I_1((\boldsymbol{\alpha} - \boldsymbol{c}) - (\boldsymbol{\beta} - \boldsymbol{c}))| \\
&\leq |I_1(\boldsymbol{\alpha} - \boldsymbol{c})| + |I_1(\boldsymbol{\beta} - \boldsymbol{c})| + \sum_{\nu=2}^{m} |I_\nu(\boldsymbol{\alpha} - \boldsymbol{c}) \cap I_\nu(\boldsymbol{\beta} - \boldsymbol{c})| \\
&\leq 2\lfloor x_1 n \rfloor + \sum_{\nu=2}^{m} \lfloor x_\nu n \rfloor.
\end{aligned}$$

Hence by the choice of the divisor G (cf. Proposition 2.4), we have

$$\deg\left(\overline{E}\right) \leq s - 1. \tag{2.12}$$

Moreover, we obtain

$$\sum_{i=1}^{n}(m+1-a_i) = (m+1)n - \sum_{i=1}^{n} a_i = (m+1)n - \deg\left(\overline{E}\right) \geq (m+1)n - s + 1,$$

where we used (2.12). Let $||\psi(f)||$ denote the Hamming weight of the vector $\psi(f) \in \mathbb{F}_q^n$. Then using Definition 2.2 and (2.1), we have

$$\begin{aligned}\sum_{i=1}^{n}(m+1-a_i) = \sum_{\substack{i=1\\ 0\le a_i\le m}}^{n}(m+1-a_i) &\le ||\psi(f)|| + \sum_{\substack{i=1\\ 0\le a_i\le m-1}}^{n}(m+1-a_i)\\ &= ||\psi(f)|| + J_m(E).\end{aligned}$$

Therefore we obtain

$$\begin{aligned}\|\psi(f)\| &\ge (m+1)n - s + 1 - J_m(E)\\ &\ge (m+1)n - s + 1 - 2\left(2\lfloor x_1 n\rfloor + 3\lfloor x_2 n\rfloor + \cdots + (m+1)\lfloor x_m n\rfloor\right).\end{aligned}$$

Using (2.9) we obtain that $d(C) \ge 1$, and so the map ψ is one-to-one on $N_{\boldsymbol{c}}$. Therefore $|C| = |N_{\boldsymbol{c}}|$, and hence the lower bound on $|C|$ follows from (2.8). This completes the proof. □

In a special case related to Theorem 2.9, we make sure to construct linear codes. Later in this paper, the following result will be used to obtain lower bounds on the function $\alpha_q^{\mathrm{lin}}(\delta)$, which is defined in (1.2).

COROLLARY 2.10. *Assume that $r \ge s \ge 0$ are integers and that $x_1 = x_2 = \cdots = x_m = 0$ satisfy* (2.5)*. Let G be a divisor of degree r obtained using* (2.5) *and Proposition* 2.4*. Assume also that*

$$|\mathcal{L}(G)| > q^{mn} \tag{2.13}$$

and that $(m+1)n \ge s$. Using the chosen divisor G and the kernel of the corresponding map $\mathbf{\Phi}$, put $C = \psi\,(\mathrm{Ker}\ \mathbf{\Phi})$. Then C is a linear code over $\mathbb{F}_q$ of length n. Moreover, for the dimension of C we have

$$\dim(C) \ge \dim\left(\mathcal{L}(G)\right) - mn$$

and for the minimum distance $d(C)$ of C we have

$$d(C) \ge (m+1)n + 1 - s.$$

Proof. The kernel of $\mathbf{\Phi}$ is an $\mathbb{F}_q$-linear subspace of $\mathcal{L}(G)$ and is the Riemann–Roch space given by

$$\mathrm{Ker}\ \mathbf{\Phi} = \mathcal{L}\left(G - m(P_1 + \cdots + P_n)\right).$$

As $\dim\left(\mathcal{L}\left(G - m(P_1 + \cdots + P_n)\right)\right) \ge \dim\left(\mathcal{L}(G)\right) - mn$, using (2.13) we obtain that Ker $\mathbf{\Phi} \ne \{0\}$. The maps $\mathbf{\Phi}$ and ψ are $\mathbb{F}_q$-linear, and hence C is a linear code over $\mathbb{F}_q$. We obtain the bounds on the dimension and the minimum distance of C using similar methods as in the proof of Theorem 2.9. □

Remark 2.11. For $x_1 = x_2 = \cdots = x_m = 0$, the conditions (2.6) and (2.13) are equivalent.

3. The cardinality of $\boldsymbol{\mathcal{V}_m(r, s; X_1, \ldots, X_m)}$. In this section we will compute the cardinality of the set $\mathcal{V}_m(r, s; X_1, \ldots, X_m)$ for integers $r \ge s \ge 0$ and nonnegative integers $X_1, \ldots, X_m$ (see Definition 2.3 for the definition of this set). We introduce a related set $\mathcal{U}(r, t; j_1, \ldots, j_m)$ in Definition 3.2. Using Lemmas 3.1 and 3.3 and Definition 3.4, we compute the cardinality of the set $\mathcal{U}(r, t; j_1, \ldots, j_m)$ in Lemma 3.5. Then the cardinality of $\mathcal{V}_m(r, s; X_1, \ldots, X_m)$ follows from (3.4) and Lemma 3.5.

The notation we introduced in section 2 remains operative.

LEMMA 3.1. *For any positive divisor D, we have*

$$\deg\left(\overline{D}\right) + \mathrm{j}_0(D) + 2\mathrm{j}_1(D) + \cdots + (m+1)\mathrm{j}_m(D) = (m+1)n.$$

Proof. For $0 \le \ell \le m$, let $S_\ell = \left\{P \in \{P_1, \dots, P_n\} : v_P(\overline{D}) = m - \ell\right\}$. Note that $|S_\ell| = \mathrm{j}_\ell(D)$ for each $0 \le \ell \le m$. We have

$$\sum_{P \in \{P_1, \dots, P_n\}} (m + 1 - v_P(\overline{D})) = (m+1)n - \deg\left(\overline{D}\right)$$

and also

$$\begin{aligned}\sum_{P \in \{P_1, \dots, P_n\}} \left(m + 1 - v_P(\overline{D})\right) &= \sum_{\ell=0}^{m} \sum_{P \in S_\ell} \left(m + 1 - v_P(\overline{D})\right) \\ &= \sum_{\ell=0}^{m} \sum_{P \in S_\ell} (\ell + 1) = \sum_{\ell=0}^{m} (\ell+1)\mathrm{j}_\ell(D).\end{aligned}$$

This completes the proof. □

DEFINITION 3.2. *For integers $r \ge t \ge 0$ and $j_1, \dots, j_m \ge 0$, let $\mathcal{U}(r, t; j_1, \dots, j_m)$ be the set of positive divisors given by*

$$\mathcal{U}(r, t; j_1, \dots, j_m) = \left\{D \ge 0 : \deg(D) = r,\ \deg\left(\overline{D}\right) = t,\ \mathrm{j}_1(D) = j_1, \dots, \mathrm{j}_m(D) = j_m\right\}.$$

LEMMA 3.3. *For integers $r \ge t \ge 0$ and $j_1, \dots, j_m \ge 0$, the set $\mathcal{U}(r, t; j_1, \dots, j_m)$ is not empty if and only if*

$$mn - (j_1 + 2j_2 + \cdots + mj_m) \le t \le (m+1)n - (2j_1 + 3j_2 + \cdots + (m+1)j_m)$$

holds and also provided that there exists a degree $r - t$ positive divisor whose support is disjoint from the set $\{P_1, \dots, P_n\}$ when $mn = t + j_1 + 2j_2 + \cdots + mj_m$ and $r > t$.

Proof. Let $D \in \mathcal{U}(r, t; j_1, \dots, j_m)$. Using Lemma 3.1 we have

$$\mathrm{j}_0(D) = (m+1)n - (2\mathrm{j}_1(D) + \cdots + (m+1)\mathrm{j}_m(D)) - t, \tag{3.1}$$

and so, in particular,

$$t \le (m+1)n - (2j_1 + 3j_2 + \cdots + (m+1)j_m).$$

Moreover, by the definition of $\overline{D}$,

$$\begin{aligned} t &\ge \mathrm{j}_{m-1}(D) + 2\mathrm{j}_{m-2}(D) + \cdots + m\mathrm{j}_0(D) \\ &= \mathrm{j}_{m-1}(D) + 2\mathrm{j}_{m-2}(D) + \cdots + (m-1)\mathrm{j}_1(D) \\ &\quad + m(m+1)n - (2m\mathrm{j}_1(D) + \cdots + (m+1)m\mathrm{j}_m(D)) - mt, \end{aligned}$$

where we used (3.1) in the second step. Therefore

$$\begin{aligned}(m+1)t &\ge (m+1)mn \\ &\quad - \big((m+1)m\mathrm{j}_m(D) + (m^2-1)\mathrm{j}_{m-1}(D) + ((m-1)m-2)\,\mathrm{j}_{m-2}(D) \\ &\qquad + \cdots + (2m-(m-1))\,\mathrm{j}_1(D)\big) \\ &= (m+1)mn - (m+1)\left(m\mathrm{j}_m(D) + (m-1)\mathrm{j}_{m-1}(D) + \cdots + \mathrm{j}_1(D)\right),\end{aligned}$$

which means that

$$t \geq mn - (j_1 + 2j_2 + \cdots + mj_m)\,. \tag{3.2}$$

Also, if this is an equality, then the set $\{P \in \{P_1, \ldots, P_n\} : v_P(D) \geq m+1\}$ is empty. Therefore, if the equality in (3.2) holds and $r > t$, then there exists a positive divisor of degree $r - t$ whose support is disjoint from $\{P_1, \ldots, P_n\}$.

Now we prove the converse. Let $S_m = \{1, \ldots, j_m\}$, $S_{m-1} = \{j_m + 1, \ldots, j_m + j_{m-1}\}, \ldots,$and $S_1 = \{(j_m + \cdots + j_2) + 1, \ldots, (j_m + \cdots + j_2) + j_1\}$, where we put $S_\ell = \emptyset$ whenever $j_\ell = 0$ for some $1 \leq \ell \leq m$. They are pairwise disjoint sets of natural numbers. We note that for each $1 \leq \ell \leq m$, we have $|S_\ell| = j_\ell$. Comparing both sides of the inequalities for t given in the statement of the lemma, we obtain that

$$j_1 + j_2 + \cdots + j_m \leq n.$$

Let

$$j_0 = (m+1)n - (2j_1 + 3j_2 + \cdots + (m+1)j_m) - t. \tag{3.3}$$

Using the upper bound on t in the statement of the lemma, we get $j_0 \geq 0$. Moreover, using $t \geq mn - (j_1 + 2j_2 + \cdots + mj_m)$ we obtain

$$j_0 + j_1 + \cdots + j_m = (m+1)n - (j_1 + 2j_2 + \cdots + mj_m) - t \leq n.$$

For $j_0 \neq 0$, let $S_0 = \{(j_m + \cdots + j_1) + 1, \ldots, (j_m + \cdots + j_1) + j_0\}$. If $j_0 = 0$, then we put $S_0 = \emptyset$. Note that $S_0, \ldots, S_m$ are pairwise disjoint subsets of $\{1, \ldots, n\}$. For each $i \in \{1, \ldots, n\}$, let

$$a_i = \begin{cases} m - \ell & \text{if } i \in S_\ell \text{ for some } 0 \leq \ell \leq m, \\ m+1 & \text{otherwise.} \end{cases}$$

Assume that $j_m + \cdots + j_1 + j_0 < n$ and put

$$D = (r-t)P_n + \sum_{i=1}^{n} a_i P_i.$$

We claim that $D \in \mathcal{U}(r, t; j_1, \ldots, j_m)$. Note that $n \notin \cup_{\ell=0}^{m} S_\ell$ by the assumption $j_m + \cdots + j_1 + j_0 < n$. Hence we have $v_{P_n}(D) = r - t + (m+1) \geq m+1$ and $v_{P_n}(\overline{D}) = m+1$, where we used $r \geq t$. Thus, for $1 \leq i \leq n$ we get $v_{P_i}(\overline{D}) = a_i$. This implies that

$$\begin{aligned} \deg\left(\overline{D}\right) &= \sum_{i=1}^{n} a_i = (m+1)\left(n - (j_0 + \cdots + j_m)\right) + \sum_{\ell=0}^{m} (m-\ell) j_\ell \\ &= (m+1)n + \sum_{\ell=0}^{m} (m - \ell - m - 1) j_\ell \\ &= (m+1)n - \sum_{\ell=0}^{m} (\ell+1) j_\ell = t, \end{aligned}$$

where we used (3.3). Moreover, $\deg(D) = \deg\left(\overline{D}\right) + (r-t) = r$, $\jmath_\ell(D) = |S_\ell| = j_\ell$ for each $1 \leq \ell \leq m$, and hence $D \in \mathcal{U}(r, t; j_1, \ldots, j_m)$.

Next we consider the case $j_m+\cdots+j_1+j_0=n$. This case implies that (cf. (3.3))

$$mn = t + j_1 + 2j_2 + \cdots + mj_m.$$

Therefore we construct $\overline{D}$ similarly and D using the existence of a degree $r-t$ positive divisor whose support is disjoint from the set $\{P_1,\ldots,P_n\}$. □

Definition 3.4. *For integers $a \geq b \geq 0$ with $b \leq n$ and a set $\{Q_1,\ldots,Q_b\}$ of rational places, let $C_{a,b}$ denote the cardinality of the set of positive divisors given by*

$$\left\{D \geq 0 : \deg(D) = a,\ \operatorname{supp}\left(\overline{D}\right) = \{Q_1,\ldots,Q_b\}\right\}.$$

Note that $C_{a,b}$ is independent of the choice of the set $\{Q_1,\ldots,Q_b\}$; only the cardinality b of this set matters.

Lemma 3.5. *For $r \geq t \geq 0$, $j_1,\ldots,j_m \geq 0$, and $mn-(j_1+\cdots+mj_m) \leq t \leq (m+1)n-(2j_1+\cdots+(m+1)j_m)$, the cardinality of $\mathcal{U}(r,t;j_1,\ldots,j_m)$ is*

$$\binom{n}{j_m}\binom{n-j_m}{j_{m-1}}\cdots\binom{n-(j_2+j_3+\cdots+j_m)}{j_1}\binom{n-(j_1+j_2+\cdots+j_m)}{t-mn+(j_1+2j_2+\cdots+mj_m)}$$

$$\times\ C_{r-mn+(j_1+2j_2+\cdots+mj_m),t-mn+(j_1+2j_2+\cdots+mj_m)}.$$

Proof. We prove the lemma for $m=2$, and the general case is similar. Assume first the degenerate subcase that $mn = t+(j_1+2j_2+\cdots+mj_m)$ and $r>t$. In this degenerate subcase we have

$$C_{r-mn+(j_1+2j_2+\cdots+mj_m),t-mn+(j_1+2j_2+\cdots+mj_m)} = C_{r-t,0},$$

and we note that $C_{r-t,0}$ is the number of positive divisors with degree $r-t$ whose support is disjoint from the set $\{P_1,\ldots,P_n\}$. By Lemma 3.3, in this degenerate subcase, the set $\mathcal{U}(r,t;j_1,j_2)$ is empty if and only if $C_{r-t,0}=0$. Thus, the formula in the current lemma holds if $C_{r-t,0}=0$. Therefore we can assume without loss of generality that we are either not in the degenerate subcase, or if we are in the degenerate subcase, then $C_{r-t,0}>0$. Hence, again by Lemma 3.3, we know that the set $\mathcal{U}(r,t;j_1,j_2)$ is nonempty. For $D \in \mathcal{U}(r,t;j_1,j_2)$, let $S_2 = \left\{P \in \{P_1,\ldots,P_n\} : v_P(D)=0\right\}$, $S_1 = \left\{P \in \{P_1,\ldots,P_n\} : v_P(D)=1\right\}$, $S_0 = \left\{P \in \{P_1,\ldots,P_n\} : v_P(D)=2\right\}$, and $S = \left\{P \in \{P_1,\ldots,P_n\} : v_P(D) \geq 3\right\}$. Note that $|S_2| = j_2$ and $|S_1| = j_1$ and that by (3.1) we get $|S_0| = \jmath_0(D) = 3n-(2j_1+3j_2)-t$. The choices of S_2, S_1, and S_0 determine S. We have $|S| = n-(j_1+j_2)-|S_0| = t-2n+(j_1+2j_2)$. Hence there are

$$\binom{n}{j_2}\binom{n-j_2}{j_1}\binom{n-(j_1+j_2)}{t-2n+(j_1+2j_2)}$$

choices for these subsets. Assume that the subsets S_2, S_1, S_0, and S are determined. For a corresponding $D \in \mathcal{U}(r,t;j_1,j_2)$, let $D_1 = b_1P_1+\cdots+b_nP_n$, where

$$b_i = \begin{cases} v_{P_i}(D) = v_{P_i}(\overline{D}) & \text{if } P_i \in S_2 \cup S_1 \cup S_0, \\ 2 = v_{P_i}(\overline{D})-1 & \text{if } P_i \in S. \end{cases}$$

Moreover, let $E = D - D_1$. Then E is a positive divisor and $\operatorname{supp}\left(\overline{E}\right) = S$. Note that

$$\deg(D_1) = t-|S|,\ \ \deg(E) = \deg(D)-\deg(D_1) = r-t+|S|.$$

Hence

$$|\operatorname{supp}(\overline{E})| = t - 2n + (j_1 + 2j_2), \quad \deg(E) = r - 2n + (j_1 + 2j_2).$$

Using Definition 3.4, we obtain that there are $C_{r-2n+(j_1+2j_2),t-2n+(j_1+2j_2)}$ choices for E, which completes the proof. □

In the following, when two sets U_1, U_2 are disjoint and we would like to emphasize that their union is the union of the disjoint subsets U_1 and U_2, then we use $U_1 \bigsqcup U_2$ to denote the disjoint union, and similarly for the disjoint union of finitely many pairwise disjoint subsets.

Recall that for integers $r \geq s \geq 0$ and nonnegative integers $X_1, \dots, X_m$, the set $\mathcal{V}_m(r, s; X_1, \dots, X_m)$ is defined in Definition 2.3. Using Definition 3.2 and Lemma 3.3, we can write the set $\mathcal{V}_m(r, s; X_1, \dots, X_m)$ as the disjoint union

$$\mathcal{V}_m(r, s; X_1, \dots, X_m) = \bigsqcup_{j_m} \bigsqcup_{j_{m-1}} \cdots \bigsqcup_{j_1} \bigsqcup_{t} \mathcal{U}(r, t; j_1, \dots, j_m), \tag{3.4}$$

where the m-tuples $(j_1, \dots, j_m)$ of indices run over the finite set of m-tuples of integers satisfying

$$\begin{aligned} 0 &\leq j_m \leq 2X_m,\ 0 \leq j_{m-1} \leq 2X_{m-1} + X_m, \dots, \\ 0 &\leq j_1 \leq 2X_1 + \sum_{\nu=2}^{m} X_\nu, \\ 2j_1 &+ 3j_2 + \cdots + (m+1)j_m \leq 2(2X_1 + 3X_2 + \cdots + (m+1)X_m), \end{aligned} \tag{3.5}$$

and for each m-tuple satisfying (3.5), the index t runs from $\max(s, mn - (j_1 + 2j_2 + \cdots + mj_m))$ to $\min(r, (m+1)n - (2j_1 + 3j_2 + \cdots + (m+1)j_m))$.

Combining (3.4) and Lemma 3.5, we can compute the cardinality of the set $\mathcal{V}_m(r, s; X_1, \dots, X_m)$.

4. Asymptotic upper bound on the cardinality of $\mathcal{V}_1(r, s; X_1)$.

In this section we obtain an asymptotic upper bound on the cardinality of $\mathcal{V}_m(r, s; X_1, \dots, X_m)$ for the case $m = 1$ in a suitable sequence of global function fields (see Corollary 4.5). The assumption $m = 1$ is made for simplicity and for the clarity of the exposition. Later in section 5 we generalize this asymptotic upper bound to the case $m \geq 1$. The bound in Corollary 4.5 is given using a real-valued function $S(\sigma, y, x, t_1)$, which is introduced in Definition 4.3. Corollary 4.5 is obtained under Assumption 1 given below. In section 7 we will show that Assumption 1 holds in many cases of interest. Under Assumption 1, we further define the real-valued function $I_{y,x_1}(\sigma)$ in Definition 4.6 and then we show that it is strictly increasing in σ. Moreover, we compute $I_{y,x_1}(\sigma)$ under some conditions, and this result will also be generalized in section 5.

The asymptotic upper bound for the cardinality of $\mathcal{V}_m(r, s; X_1, \dots, X_m)$ will be used later to prove the existence of a sequence of distinguished divisors on the basis of Proposition 2.4.

Definition 4.1. *For any prime power q, let E_q be the real-valued function defined on the interval $[0, 1]$ as follows: for $0 < x < 1$ we put $E_q(x) = -x \log_q x - (1 - x) \log_q(1 - x)$ and for $x \in \{0, 1\}$ we put $E_q(0) = E_q(1) = \lim_{x \to 0^+} E_q(x) = \lim_{x \to 1^-} E_q(x) = 0$.*

Using Stirling's formula, we obtain the following well-known results. For any real number $0 \leq \alpha \leq 1$, we have

$$\lim_{n \to \infty} \frac{\log_q \binom{n}{\lfloor \alpha n \rfloor}}{n} = E_q(\alpha). \tag{4.1}$$

For any real numbers $0 \le \alpha_1 \le 1$ and $0 \le \alpha_2 < 1$ with $\alpha_1 + \alpha_2 \le 1$, we have

$$\lim_{n\to\infty} \frac{\log_q \binom{n-\lfloor \alpha_2 n \rfloor}{\lfloor \alpha_1 n \rfloor}}{n} = (1-\alpha_2) E_q\left(\frac{\alpha_1}{1-\alpha_2}\right). \tag{4.2}$$

Now we state an important assumption and introduce related notation.

Assumption 1. Assume that $(F_i/\mathbb{F}_q)_{i=1}^{\infty}$ is a sequence of global function fields with full constant field $\mathbb{F}_q$, with $g_i \to \infty$ as $i \to \infty$, and with $\limsup_{i\to\infty} \frac{n_i}{g_i} = \gamma > 0$, where n_i and g_i denote the number of rational places and the genus of F_i, respectively. Using a suitable subsequence of $(F_i/\mathbb{F}_q)_{i=1}^{\infty}$, we may assume that $\lim_{i\to\infty} \frac{n_i}{g_i} = \gamma > 0$.

We will use the following proposition in our upper bounds.

Proposition 4.2. *Under Assumption* 1, *let* $(a_i)_{i=1}^{\infty}$ *and* $(b_i)_{i=1}^{\infty}$ *be sequences of integers such that* $a_i \ge b_i \ge 0$ *and* $b_i \le n_i$ *for all* $i \ge 1$. *We also assume that there exist the limits*

$$\lim_{i\to\infty} \frac{a_i}{n_i} = a, \quad \lim_{i\to\infty} \frac{b_i}{n_i} = b \quad \text{with } 0 < b \le a < \infty. \tag{4.3}$$

For each $i \ge 1$, *let* $C_{a_i,b_i}^{(i)}$ *denote the cardinality of the set of positive divisors given in Definition* 3.4 *for a suitable set* $\{Q_1^{(i)}, \ldots, Q_{b_i}^{(i)}\}$ *of rational places of* F_i. *Then we have*

$$\limsup_{i\to\infty} \frac{\log_q C_{a_i,b_i}^{(i)}}{n_i} \le \begin{cases} aE_q\left(\frac{b}{a}\right) & \text{if } \frac{b}{a} \ge 1 - \frac{1}{q}, \\ a - b\log_q(q-1) & \text{if } \frac{b}{a} \le 1 - \frac{1}{q}. \end{cases}$$

Proof. This follows from Definition 3.4 and the proof of [10, Lemma 3.4.10]. □

Let $y, \sigma, x_1 \ge 0$ be real numbers. Under Assumption 1, for each $i \ge 1$ we define the integers

$$r_i = \left\lfloor \left(1 + y + \frac{\sigma}{\gamma}\right) n_i \right\rfloor, \; s_i = \lfloor (1+y) n_i \rfloor, \; X_1^{(i)} = \lfloor x_1 n_i \rfloor. \tag{4.4}$$

Let $\mathcal{V}_1^{(i)}(r_i, s_i; X_1^{(i)})$ be the set of positive divisors of degree r_i of F_i, which is defined using Definition 2.3. We note that for each real number $0 \le t_1 \le 2x_1$ and each integer $i \ge 1$, we have

$$\max\{s_i, n_i - \lfloor t_1 n_i \rfloor\} = s_i.$$

Moreover, if

$$1 + y + \frac{\sigma}{\gamma} < 2 - 4x_1 \quad \text{or equivalently} \quad y + 4x_1 + \frac{\sigma}{\gamma} < 1 \tag{4.5}$$

holds, then for each real number $0 \le t_1 \le 2x_1$ and integer $i \ge 1$ we also have

$$\min\{r_i, 2n_i - 2\lfloor t_1 n_i \rfloor\} = r_i.$$

DEFINITION 4.3. *For real numbers $y > 0$, $x_1, \sigma \geq 0$ satisfying* (4.5) *and real numbers $0 \leq t_1 \leq 2x_1$, $0 \leq x \leq \frac{\sigma}{\gamma}$, let $S(\sigma, y, x, t_1)$ be the real-valued function*

$$S(\sigma, y, x, t_1) = E_q(t_1) + (1 - t_1)E_q\left(\frac{y + x + t_1}{1 - t_1}\right)$$
$$+ \begin{cases} \left(y + \frac{\sigma}{\gamma} + t_1\right) E_q\left(\frac{y+x+t_1}{y+\frac{\sigma}{\gamma}+t_1}\right) & \text{if } \frac{y+x+t_1}{y+\frac{\sigma}{\gamma}+t_1} \geq 1 - \frac{1}{q}, \\ \left(y + \frac{\sigma}{\gamma} + t_1\right) - (y + x + t_1)\log_q(q-1) & \text{if } \frac{y+x+t_1}{y+\frac{\sigma}{\gamma}+t_1} \leq 1 - \frac{1}{q}. \end{cases}$$

Note that by (4.5) *we have $4x_1 < 1$ and hence $t_1 < \frac{1}{2}$.*

PROPOSITION 4.4. *Under Assumption* 1, *let $y > 0$ and $x_1, \sigma \geq 0$ be real numbers satisfying* (4.5). *For each integer $i \geq 1$ and real numbers $0 \leq t_1 \leq 2x_1$, $0 \leq x \leq \frac{\sigma}{\gamma}$, let $\mathcal{U}^{(i)}(\lfloor(1 + y + \frac{\sigma}{\gamma})n_i\rfloor, \lfloor(1 + y + x)n_i\rfloor; \lfloor t_1 n_i\rfloor)$ be the set of positive divisors of F_i defined in Definition* 3.2 *for $m = 1$. Then for the cardinalities of these sets we have*

$$\limsup_{i\to\infty} \frac{\log_q \left|\mathcal{U}^{(i)}\left(\left\lfloor(1 + y + \frac{\sigma}{\gamma})n_i\right\rfloor, \lfloor(1 + y + x)n_i\rfloor; \lfloor t_1 n_i\rfloor\right)\right|}{n_i} \leq S(\sigma, y, x, t_1).$$

Proof. Note that $n_i - \lfloor t_1 n_i\rfloor \leq \lfloor(1 + y + x)n_i\rfloor$ and using (4.5) we get $\lfloor(1 + y + x)n_i\rfloor \leq 2n_i - 2\lfloor t_1 n_i\rfloor$ for each x and t_1 in the range under consideration. Hence using Lemma 3.5, we obtain

$$\begin{aligned} &\left|\mathcal{U}^{(i)}\left(\left\lfloor\left(1 + y + \frac{\sigma}{\gamma}\right) n_i\right\rfloor, \lfloor(1 + y + x)n_i\rfloor; \lfloor t_1 n_i\rfloor\right)\right| \\ &\quad = \binom{n_i}{\lfloor t_1 n_i\rfloor}\binom{n_i - \lfloor t_1 n_i\rfloor}{\lfloor(1+y+x)n_i\rfloor - n_i + \lfloor t_1 n_i\rfloor} \\ &\quad \times C^{(i)}_{\lfloor(1+y+\frac{\sigma}{\gamma})n_i\rfloor - n_i + \lfloor t_1 n_i\rfloor, \lfloor(1+y+x)n_i\rfloor - n_i + \lfloor t_1 n_i\rfloor}. \end{aligned} \tag{4.6}$$

Using (4.1) and (4.2), we obtain

$$\begin{aligned} &\lim_{i\to\infty} \frac{\log_q \binom{n_i}{\lfloor t_1 n_i\rfloor}}{n_i} = E_q(t_1), \\ &\lim_{i\to\infty} \frac{\log_q \binom{n_i - \lfloor t_1 n_i\rfloor}{\lfloor(1+y+x)n_i\rfloor - n_i + \lfloor t_1 n_i\rfloor}}{n_i} = (1 - t_1)E_q\left(\frac{y + x + t_1}{1 - t_1}\right). \end{aligned} \tag{4.7}$$

Note that $\lim_{i\to\infty} \frac{\lfloor(1+y+\frac{\sigma}{\gamma})n_i\rfloor - n_i + \lfloor t_1 n_i\rfloor}{n_i} = y + \frac{\sigma}{\gamma} + t_1$ and $\lim_{i\to\infty} \frac{\lfloor(1+y+x)n_i\rfloor - n_i + \lfloor t_1 n_i\rfloor}{n_i} = y + x + t_1$. Hence from Proposition 4.2 we get

$$\begin{aligned} &\limsup_{i\to\infty} \frac{\log_q C^{(i)}_{\lfloor(1+y+\frac{\sigma}{\gamma})n_i\rfloor - n_i + \lfloor t_1 n_i\rfloor, \lfloor(1+y+x)n_i\rfloor - n_i + \lfloor t_1 n_i\rfloor}}{n_i} \\ &\leq \begin{cases} (y + \frac{\sigma}{\gamma} + t_1)E_q\left(\frac{y+x+t_1}{y+\frac{\sigma}{\gamma}+t_1}\right) & \text{if } \frac{y+x+t_1}{y+\frac{\sigma}{\gamma}+t_1} \geq 1 - \frac{1}{q}, \\ (y + \frac{\sigma}{\gamma} + t_1) - (y + x + t_1)\log_q(q-1) & \text{if } \frac{y+x+t_1}{y+\frac{\sigma}{\gamma}+t_1} \leq 1 - \frac{1}{q}. \end{cases} \end{aligned} \tag{4.8}$$

Using (4.6), (4.7), (4.8), and Definition 4.3, we complete the proof. □

Corollary 4.5. *Under Assumption* 1, *let $y > 0$ and $x_1, \sigma \geq 0$ be real numbers satisfying* (4.5). *For each integer $i \geq 1$, let r_i, s_i, and $X_1^{(i)}$ be the integers defined in* (4.4), *and let $\mathcal{V}_1^{(i)}(r_i, s_i, X_1^{(i)})$ be the set of positive divisors of F_i defined in Definition* 2.3 *for $m = 1$. Then for the cardinalities of these sets we have*

$$\limsup_{i \to \infty} \frac{\log_q |\mathcal{V}_1^{(i)}(r_i, s_i; X_1^{(i)})|}{n_i} \leq \max S(\sigma, y, x, t_1),$$

where the maximum is over all real numbers x and t_1 satisfying $0 \leq x \leq \frac{\sigma}{\gamma}$ and $0 \leq t_1 \leq 2x_1$.

Proof. Using (3.4) and Lemma 3.5 for each $i \geq 1$, we obtain that

$$|\mathcal{V}_1^{(i)}(r_i, s_i; X_1^{(i)})| = \sum_{j_1=0}^{2X_1^{(i)}} \sum_t |\mathcal{U}^{(i)}(r_i, t; j_1)|, \tag{4.9}$$

where t runs from $\max\{s_i, n_i - j_1\}$ to $\min\{r_i, 2n_i - 2j_1\}$. Note that $s_i \geq n_i - j_1$ and $r_i \leq 2n_i - 2j_1$ for each $i \geq 1$ and $0 \leq j_1 \leq 2X_1^{(i)}$. Moreover, for the number of terms $(2X_1^{(i)} + 1)\,(r_i - s_i + 1)$ in the summation in (4.9) we have

$$\begin{aligned} &\lim_{i \to \infty} \frac{\log_q \left(\left(2X_1^{(i)} + 1 \right) (r_i - s_i + 1) \right)}{n_i} \\ &= \lim_{i \to \infty} \left\{ \frac{\log_q (2x_1 + 1/n_i) + \log_q \left(\frac{\sigma}{\gamma} + 1/n_i \right)}{n_i} + 2\frac{\log_q n_i}{n_i} \right\} = 0. \end{aligned} \tag{4.10}$$

Since the summands on the right-hand side of (4.9) are nonnegative, we get

$$|\mathcal{V}_1^{(i)}(r_i, s_i; X_1^{(i)})| \leq \left(2X_1^{(i)} + 1 \right) (r_i - s_i + 1) \max |\mathcal{U}^{(i)}(r_i, t; j_1)|, \tag{4.11}$$

where $\max |\mathcal{U}^{(i)}(r_i, t; j_1)|$ is over the set of ordered pairs (t, j_1) with $s_i \leq t \leq r_i$ and $0 \leq j_1 \leq 2X_1^{(i)}$. Taking the logarithm of both sides of (4.11) and using (4.10) and Proposition 4.4, we complete the proof. □

Definition 4.6. *Under Assumption* 1, *let $y > 0$ and $x_1 \geq 0$ be real numbers such that $y + 4x_1 < 1$. For $\sigma \geq 0$ and $y + 4x_1 + \frac{\sigma}{\gamma} < 1$, let $I_{y,x_1}(\sigma)$ be the real-valued function of σ defined by*

$$I_{y,x_1}(\sigma) = \max S(\sigma, y, x, t_1),$$

where the maximum is over all real numbers x and t_1 such that $0 \leq t_1 \leq 2x_1$ and $0 \leq x \leq \frac{\sigma}{\gamma}$.

By straightforward manipulations, the expression for $S(\sigma, y, x, t_1)$ is simplified to

$$\begin{aligned} S(\sigma, y, x, t_1) &\\ = -t_1 \log_q t_1 &\\ - (y + x + t_1) \log_q (y + x + t_1) &\\ - (1 - y - x - 2t_1) \log_q (1 - y - x - 2t_1) &\\ + \begin{cases} -(y + x + t_1) \log_q (y + x + t_1) - \left(\frac{\sigma}{\gamma} - x\right) \log_q \left(\frac{\sigma}{\gamma} - x\right) & \\ + \left(y + \frac{\sigma}{\gamma} + t_1\right) \log_q \left(y + \frac{\sigma}{\gamma} + t_1\right) & \text{if } \dfrac{y + x + t_1}{y + \frac{\sigma}{\gamma} + t_1} \geq 1 - \dfrac{1}{q}, \\ \left(y + \frac{\sigma}{\gamma} + t_1\right) - (y + x + t_1) \log_q (q - 1) & \text{if } \dfrac{y + x + t_1}{y + \frac{\sigma}{\gamma} + t_1} \leq 1 - \dfrac{1}{q}. \end{cases} \end{aligned} \tag{4.12}$$

We first show that $I_{y,x_1}(\sigma)$ is a strictly increasing function of σ.

LEMMA 4.7. *Under the assumptions of Definition* 4.6, *the real-valued function* $I_{y,x_1}(\sigma)$ *is a strictly increasing function of* σ *on its domain of definition, which is the interval of* σ *such that* $\sigma \geq 0$ *and* $y + 4x_1 + \frac{\sigma}{\gamma} < 1$.

Proof. Using the expression (4.12), for the partial derivative of $S(\sigma, y, x, t_1)$ with respect to σ we obtain

$$\frac{\partial S}{\partial \sigma}(\sigma, y, x, t_1) = \begin{cases} \frac{1}{\gamma} \log_q \frac{y + \frac{\sigma}{\gamma} + t_1}{\frac{\sigma}{\gamma} - x} & \text{if } \frac{y + x + t_1}{y + \frac{\sigma}{\gamma} + t_1} \geq 1 - \frac{1}{q}, \\ \frac{1}{\gamma} & \text{if } \frac{y + x + t_1}{y + \frac{\sigma}{\gamma} + t_1} \leq 1 - \frac{1}{q}. \end{cases}$$

Therefore $\frac{\partial S}{\partial \sigma}(\sigma, y, x, t_1) > 0$ for each $0 \leq x < \frac{\sigma}{\gamma}$ and $0 \leq t_1 \leq 2x_1$. Moreover, $\lim_{x \to \frac{\sigma}{\gamma}-} \frac{\partial S}{\partial \sigma}(\sigma, y, x, t_1) = +\infty$ for $0 \leq t_1 \leq 2x_1$. This completes the proof. □

LEMMA 4.8. *Under the assumptions of Definition* 4.6, *for the partial derivatives* $\frac{\partial S}{\partial t_1}(\sigma, y, x, t_1)$ *and* $\frac{\partial S}{\partial x}(\sigma, y, x, t_1)$ *of* $S(\sigma, y, x, t_1)$ *with respect to* t_1 *and* x *we obtain*

$$\frac{\partial S}{\partial t_1}(\sigma, y, x, t_1) = \log_q \frac{(1 - y - x - 2t_1)^2}{t_1(y + x + t_1)} + \begin{cases} \log_q \frac{y + \frac{\sigma}{\gamma} + t_1}{y + x + t_1} & \textit{if } \frac{y + x + t_1}{y + \frac{\sigma}{\gamma} + t_1} \geq 1 - \frac{1}{q}, \\ \log_q \frac{q}{q - 1} & \textit{if } \frac{y + x + t_1}{y + \frac{\sigma}{\gamma} + t_1} \leq 1 - \frac{1}{q}, \end{cases}$$

and

$$\frac{\partial S}{\partial x}(\sigma, y, x, t_1) = \log_q \frac{1 - y - x - 2t_1}{y + x + t_1} + \begin{cases} \log_q \frac{\frac{\sigma}{\gamma} - x}{y + x + t_1} & \textit{if } \frac{y + x + t_1}{y + \frac{\sigma}{\gamma} + t_1} \geq 1 - \frac{1}{q}, \\ -\log_q (q - 1) & \textit{if } \frac{y + x + t_1}{y + \frac{\sigma}{\gamma} + t_1} \leq 1 - \frac{1}{q}. \end{cases}$$

Proof. Let S_1, T_1, and T_2 denote the following expressions from (4.12):

$$\begin{aligned} S_1 = -t_1 \log_q t_1 - (y + x + t_1) \log_q (y + x + t_1) \\ - (1 - y - x - 2t_1) \log_q (1 - y - x - 2t_1), \end{aligned}$$

$$\begin{aligned} T_1 &= -(y+x+t_1)\log_q(y+x+t_1) - \left(\frac{\sigma}{\gamma} - x\right)\log_q\left(\frac{\sigma}{\gamma} - x\right) \\ &\quad + \left(y + \frac{\sigma}{\gamma} + t_1\right)\log_q(y + \frac{\sigma}{\gamma} + t_1), \\ T_2 &= \left(y + \frac{\sigma}{\gamma} + t_1\right) - (y+x+t_1)\log_q(q-1). \end{aligned}$$

For their partial derivatives with respect to t_1 and x we obtain

$$\begin{aligned} \frac{\partial S_1}{\partial t_1} &= -\log_q t_1 - \log_q(y+x+t_1) + 2\log_q(1-y-x-2t_1), \\ \frac{\partial T_1}{\partial t_1} &= -\log_q(y+x+t_1) + \log_q\left(y + \frac{\sigma}{\gamma} + t_1\right), \\ \frac{\partial T_2}{\partial t_1} &= 1 - \log_q(q-1) = \log_q\frac{q}{q-1}, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial S_1}{\partial x} &= -\log_q(y+x+t_1) + \log_q(1-y-x-2t_1), \\ \frac{\partial T_1}{\partial x} &= -\log_q(y+x+t_1) + \log_q\left(\frac{\sigma}{\gamma} - x\right), \\ \frac{\partial T_2}{\partial x} &= -\log_q(q-1). \end{aligned}$$

Using (4.12) and combining the partial derivatives above, we get the desired formulas. $\square$

COROLLARY 4.9. *Under the assumptions of Definition* 4.6, *if all of the conditions*

- $C1$: $\frac{\sigma}{\gamma} \le \frac{y}{q-1}$,
- $C2$: $2x_1(y + \frac{\sigma}{\gamma} + 2x_1)^2 < (1 - y - \frac{\sigma}{\gamma} - 4x_1)^2(y + \frac{\sigma}{\gamma})$,
- $C3$: $\frac{\sigma}{\gamma}(1-y) < y^2$

hold, then we have

$$\begin{aligned} I_{y,x_1}(\sigma) &= S(\sigma, y, 0, 2x_1) \\ &= E_q(2x_1) + (1-2x_1)\,E_q\left(\tfrac{y+2x_1}{1-2x_1}\right) + \left(y + \tfrac{\sigma}{\gamma} + 2x_1\right)E_q\left(\tfrac{y+2x_1}{y+\frac{\sigma}{\gamma}+2x_1}\right). \end{aligned}$$

Proof. Assume that $0 \le x \le \frac{\sigma}{\gamma}$ and $0 \le t_1 \le 2x_1$. First we observe that

$$\frac{y+x+t_1}{y + \frac{\sigma}{\gamma} + t_1} \ge \frac{y}{y + \frac{\sigma}{\gamma}}.$$

Using condition C1 we obtain

$$\frac{y+x+t_1}{y + \frac{\sigma}{\gamma} + t_1} \ge \frac{y}{y + \frac{\sigma}{\gamma}} \ge 1 - \frac{1}{q}. \tag{4.13}$$

Moreover, using condition C2 we also get

$$t_1(y+x+t_1)^2 \le 2x_1\left(y+\tfrac{\sigma}{\gamma}+2x_1\right)^2 < \left(y+\tfrac{\sigma}{\gamma}\right)\left(1-y-\tfrac{\sigma}{\gamma}-4x_1\right)^2$$
$$\le \left(y+\tfrac{\sigma}{\gamma}+t_1\right)(1-y-x-2t_1)^2.$$

Therefore by Lemma 4.8 and (4.13) we have $\frac{\partial S}{\partial t_1}(\sigma,y,x,t_1)>0$. Similarly, condition C3 implies

$$\left(\frac{\sigma}{\gamma}-x\right)(1-y-x-2t_1)\le\frac{\sigma}{\gamma}(1-y)<y^2\le(x+y+t_1)^2,$$

and by Lemma 4.8 we also have $\frac{\partial S}{\partial x}(\sigma,y,x,t_1)<0$. Hence we obtain $I_{y,x_1}(\sigma)=S(\sigma,y,0,2x_1)$. Using Definition 4.3 we complete the proof. □

5. Asymptotic upper bound on the cardinality of $\mathcal{V}_m(r,s;X_1,\dots,X_m)$ for the general case $m\ge1$. In this section we obtain generalizations of the results of section 4 to the general case $m\ge1$. In particular, we derive the asymptotic upper bound on the cardinality of $\mathcal{V}_m(r,s;X_1,\dots,X_m)$ in Proposition 5.7. The bound in Proposition 5.7 uses a real-valued function $S(\sigma,y,x,t_1,t_2,\dots,t_m)$ that is introduced in Definition 5.6, which generalizes Definition 4.3. The generalization $I_{y,x_1,x_2,\dots,x_m}(\sigma)$ of the real-valued function $I_{y,x_1}(\sigma)$ of section 4 is given in Definition 5.8. Finally, we compute $I_{y,x_1,x_2,\dots,x_m}(\sigma)$ under some conditions in Proposition 5.10 that will be used in section 7.

For the clarity of exposition in this rather technical part, we begin with the case $m=2$ which corresponds to the two-variable case t_1,t_2.

Definition 5.1. *Let $\gamma>0$ be as in Assumption 1 (cf. section 4). Let $y>0$, $x_1,x_2,\sigma\ge0$ be real numbers satisfying*

$$y+2(2x_1+3x_2)+\frac{\sigma}{\gamma}<1. \tag{5.1}$$

For real numbers $0\le x\le\frac{\sigma}{\gamma}$ and $0\le t_1,t_2$ satisfying $t_2\le 2x_2$, $t_1\le 2x_1+x_2$, and $2t_1+3t_2\le2(2x_1+3x_2)$, let $S(\sigma,y,x,t_1,t_2)$ be the real-valued function

$$S(\sigma,y,x,t_1,t_2)=E_q(t_2)+(1-t_2)E_q\left(\frac{t_1}{1-t_2}\right)$$
$$+(1-t_1-t_2)E_q\left(\frac{y+x+t_1+2t_2}{1-t_1-t_2}\right)$$
$$+\begin{cases}\left(y+\frac{\sigma}{\gamma}+t_1+2t_2\right)E_q\left(\frac{y+x+t_1+2t_2}{y+\frac{\sigma}{\gamma}+t_1+2t_2}\right) & \textit{if } \frac{y+x+t_1+2t_2}{y+\frac{\sigma}{\gamma}+t_1+2t_2}\ge1-\frac{1}{q},\\ \left(y+\frac{\sigma}{\gamma}+t_1+2t_2\right)-(y+x+t_1+2t_2)\log_q(q-1) & \textit{if } \frac{y+x+t_1+2t_2}{y+\frac{\sigma}{\gamma}+t_1+2t_2}\le1-\frac{1}{q}.\end{cases}$$

Note that by (5.1) we have $2(2x_1+3x_2)<1$ and hence $t_1+t_2\le t_1+\frac{3}{2}t_2<\frac{1}{2}$.

Instead of stating a generalization of Proposition 4.4 explicitly, we prefer to give a generalization of Corollary 4.5 directly in the following proposition, whose proof includes a generalization of Proposition 4.4.

Proposition 5.2. *Under Assumption* 1 *(cf. section* 4*), let $y > 0$ and $x_1, x_2, \sigma \geq 0$ be real numbers satisfying* (5.1)*. For each integer $i \geq 1$, let $r_i = \lfloor (2 + y + \frac{\sigma}{\gamma})n_i \rfloor$, $s_i = \lfloor (2+y)n_i \rfloor$, $X_1^{(i)} = \lfloor x_1 n_i \rfloor$, $X_2^{(i)} = \lfloor x_2 n_i \rfloor$, and $\mathcal{V}_2^{(i)}(r_i, s_i; X_1^{(i)}, X_2^{(i)})$ be the set of positive divisors of F_i defined in Definition* 2.3 *for $m = 2$. Then for the cardinalities of these sets we have*

$$\limsup_{i\to\infty} \frac{\log_q \left| \mathcal{V}_2^{(i)}(r_i, s_i; X_1^{(i)}, X_2^{(i)}) \right|}{n_i} \leq \max\, S(\sigma, y, x, t_1, t_2),$$

where the maximum is over all real numbers x and t_1, t_2 satisfying $0 \leq x \leq \frac{\sigma}{\gamma}$ and $0 \leq t_2 \leq 2x_2$, $0 \leq t_1 \leq 2x_1 + x_2$, and $2t_1 + 3t_2 \leq 2(2x_1 + 3x_2)$.

Proof. We follow similar methods as in the proofs of Proposition 4.4 and Corollary 4.5. First note that for each integer $i \geq 1$ and real numbers $0 \leq t_1, t_2$ with $2t_1 + 3t_2 \leq 2(2x_1 + 3x_2)$, using (5.1) we obtain $r_i \leq 3n_i - (2\lfloor t_1 n_i \rfloor + 3\lfloor t_2 n_i \rfloor)$. Moreover, it is also clear that $s_i \geq 2n_i - (\lfloor t_1 n_i \rfloor + 2\lfloor t_2 n_i \rfloor)$ for each integer $i \geq 1$ and real numbers $t_1, t_2 \geq 0$. Hence using (3.4) and Lemma 3.5 as in the proof of Corollary 4.5, for integers $i \geq 1$ and real numbers $0 \leq x, t_1, t_2$ such that $x \leq \frac{\sigma}{\gamma}$, $t_2 \leq 2x_2$, $t_1 \leq 2x_1 + x_2$, and $2t_1 + 3t_2 \leq 2(2x_1 + 3x_2)$, we need to consider the cardinality $\left|\mathcal{U}^{(i)}\left(r_i, \lfloor (2+y+x)n_i \rfloor; \lfloor t_1 n_i \rfloor, \lfloor t_2 n_i \rfloor\right)\right|$ of the set of positive divisors of F_i defined in Definition 3.2 for $m = 2$. By Lemma 3.5 we have

$$\begin{aligned} &\left|\mathcal{U}^{(i)}\left(r_i, \lfloor (2+y+x)n_i \rfloor; \lfloor t_1 n_i \rfloor, \lfloor t_2 n_i \rfloor\right)\right| \\ &\quad = \binom{n_i}{\lfloor t_2 n_i \rfloor}\binom{n_i - \lfloor t_2 n_i \rfloor}{\lfloor t_1 n_i \rfloor}\binom{n_i - (\lfloor t_1 n_i \rfloor + \lfloor t_2 n_i \rfloor)}{\lfloor (2+y+x)n_i \rfloor - 2n_i + (\lfloor t_1 n_i \rfloor + 2\lfloor t_2 n_i \rfloor)} \\ &\quad\quad \times C^{(i)}_{r_i - 2n_i + \lfloor t_1 n_i \rfloor + 2\lfloor t_2 n_i \rfloor,\, \lfloor (2+y+x)n_i \rfloor - 2n_i + \lfloor t_1 n_i \rfloor + 2\lfloor t_2 n_i \rfloor}. \end{aligned}$$

We complete the proof using similar arguments as in the proofs of Proposition 4.4 and Corollary 4.5. □

Now we generalize Definition 4.6.

Definition 5.3. *Under Assumption* 1 *(cf. section* 4*), let $y > 0$ and $x_1, x_2 \geq 0$ be real numbers such that $y + 2(2x_1 + 3x_2) < 1$. For $\sigma \geq 0$ and $y + 2(2x_1 + 3x_2) + \frac{\sigma}{\gamma} < 1$, let $I_{y,x_1,x_2}(\sigma)$ be the real-valued function of σ defined by*

$$I_{y,x_1,x_2}(\sigma) = \max\, S(\sigma, y, x, t_1, t_2),$$

where the maximum is over all real numbers x, t_1, and t_2 such that $0 \leq x \leq \frac{\sigma}{\gamma}$ and $0 \leq t_2 \leq 2x_2$, $0 \leq t_1 \leq 2x_1 + x_2$, and $2t_1 + 3t_2 \leq 2(2x_1 + 3x_2)$.

The following lemma generalizes Lemma 4.7.

Lemma 5.4. *Under the assumptions of Definition* 5.3*, the real-valued function $I_{y,x_1,x_2}(\sigma)$ is a strictly increasing function of σ on its domain of definition, which is the interval of σ such that $\sigma \geq 0$ and $y + 2(2x_1 + 3x_2) + \frac{\sigma}{\gamma} < 1$.*

Proof. For the partial derivative of $S(\sigma, y, x, t_1, t_2)$ with respect to σ we obtain

$$\frac{\partial S}{\partial \sigma}(\sigma, y, x, t_1, t_2) = \begin{cases} \frac{1}{\gamma}\log_q\left(\frac{y + \frac{\sigma}{\gamma} + t_1 + 2t_2}{\frac{\sigma}{\gamma} - x}\right) & \text{if } \frac{y + x + t_1 + 2t_2}{y + \frac{\sigma}{\gamma} + t_1 + 2t_2} \geq 1 - \frac{1}{q}, \\ \frac{1}{\gamma} & \text{if } \frac{y + x + t_1 + 2t_2}{y + \frac{\sigma}{\gamma} + t_1 + 2t_2} \leq 1 - \frac{1}{q}. \end{cases}$$

Then the proof is similar to the proof of Lemma 4.7. □

Now we give a generalization of Corollary 4.9 in the following proposition.

PROPOSITION 5.5. *Under the assumptions of Definition* 5.3, *assume also that all of the following conditions hold*:

- $C1$: $\frac{\sigma}{\gamma} \le \frac{y}{q-1}$;
- $C2.1$:

$$(2x_1+x_2)\left(y+\frac{\sigma}{\gamma}+2x_1+4x_2\right)^2 < \left(1-y-\frac{\sigma}{\gamma}-2(2x_1+3x_2)\right)^2\left(y+\frac{\sigma}{\gamma}\right);$$

- $C2.2$:

$$2x_2\left(y+\frac{\sigma}{\gamma}+2x_1+4x_2\right)^4 < \left(1-y-\frac{\sigma}{\gamma}-2(2x_1+3x_2)\right)^3\left(y+\frac{\sigma}{\gamma}\right)^2;$$

- $C3$: $\frac{\sigma}{\gamma}(1-y) < y^2$;
- $C4$:

$$x_2^2\left(y+\frac{\sigma}{\gamma}+2x_1+4x_2\right) \le 2x_1^3.$$

Then we have $I_{y,x_1,x_2}(\sigma) = S(\sigma, y, 0, 2x_1, 2x_2)$.

Proof. As in the proof of Corollary 4.9, we first observe that for $0 \le x \le \frac{\sigma}{\gamma}$ and $0 \le t_1, t_2$ with $t_2 \le 2x_2$, $t_1 \le 2x_1 + x_2$, and $2t_1 + 3t_2 \le 2(2x_1 + 3x_2)$, using condition C1 we obtain

$$\frac{y+x+t_1+2t_2}{y+\frac{\sigma}{\gamma}+t_1+2t_2} \ge \frac{y}{y+\frac{\sigma}{\gamma}} \ge 1-\frac{1}{q}. \tag{5.2}$$

For the partial derivative $\frac{\partial S}{\partial x}(\sigma, y, x, t_1, t_2)$ of $S(\sigma, y, x, t_1, t_2)$ with respect to x, by using (5.2) and some straightforward manipulations we get

$$\frac{\partial S}{\partial x}(\sigma,y,x,t_1,t_2) = \log_q \frac{(1-y-x-2t_1-3t_2)\left(\frac{\sigma}{\gamma}-x\right)}{(y+x+t_1+2t_2)^2}.$$

By condition C3 we have

$$\left(\frac{\sigma}{\gamma}-x\right)(1-y-x-2t_1-3t_2) \le \frac{\sigma}{\gamma}(1-y) < y^2 \le (y+x+t_1+2t_2)^2,$$

and hence

$$\frac{\partial S}{\partial x}(\sigma,y,x,t_1,t_2) \le \log_q \frac{\frac{\sigma}{\gamma}(1-y)}{y^2} < 0$$

for $0 < x < \frac{\sigma}{\gamma}$ and $0 \le t_1, t_2$ with $t_2 \le 2x_2$, $t_1 \le 2x_1+x_2$, and $2t_1+3t_2 \le 2(2x_1+3x_2)$.

Now we assume that

$$x_1 > 0 \quad \text{and} \quad x_2 > 0. \tag{5.3}$$

For the partial derivatives $\frac{\partial S}{\partial t_1}(\sigma, y, x, t_1, t_2)$ and $\frac{\partial S}{\partial t_2}(\sigma, y, x, t_1, t_2)$ of $S(\sigma, y, x, t_1, t_2)$ with respect to t_1 and t_2, again using (5.2) and some straightforward manipulations we get

$$\frac{\partial S}{\partial t_1} = \log_q \frac{(1-y-x-2t_1-3t_2)^2\left(y+\frac{\sigma}{\gamma}+t_1+2t_2\right)}{(y+x+t_1+2t_2)^2 t_1}$$

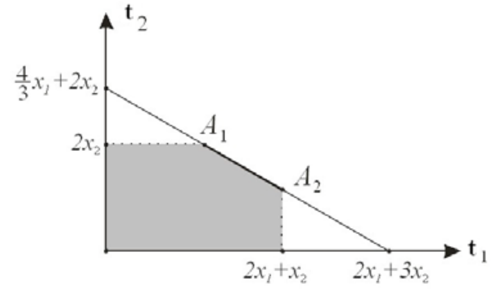


Fig. 1.

and

$$\frac{\partial S}{\partial t_2} = \log_q \frac{(1-y-x-2t_1-3t_2)^3\left(y+\frac{\sigma}{\gamma}+t_1+2t_2\right)^2}{(y+x+t_1+2t_2)^4 t_2}.$$

Note that t_1+2t_2 assumes its maximum over the region

$$0 \le t_2 \le 2x_2,\ 0 \le t_1 \le 2x_1+x_2, \text{ and } 2t_1+3t_2 \le 2(2x_1+3x_2) \tag{5.4}$$

when $t_1 = 2x_1$ and $t_2 = 2x_2$ (see Figure 1). Therefore we have

$$t_1+2t_2 \le 2x_1+4x_2 \tag{5.5}$$

over the region (5.4).

Using (5.5) and condition C2.1, we obtain

$$\begin{aligned} t_1(y+x+t_1+2t_2)^2 &\le (2x_1+x_2)\left(y+\frac{\sigma}{\gamma}+2x_1+4x_2\right)^2 \\ &< \left(y+\frac{\sigma}{\gamma}\right)\left(1-y-\frac{\sigma}{\gamma}-2(2x_1+3x_2)\right)^2 \\ &\le \left(y+\frac{\sigma}{\gamma}+t_1+2t_2\right)(1-y-x-2t_1-3t_2)^2. \end{aligned}$$

Similarly, using (5.5) and condition C2.2, we obtain

$$\begin{aligned} t_2(y+x+t_1+2t_2)^4 &\le 2x_2\left(y+\frac{\sigma}{\gamma}+2x_1+4x_2\right)^4 \\ &< \left(y+\frac{\sigma}{\gamma}\right)^2\left(1-y-\frac{\sigma}{\gamma}-2(2x_1+3x_2)\right)^3 \\ &\le \left(y+\frac{\sigma}{\gamma}+t_1+2t_2\right)^2(1-y-x-2t_1-3t_2)^3. \end{aligned}$$

Hence we have

$$\frac{\partial S}{\partial t_1} \ge \log_q \frac{\left(1-y-\frac{\sigma}{\gamma}-2(2x_1+3x_2)\right)^2\left(y+\frac{\sigma}{\gamma}\right)}{(2x_1+x_2)\left(y+\frac{\sigma}{\gamma}+2x_1+4x_2\right)^2} > 0$$

and

$$\frac{\partial S}{\partial t_2} \geq \log_q \frac{\left(1 - y - \frac{\sigma}{\gamma} - 2(2x_1 + 3x_2)\right)^3 \left(y + \frac{\sigma}{\gamma}\right)^2}{2x_2 \left(y + \frac{\sigma}{\gamma} + 2x_1 + 4x_2\right)^4} > 0$$

for $0 \leq x \leq \frac{\sigma}{\gamma}$ and $0 < t_1, t_2$ with $t_2 \leq 2x_2$, $t_1 \leq 2x_1 + x_2$, and $2t_1 + 3t_2 \leq 2(2x_1 + 3x_2)$.

Then for fixed σ, y, and $0 \leq x \leq \frac{\sigma}{\gamma}$, the function $S(\sigma, y, x, t_1, t_2)$ assumes its maximum over the region (5.4) on the part of the boundary formed by the closed line connecting the two points (see Figure 1)

$$A_1 = (2x_1, 2x_2) \quad \text{and} \quad A_2 = \left(2x_1 + x_2, \frac{4}{3}x_2\right).$$

The direction vector $\overrightarrow{A_2A_1}$ from A_2 to A_1 is parallel to the vector $(-3, 2)$. Hence for fixed σ, y, and $0 \leq x \leq \frac{\sigma}{\gamma}$, the function $S(\sigma, y, x, t_1, t_2)$ is nondecreasing on the closed line from A_2 to A_1 if

$$-3\frac{\partial S}{\partial t_1}(\sigma, y, x, t_1, t_2) + 2\frac{\partial S}{\partial t_2}(\sigma, y, x, t_1, t_2) \geq 0 \tag{5.6}$$

holds for fixed σ, y, and $0 \leq x \leq \frac{\sigma}{\gamma}$ and for each point (t_1, t_2) on the closed line from A_2 to A_1. By straightforward manipulations, we obtain that (5.6) is equivalent to

$$t_1^3\left(y + \frac{\sigma}{\gamma} + t_1 + 2t_2\right) \geq t_2^2(y + x + t_1 + 2t_2)^2. \tag{5.7}$$

We have $t_1 \geq 2x_1$, $t_2 \leq 2x_2$, and $t_1 + 2t_2 \leq 2x_1 + 4x_2$ on the closed line from A_2 to A_1. Therefore using $y + \frac{\sigma}{\gamma} + t_1 + 2t_2 \geq y + x + t_1 + 2t_2$ and condition C4, we see that (5.7) holds. Hence $S(y, \sigma, x, t_1, t_2)$ assumes its maximum at $x = 0$ and $(t_1, t_2) = A_1 = (2x_1, 2x_2)$. It is easy to check that if the assumption (5.3) does not hold, but the assumptions of the proposition do hold, then similar methods also apply and we again have $I_{t,x_1,x_2}(\sigma) = S(\sigma, y, 0, 2x_1, 2x_2)$. This completes the proof. □

Now that we have dealt with the cases $m = 1$ and $m = 2$, we present the generalizations for any $m \geq 1$.

Definition 5.6. *Under Assumption* 1 *(cf. section* 4*), let* $y > 0$, $x_1, x_2, \ldots, x_m$, $\sigma \geq 0$ *be real numbers satisfying*

$$y + 2(2x_1 + 3x_2 + \cdots + (m+1)x_m) + \frac{\sigma}{\gamma} < 1. \tag{5.8}$$

For real numbers $0 \leq x \leq \frac{\sigma}{\gamma}$ *and* $t_1, t_2, \ldots, t_m$ *satisfying*

$$\begin{aligned} &0 \leq t_m \leq 2x_m,\ 0 \leq t_{m-1} \leq 2x_{m-1} + x_m, \ldots, \\ &0 \leq t_1 \leq 2x_1 + (x_2 + x_3 + \cdots + x_m), \end{aligned} \tag{5.9}$$

and

$$2t_1 + 3t_2 + \cdots + (m+1)t_m \leq 2(2x_1 + 3x_2 + \cdots + (m+1)x_m), \tag{5.10}$$

let $S(\sigma, y, x, t_1, t_2, \ldots, t_m)$ be the real-valued function

$$\begin{aligned}
&S(\sigma, y, x, t_1, t_2, \ldots, t_m)\\
&= E_q(t_m) + (1-t_m)E_q\left(\tfrac{t_{m-1}}{1-t_m}\right) + \cdots + (1-(t_2+\cdots+t_m))\,E_q\left(\tfrac{t_1}{1-(t_2+\cdots+t_m)}\right)\\
&+ (1-(t_1+t_2+\cdots+t_m))\,E_q\left(\frac{y+x+(t_1+2t_2+\cdots+mt_m)}{1-(t_1+t_2+\cdots+t_m)}\right)\\
&+\begin{cases}
\left(y+\frac{\sigma}{\gamma}+(t_1+2t_2+\cdots+mt_m)\right)E_q\left(\frac{y+x+(t_1+2t_2+\cdots+mt_m)}{y+\frac{\sigma}{\gamma}+(t_1+2t_2+\cdots+mt_m)}\right) & \\
\qquad\qquad \textit{if } \frac{y+x+(t_1+2t_2+\cdots+mt_m)}{y+\frac{\sigma}{\gamma}+(t_1+2t_2+\cdots+mt_m)} \geq 1-\frac{1}{q}, & \\
\left(y+\frac{\sigma}{\gamma}+(t_1+2t_2+\cdots+mt_m)\right) - (y+x+(t_1+2t_2+\cdots+mt_m))\log_q(q-1) & \\
\qquad\qquad \textit{if } \frac{y+x+(t_1+2t_2+\cdots+mt_m)}{y+\frac{\sigma}{\gamma}+(t_1+2t_2+\cdots+mt_m)} \leq 1-\frac{1}{q}. &
\end{cases}
\end{aligned}$$

Note that by (5.8) *we have* $2\,(2x_1+3x_2+\cdots(m+1)x_m) < 1$, *and hence using* (5.10) *we obtain* $t_1+t_2+\cdots+t_m \leq t_1+\frac{3}{2}t_2+\cdots+\frac{m+1}{2}t_m < \frac{1}{2}$.

We state the generalization of Proposition 5.2, whose proof is similar.

PROPOSITION 5.7. *Under Assumption* 1 *(cf. section* 4*), let* $y > 0$ *and* $x_1, x_2, \ldots, x_m, \sigma \geq 0$ *be real numbers satisfying* (5.8). *For each integer* $i \geq 1$, *let* $r_i = \lfloor (m+y+\frac{\sigma}{\gamma})n_i \rfloor$, $s_i = \lfloor (m+y)n_i \rfloor$, $X_1^{(i)} = \lfloor x_1 n_i \rfloor$, $X_2^{(i)} = \lfloor x_2 n_i \rfloor, \ldots, X_m^{(i)} = \lfloor x_m n_i \rfloor$, *and* $\mathcal{V}_m^{(i)}(r_i, s_i; X_1^{(i)}, X_2^{(i)}, \ldots, X_m^{(i)})$ *be the set of positive divisors of* F_i *defined in Definition* 2.3. *Then for the cardinalities of these sets we have*

$$\limsup_{i\to\infty} \frac{\log_q \left|\mathcal{V}_m^{(i)}(r_i, s_i; X_1^{(i)}, X_2^{(i)}, \ldots, X_m^{(i)})\right|}{n_i} \leq \max\, S(\sigma, y, x, t_1, t_2, \ldots, t_m),$$

where the maximum is over all real numbers x *and* $t_1, t_2, \ldots, t_m$ *satisfying* $0 \leq x \leq \frac{\sigma}{\gamma}$ *and the conditions in* (5.9) *and* (5.10).

Now we generalize Definition 5.3.

DEFINITION 5.8. *Under Assumption* 1 *(cf. section* 4*), let* $y > 0$ *and* $x_1, x_2, \ldots, x_m \geq 0$ *be real numbers such that* $y + 2\,(2x_1+3x_2+\cdots+(m+1)x_m) < 1$. *For* $\sigma \geq 0$ *and* $y + 2\,(2x_1+3x_2+\cdots+(m+1)x_m) + \frac{\sigma}{\gamma} < 1$, *let* $I_{y,x_1,x_2,\ldots,x_m}(\sigma)$ *be the real-valued function of* σ *defined by*

$$I_{y,x_1,x_2,\ldots,x_m}(\sigma) = \max\, S(\sigma, y, x, t_1, t_2, \ldots, t_m),$$

where the maximum is over all real numbers $x, t_1, t_2, \ldots, t_m$ *with* $0 \leq x \leq \frac{\sigma}{\gamma}$ *and* $t_1, t_2, \ldots, t_m$ *satisfying conditions* (5.9) *and* (5.10).

The proof of the next lemma generalizing Lemma 5.4 is also similar.

LEMMA 5.9. *Under the assumptions of Definition* 5.8, *the real-valued function* $I_{y,x_1,x_2,\ldots,x_m}(\sigma)$ *is a strictly increasing function of* σ *on its domain of definition, which is the interval of* σ *such that* $\sigma \geq 0$ *and* $y + 2\,(2x_1+3x_2+\cdots+(m+1)x_m) + \frac{\sigma}{\gamma} < 1$.

Now we are ready to compute $I_{y,x_1,x_2,\ldots,x_m}(\sigma)$ for general m under some conditions. We note that since the region defined by the conditions (5.9) and (5.10) is more complicated in the general case than the one in the case $m = 2$, we need to define new parameters in the following proposition in order to state the result.

PROPOSITION 5.10. *Under the assumptions of Definition* 5.8, *let*

$$\bar{t}_m = 2x_m \quad \textit{and} \quad \bar{t}_\ell = 2x_\ell + \sum_{\nu=\ell+1}^{m} x_\nu \qquad \textit{for } 1 \le \ell \le m-1.$$

Let t_1^* be the real number defined by

$$2t_1^* + \sum_{\ell=2}^{m} (\ell+1)\bar{t}_\ell = 2\sum_{\ell=1}^{m} (\ell+1)x_\ell,$$

and for each $2 \le \ell \le m$, let t_ℓ^* be the real number defined inductively using $t_{\ell-1}^*$ by

$$(\ell+1)t_\ell^* - (\ell+1)\bar{t}_\ell = \ell t_{\ell-1}^* - \ell\bar{t}_{\ell-1}. \tag{5.11}$$

Moreover, let u be the real number depending on $x_1, \dots, x_m$ defined by

$$u = t_1^* + \sum_{\ell=2}^{m} \ell\bar{t}_\ell.$$

Assume also that all of the following conditions hold:

- *C*1: $\frac{\sigma}{\gamma} \le \frac{y}{q-1}$;
- *C*2: *for each* $1 \le \ell \le m$,

$$\bar{t}_\ell \left(y + \frac{\sigma}{\gamma} + u\right)^{2\ell} < \left(1 - y - \frac{\sigma}{\gamma} - 2\sum_{\nu=1}^{m}(\nu+1)x_\nu\right)^{\ell+1} \left(y + \frac{\sigma}{\gamma}\right)^{\ell};$$

- *C*3: $\frac{\sigma}{\gamma}(1-y) < y^2$;
- *C*4: *for each* $1 \le \ell \le m-1$,

$$\left(\bar{t}_{\ell+1}\right)^{\ell+1} \left(y + \frac{\sigma}{\gamma} + u\right) \le \left(t_\ell^*\right)^{\ell+2}.$$

Then we have $I_{y,x_1,x_2,x_3,\dots,x_m}(\sigma) = S(\sigma, y, 0, t_1^*, \bar{t}_2, \bar{t}_3, \dots, \bar{t}_m)$.

Proof. By condition C1 we have

$$\frac{y + x + t_1 + 2t_2 + \cdots + mt_m}{y + \frac{\sigma}{\gamma} + t_1 + 2t_2 + \cdots + mt_m} \ge \frac{y}{y + \frac{\sigma}{\gamma}} \ge 1 - \frac{1}{q}. \tag{5.12}$$

The following identities for partial derivatives hold:

$$\frac{\partial}{\partial x}\left\{(1 - t_1 - t_2 - \cdots - t_m)\, E_q\left(\frac{y + x + t_1 + 2t_2 + \cdots + mt_m}{1 - t_1 - t_2 - \cdots - t_m}\right)\right\}$$
$$= \log_q \frac{1 - y - x - 2t_1 - 3t_2 - \cdots - (m+1)t_m}{y + x + t_1 + 2t_2 + \cdots + mt_m}$$

and

$$\frac{\partial}{\partial x}\left\{\left(y + \frac{\sigma}{\gamma} + t_1 + 2t_2 + \cdots + mt_m\right) E_q\left(\frac{y + x + t_1 + 2t_2 + \cdots + mt_m}{y + \frac{\sigma}{\gamma} + t_1 + 2t_2 + \cdots + mt_m}\right)\right\}$$
$$= \log_q \frac{\frac{\sigma}{\gamma} - x}{y + x + t_1 + 2t_2 + \cdots + mt_m}.$$

Hence using Definition 5.6 and (5.12), we obtain that

$$\frac{\partial S}{\partial x} = \log_q \frac{(1-y-x-2t_1-3t_2-\cdots-(m+1)t_m)\left(\frac{\sigma}{\gamma}-x\right)}{(y+x+t_1+2t_2+\cdots+mt_m)^2}.$$

Therefore if conditions C1 and C3 hold, then

$$\frac{\partial S}{\partial x}(\sigma, y, x, t_1, t_2, \ldots, t_m) < 0 \tag{5.13}$$

for each $0 < x < \frac{\sigma}{\gamma}$ and $t_1, \ldots, t_m$ in the region defined by (5.9) and (5.10).

Now we further assume that

$$x_1 > 0,\ x_2 > 0, \ldots, x_m > 0. \tag{5.14}$$

For $1 \le \ell \le m$, by straightforward manipulations we also obtain the following identities for partial derivatives:

$$\begin{aligned}
&\frac{\partial}{\partial t_\ell}\left\{E_q(t_m) + (1-t_m)E_q\left(\frac{t_{m-1}}{1-t_m}\right) + \cdots + (1-t_2-\cdots-t_m)\right.\\
&\quad\left.\times E_q\left(\frac{t_1}{1-t_2-\cdots-t_m}\right)\right\}\\
&\quad= \log_q(1-t_1-t_2-\cdots-t_m) - \log_q(t_\ell),\\
&\quad\frac{\partial}{\partial t_\ell}\left\{(1-t_1-t_2-\cdots-t_m)E_q\left(\frac{y+x+t_1+2t_2+\cdots+mt_m}{1-t_1-t_2-\cdots-t_m}\right)\right\}\\
&\quad= (\ell+1)\log_q(1-y-x-2t_1-3t_2-\cdots-(m+1)t_m)\\
&\qquad -\log_q(1-t_1-t_2-\cdots-t_m) - \ell\log_q(y+x+t_1+2t_2+\cdots+mt_m),
\end{aligned}$$

and

$$\begin{aligned}
&\frac{\partial}{\partial t_\ell}\left\{\left(y+\tfrac{\sigma}{\gamma}+t_1+2t_2+\cdots+mt_m\right)E_q\left(\frac{y+x+t_1+2t_2+\cdots+mt_m}{y+\frac{\sigma}{\gamma}+t_1+2t_2+\cdots+mt_m}\right)\right\}\\
&\quad= \ell\log_q\left(y+\frac{\sigma}{\gamma}+t_1+2t_2+\cdots+mt_m\right) - \ell\log_q(y+x+t_1+2t_2+\cdots+mt_m).
\end{aligned}$$

Hence using Definition 5.6 and (5.12), for $1 \le \ell \le m$ we obtain that

$$\begin{aligned}
\frac{\partial S}{\partial t_\ell} &= \log_q(1-y-x-2t_1-3t_2-\cdots-(m+1)t_m)^{\ell+1}\\
&\quad+\log_q\left(y+\frac{\sigma}{\gamma}+t_1+2t_2+\cdots+mt_m\right)^{\ell}\\
&\quad-\log_q(y+x+t_1+2t_2+\cdots+mt_m)^{2\ell} - \log_q t_\ell.
\end{aligned} \tag{5.15}$$

Now we also assume that for the real number u defined in the statement of the proposition we have

$$u = \max\left(t_1 + 2t_2 + \cdots + mt_m\right), \tag{5.16}$$

where the maximum is over the region defined by the conditions (5.9) and (5.10). Later in this proof, we will show that the assumption (5.16) holds.

Using (5.12), (5.15), (5.16), and condition C2, as in the proof of Proposition 5.5, we obtain that for each $1 \leq \ell \leq m$,

$$\frac{\partial S}{\partial t_\ell}(\sigma, y, x, t_1, \ldots, t_m) > 0$$

holds for $0 \leq x \leq \frac{\sigma}{\gamma}$ and the real numbers $0 < t_1, \ldots, t_m$ satisfying the conditions (5.9) and (5.10). This implies that for each $0 \leq x \leq \frac{\sigma}{\gamma}$, $S(\sigma, y, x, t_1, \ldots, t_m)$ assumes its maximum over the region defined by (5.9) and (5.10) on the closed set, forming a part of the boundary of the region, defined by the conditions

$$0 \leq t_\ell \leq \bar{t}_\ell \qquad \text{for } 1 \leq \ell \leq m \tag{5.17}$$

and

$$\sum_{\ell=1}^{m} (\ell+1) t_\ell = 2 \sum_{\ell=1}^{m} (\ell+1) x_\ell, \tag{5.18}$$

where $\bar{t}_\ell$ is defined in the statement of the proposition.

For each $1 \leq \ell \leq m$, it follows from the definition of t_ℓ^* in the statement of the proposition that t_ℓ^* is the smallest value of the parameter t_ℓ over the closed set defined by the conditions (5.17) and (5.18). For each $1 \leq \ell \leq m$, let A_ℓ be the point of the $(t_1, \ldots, t_m)$-space given by

$$A_\ell = (t_1, \ldots, t_m), \qquad \text{where } t_\ell = t_\ell^* \text{ and } t_\nu = \bar{t}_\nu \text{ for } \nu \in \{1, \ldots, m\} \setminus \{\ell\}.$$

We observe that the points $A_1, A_2, \ldots, A_m$ are the corners of the closed set given by (5.17) and (5.18).

For each $1 \leq \ell \leq m-1$, the direction vector $\overrightarrow{A_{\ell+1}A_\ell}$ from $A_{\ell+1}$ to A_ℓ in the $(t_1, \ldots, t_m)$-space is

$$\overrightarrow{A_{\ell+1}A_\ell} = \left(\underbrace{0, \ldots, 0}_{\ell-1 \text{ times}}, t_\ell^* - \bar{t}_\ell, \bar{t}_{\ell+1} - t_{\ell+1}^*, \underbrace{0, \ldots, 0}_{m-\ell-1 \text{ times}} \right).$$

Using (5.11) we observe that for each $1 \leq \ell \leq m-1$, the direction vector $\overrightarrow{A_{\ell+1}A_\ell}$ is parallel to the vector

$$\left(\underbrace{0, \ldots, 0}_{\ell-1 \text{ times}}, -(\ell+2), \ell+1, \underbrace{0, \ldots, 0}_{m-\ell-1 \text{ times}} \right) \tag{5.19}$$

in the $(t_1, \ldots, t_m)$-space.

If for each $1 \leq \ell \leq m-1$ the inequality

$$\overrightarrow{A_{\ell+1}A_\ell} \cdot \left(\frac{\partial S}{\partial t_1}, \ldots, \frac{\partial S}{\partial t_m} \right)(\sigma, y, x, t_1, \ldots, t_m) \geq 0 \tag{5.20}$$

for the standard inner product of vectors in the $(t_1, \ldots, t_m)$-space holds for each $0 \leq x \leq \frac{\sigma}{\gamma}$ and $t_1, \ldots, t_m$ satisfying (5.17) and (5.18), then $S(\sigma, y, x, t_1, \ldots, t_m)$ is non-decreasing in the directions from A_m to A_{m-1}, from A_{m-1} to A_{m-2}, ..., and from A_2

to A_1. This implies that if (5.20) holds, then for each $0 \le x \le \frac{\sigma}{\gamma}$, $S(\sigma, y, x, t_1, \dots, t_m)$ assumes its maximum at A_1. Using (5.19), we obtain that (5.20) is equivalent to

$$(\ell+1)\frac{\partial S}{\partial t_{\ell+1}}(\sigma, y, x, t_1, \dots, t_m) \ge (\ell+2)\frac{\partial S}{\partial t_\ell}(\sigma, y, x, t_1, \dots, t_m). \tag{5.21}$$

Using (5.15), (5.21), and some straightforward manipulations, we observe that (5.20) holds if

$$\begin{aligned}&(t_\ell)^{\ell+2}\left(y + \tfrac{\sigma}{\gamma} + t_1 + 2t_2 + \cdots + mt_m\right)\\ &\quad \ge (t_{\ell+1})^{\ell+1}\left(y + x + t_1 + 2t_2 + \cdots + mt_m\right)^2.\end{aligned} \tag{5.22}$$

Using the fact that $y + \frac{\sigma}{\gamma} + t_1 + 2t_2 + \cdots + mt_m \ge y + x + t_1 + 2t_2 + \cdots + mt_m$, the assumption (5.16), and the condition C4, as in the proof of Proposition 5.5, we obtain that (5.22) holds, and hence for each $0 \le x \le \frac{\sigma}{\gamma}$, $S(\sigma, y, x, t_1, \dots, t_m)$ assumes its maximum at A_1.

Next we prove the claim (5.16). Note that the gradient of the m-variable function $f(t_1, t_2, \dots, t_m) = t_1 + 2t_2 + \cdots + mt_m$ is $(1, 2, \dots, m)$ at any point of the $(t_1, \dots, t_m)$-space. For each $1 \le \ell \le m-1$, from the standard inner product with the vector in (5.19) we obtain

$$\begin{aligned}&(1, 2, \dots, m) \cdot \left(\underbrace{0, \dots, 0}_{\ell-1 \text{ times}}, -(\ell+2), \ell+1, \underbrace{0, \dots, 0}_{m-\ell-1 \text{ times}}\right)\\ &= -(\ell+2)\ell + (\ell+1)^2 = 1 > 0.\end{aligned}$$

Then, as the function $S(\sigma, y, x, t_1, \dots, t_m)$, the function $f(t_1, \dots, t_m)$ assumes its maximum at A_1 and hence the claim (5.16) holds. Finally, using (5.13) we complete the proof under the assumption (5.14). As in the proof of Proposition 5.5, we observe that if the assumption (5.14) does not hold, but the assumptions of the proposition do hold, then similar methods also apply and we again have $I_{y,x_1,x_2,x_3,\dots,x_m}(\sigma) = S(\sigma, y, 0, t_1^*, \bar{t}_2, \bar{t}_3, \dots, \bar{t}_m)$. This completes the proof. □

Remark 5.11. We note that Proposition 5.10 reduces to Proposition 5.5 and Corollary 4.9 if $m = 2$ and $m = 1$, respectively.

6. Asymptotic bounds for codes. In this section we prove our main results (Theorem 6.3 and Corollary 6.4) which establish improved lower bounds on $\alpha_q(\delta)$ and $\alpha_q^{\text{lin}}(\delta)$. We use a (nonnegative) real-valued function $\Psi(y, x_1, \dots, x_m)$ given in Definition 6.2. Moreover, a well-known result stated in Proposition 6.1 is used in the proof of Theorem 6.3 for the existence of a sequence of distinguished divisors on the basis of Proposition 2.4.

We first state our main assumption, which is like Assumption 1 in section 4, but introduces more notation.

Assumption 1′. Assume that $(F_i/\mathbb{F}_q)_{i=1}^{\infty}$ is a sequence of global function fields with full constant field $\mathbb{F}_q$, with $g_i \to \infty$ as $i \to \infty$, and with $\limsup_{i\to\infty} \frac{n_i}{g_i} = \gamma > 0$, where n_i and g_i denote the number of rational places and the genus of F_i, respectively. For each $l \ge 1$, let $\gamma_l \ge 0$ be a real number with $\liminf_{i\to\infty} \frac{B_{i,l}}{g_i} \ge \gamma_l$, where $B_{i,l}$ is the number of degree l places of F_i. Using a suitable subsequence of $(F_i/\mathbb{F}_q)_{i=1}^{\infty}$, we can take $\gamma_1 = \gamma$.

The following well-known result will be useful.

PROPOSITION 6.1. *Under Assumption* $1'$ *we have*

$$\liminf_{i\to\infty} \frac{\log_q h_i}{n_i} \geq \frac{1}{\gamma}\left[1+\sum_{l=1}^{\infty}\gamma_l \log_q \frac{q^l}{q^l-1}\right],$$

where h_i *is the class number of* F_i.

Proof. This follows from [9, Corollary 2] (see also [10, Exercise 2.3.27]). □

Now we introduce an important function based on the function $I_{y,x_1,\ldots,x_m}(\sigma)$ defined in Definition 5.8. In the next definition we use the fact that $I_{y,x_1,\ldots,x_m}(\sigma)$ is an increasing function on its domain of definition; see Lemma 5.9.

DEFINITION 6.2. *Under Assumption* $1'$ *and for real numbers* $y > 0$ *and* $x_1, \ldots, x_m \geq 0$ *with* $y + 2(2x_1 + 3x_2 + \cdots + (m+1)x_m) < 1$, *let* $\Psi(y, x_1, \ldots, x_m)$ *be the real-valued function of* $y, x_1, \ldots, x_m$ *defined by*

$$\Psi(y,x_1,\ldots,x_m)=\begin{cases} I^{-1}_{y,x_1,\ldots,x_m}\left(\dfrac{1}{\gamma}\left[1+\displaystyle\sum_{l=1}^{\infty}\gamma_l\log_q\frac{q^l}{q^l-1}\right]\right) & \\ \qquad\qquad \textit{if } \displaystyle\lim_{\sigma\to\theta^-} I_{y,x_1,\ldots,x_m}(\sigma) > \frac{1}{\gamma}\left[1+\sum_{l=1}^{\infty}\gamma_l\log_q\frac{q^l}{q^l-1}\right], & \\ 0 \qquad\qquad \textit{otherwise}, & \end{cases}$$

where $\theta = \gamma\,(1 - y - 2(2x_1 + 3x_2 + \cdots + (m+1)x_m))$.

Now we are ready to establish our main results. We recall that the functions $\alpha_q(\delta)$ and $\alpha_q^{\text{lin}}(\delta)$ are defined in (1.1) and (1.2), respectively.

THEOREM 6.3. *Under Assumption* $1'$, *let* $x_1, \ldots, x_m \geq 0$ *be real numbers with* $2(2x_1 + 3x_2 + \cdots + (m+1)x_m) < 1$. *For each real number* $0 < \delta < 1 - 2(2x_1 + 3x_2 + \cdots + (m+1)x_m)$ *we have*

$$\begin{aligned}\alpha_q(\delta) \geq R_{\{\gamma_l\},x_1,\ldots,x_m}(\delta) &:= 1-\delta-\frac{1}{\gamma}+(x_1+\cdots+x_m)\log_q(q-1)\\ &-(x_1\log_q x_1+\cdots+x_m\log_q x_m)-(1-(x_1+\cdots+x_m))\log_q(1-(x_1+\cdots+x_m))\\ &-(4x_1+5x_2+\cdots+(m+3)x_m)\\ &+\frac{1}{\gamma}\Psi\Big(1-\delta-2(2x_1+3x_2+\cdots+(m+1)x_m),x_1,x_2,\ldots,x_m\Big).\end{aligned}$$

Proof. Let $y = 1 - \delta - 2(2x_1 + 3x_2 + \cdots + (m+1)x_m)$ and $\sigma = \Psi(y, x_1, \ldots, x_m)$. If $R_{\{\gamma_l\},x_1,\ldots,x_m}(\delta) \leq 0$, then the statement of the theorem is trivial. If $\sigma = 0$ and $R_{\{\gamma_l\},x_1,\ldots,x_m}(\delta) > 0$, then the theorem follows from [4, Theorem 5.1]. Indeed, in this case let $r_i = \lfloor (m+y)n_i \rfloor$ for $i \geq 1$. As $R_{\{\gamma_l\},x_1,\ldots,x_m}(\delta) > 0$, the conditions of [4, Theorem 5.1] are satisfied. Then using [4, Theorem 5.1] for sufficiently large i with a divisor of degree r_i of F_i, we obtain a sequence of q-ary codes proving the theorem in this case. The computation of the parameters is similar to the case where $\sigma > 0$ and $R_{\{\gamma_l\},x_1,\ldots,x_m}(\delta) > 0$, which is explained in detail below.

Now we consider the remaining case where $\sigma > 0$ and $R_{\{\gamma_l\},x_1,\ldots,x_m}(\delta) > 0$. Let

$0 < \epsilon < \sigma$ be a real number satisfying

$$\begin{aligned} & y + (x_1 + \cdots + x_m)\log_q(q-1) - (x_1 \log_q x_1 + \cdots + x_m \log_q x_m) \\ & \quad - (1 - (x_1 + \cdots + x_m))\log_q(1 - (x_1 + \cdots + x_m)) \\ & \quad +(x_2 + 2x_3 + \cdots + (m-1)x_m) \\ & \quad > \frac{1 - (\sigma - \epsilon)}{\gamma}. \end{aligned} \tag{6.1}$$

For $i \geq 1$, let

$$\begin{aligned} & r_i = \left\lfloor \left(m + y + \tfrac{\sigma - \epsilon}{\gamma} \right) n_i \right\rfloor, s_i = \lfloor (m+y) n_i \rfloor, \\ & X_1^{(i)} = \lfloor x_1 n_i \rfloor, X_2^{(i)} = \lfloor x_2 n_i \rfloor, \ldots, X_m^{(i)} = \lfloor x_m n_i \rfloor. \end{aligned} \tag{6.2}$$

For sufficiently large i, by Propositions 5.7 and 6.1, the hypotheses of Proposition 2.4 for the global function field F_i with r_i, s_i, and $X_1^{(i)}, \ldots, X_m^{(i)}$ as in (6.2) are satisfied. Let G_i be the divisor of F_i given by Proposition 2.4 with these parameters for sufficiently large i.

Note that

$$\begin{aligned} & \liminf_{i \to \infty} \frac{\log_q |M(x_1, \ldots, x_m; \mathbf{0})|}{n_i} \\ & \quad \geq (x_1 + \cdots + x_m)\log_q(q-1) - (x_1 \log_q x_1 + \cdots + x_m \log_q x_m) \\ & \qquad - (1 - (x_1 + \cdots + x_m))\log_q(1 - (x_1 + \cdots + x_m)) \\ & \qquad + (x_2 + 2x_3 + \cdots + (m-1)x_m) \end{aligned} \tag{6.3}$$

(see [4, section 4]). Since we have (6.1), using the divisor G_i of the global function field F_i for sufficiently large i, Theorem 2.9, and (6.3), we obtain a sequence of q-ary codes $\{C_i\}_{i=1}^{\infty}$ of lengths $\{n_i\}_{i=1}^{\infty}$, respectively, such that $n_i \to \infty$ as $i \to \infty$ by Assumption $1'$ as well as

$$\begin{aligned} & \liminf_{i \to \infty} \frac{\log_q |C_i|}{n_i} \\ & \quad \geq y + \frac{\sigma - \epsilon}{\gamma} - \frac{1}{\gamma} \\ & \qquad + (x_1 + \cdots + x_m)\log_q(q-1) - (x_1 \log_q x_1 + \cdots + x_m \log_q x_m) \\ & \qquad - (1 - (x_1 + \cdots + x_m))\log_q(1 - (x_1 + \cdots + x_m)) \\ & \qquad + (x_2 + 2x_3 + \cdots + (m-1)x_m) \\ & \quad = 1 - \delta - 2(2x_1 + 3x_2 + \cdots + (m+1)x_m) + \frac{\sigma - \epsilon}{\gamma} - \frac{1}{\gamma} \\ & \qquad + (x_1 + \cdots + x_m)\log_q(q-1) - (x_1 \log_q x_1 + \cdots + x_m \log_q x_m) \\ & \qquad - (1 - (x_1 + \cdots + x_m))\log_q(1 - (x_1 + \cdots + x_m)) \\ & \qquad + (x_2 + 2x_3 + \cdots + (m-1)x_m) \\ & \quad = R_{\{\gamma_l\}, x_1, \ldots, x_m}(\delta) - \frac{\epsilon}{\gamma} \end{aligned}$$

and

$$\liminf_{i \to \infty} \frac{d(C_i)}{n_i} \geq \delta.$$

Using the fact that $\alpha_q(\delta)$ is a nonincreasing function of δ, we get

$$\alpha_q(\delta) \geq R_{\{\gamma_l\},x_1,\ldots,x_m}(\delta) - \frac{\epsilon}{\gamma}.$$

Letting $\epsilon \to 0^+$ completes the proof. □

COROLLARY 6.4. *Under Assumption $1'$, for each real number $0 < \delta < 1$ we have*

$$\alpha_q^{\text{lin}}(\delta) \geq R_{\{\gamma_l\}}^{\text{lin}}(\delta) := 1 - \delta - \frac{1}{\gamma} + \frac{1}{\gamma}\Psi(1-\delta, 0).$$

Proof. Taking $m = 1$ and using similar methods as in the proof of Theorem 6.3, but applying Corollary 2.10 instead of Theorem 2.9, we obtain the desired result. □

7. Examples. In this section we demonstrate that Theorem 6.3 and Corollary 6.4 yield improvements on the lower bounds for $\alpha_q(\delta)$ and $\alpha_q^{\text{lin}}(\delta)$ at least for certain values of q and certain values of δ. In our examples we use well-known values for $\gamma = \gamma_1$ and take $\gamma_l = 0$ for $l \geq 2$ for the parameters defined in Assumption $1'$. Nevertheless, we note that there is a potential for the demonstration of further improvements by Theorem 6.3 and Corollary 6.4 using $\gamma_l > 0$ for $l = 1$ and some $l \geq 2$ when q is not a square (the situation is different when q is a square; cf. [9, Corollary 1]).

For simplicity of notation, for $\gamma = \gamma_1$ and $\gamma_l = 0$ for $l \geq 2$, we denote the lower bounds of Theorem 6.3 and Corollary 6.4 by $R_{\gamma,x_1,\ldots,x_m}(\delta)$ and $R_\gamma^{\text{lin}}(\delta)$, respectively. In the examples below, the required values of these two functions are computed by using Definition 6.2 and Proposition 5.10.

Let $R_{NO2,\gamma,x}(\delta)$ denote the lower bound in [5, Theorem 5.1]. Moreover, let $R_{X,\gamma}^{\text{lin}}(\delta)$ denote Xing's lower bound for $\alpha_q^{\text{lin}}(\delta)$ in [13] (see also [5, Theorem 4.6]).

Example 7.1. Let $q = 2^6$, $\gamma = \gamma_1 = \sqrt{q} - 1$, $\gamma_l = 0$ for $l \geq 2$, and

$$\begin{aligned}\delta &= \frac{1376386844325023892952150398483338159773141 2559044}{46065097831342932365531985486767649347321318605709}\\ &= 0.29879169026501515839\ldots.\end{aligned}$$

In [5, Example 5.2], using $x = 10^{-13}$ it has been obtained that

$$\alpha_q(\delta) \geq R_{NO2,\gamma,x}(\delta) = 0.55835371587781529071\ldots,$$

and has been demonstrated that $R_{NO2,\gamma,x}(\delta) - R_{X,\gamma}^{\text{lin}}(\delta) \geq 7.3387 \cdot 10^{-15}$.

By Corollary 6.4 we obtain that

$$\alpha_q^{\text{lin}}(\delta) \geq R_\gamma^{\text{lin}}(\delta) = 0.55835395724081743804\ldots.$$

Note that $R_\gamma^{\text{lin}}(\delta) - R_{NO2,\gamma,x}(\delta) \geq 2.4136300214732 \cdot 10^{-7}$, and $R_\gamma^{\text{lin}}(\delta)$ is better than $R_{X,\gamma}^{\text{lin}}(\delta)$. Hence we have an improvement on the lower bound for $\alpha_q^{\text{lin}}(\delta)$ compared to Xing's bound in [13].

By Theorem 6.3 with $x_1 = 3.41 \cdot 10^{-16}$, $x_2 = 1.0634 \cdot 10^{-23}$, and $x_3 = 1.93 \cdot 10^{-31}$, we obtain $\alpha_q(\delta) \geq R_{\gamma,x_1,x_2,x_3}(\delta)$, where

$$R_{\gamma,x_1,x_2,x_3}(\delta) - R_\gamma^{\mathrm{lin}}(\delta) \geq 2.711029 \cdot 10^{-17}.$$

Hence $R_{\gamma,x_1,x_2,x_3}(\delta)$ gives a further improvement on the lower bound for $\alpha_q(\delta)$. Note that $R_{\gamma,x_1,x_2,x_3}(\delta)$ yields an improvement on $R_{NO2,\gamma,x}(\delta)$ of the order 10^{-7}, whereas Maharaj [3] obtained only an improvement of the order 10^{-15}.

Now let

$$\begin{aligned}\delta &= \frac{32301229388092693436010481501934267749589906046665}{46065097831342932365531985486767649347321318605709} \\ &= 0.70120830973498484160 \ldots.\end{aligned}$$

In [5, Example 5.2], using $x = 10^{-13}$ it has been obtained that

$$\alpha_q(\delta) \geq R_{NO2,\gamma,x}(\delta) = 0.15593709640785805503 \ldots,$$

and has been demonstrated that $R_{NO2,\gamma,x}(\delta) - R_{X,\gamma}^{\mathrm{lin}}(\delta) \geq 1.97862 \cdot 10^{-14}$.

By Corollary 6.4 we obtain that

$$\alpha_q^{\mathrm{lin}}(\delta) \geq R_\gamma^{\mathrm{lin}}(\delta) = 0.15593754394482448829 \ldots.$$

Note that $R_\gamma^{\mathrm{lin}}(\delta) - R_{NO2,\gamma,x}(\delta) \geq 4.4753696643325 \cdot 10^{-7}$, hence $R_\gamma^{\mathrm{lin}}(\delta)$ is better than $R_{X,\gamma}^{\mathrm{lin}}(\delta)$. Hence we have an improvement on the lower bound for $\alpha_q^{\mathrm{lin}}(\delta)$ compared to Xing's bound in [13].

By Theorem 6.3 with $x_1 = 3.89 \cdot 10^{-18}$, $x_2 = 1.98 \cdot 10^{-26}$, and $x_3 = 5.87 \cdot 10^{-35}$, we obtain $\alpha_q(\delta) \geq R_{\gamma,x_1,x_2,x_3}(\delta)$, where

$$R_{\gamma,x_1,x_2,x_3}(\delta) - R_\gamma^{\mathrm{lin}}(\delta) \geq 2.592642 \cdot 10^{-19}.$$

Hence $R_{\gamma,x_1,x_2,x_3}(\delta)$ gives a further improvement on the lower bound for $\alpha_q(\delta)$. Note that $R_{\gamma,x_1,x_2,x_3}(\delta)$ yields again an improvement on $R_{NO2,\gamma,x}(\delta)$ of the order 10^{-7}, whereas Maharaj [3] obtained only an improvement of the order 10^{-14}.

Example 7.2. Let $q = 7^2$, $\gamma = \gamma_1 = \sqrt{q} - 1$, $\gamma_l = 0$ for $l \geq 2$, and

$$\begin{aligned}\delta &= \frac{7334559589562321721169749749908497945081695123431}{18755194537338788993696079784908084949457099261873} \\ &= 0.39106816913897159912 \ldots.\end{aligned}$$

In [5, Example 5.3], using $x = 10^{-13}$ it has been obtained that

$$\alpha_q(\delta) \geq R_{NO2,\gamma,x}(\delta) = 0.44226734872224546020 \ldots,$$

and has been demonstrated that $R_{NO2,\gamma,x}(\delta) - R_{X,\gamma}^{\mathrm{lin}}(\delta) \geq 6.57561 \cdot 10^{-14}$.

By Corollary 6.4 we obtain that

$$\alpha_q^{\mathrm{lin}}(\delta) \geq R_\gamma^{\mathrm{lin}}(\delta) = 0.44226758374884970747 \ldots.$$

Note that $R_\gamma^{\mathrm{lin}}(\delta) - R_{NO2,\gamma,x}(\delta) \geq 2.3502660424726 \cdot 10^{-7}$, and $R_\gamma^{\mathrm{lin}}(\delta)$ is better than $R_{X,\gamma}^{\mathrm{lin}}(\delta)$. Hence we have an improvement on the lower bound for $\alpha_q^{\mathrm{lin}}(\delta)$ compared to Xing's bound in [13].

By Theorem 6.3 with $x_1 = 1.93 \cdot 10^{-13}$, $x_2 = 1.53 \cdot 10^{-19}$, and $x_3 = 7.08 \cdot 10^{-26}$, we obtain $\alpha_q(\delta) \geq R_{\gamma,x_1,x_2,x_3}(\delta)$, where

$$R_{\gamma,x_1,x_2,x_3}(\delta) - R_{\gamma}^{\mathrm{lin}}(\delta) \geq 1.857062 \cdot 10^{-14}.$$

Hence $R_{\gamma,x_1,x_2,x_3}(\delta)$ gives a further improvement on the lower bound for $\alpha_q(\delta)$. Note that $R_{\gamma,x_1,x_2,x_3}(\delta)$ yields an improvement on $R_{NO2,\gamma,x}(\delta)$ of the order 10^{-7}, whereas Maharaj [3] obtained only an improvement of the order 10^{-12}.

Now let

$$\begin{aligned}\delta &= \frac{11420634947776467272526330034999587004375404138442}{18755194537338788993696079784908084949457099261873} \\ &= 0.60893183086102840087\ldots.\end{aligned}$$

In [5, Example 5.3], using $x = 10^{-13}$ it has been obtained that

$$\alpha_q(\delta) \geq R_{NO2,\gamma,x}(\delta) = 0.22440368700019503856\ldots,$$

and has been demonstrated that $R_{NO2,\gamma,x}(\delta) - R_{X,\gamma}^{\mathrm{lin}}(\delta) \geq 7.21362 \cdot 10^{-14}$.

By Corollary 6.4 we obtain that

$$\alpha_q^{\mathrm{lin}}(\delta) \geq R_{\gamma}^{\mathrm{lin}}(\delta) = 0.22440401150099750683\ldots.$$

Note that $R_{\gamma}^{\mathrm{lin}}(\delta) - R_{NO2,\gamma,x}(\delta) \geq 3.2450080246826 \cdot 10^{-7}$, and $R_{\gamma}^{\mathrm{lin}}(\delta)$ is better than $R_{X,\gamma}^{\mathrm{lin}}(\delta)$. Hence we have an improvement on the lower bound for $\alpha_q^{\mathrm{lin}}(\delta)$ compared to Xing's bound in [13].

By Theorem 6.3 with $x_1 = 5.86 \cdot 10^{-14}$, $x_2 = 3.207 \cdot 10^{-20}$, and $x_3 = 1.02 \cdot 10^{-26}$, we obtain $\alpha_q(\delta) \geq R_{\gamma,x_1,x_2,x_3}(\delta)$, where

$$R_{\gamma,x_1,x_2,x_3}(\delta) - R_{\gamma}^{\mathrm{lin}}(\delta) \geq 5.258306 \cdot 10^{-15}.$$

Hence $R_{\gamma,x_1,x_2,x_3}(\delta)$ gives a further improvement on the lower bound for $\alpha_q(\delta)$. Note that $R_{\gamma,x_1,x_2,x_3}(\delta)$ yields again an improvement on $R_{NO2,\gamma,x}(\delta)$ of the order 10^{-7}, whereas Maharaj [3] obtained only an improvement of the order 10^{-12}.

Example 7.3. Let $q = 2^{21}$, $\gamma = \gamma_1 = \frac{2(q^{2/3}-1)}{q^{1/3}+2}$ (see (1.5)), $\gamma_l = 0$ for $l \geq 2$, and

$$\begin{aligned}\delta &= \frac{10343234848654524734637261103098140324984446010098}{99621193732964014413326435515634059733734238550355} \\ &= 0.01038256465424386359\ldots.\end{aligned}$$

In [5, Example 5.4], using $x = 10^{-60}$ it has been obtained that

$$\alpha_q(\delta) \geq R_{NO2,\gamma,x}(\delta) = 0.98564990803085654665\ldots,$$

and has been demonstrated that $R_{NO2,\gamma,x}(\delta) - R_{X,\gamma}^{\mathrm{lin}}(\delta) \geq 2.1335699248 \cdot 10^{-61}$.

By Corollary 6.4 we obtain that

$$\alpha_q^{\mathrm{lin}}(\delta) \geq R_{\gamma}^{\mathrm{lin}}(\delta) = 0.98564990803085654673\ldots.$$

Note that $R_{\gamma}^{\mathrm{lin}}(\delta) - R_{NO2,\gamma,x}(\delta) \geq 7 \cdot 10^{-20}$, and $R_{\gamma}^{\mathrm{lin}}(\delta)$ is better than $R_{X,\gamma}^{\mathrm{lin}}(\delta)$. Hence we have an improvement on the lower bound for $\alpha_q^{\mathrm{lin}}(\delta)$ compared to Xing's bound in [13].

By Theorem 6.3 with $x_1 = 6.29 \cdot 10^{-65}$ and $x_2 = 7.09 \cdot 10^{-97}$, we obtain $\alpha_q(\delta) \geq R_{\gamma,x_1,x_2}(\delta)$, where

$$R_{\gamma,x_1,x_2}(\delta) - R_{\gamma}^{\text{lin}}(\delta) \geq 1.261672 \cdot 10^{-66}.$$

Hence $R_{\gamma,x_1,x_2}(\delta)$ gives a further improvement on the lower bound for $\alpha_q(\delta)$.

Now let

$$\delta = \frac{985868702480985619398627094053242457012357925402 57}{99621193732964014413326435515634059733734238550355}$$

$$= 0.98961743534575613640\ldots.$$

In [5, Example 5.4], using $x = 10^{-60}$ it has been obtained that

$$\alpha_q(\delta) \geq R_{NO2,\gamma,x}(\delta) = 0.00641503733934427385\ldots,$$

and has been demonstrated that $R_{NO2,\gamma,x}(\delta) - R_{X,\gamma}^{\text{lin}}(\delta) \geq 4.2225689802 \cdot 10^{-61}$.

By Corollary 6.4 we obtain that

$$\alpha_q^{\text{lin}}(\delta) \geq R_{\gamma}^{\text{lin}}(\delta) = 0.00641503733934427410\ldots.$$

Note that $R_{\gamma}^{\text{lin}}(\delta) - R_{NO2,\gamma,x}(\delta) \geq 2.4 \cdot 10^{-19}$, and $R_{\gamma}^{\text{lin}}(\delta)$ is better than $R_{X,\gamma}^{\text{lin}}(\delta)$. Hence we have an improvement on the lower bound for $\alpha_q^{\text{lin}}(\delta)$ compared to Xing's bound in [13].

By Theorem 6.3 with $x_1 = 6.5 \cdot 10^{-86}$ and $x_2 = 2.4 \cdot 10^{-127}$, we obtain $\alpha_q(\delta) \geq R_{\gamma,x_1,x_2}(\delta)$, where

$$R_{\gamma,x_1,x_2}(\delta) - R_{\gamma}^{\text{lin}}(\delta) \geq 9.103449 \cdot 10^{-88}.$$

Hence $R_{\gamma,x_1,x_2}(\delta)$ gives a further improvement on the lower bound for $\alpha_q(\delta)$.

8. Conclusions. In this paper we improved on various lower bounds for $\alpha_q(\delta)$ and $\alpha_q^{\text{lin}}(\delta)$ that were established by several authors in recent years. These improvements were obtained by combining, for the first time, three important methods in the area: (i) distinguished divisors of global function fields (see Proposition 2.4); (ii) local expansions of arbitrary length for elements of global function fields (see (2.2) and (2.3)); (iii) averaging arguments such as those leading to (2.8). This has come at the cost of considerable complications in the analysis. To get the most out of this combination of methods, several technical innovations had to be introduced (see sections 3, 4, and 5).

Since we tried our best to optimize the combination of the three methods above, we tend to believe that further progress along these lines will have to be based on completely new ideas. Stichtenoth and Xing [8] introduced a different approach that led to an alternative proof of (1.6) and Maharaj [3] pursued this approach further. However, for all the examples in [3] the bounds in the present paper are better (see section 7). It remains to be seen whether the approach in [8] and [3], if refined further, has the potential to improve the bounds in the present paper.

Acknowledgments. The authors would like to thank the anonymous referees for their useful suggestions. The second author would like to thank Temesek Laboratories and the Department of Mathematics at the National University of Singapore for their hospitality during the writing of this paper.

REFERENCES

[1] J. Bezerra, A. Garcia, and H. Stichtenoth, *An explicit tower of function fields over cubic finite fields and Zink's lower bound*, J. Reine Angew. Math., 589 (2005), pp. 159–199.

[2] N. D. Elkies, *Excellent nonlinear codes from modular curves*, in STOC'01, Proceedings of the 33rd Annual ACM Symposium on Theory of Computing (Hersonissos, Greece, 2001), ACM Press, New York, 2001, pp. 200–208.

[3] H. Maharaj, *A note on further improvements of the TVZ-bound*, IEEE Trans. Inform. Theory, 53 (2007), pp. 1210–1214.

[4] H. Niederreiter and F. Özbudak, *Constructive asymptotic codes with an improvement on the Tsfasman-Vlăduţ-Zink and Xing bounds*, in Coding, Cryptography and Combinatorics K. Q. Feng, H. Niederreiter, and C. P. Xing, eds., Progr. Comput. Sci. Appl. Logic 23, Birkhäuser, Basel, 2004, pp. 259–275.

[5] H. Niederreiter and F. Özbudak, *Further improvements on asymptotic bounds for codes using distinguished divisors*, Finite Fields Appl., 13 (2007), pp. 423–443.

[6] H. Niederreiter and C. P. Xing, *Rational Points on Curves over Finite Fields: Theory and Applications*, Cambridge University Press, Cambridge, UK, 2001.

[7] H. Stichtenoth, *Algebraic Function Fields and Codes*, Springer-Verlag, Berlin, 1993.

[8] H. Stichtenoth and C. P. Xing, *Excellent nonlinear codes from algebraic function fields*, IEEE Trans. Inform. Theory, 51 (2005), pp. 4044–4046.

[9] M. A. Tsfasman, *Some remarks on the asymptotic number of points*, in Coding Theory and Algebraic Geometry, H. Stichtenoth and M. A. Tsfasman, eds., Lecture Notes in Math. 1518, Springer, Berlin, 1992, pp. 178–192.

[10] M. A. Tsfasman and S. G. Vlăduţ, *Algebraic-Geometric Codes*, Kluwer, Dordrecht, The Netherlands, 1991.

[11] M. A. Tsfasman, S. G. Vlăduţ, and T. Zink, *Modular curves, Shimura curves, and Goppa codes, better than Varshamov-Gilbert bound*, Math. Nachr., 109 (1982), pp. 21–28.

[12] S. G. Vlăduţ, *An exhaustion bound for algebro-geometric "modular" codes*, Problemy Peredachi Informatsii, 23 (1987), pp. 28–41.

[13] C. P. Xing, *Algebraic-geometry codes with asymptotic parameters better than the Gilbert-Varshamov and the Tsfasman-Vlăduţ-Zink bounds*, IEEE Trans. Inform. Theory, 47 (2001), pp. 347–352.

[14] C. P. Xing, *Nonlinear codes from algebraic curves improving the Tsfasman-Vlăduţ-Zink bound*, IEEE Trans. Inform. Theory, 49 (2003), pp. 1653–1657.

SIAM J. Discrete Math.
Vol. 21, No. 4, pp. 900–929

THE COMPLEXITY OF THE LIST PARTITION PROBLEM FOR GRAPHS[*]

KATHIE CAMERON[†], ELAINE M. ESCHEN[‡], CHíNH T. HOÀNG[§], AND R. SRITHARAN[¶]

Abstract. The k-partition problem is as follows: Given a graph G and a positive integer k, partition the vertices of G into at most k parts $A_1, A_2, \ldots, A_k$, where it may be specified that A_i induces a stable set, a clique, or an arbitrary subgraph, and pairs A_i, A_j $(i \neq j)$ be completely nonadjacent, completely adjacent, or arbitrarily adjacent. The list k-partition problem generalizes the k-partition problem by specifying for each vertex x, a list $L(x)$ of parts in which it is allowed to be placed. Many well-known graph problems can be formulated as list k-partition problems: e.g., 3-colorability, clique cutset, stable cutset, homogeneous set, skew partition, and 2-clique cutset. We classify, with the exception of two polynomially equivalent problems, each list 4-partition problem as either solvable in polynomial time or NP-complete. In doing so, we provide polynomial-time algorithms for many problems whose polynomial-time solvability was open, including the list 2-clique cutset problem. This also allows us to classify each list generalized 2-clique cutset problem and list generalized skew partition problem as solvable in polynomial time or NP-complete.

Key words. graph partition, list partition, complexity, algorithm

AMS subject classifications. 05C85, 05C69, 68R10

DOI. 10.1137/060666238

1. Introduction.

The problem of partitioning the vertex-set of a graph subject to a given set of constraints on adjacencies between vertices in two distinct parts, or among vertices within a part, is fundamental and ubiquitous in algorithmic graph theory. For example, the problem of testing whether graph G is bipartite is equivalent to testing whether the vertex-set of G can be partitioned into parts A_1 and A_2 such that each A_i is a stable set; here we have no constraint on the adjacencies between vertices in A_1 and vertices in A_2. A graph is a *split graph* [28] if its vertex-set can be partitioned into a clique and a stable set. As the definition itself suggests, testing whether graph G is a split graph is another partition problem where we do not restrict the adjacencies between vertices placed in different parts of the partition. On the other hand, testing whether graph G is a complete tripartite graph is equivalent to testing whether the vertex-set of G can be partitioned into parts A_1, A_2, and A_3 such that each A_i induces a stable set, and between vertices in parts A_i, A_j, $i \neq j$, we have all possible edges; hence, the relationship between vertices placed in distinct parts is relevant here.

[*] Received by the editors July 27, 2006; accepted for publication (in revised form) January 5, 2007; published electronically December 7, 2007. A preliminary version of this paper appeared in *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2004.
http://www.siam.org/journals/sidma/21-4/66623.html

[†] Corresponding author. Department of Mathematics, Wilfrid Laurier University, Waterloo, Canada (kcameron@wlu.ca). This author acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC).

[‡] Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506 (Elaine.Eschen@wvu.edu). This author acknowledges support from NSERC of Canada.

[§] Department of Physics and Computer Science, Wilfrid Laurier University, Waterloo, Canada (choang@wlu.ca). This author acknowledges support from NSERC of Canada.

[¶] Computer Science Department, The University of Dayton, Dayton, OH 45469 (srithara@notes.udayton.edu). This author acknowledges support from the Research Council of The University of Dayton and NSERC of Canada.

1.1. The problem. In general, we can ask whether the vertex-set of a graph can be partitioned into at most k parts, $A_1, A_2, \ldots, A_k$, subject to constraints that require "no edges," "all edges," or "no restriction" between vertices placed in parts A_i and A_j; when $i = j$, the resulting constraint is on the subgraph induced by A_i. We can specify the required constraints on the partition via a symmetric $k \times k$ matrix M over $\{0, 1, *\}$. The natural interpretation is as follows: for $i \neq j$, if $M_{i,j} = 0$ (resp., 1, *), then we require "no edges" (resp., "all edges", "no restriction") between vertices placed in part A_i and vertices placed in part A_j; if $M_{i,i} = 0$ (resp., 1, *), then we require A_i to be a stable set (resp., clique, arbitrary subgraph). An *M-partition* of graph G then is a partition of the vertex-set of G into at most k parts so that all the constraints specified by M are respected. The *M-partition problem* asks the following: "Given G and a symmetric $k \times k$ matrix M over $\{0, 1, *\}$, does G admit an M-partition?". Many well-known graph theoretic problems are specific instances of the M-partition problem. For example, the 3-colorability problem is an M-partition problem where M is a 3×3 matrix with zeros on the main diagonal and asterisks everywhere else. Testing whether a graph is a split graph is asking whether the graph has an M-partition where M is a 2×2 matrix with a zero and one on the diagonal and asterisks everywhere else.

Feder et al. [22] introduced the M-partition problem and also generalized it to the *list M-partition problem.* In the list M-partition problem, in addition to being given graph G and a symmetric $k \times k$ matrix M over $\{0, 1, *\}$, for each vertex v of G, we are given a list $\mathcal{L}(v)$ that is a nonempty subset of $\{A_1, A_2, \ldots, A_k\}$. The problem asks the following: "Does G admit an M-partition in which each vertex v of G is assigned to a part in $\mathcal{L}(v)$?".

Many well-known graph problems can be formulated as list M-partition problems: e.g., list k-coloring, clique cutset, stable cutset, homogeneous set, skew partition, and 2-clique cutset. We study the list M-partition problems when M has dimension 4 with the goal of classifying them according to their complexity. Figure 1.1 illustrates the matrices corresponding to some of the problems we discuss.

1.2. Main results. In the following discussion, we use A, B, C, D to denote the parts of the M-partition problem. Let the *stubborn problem* be the list M-partition problem where $M_{A,A} = 0$, $M_{B,B} = 0$, $M_{D,D} = 1$, $M_{A,C} = M_{C,A} = 0$, and all other entries are asterisks (see Figure 1.1). The complement problem is obtained by interchanging the zeros and ones in the matrix. When M has dimension 4, we classify, with the sole exception of the stubborn problem and its complement, each list M-partition problem as either solvable in polynomial time or NP-complete. In doing so, we provide polynomial-time algorithms for many problems whose polynomial-time solvability was open. For example, we settle the open problem posed by Feder et al. [22] as to the existence of a polynomial-time algorithm to find a 2-clique cutset in a graph by providing a polynomial-time algorithm for the list 2-clique cutset problem. A 2-*clique cutset* is a cutset that induces the union of two cliques (or, equivalently, induces a bipartite graph in the complement).

Suppose $\mathcal{P}$ is an M-partition problem. A *generalized* $\mathcal{P}$ *problem* is an M'-partition problem where M' is obtained from M by changing some asterisks to either 0 or 1. Among other results, we prove that *any* list generalized 2-clique cutset problem (i.e., $M'_{A,A} = 1$, $M'_{B,B} = 1$, $M'_{C,D} = M'_{D,C} = 0$, and the other entries are 0, 1, or $*$) is solvable in polynomial time, unless it contains the complement of the 3-colorability problem, in which case it is NP-complete. This implies that the list strict 2-clique cutset problem is polynomial-time solvable, and via this we provide

$$\begin{array}{ccccc}
\begin{bmatrix} 0 & \ast \\ \ast & 1 \end{bmatrix} &
\begin{bmatrix} 0 & \ast & \ast \\ \ast & 0 & \ast \\ \ast & \ast & 0 \end{bmatrix} &
\begin{bmatrix} \ast & 0 & \ast \\ 0 & \ast & \ast \\ \ast & \ast & 1 \end{bmatrix} &
\begin{bmatrix} \ast & 0 & \ast \\ 0 & \ast & \ast \\ \ast & \ast & 0 \end{bmatrix} &
\begin{bmatrix} \ast & \ast & 1 \\ \ast & \ast & 0 \\ 1 & 0 & \ast \end{bmatrix} \\
\text{split graph} & \text{3-colorability} & \text{clique cutset} & \text{stable cutset} & \text{homogeneous set}
\end{array}$$

$$\begin{array}{ccc}
\begin{bmatrix} \ast & \ast & 0 & 0 \\ \ast & 0 & \ast & 0 \\ 0 & \ast & 0 & \ast \\ 0 & 0 & \ast & \ast \end{bmatrix} &
\begin{bmatrix} \ast & 1 & \ast & \ast \\ 1 & \ast & \ast & \ast \\ \ast & \ast & \ast & 1 \\ \ast & \ast & 1 & \ast \end{bmatrix} &
\begin{bmatrix} \ast & 1 & \ast & \ast \\ 1 & \ast & \ast & \ast \\ \ast & \ast & \ast & 0 \\ \ast & \ast & 0 & \ast \end{bmatrix} \\
\text{stable cutset pair} & 2K_2 & \text{skew partition}
\end{array}$$

$$\begin{array}{ccc}
\begin{bmatrix} 1 & 0 & \ast & \ast \\ 0 & 1 & \ast & \ast \\ \ast & \ast & \ast & 0 \\ \ast & \ast & 0 & \ast \end{bmatrix} &
\begin{bmatrix} 1 & \ast & \ast & \ast \\ \ast & 1 & \ast & \ast \\ \ast & \ast & \ast & 0 \\ \ast & \ast & 0 & \ast \end{bmatrix} &
\begin{bmatrix} 0 & \ast & 0 & \ast \\ \ast & 0 & \ast & \ast \\ 0 & \ast & \ast & \ast \\ \ast & \ast & \ast & 1 \end{bmatrix} \\
\begin{array}{c}\text{strict} \\ \text{2-clique cutset}\end{array} & \text{2-clique cutset} & \begin{array}{c}\text{stubborn} \\ \text{problem}\end{array}
\end{array}$$

FIG. 1.1. *Some M-partition problems.*

a polynomial-time algorithm to find a strict 2-clique cutset. A *strict* 2-*clique cutset* is a cutset that induces the disjoint union of two cliques (or, equivalently, induces a complete bipartite graph in the complement). We also classify each list generalized skew partition problem as solvable in polynomial time or NP-complete.

1.3. Significance. Many important graph decomposition problems can be formulated as M-partition problems with additional constraints imposed on the parts. Indeed, the eventual resolution of the Strong Perfect Graph Conjecture by Chudnovsky et al. [5] relies in part on three types of decompositions (a type of skew cutset partition and two generalizations of the homogeneous set partition) that can be formulated as M-partition problems with constraints. Such extra constraints typically are that certain parts be nonempty, have at least a given number of vertices, induce subgraphs that have at least one edge, etc. As discussed later, an instance of the M-partition problem with additional constraints can be reduced to a set of instances of the list M-partition problem. Thus, the list M-partition problem provides a flexible model to capture extra constraints placed on the required partition.

Every list M-partition problem with M of dimension 4 was classified by Feder et al. [22] as either 'solvable in quasi-polynomial time' or NP-complete. Here, quasi-polynomial time is complexity of $\mathrm{O}(n^{c \log^t n})$, where t and c are positive constants and n is the number of vertices in the input graph. Complete classification into polynomial-time solvable and NP-complete problems has been obtained for the list M-partition problem under several restrictions on M: when M is a matrix over $\{0, \ast\}, \{1, \ast\}$, or $\{0, 1\}$ [16, 19, 20, 22], has dimension 4 and does not contain an asterisk on the main diagonal [22], is the matrix for skew partition [15], has dimension 3 [22], and, trivially, when M has dimension 2. We complete this dichotomy classification (polynomial-time solvable and NP-complete) for all problems when M has dimension 4, with the exception of the stubborn problem (see Figure 1.1) and its complement. Further, when M has dimension 4, we give polynomial-time algorithms for many list M-partition problems that were previously not known to be solvable in polynomial time [22]. The techniques we employ, obtained by strengthening the techniques used in [15], are general enough that they may prove useful in solving other decomposition problems. For instance, we develop tools that are applicable to list M-partition problems of any dimension.

In general, such dichotomy (into polynomial-time solvable and NP-complete problems) results are uncommon. However, Feder and Vardi [26] have made a dichotomy conjecture in the context of constraint-satisfaction problems which has generated considerable interest and has been proven in several special cases [17]. It is noted in [17, 22] that general list M-partition problems are similar to, but not exactly the same as, list constraint-satisfaction problems. It was conjectured in [22] that every list M-partition problem (with no restriction on dimension of M) is either solvable in quasi-polynomial time or NP-complete. This "quasi-dichotomy" has since been established by Feder and Hell [17].

We show that all the quasi-polynomial-time cases of the Feder et al. [22] quasi-dichotomy result for the list M-partition problem when M has dimension 4 are actually polynomial-time solvable, with the single exception of the stubborn problem (and its complement), for which the best known complexity remains quasi-polynomial time. There is no NP-complete problem that is known to have a quasi-polynomial-time solution, and it is generally believed that problems solvable in quasi-polynomial time are unlikely to be NP-complete. A polynomial-time solution for the stubborn problem, if one exists, appears to be difficult and to require methods different from those presented here and those in [17, 22].

Next, we remark on the attention that the stubborn problem has received subsequent to the appearance of a preliminary version of this paper in [4]. Feder and Hell have independently identified the so-called "edge-free three-coloring problem" (see [17]), in their attempt to classify certain list partition and list constraint satisfaction problems, whose complexity has also eluded classification. Further, it is shown in [17] that the two problems are closely related and also that the latter problem is at least as hard as the stubborn problem. Finally, in a recent work in [24], it was shown that each of these two problems can be solved in $\mathrm{O}(n^{O(\frac{\log n}{\log\log n})})$ time, thus improving the bound of $\mathrm{O}(n^{O(\log n)})$ established in [22]. This remains the current best complexity for solving the stubborn problem.

1.4. Background and previous work. Feder et al. [22] introduced the M-partition problem and, motivated by the need to capture additional restrictions on the contents of individual parts or the connections between parts, generalized it to the list M-partition problem. Lists also facilitate solving problems by recursing to subproblems with modified lists. We use this technique, which was also employed in the algorithms of [15, 22]. The list M-partition problem generalizes the M-partition, list k-coloring, and list homomorphism (cf. below) problems. An instance of the M-partition problem with certain additional constraints (that certain parts be nonempty, have at least a given number of vertices, induce subgraphs that have at least one edge, etc.) can be reduced to a set of instances of the list M-partition problem. In this manner, the list M-partition problem provides a flexible model to capture extra constraints on the required partition. Many well-known graph theoretic problems correspond to M-partitions with additional constraints. We elaborate on this notion next.

A *clique cutset* in a graph is a cutset that induces a clique. It is easy to see that a connected graph has a clique cutset if and only if its vertex-set can be partitioned into parts A, B, and C, such that C is a clique, there are no edges between parts A and B, and, *further*, each part is nonempty. For a graph G on n vertices, the clique cutset problem can be reduced to $\mathrm{O}(n^3)$ instances of the list M-partition problem, where M is the matrix corresponding to the clique cutset problem, as follows: in order to handle the restriction that each of the parts A, B, and C be nonempty, for each triple x, y, z of vertices, we construct an instance with $\mathcal{L}(x) = \{A\}$, $\mathcal{L}(y) = \{B\}$, $\mathcal{L}(z) = \{C\}$, and the list for any other vertex is $\{A, B, C\}$. G has a clique cutset if and only if some such instance has a valid list M-partition. We note that finding a clique cutset and decomposing a graph via clique cutsets have applications in algorithmic graph theory [7, 28], and efficient algorithms exist for these problems [28, 33, 35, 36].

A 2-*clique cutset* is a cutset that is the union of two cliques (equivalently, the set of vertices in the cutset induces a bipartite graph in the complement). As illustrated in Figure 1.1, if parts A and B correspond to the two cliques whose union disconnects part C from part D, then whether a graph admits a 2-clique cutset is again an instance of the M-partition problem with the extra stipulation that each part be nonempty. Hayward and Reed [29] conjectured that every (even hole)-free graph (a graph that does not contain any induced cycle on an even number of vertices ≥ 4) that is not a complete graph contains a vertex whose neighborhood can be partitioned into two cliques. This conjecture implies that an (even hole)-free graph G has chromatic number at most $2\omega(G)$, where $\omega(G)$ is the clique number of G. Hoàng [31] proposed the weaker conjecture that (even hole)-free graphs different from a clique have a 2-clique cutset. Feder et al. [22] provided the first subexponential-time (but, not polynomial-time) algorithm to solve the list M-partition problem where M is the matrix for a 2-clique cutset, and hence, they also solved the 2-clique cutset problem in subexponential time. They posed the question [22] of the existence of a polynomial-time algorithm for the problem, which is answered in the affirmative here. We note that (even hole)-free graphs can be recognized in polynomial time [8, 9].

Analogous to a clique cutset, if we require the cutset to induce a stable set, then we get the stable cutset problem. A *skew partition* of a graph is a partition of its vertex-set into nonempty parts A, B, C, and D such that there are all possible edges between parts A and B and there are no edges between parts C and D. These problems are M-partition problems with the added constraint that each part be nonempty. Both the stable cutset and skew partition problems play prominent roles in the area of perfect graph theory. The interest in the stable cutset problem was motivated by Tucker's result [34] that a minimal imperfect graph, other than a chordless odd cycle, cannot contain a stable cutset. Chvátal conjectured [6] that a minimal imperfect graph does not admit a skew partition. Skew partitions played an important role in the proof of the Strong Perfect Graph Conjecture by Chudnovsky et al. [5]; this work also proved Chvátal's conjecture. Testing whether a graph has a stable cutset is known to be NP-complete [14]. However, Feder et al. [22] gave the first subexponential-time algorithm for the (list) skew partition problem. A polynomial-time algorithm for the (list) skew partition problem was developed subsequently by de Figueiredo et al. [15].

In certain other M-partition problems, there are constraints that there be at least a certain number of vertices in some parts. A *homogeneous set* or *module* in a graph is a set C of vertices such that C has at least two, but not all, of the vertices of the graph, and every vertex not in C is either adjacent to all the vertices in C, or none of the vertices in C. Among vertices not in C, if A is the set of vertices that are adjacent to all the vertices in C, and B is the set of vertices that are adjacent to none of the vertices in C, then testing for the presence of module is an M-partition problem with the additional requirements that $|C| \geq 2$ and $A \cup B$ is nonempty. We can reduce the homogeneous set problem for a graph G on n vertices to $\mathrm{O}(n^3)$ instances of the list M-partition problem, where M is the matrix corresponding to the homogeneous set problem, as follows: for each triple x, y, z of vertices, we set $\mathcal{L}(x) = \{C\}$, $\mathcal{L}(y) = \{C\}$, and $\mathcal{L}(z) = \{A, B\}$, the list of any other vertex to $\{A, B, C\}$, and check if any such instance has a valid list M-partition. Testing for the presence of modules and decomposition of a graph via modules have important applications in algorithmic graph theory, and efficient algorithms exist for these problems [10, 28, 32].

Feder et al. [22] studied the list M-partition problem with the goal of classifying matrices M into those for which the problem is efficiently solvable and those for which an efficient solution is perhaps unlikely. Next, we present results known on restricted versions of the list M-partition problem and then results known on the general list M-partition problem.

A k-coloring of graph G is the same as an M-partition of G where M (with dimension k) has zeros along the main diagonal and all other entries are asterisks. Therefore, the k-colorability problem is an M-partition problem where M is obtained from the 0-1 adjacency matrix of a complete (loopless) graph on k vertices by replacing every 1 with an asterisk. The more general H-coloring problem [30] is derived when M is obtained from the adjacency matrix of an arbitrary graph in the same way. More precisely, in the *H-coloring problem* [30], also called the *homomorphism problem*, given graph G and a specific graph H (possibly containing loops), we are asked whether it is possible to partition $V(G)$ into parts $A_u, u \in V(H)$, such that A_u is a stable set when u does not have a loop in H, and there are no edges between parts A_x and A_y whenever $xy \notin E(H)$. The H-coloring problem is solvable in polynomial time when H is bipartite or when H contains a loop, and is NP-complete otherwise [30].

The list H-coloring problem [16, 19, 20] is the list version of the H-coloring problem where, in addition to being given G and H, for each vertex v of G we are given a list, $\mathcal{L}(v)$ which is a subset of $V(H)$. The problem then asks whether there is an H-coloring subject to the additional restriction that each vertex v of G is placed in a part A_y such that $y \in \mathcal{L}(v)$. Just as the list coloring is a special case of list H-coloring (when H is a complete graph with no loops), list H-coloring is a special case of list M-partition where the matrix M is obtained from the adjacency matrix of the graph H by replacing every 1 with an asterisk.

In a sequence of papers [16, 19, 20], it was established that every list H-coloring problem (namely, every list M-partition problem where M is a matrix over $\{0, *\}$) is either solvable in polynomial time or NP-complete. The complement $\overline{M}$ of a matrix M over $\{0, 1, *\}$ is obtained from M by interchanging the zeros and ones and leaving the asterisks unchanged. Since the list M-partition problem for G, where M is a matrix over $\{1, *\}$, is essentially the same as the list $\overline{M}$-partition problem for the complement of G, it follows that every list M-partition problem, where M is a matrix over $\{1, *\}$, is also either solvable in polynomial time or NP-complete. See Figure 1.1 for definitions of the problems in the following theorems.

THEOREM 1.1 (see [16, 19, 20]). *If M is a matrix over $\{0, *\}$ or $\{1, *\}$, then the list M-partition problem is either solvable in polynomial time or NP-complete.*

The following corollary can be derived from [16, 19, 20].

COROLLARY 1.2 (see [16, 19, 20, 21]). *If M is a matrix over $\{0, *\}$ or $\{1, *\}$ and has dimension 4, then the list M-partition problem is solvable in polynomial time,* *except when M contains the matrix for* 3*-coloring, stable cutset, or their complements, or M is the matrix for stable cutset pair, $2K_2$, or their complements, in which cases the problem is NP-complete.*

Feder et al. [22] proved the following three theorems.

THEOREM 1.3 (see [22]). *When M has dimension* 3*, the list M-partition problem is solvable in polynomial time, except when M is the matrix for* 3*-coloring, stable cutset, or their complements, in which cases the problem is NP-complete.*

THEOREM 1.4 (see [22]). *When M has dimension* 4 *and does not contain a $*$ on the main diagonal, the list M-partition problem is solvable in polynomial time, except when M contains the matrix for* 3*-coloring, or its complement, in which cases the problem is NP-complete.*

THEOREM 1.5 (see [22]). *When M has dimension* 4*, the list M-partition problem is solvable in quasi-polynomial time or NP-complete.*

Feder et al. [22] also showed that if M is a matrix over $\{0, 1\}$, then the list M-partition problem is polynomial-time solvable. When M has dimension 2, the problem can be reduced to the 2-satisfiability problem and solved in polynomial time using the algorithm of [1].

It was conjectured in [22] that every list M-partition problem (with no restriction on dimension of M) is either solvable in quasi-polynomial time or NP-complete, and this now has been shown to be the case by Feder and Hell [17]. In a recent work [17], it has been shown that every list M-partition problem for directed graphs is either solvable in quasi-polynomial time or NP-complete. Further, when M has dimension at most 3, the quasi-polynomial cases of the list M-partition problem for directed graphs are now known to be polynomial-time solvable [25].

We close this section by referring the reader to [22] for a fine exposition on other graph theoretic problems that can be modeled as list M-partition problems.

2. Tools.

We borrow some tools from [15] and [22]. For a vertex v of graph G, $N(v)$ denotes the set of vertices adjacent to v in G, i.e., $N(v)$ is the set of *neighbors of* v in G.

A basic strategy that we employ, much akin to [22] and [15], is replacing an instance $\mathcal{I}$ of the list M-partition problem on graph G by a polynomially bounded number of instances $\mathcal{I}_1, \mathcal{I}_2, \dots, \mathcal{I}_p$ such that

- The answer to $\mathcal{I}$ is "yes" if and only if the answer to *some* $\mathcal{I}_k$ is "yes."

Moreover, each instance $\mathcal{I}_k$ satisfies *at least one* of the following:

- The longest list of $\mathcal{I}$ is missing in $\mathcal{I}_k$.
- The number of distinct lists in $\mathcal{I}_k$ is fewer than the number of distinct lists in $\mathcal{I}$.
- $\mathcal{I}_k$ is an instance of the list M'-partition problem for graph H where H is an induced subgraph of G and M' is a principal submatrix of M.
- $\mathcal{I}_k$ is easy to resolve.

Next we reproduce and summarize the tools from [22] that we use in this regard.

TOOL 1. *An instance of the list M-partition problem in which the list for every vertex of the input graph has size at most two, is solvable in polynomial time.*

Justification. Such a problem can easily be modeled as an instance of the 2-satisfiability problem (2-SAT) and solved using the algorithm in [1]. □

In the course of dealing with an instance of the list M-partition problem, our methods might decide to place a particular vertex in a specific part of the partition (either because the list of the vertex has size one, or this is one of the many possibilities that will be tried). The following tool addresses how the instance can then be "cleaned up" to account for the placement of the vertex without altering the outcome.

TOOL 2. *Suppose we have an instance of the list M-partition problem on graph G with lists L, and suppose we decide to place vertex v in part X. Let L′ be the lists obtained from L as follows: for all parts Y such that* $M_{X,Y} = 0$*, remove Y from the lists of neighbors of v. For all parts Y such that* $M_{X,Y} = 1$*, remove Y from the lists of nonneighbors of v. Then there is a list M-partition of G with respect to lists L and with v in X if and only if there is a list M-partition of G-v with respect to lists L′.*

TOOL 3. *Suppose we have an instance of the list M-partition problem for a graph on n vertices where* $M_{X,Y} = 0$ *and* $M_{X,Z} = 1$*. Then we can replace the instance with a set of instances consisting of one instance in which no vertex has X in its list, and at most n other instances in each of which no vertex has both Y and Z in its list such that the original instance admits a list M-partition if and only if some new instance does.*

Justification. If the original instance were to admit a list M-partition, then the possibilities are that either some vertex that had X in its list is placed in part X, or no vertex that had X in its list is placed in part X. The latter case can be covered by creating an instance by deleting X from every list. The former case can be covered by creating, for each vertex v that has X in its list, an instance by placing v in X and then applying Tool 2. □

Following the terminology used in [22], we say part X dominates part Y in matrix M, if for every part Z (including X and Y), we have $M_{X,Z} = M_{Y,Z}$ or $M_{X,Z} = *$.

TOOL 4. *Suppose we have an instance of the list M-partition problem on graph G with lists L, and part X dominates part Y in M. Let L′ be the lists obtained from L by removing Y from any list that also contains X. Then there is a list M-partition of G with respect to lists L if and only if there is a list M-partition of G with respect to lists L′.*

Justification. If part X dominates part Y in matrix M, then in any list M-partition of G, a vertex in part Y can also be placed in part X. □

Again, following the terminology in [22], we say that a $k \times k$ matrix M *contains* a $p \times p$ matrix M', $p \leq k$, if M' is a principal submatrix of M.

TOOL 5. *If M contains M′ and the list M′-partition problem is NP-complete, then the list M-partition problem is also NP-complete.*

Justification. Clearly, any polynomial-time algorithm for the list M-partition problem can be used, without any changes, to solve the list M'-partition problem in polynomial time. □

Recall that the complement $\overline{M}$ of matrix M is obtained from M by replacing every 0 with a 1, every 1 with a 0, and leaving the asterisks unchanged.

TOOL 6. *Graph G admits a list M-partition with respect to lists L if and only if the complement of G admits a list* $\overline{M}$*-partition with respect to the lists L.*

The following lemmata can be extracted from the details in [15]; however, they are not explicitly presented as lemmata there. We state them explicitly and present their proofs in their entirety for the sake of completeness. For simplicity of exposition (as was done in [15]) we use the constant 1/10 (and the related constants 7/10, 8/10, and 9/10) in the following lemmata. However, this can be replaced by any constant $1/c$ (and the related constants replaced by $(c-3)/c$, etc.) such that $c \geq 5$.

With respect to graph G and vertex-subset O of G, $\overline{O}$ denotes the subgraph induced by O in $\overline{G}$, the complement of G.

LEMMA 2.1 (see [15]). *Let G be a graph on n vertices and W be the set of those* *vertices of G whose degree is more than $\frac{9n}{10}$. If $|W| > \frac{9n}{10}$, then there is a linear time algorithm that*

- *either finds pairwise disjoint vertex subsets O, T, and NT of G such that $|O| + |NT| \geq \frac{n}{10}$, $|T| \geq \frac{n}{10}$, $\overline{O}$ is connected, there are all possible edges between O and T, and each vertex in NT is nonadjacent to a vertex of O,*
- *or finds disjoint vertex subsets O^*, T^* of G such that $|O^*| \geq \frac{n}{10}$, $|T^*| \geq \frac{7n}{10}$, and there are all possible edges between O^* and T^*.*

Proof. Consider the following algorithm that partitions a subset W' of W into sets O, T, and NT, where $|W'| > \frac{8n}{10}$. The algorithm starts with a single vertex in set O and attempts to grow the set.

```
**Algorithm** $\alpha$.
**Input:**
  $W' \subseteq W$ such that $|W'| > \frac{8n}{10}$.

  pick vertex $u \in W'$;
  $O = \{u\}$;
  $T = N(u) \cap W'$;
  $NT = W' - T - \{u\}$;
  **repeat**
      pick $v \in NT$;
      move $v$ from $NT$ to $O$;
      move $T \backslash N(v)$ from $T$ to $NT$
  **until** $(|O| + |NT| \geq \frac{n}{10})$ **or** $(NT = \emptyset)$
```

We first set $W' = W$ and invoke Algorithm α. As u is nonadjacent to fewer than $\frac{n}{10}$ vertices of G (hence, of W'), initially $|NT| < \frac{n}{10}$. Suppose the algorithm stops with $|O| + |NT| \geq \frac{n}{10}$. As v is nonadjacent to fewer than $\frac{n}{10}$ vertices of G (and hence, of W'), fewer than $\frac{n}{10}$ new vertices were moved into $O \cup NT$ during the final iteration. Therefore, $\frac{n}{10} \leq |O| + |NT| < \frac{2n}{10}$ and $|T| \geq (|W'| - \frac{2n}{10}) \geq (\frac{8n}{10} - \frac{2n}{10}) \geq \frac{n}{10}$. Further, as any vertex v moved into O is nonadjacent to some vertex of O, $\overline{O}$ remains connected. Clearly, there are all possible edges between O and T and every vertex in NT is nonadjacent to some vertex in O. Therefore, the sets O, T, and NT meet the conditions of the lemma.

On the other hand, suppose the algorithm stops with $|O|+|NT| < \frac{n}{10}$ and $NT = \emptyset$; clearly, $|O| < \frac{n}{10}$ and W was partitioned into O and T, and there are all possible edges between O and $W\backslash O$. We then apply the following algorithm to find the desired sets.

```
**Algorithm** $\beta$
**Input:**
  $O \subseteq W$ such that $|O| < \frac{n}{10}$ and there are all possible edges between $O$
  and $W \backslash O$.

  $O^* = O$;
  $W' = W \backslash O^*$;
  **repeat**
      Apply Algorithm $\alpha$ to $W'$ to partition it into sets $O$, $T$, and $NT$;
      **if** $(|O| + |NT| \geq \frac{n}{10})$ **then**
          **stop** /* $O$, $T$, and $NT$ are as desired */
```

```
        else
        {
            $O^* = O^* \cup O$;
            $W' = W \backslash O^*$
        }
    until ($|O^*| \geq \frac{n}{10}$);
    $T^* = W \backslash O^*$
```

Note that as Algorithm β begins, $|W'| > \frac{8n}{10}$; also, there are all possible edges between O^* and $W \backslash O^*$. If the algorithm stops with $|O| + |NT| \geq \frac{n}{10}$, then we have found appropriate sets O, T, and NT. Otherwise, $|O| < \frac{n}{10}$ and W' is partitioned into O and T. This implies that at the end of each iteration, there are all possible edges between O^* and $W \backslash O^*$. If $|O^*| < \frac{n}{10}$ (and hence, the loop does not terminate), then $|W'| > \frac{8n}{10}$ for the next iteration, satisfying the precondition for Algorithm α. Suppose Algorithm β stops with $|O^*| \geq \frac{n}{10}$; then, at the end of the penultimate iteration, $|O^*| < \frac{n}{10}$. Since the set O of vertices added to O^* during the final iteration has fewer than $\frac{n}{10}$ vertices, when the algorithm stops, $|O^*| < \frac{2n}{10}$. Taking $T^* = W \backslash O^*$ then guarantees that $|T^*| \geq (|W| - |O^*|) \geq (\frac{9n}{10} - \frac{2n}{10}) \geq \frac{7n}{10}$ and there are all possible edges between O^* and T^*. Finally, the algorithms can easily be implemented to run in linear time. □

LEMMA 2.2 (see [15]). *Let G be a graph on n vertices with a partition of its vertex set into sets S_1, S_2 with $|S_1| = n_1$ and $|S_2| = n_2$. Let X_1 be the set of those vertices in S_1 each of which has fewer than $\frac{n_2}{10}$ neighbors in S_2. If $|X_1| \geq \frac{n_1}{2}$, then there is a linear time algorithm that finds vertex subsets O, M, and NM of G such that*

1. *$O \subseteq X_1$,*
2. *S_2 is partitioned into M and NM,*
3. *there are no edges between O and M,*
4. *every $u \in NM$ has a neighbor $u' \in O$, and*
5. *either $\frac{2n_2}{5} \leq |M| \leq \frac{n_2}{2}$ and $|NM| \geq \frac{n_2}{2}$, or $|O| \geq \frac{n_1}{10}$ and $|M| > \frac{n_2}{2}$.*

Proof. We apply the following linear time algorithm to grow the set $O \subseteq X_1$ starting with a single vertex in O while partitioning S_2 into sets M and NM.

```
Algorithm $\gamma$
Input:
    Sets $S_1$, $S_2$, and $X_1$ as specified in Lemma 2.2.

    pick vertex $u \in X_1$;
    $O = \{u\}$;
    $NM = N(u) \cap S_2$;
    $M = S_2 \backslash NM$;
    repeat
        pick $v \in (X_1 \backslash O)$;
        move $v$ to $O$;
        move $N(v) \cap M$ from $M$ to $NM$
    until ($|M| \leq \frac{n_2}{2}$) or ($|O| \geq \frac{n_1}{10}$)
```

It is evident from Algorithm γ that there are no edges between O and M, and every vertex in NM has a neighbor in O. As u is adjacent to fewer than $\frac{n_2}{10}$ vertices of S_2, initially $|M| > \frac{9n_2}{10}$. Suppose $|M| \leq \frac{n_2}{2}$ when the algorithm stops. Since v is adjacent to fewer than $\frac{n_2}{10}$ vertices of S_2 (hence, of M), fewer than $\frac{n_2}{10}$ vertices were moved from M to NM during the final iteration. Therefore, $|M| > (\frac{n_2}{2} - \frac{n_2}{10})$, and we have $\frac{2n_2}{5} \leq |M| \leq \frac{n_2}{2}$. As M and NM partition the set S_2, we also have $|NM| \geq \frac{n_2}{2}$, as desired. On the other hand, suppose the algorithm stops with $|M| > \frac{n_2}{2}$ and $|O| \geq \frac{n_1}{10}$. The conditions of the lemma are then trivially met. □

Lemma 2.3 (see [15]). *Let G be a graph on n vertices with a partition of its vertex set into sets S_1, S_2 with $|S_1| = n_1$ and $|S_2| = n_2$. Let W_1 be the set of those vertices in S_1 each of which has more than $\frac{9n_1}{10}$ neighbors in S_1 and more than $\frac{9n_2}{10}$ neighbors in S_2. Let W_2 be the set of those vertices in S_2 each of which has more than $\frac{9n_2}{10}$ neighbors in S_2 and more than $\frac{9n_1}{10}$ neighbors in S_1. If $|W_1| > \frac{9n_1}{10}$ and $|W_2| > \frac{9n_2}{10}$, then there is a linear time algorithm that*

- *either finds pairwise disjoint vertex subsets O, T, and NT of G such that*
  1. *$\overline{O}$ is connected,*
  2. *there are all possible edges between O and T,*
  3. *each vertex in NT is nonadjacent to a vertex in O,*
  4. *$|T \cap S_1| \geq \frac{n_1}{10}$,*
  5. *$|T \cap S_2| \geq \frac{n_2}{10}$, and*
  6. *either $|O \cap S_1| + |NT \cap S_1| \geq \frac{n_1}{10}$, or $|O \cap S_2| + |NT \cap S_2| \geq \frac{n_2}{10}$,*
- *or finds disjoint vertex subsets O^*, T^* of G such that*
  1. *either $O^* \subseteq S_1$ and $|O^*| \geq \frac{n_1}{10}$, or $O^* \subseteq S_2$ and $|O^*| \geq \frac{n_2}{10}$,*
  2. *$|T^* \cap S_1| \geq \frac{n_1}{10}$,*
  3. *$|T^* \cap S_2| \geq \frac{n_2}{10}$, and*
  4. *there are all possible edges between O^* and T^*.*

Proof. We begin by noting that the proof of Lemma 2.3 is similar in principle to that of Lemma 2.1. Let $W = (W_1 \cup W_2)$, and therefore, $|W \cap S_1| > \frac{9n_1}{10}$ and $|W \cap S_2| > \frac{9n_2}{10}$.

Consider the following algorithm that partitions a subset W' of W into sets O, T, and NT, where $|W' \cap S_1| > \frac{8n_1}{10}$ and $|W' \cap S_2| > \frac{8n_2}{10}$. The algorithm starts with a single vertex in set O and attempts to grow the set.

```
Algorithm $\delta$
Input:
  $W' \subseteq W$ such that $|W' \cap S_1| > \frac{8n_1}{10}$ and $|W' \cap S_2| > \frac{8n_2}{10}$.

  pick vertex $u \in W'$;
  $O = \{u\}$;
  $T = N(u) \cap W'$;
  $NT = W' - T - \{u\}$;
  repeat
      pick $v \in NT$;
      move $v$ from $NT$ to $O$;
      move $T \backslash N(v)$ from $T$ to $NT$
  until $(|O \cap S_1| + |NT \cap S_1| \geq \frac{n_1}{10})$ or $(|O \cap S_2| + |NT \cap S_2| \geq \frac{n_2}{10})$ or $(NT = \emptyset)$
```

We first set $W' = W$ and invoke Algorithm δ. As u is nonadjacent to fewer than $\frac{n_1}{10}$ vertices of S_1 (hence, of $W' \cap S_1$) and fewer than $\frac{n_2}{10}$ of vertices of S_2 (hence, of $W' \cap S_2$), initially $|NT \cap S_1| < \frac{n_1}{10}$ and $|NT \cap S_2| < \frac{n_2}{10}$.

Suppose when the algorithm stops, $(((|O \cap S_1| + |NT \cap S_1| \geq \frac{n_1}{10})$ or $(|O \cap S_2| + |NT \cap$

$S_2| \geq \frac{n_2}{10}$)) is true; without loss of generality, assume that $|O \cap S_1| + |NT \cap S_1| \geq \frac{n_1}{10}$. As v is nonadjacent to fewer than $\frac{n_1}{10}$ vertices of S_1 (hence, of $W' \cap S_1$), fewer than $\frac{n_1}{10}$ new vertices were moved into $(O \cap S_1) \cup (NT \cap S_1)$ during the final iteration. Therefore, $|O \cap S_1| + |NT \cap S_1| < \frac{2n_1}{10}$. For similar reasons, $|O \cap S_2| + |NT \cap S_2| < \frac{2n_2}{10}$. Hence, $|T \cap S_1| \geq (|W' \cap S_1| - (|O \cap S_1| + |NT \cap S_1|)) \geq (\frac{8n_1}{10} - \frac{2n_1}{10}) \geq \frac{n_1}{10}$ and $|T \cap S_2| \geq (|W' \cap S_2| - (|O \cap S_2| + |NT \cap S_2|)) \geq (\frac{8n_2}{10} - \frac{2n_2}{10}) \geq \frac{n_2}{10}$. Further, as any vertex v moved into O is nonadjacent to some vertex of O, $\overline{O}$ remains connected. Clearly, there are all possible edges between O and T and every vertex in NT is nonadjacent to some vertex in O. Therefore, the sets O, T, and NT meet the conditions of the lemma.

On the other hand, suppose the algorithm stops with $|O \cap S_1| + |NT \cap S_1| < \frac{n_1}{10}$, $|O \cap S_2| + |NT \cap S_2| < \frac{n_2}{10}$, and $NT = \emptyset$; clearly, $|O \cap S_1| < \frac{n_1}{10}$, $|O \cap S_2| < \frac{n_2}{10}$, W is partitioned into O and T, and there are all possible edges between O and $W \backslash O$. We then apply the following algorithm to find the desired sets.

```
**Algorithm $\epsilon$**
**Input:**
  $O \subseteq W$ such that $|O \cap S_1| < \frac{n_1}{10}$ and $|O \cap S_2| < \frac{n_2}{10}$
  and there are all possible edges between $O$ and $W \backslash O$.

  $J^* = O$;
  $W' = W \backslash J^*$;
  repeat
      Apply Algorithm $\delta$ to $W'$ to partition it into sets $O$, $T$, and $NT$;
      if ($|O \cap S_1| + |NT \cap S_1| \geq \frac{n_1}{10}$) or ($|O \cap S_2| + |NT \cap S_2| \geq \frac{n_2}{10}$) then
          stop /* $O$, $T$, and $NT$ are as desired */
      else
      {
          $J^* = J^* \cup O$;
          $W' = W \backslash J^*$
      }
  until ($|J^* \cap S_1| \geq \frac{n_1}{10}$) or ($|J^* \cap S_2| \geq \frac{n_2}{10}$);
  if ($|J^* \cap S_1| \geq \frac{n_1}{10}$) then
      $O^* = J^* \cap S_1$
  else
      $O^* = J^* \cap S_2$;
  $T^* = W \backslash J^*$
```

Note that as Algorithm ϵ begins, $|W' \cap S_1| > \frac{8n_1}{10}$ and $|W' \cap S_2| > \frac{8n_2}{10}$; also, there are all possible edges between J^* and $W \backslash J^*$. If the algorithm stops with (($|O \cap S_1| + |NT \cap S_1| \geq \frac{n_1}{10}$) or ($|O \cap S_2| + |NT \cap S_2| \geq \frac{n_2}{10}$)) being true, then we have found appropriate sets O, T, and NT. Otherwise, $|O \cap S_1| < \frac{n_1}{10}$, $|O \cap S_2| < \frac{n_2}{10}$, and W' is partitioned into O and T. This implies that at the end of each iteration, there are all possible edges between J^* and $W \backslash J^*$. If $|J^* \cap S_1| < \frac{n_1}{10}$ and $|J^* \cap S_2| < \frac{n_2}{10}$ (and hence, the loop does not terminate), then $|W' \cap S_1| > \frac{8n_1}{10}$ and $|W' \cap S_2| > \frac{8n_2}{10}$ for the next iteration, satisfying the precondition for Algorithm δ. Without loss of generality, suppose the loop in Algorithm ϵ terminates with $|J^* \cap S_1| \geq \frac{n_1}{10}$; then, at the end of the penultimate iteration, $|J^* \cap S_1| < \frac{n_1}{10}$ and $|J^* \cap S_2| < \frac{n_2}{10}$. Since the set O of vertices added to J^* during the final iteration has fewer than $\frac{n_i}{10}$ vertices of S_i, $i = 1, 2$,

$|J^* \cap S_i| < \frac{2n_i}{10}$ for $i = 1, 2$. Taking $T^* = W \backslash J^*$ and $O^* = (J^* \cap S_1)$ then guarantees that $|T^* \cap S_1| \geq (|W \cap S_1| - |J^* \cap S_1|) \geq \frac{n_1}{10}$, $|T^* \cap S_2| \geq (|W \cap S_2| - |J^* \cap S_2|) \geq \frac{n_2}{10}$, and there are all possible edges between O^* and T^*. □

3. Three procedures.

We assume the input is a graph $G = (V, E)$ with the adjacency requirements on the parts A_i and a set Φ of lists $\mathcal{L}(v)$. We consider the instance Φ as a partition of V into at most $2^k - 1$ sets $S_{\mathcal{L}}$, indexed by the nonempty subsets $\mathcal{L}$ of $\mathcal{Z} = \{A_1, A_2, \ldots, A_k\}$. That is, $S_{\mathcal{L}}$ is the set of vertices with list $\mathcal{L}$. For example, if $\mathcal{L}(v) = \{A_1, A_2\}$, then $v \in S_{\{A_1,A_2\}}$. For simplicity we will drop the set brackets in the subscript, i.e., $S_{A_1A_2} = S_{\{A_1,A_2\}}$. $S_{\mathcal{L}}(\Phi)$ refers to the set $S_{\mathcal{L}}$ defined by Φ. When the context is clear, we write $S_{\mathcal{L}} = S_{\mathcal{L}}(\Phi)$. When we say Φ has a solution, it is assumed the parts are $A_1, A_2, \ldots, A_k$.

Throughout the algorithms used in the proof of the main theorem in section 4, Properties 1 and 2 below are always satisfied by the partition of V according to the sets $S_{\mathcal{L}}$.

Property 1. If the algorithm returns a partition and v is in $S_{\mathcal{L}}$, then the returned part A_i containing v is a part in $\mathcal{L}$.

Property 2. If $v \in S_{\mathcal{L}}$ for some $\mathcal{L}$, then for each $A_i \in \mathcal{L}$ and each S_{A_j}, v is adjacent (resp., nonadjacent) to all vertices in S_{A_j} whenever $M_{A_i,A_j} = 1$ (resp., $M_{A_i,A_j} = 0$). (It is possible that $i = j$.)

Often, we replace an instance Φ by a set of instances $\{\Phi_1, \Phi_2, \ldots, \Phi_p\}$ such that Φ has a solution if and only if some Φ_i has a solution. In this case, we say the set of instances $\{\Phi_1, \Phi_2, \ldots, \Phi_p\}$ is *equivalent* to Φ.

Let $X \subseteq S_{\mathcal{L}}$ and $A_i \in \mathcal{L}$. In creating a new instance Φ_j from Φ, we often say X *drops (part)* A_i. By this we mean for each vertex $v \in X$, $\mathcal{L}(v) = \mathcal{L} - \{A_i\}$, and, consequently, $S_{\mathcal{L}}(\Phi_j) = S_{\mathcal{L}}(\Phi) - X, S_{\mathcal{L}-\{A_i\}}(\Phi_j) = S_{\mathcal{L}-\{A_i\}}(\Phi) \cup X$ and $S_{\mathcal{L}'}(\Phi_j) = S_{\mathcal{L}'}(\Phi)$ for all other subsets $\mathcal{L}'$ of $\mathcal{Z}$. When we say X *gets the list* A_i we mean X drops all parts except A_i (i.e., $X \subseteq S_{A_i}(\Phi_j)$).

The reduction operation. Whenever a new instance Φ_j is created, a set $S_{A_i}(\Phi_j)$ may be a proper superset of $S_{A_i}(\Phi)$, and in any solution of Φ_j we must have $S_{A_i}(\Phi_j) \subseteq A_i$ for all i. If some $v \in S_{\mathcal{L}}(\Phi_j)$, where $A_i \in \mathcal{L}$, is not adjacent to all vertices in $S_{A_j}(\Phi_j)$ and $M_{A_i,A_j} = 1$, then v cannot be in part A_i in any solution. So we can *reduce* to a new problem where v drops the part A_i. In the case that $\mathcal{L}$ is a singleton set, Φ_j has no solution. The case where $M_{A_i,A_j} = 0$ is handled in a similar way. It is easy to see that after $O(n)$ similar reductions, we obtain an equivalent instance satisfying Property 2, or halt because Φ_j has no solution.

We refer to parts A_i, A_j such that $M_{A_i,A_j} = 1$ ($M_{A_i,A_j} = 0$) as *true partners* (*false partners*). We use *partner* without qualification to refer to a true or false partner. Note that a part can be its own partner.

The following two procedures (1 and 2) generalize two procedures in [15]. These generalizations are necessary for the proof of our main result in section 4. Also, these procedures are applicable to more general list partition problems than 4-part problems.

Remark. As in the lemmata of section 2, we assume $k \leq 10$ and use the corresponding constant 1/10 (and the related constants 7/10, 8/10, and 9/10) in the following procedures. However, for arbitrary dimension k, the constant 1/10 can be replaced by any constant $1/c$ (and the related constants replaced by $(c-3)/c$, etc.) such that $c \geq \max\{5, k\}$. Thus, these procedures are applicable to partition problems of any dimension k. The procedures are applied recursively to a given instance Φ to generate an equivalent set of instances (cf. Notes 1, 2, and 3). Taking $c = \max\{5, k\}$

minimizes the number of instances generated for any k. □

Procedure 1.

Input: An instance Φ of the list M-partition problem with set $\mathcal{Z}$ of parts $A_1, A_2, \ldots, A_k$, and a set $\mathcal{L} \subseteq \mathcal{Z}$ such that $S_{\mathcal{L}} \neq \emptyset$ and the parts $A_i \in \mathcal{L}$ can be put into sets $\mathcal{U}$ and $\mathcal{F}$ such that $\mathcal{U} \neq \emptyset, \mathcal{F} \neq \emptyset, \mathcal{U} \cup \mathcal{F} = \mathcal{L}$, but $\mathcal{U} \cap \mathcal{F}$ may or may not be empty, and the following properties hold:

(a) **Clique structure.** $\mathcal{U} = \{U_1, U_2, \ldots, U_u\}$. If $|\mathcal{U}| = 1$, then $M_{U_1,U_1} = 1$; otherwise $M_{U_iU_j} = 1$ for all i and j, $i \neq j$, except possibly when $i = u-1$ and $j = u$. If $M_{U_{u-1},U_u} \neq 1$, then $M_{U_{u-1},U_{u-1}} = M_{U_u,U_u} = 1$.

(b) $\mathcal{F} = \{F_1, \ldots, F_f\}$. If $|\mathcal{F}| = 1$, then $M_{F_1,F_1} = 0$; otherwise $M_{F_i,F_j} = 0$ for all i, j, $i \neq j$, except possibly when $i = f-1$ and $j = f$. If $M_{F_{f-1},F_f} \neq 0$, then $M_{F_{f-1},F_{f-1}} = M_{F_f,F_f} = 0$.

As noted above, lists satisfying property (a) are said to have the *clique structure.*

Output: A set of at most k instances, $\{\Phi_1, \Phi_2, \ldots\}$, that is equivalent to Φ, and such that for each i, $|S_{\mathcal{L}}(\Phi_i)| \leq \frac{9}{10}|S_{\mathcal{L}}(\Phi)|$, or a proof that Φ has no solution.

Note 1. Given an instance Φ, applying Procedure 1 to Φ produces at most k instances Φ_i with $|S_{\mathcal{L}}(\Phi_i)| \leq \frac{9}{10}|S_{\mathcal{L}}(\Phi)|$. Thus, given an instance Φ on a graph G with n vertices (with $k \leq 10$), recursively applying Procedure 1 produces a polynomial number of instances Φ' for which $S_{\mathcal{L}}(\Phi') = \emptyset$, and the set of instances produced is equivalent to Φ. It is easy to see that the number of instances Φ' is at most $k^{\log_{\frac{10}{9}} n} = n^{\log_{\frac{10}{9}} k}$. We shall refer to this process as *eliminating* the set $\mathcal{S}_{\mathcal{L}}$. □

Details of Procedure 1. Let $n = |S_{\mathcal{L}}(\Phi)|$. Any partner referred to here is a partner in $\mathcal{L}$.

Case 1. There is a vertex v in $S_{\mathcal{L}}$ such that $\frac{n}{10} \leq |S_{\mathcal{L}} \cap N(v)| \leq \frac{9n}{10}$.

To cover the possibility that v is placed in part A_i in the solution, we generate instances Φ_i, $i = 1, \ldots, k$, by setting $S_{A_i}(\Phi_i) = \{v\} \cup S_{A_i}(\Phi)$ and reducing so that Property 2 holds. If $A_i \in \mathcal{U}$, then the nonneighbors of v must drop the part $p(A_i)$ (hence, they cannot remain in $S_{\mathcal{L}}$) where $p(A_i)$ is the true partner of A_i. Since there are at least $\frac{n}{10}$ nonneighbors of v, $|S_{\mathcal{L}}(\Phi_i)| \leq \frac{9n}{10}$. Similarly, if $A_i \in \mathcal{F}$, then the neighbors of v must drop the part $p(A_i)$ where $p(A_i)$ is the false partner of A_i; hence, $|S_{\mathcal{L}}(\Phi_i)| \leq \frac{9n}{10}$. Clearly, the set of instances $\{\Phi_1, \ldots, \Phi_k\}$ is equivalent to Φ.

We may now assume that every vertex in $S_{\mathcal{L}}$ has more than $\frac{9n}{10}$ neighbors or fewer than $\frac{n}{10}$ neighbors in $S_{\mathcal{L}}$.

Let $W = \{v \in S_{\mathcal{L}} : |S_{\mathcal{L}} \cap N(v)| > \frac{9n}{10}\}$ and $X = \{v \in S_{\mathcal{L}} : |S_{\mathcal{L}} \cap N(v)| < \frac{n}{10}\}$.

Case 2. $|X| \geq \frac{n}{10}$ and $|W| \geq \frac{n}{10}$.

In any solution to Φ, $|A_i \cap S_{\mathcal{L}}(\Phi)| \geq \frac{n}{k} \geq \frac{n}{10}$ for some A_i; thus, we generate an instance for each A_i to cover the possibility that A_i is such a part. If $|A_i \cap S_{\mathcal{L}}(\Phi)| \geq \frac{n}{10}$ and A_i has a true (false) partner $p(A_i)$, then $p(A_i) \cap X = \emptyset$ ($p(A_i) \cap W = \emptyset$). Properties (a) and (b) ensure that each A_i has either a true or false partner $p(A_i)$. Thus, for $i = 1, \ldots, k$, generate Φ_i in which X drops $p(A_i)$, if $p(A_i)$ is a true partner; otherwise, generate Φ_i in which W drops $p(A_i)$. For each i, $|S_{\mathcal{L}}(\Phi_i)| \leq \frac{9n}{10}$, and the set of instances $\{\Phi_1, \ldots, \Phi_k\}$ is equivalent to Φ.

Case 3. $|W| > \frac{9n}{10}$.

By Lemma 2.1, we can either

(i) find pairwise disjoint subsets O, T, NT of $S_{\mathcal{L}}$ such that $\overline{O}$ is connected, $|O| + |NT| \geq \frac{n}{10}, |T| \geq \frac{n}{10}$, there are all possible edges between O and T, and each vertex in NT is nonadjacent to some vertex in O, or

(ii) find disjoint subsets O^* and T^* of $S_{\mathcal{L}}$ such that $|O^*| \geq \frac{n}{10}, |T^*| \geq \frac{7n}{10}$, and there are all possible edges between O^* and T^*.

Case (i). We create an instance Φ_{A_i} for each part A_i of $\mathcal{L}$ as follows. First, for each A_i of $\mathcal{F} - \mathcal{U}$ with false partner $p(A_i)$, construct Φ_{A_i} by making T drop the part $p(A_i)$.

Now, we may assume that the remaining parts in $\mathcal{L}$ can be named $U_1, U_2, \dots, U_l$ so that they have the clique structure. We create instances as follows:

1. If $l = 1$, then $M_{U_1,U_1} = 1$. If $|O| > 1$, do not create a new instance. Otherwise, create instance Φ_{U_1} by placing the only vertex of O in part U_1 and making NT drop part U_1.
2. If $l \geq 2$ and $M_{U_i,U_j} = 1$ for all $j \neq i$, create, for each i, Φ_{U_i} from Φ by making $O \cup NT$ drop every part U_j, $j \neq i$.
3. If $l \geq 2$ and $M_{U_iU_j} \neq 1$ for some i, j, then we must have $\{i, j\} = \{l-1, l\}$ and $M_{U_{l-1},U_{l-1}} = M_{U_l,U_l} = 1$. Test whether O has a unique partition into two cliques K_1, K_2. If not, do not create a new instance (see the explanation below). Otherwise, create two instances Φ_1, Φ_2 as follows. In Φ_1, K_1 gets the list U_{l-1} (it drops all other parts) and K_2 gets the list U_l; for each vertex x in NT, x drops part U_{l-1} if x is nonadjacent to some vertex in K_1, or x drops part U_l if x is nonadjacent to some vertex in K_2. The instance Φ_2 is defined similarly with K_1 getting list U_l and K_2 getting list U_{l-1}.

We now show that the set of new instances is equivalent to Φ. Suppose there is a solution $A_1, \dots, A_k$ to Φ. It must be the case that for some i, $O \cap A_i \neq \emptyset$. If there is an A_i in $\mathcal{F} - \mathcal{U}$ with a false partner $p(A_i)$ such that $O \cap A_i \neq \emptyset$, then $T \cap p(A_i) = \emptyset$; this eventuality is covered by Φ_{A_i}.

Now suppose there is no part in $\mathcal{F} - \mathcal{U}$ that has nonempty intersection with O. Let the parts not in $\mathcal{F} - \mathcal{U}$ be $U_1, \dots, U_l$ (if they exist). These parts must have the clique structure. If $l = 1$, then we have $M_{U_1,U_1} = 1$ and $O \subseteq U_1$ in the solution. Since $\overline{O}$ is connected, it follows that when $|O| > 1$, there is no solution. Otherwise, the only vertex in O must go to part U_1. As no vertex in NT can now be in part U_1, NT must drop the part U_1; this eventuality is covered by the instance Φ_{U_1}.

Now suppose $l \geq 2$. For any U_i that is a true partner of all U_j with j different from i, if $O \cap U_i \neq \emptyset$, then (as $\overline{O}$ is connected) $O \subseteq U_i$. Since no member of NT can now be placed in a part that is a true partner of U_i, it follows that NT must drop all parts U_j with $i \neq j$; this eventuality is covered by Φ_{U_i}.

Last, we consider the case $M_{U_{l-1},U_{l-1}} = M_{U_l,U_1} = 1$ and every vertex in O belongs to $U_{l-1} \cup U_l$. Since $\overline{O}$ is connected, O must be partitioned uniquely into two cliques K_1, K_2; otherwise, there is no solution. We see that every vertex in NT must drop a part (U_{l-1} or U_l); this eventuality is covered by Φ_1 and Φ_2.

Case (ii). We construct two new instances from Φ as follows. Choose an A_i that has a false partner $p(A_i)$ and create Φ_1 by making T^* drop $p(A_i)$; then create Φ_2 by making O^* drop A_i. This can be justified as follows. In any solution to Φ, if $A_i \cap O^* \neq \emptyset$, then $T^* \cap p(A_i) = \emptyset$; otherwise, $O^* \cap A_i = \emptyset$.

Case 4. $|X| > \frac{9n}{10}$.

This case is similar to Case 3 with G replaced by $\overline{G}$ and M replaced by $\overline{M}$.

It is easily verified that in each instance Γ created, $|S_{\mathcal{L}}(\Gamma)| \leq \frac{9}{10}|S_{\mathcal{L}}(\Phi)|$. If no new instances are produced by the above analysis, then Φ has no solution. This completes the description of Procedure 1. □

Procedure 2.

Input: Instance Φ of the list M-partition problem with set $\mathcal{Z}$ of parts $A_1, \dots, A_k$, and two sets $\mathcal{L}$ and $\mathcal{R}$, which are subsets of $\mathcal{Z}$, such that $S_{\mathcal{L}} \neq \emptyset, S_{\mathcal{R}} \neq \emptyset$, $\mathcal{L} \not\subseteq \mathcal{R}$, $\mathcal{R} \not\subseteq \mathcal{L}$, and we can write $\mathcal{L} = \{L_1, L_2, \dots L_p\}$ $(p \geq 3)$ and $\mathcal{R} = \{R_1, R_2, \dots, R_q\}$ $(q \geq 3)$ so that

1. each L_i has a partner in $\mathcal{R}$,
2. each R_i has a partner in $\mathcal{L}$,
3. some L_i has a true partner in $\mathcal{R}$ (equivalently, some R_i has a true partner in $\mathcal{L}$),
4. some L_i has a false partner in $\mathcal{L} \cup \mathcal{R}$,
5. some R_j has a false partner in $\mathcal{L} \cup \mathcal{R}$,
6. for each i, if L_i has no false partner in $\mathcal{R}$, then L_i has a true partner in $\mathcal{L}$,
7. for each i, if R_i has no false partner in $\mathcal{L}$, then R_i has a true partner in $\mathcal{R}$,
8. if the set $\mathcal{F}$ of parts in $\mathcal{L}$ that have no true partners in $\mathcal{R}$ is not empty, then there is a part R_j that is a false partner of all parts in $\mathcal{F}$,
9. if the set $\mathcal{H}$ of parts in $\mathcal{R}$ that have no true partners in $\mathcal{L}$ is not empty, then there is a part L_i that is a false partner of all parts in $\mathcal{H}$,
10. if the set $\mathcal{U}$ of parts of $\mathcal{L} \cup \mathcal{R}$ that have no false partners in $\mathcal{L} \cup \mathcal{R}$ is not empty, then the parts in $\mathcal{U}$ must have the clique structure, each of them has a true partner in $\mathcal{L}$ and in $\mathcal{R}$, and the two parts in $\mathcal{U}$ that are not true partners (if they exist) must belong to $\mathcal{L} \cap \mathcal{R}$.

Output: A set of at most $2k$ instances $\{\Phi_1, \Phi_2, \ldots\}$ that is equivalent to Φ, and such that for each i, $|S_{\mathcal{L}}(\Phi_i)|\,|S_{\mathcal{R}}(\Phi_i)| \le \frac{9}{10}|S_{\mathcal{L}}(\Phi)|\,|S_{\mathcal{R}}(\Phi)|$, or a proof that Φ has no solution.

Note 2. Given an instance Φ on a graph G with n vertices (with $k \le 10$) that satisfies the conditions of Procedure 2, recursively applying Procedure 2 produces a polynomial number of instances Φ' for which $S_{\mathcal{L}}(\Phi') = \emptyset$ or $S_{\mathcal{R}}(\Phi') = \emptyset$, and the set of instances produced is equivalent to Φ. It is easy to see that the number of instances Φ' is at most $(2k)^{\log_{\frac{10}{9}} n^2} = n^{2\log_{\frac{10}{9}} 2k}$. □

Details of Procedure 2. Write $S_1 = S_{\mathcal{L}}, S_2 = S_{\mathcal{R}}$. Let $n_1 = |S_1|$ and $n_2 = |S_2|$. For a vertex $v \in S_1 \cup S_2$, let $d_i(v) = |N(v) \cap S_i|$, $i = 1,2$.

Case 1. There is a vertex v in S_1 with $\frac{n_2}{10} \le d_2(v) \le \frac{9n_2}{10}$.

For each $L_i \in \mathcal{L}$, let $p(L_i)$ be a partner of L_i in $\mathcal{R}$. For each $L_i \in \mathcal{L}$, construct an instance Φ_i from Φ as follows. If L_i is a true partner of $p(L_i)$, then $S_2 - N(v)$ drops the part $p(L_i)$; otherwise, $S_2 \cap N(v)$ drops part $p(L_i)$. It is a routine matter to verify that the set of new instances is equivalent to Φ.

Case 1$'$. There is a vertex v in S_2 with $\frac{n_1}{10} \le d_1(v) \le \frac{9n_1}{10}$.

This case is symmetric to Case 1.

Case 2. Every vertex v in S_1 satisfies $d_2(v) < \frac{n_2}{10}$ or $d_2(v) > \frac{9n_2}{10}$. Every vertex v in S_2 satisfies $d_1(v) < \frac{n_1}{10}$ or $d_1(v) > \frac{9n_1}{10}$.

Define four sets as follows:

$$X_1 = \left\{ v \in S_1 | d_2(v) < \frac{n_2}{10} \right\},\ X_2 = \left\{ v \in S_2 | d_1(v) < \frac{n_1}{10} \right\},$$

$$W_1 = \left\{ v \in S_1 | d_2(v) > \frac{9n_2}{10} \right\},\ W_2 = \left\{ v \in S_2 | d_1(v) > \frac{9n_1}{10} \right\}.$$

There are three cases to consider.

Case 2.1. $|X_1|, |W_1| \ge \frac{n_1}{10}$.

Create q new instances from Φ as follows. For each $R_j \in \mathcal{R}$, let $p(R_j)$ be a partner of R_j in $\mathcal{L}$. If $p(R_j)$ is a true (resp., false) partner of R_j, then Φ_j is obtained from Φ by making X_1 (resp., W_1) drop the part $p(R_j)$. This is justified as follows. In any solution to Φ some R_j must have $|R_j \cap S_2| \geq \frac{n_2}{q} \geq \frac{n_2}{k} \geq \frac{n_2}{10}$; if $M_{R_j,p(R_j)} = 1$ (resp., 0), then $X_1 \cap p(R_j) = \emptyset$ (resp., $W_1 \cap p(R_j) = \emptyset$). Thus, the q new instances cover all the eventualities.

Case 2.1′. $|X_2|, |W_2| \geq \frac{n_2}{10}$.

This case is symmetric to Case 2.1.

Case 2.2. $|X_1| > \frac{9n_1}{10}$.

Find the sets O, M, and NM as defined by Lemma 2.2.

Suppose first that $|O| \geq \frac{n_1}{10}$ and $|M| > \frac{n_2}{2}$. Replace Φ by two new instances Φ_1, Φ_2 as follows. Let L_i be a part with a true partner $p(L_i)$ in $\mathcal{R}$. Φ_1 is obtained from Φ by making M drop the part $p(L_i)$ and Φ_2 is obtained from Φ by making O drop the part L_i. This can be justified as follows. Consider any solution of Φ. If $O \cap L_i \neq \emptyset$, then no vertex of M can be in part $p(L_i)$; otherwise, no vertex of O is in part L_i. Thus, the two new instances Φ_1, Φ_2 cover all the eventualities.

Now, we may assume that $\frac{2n_2}{5} \leq |M| \leq \frac{n_2}{2}$ and $|NM| \geq \frac{n_2}{2}$. Let $\mathcal{L}^+$ be the set of parts in $\mathcal{L}$ that have a true partner $p(L_i)$ in $\mathcal{R}$. Construct at most $|\mathcal{L}^+| + 1$ new instances as follows. For each $L_i \in \mathcal{L}^+$, construct Φ_{L_i} from Φ by making M drop the part $p(L_i)$. If $\mathcal{L} - \mathcal{L}^+ \neq \emptyset$, then there is a part R_j in $\mathcal{R}$ that is a false partner of each part in $\mathcal{L} - \mathcal{L}^+$; construct a new instance Φ' from Φ by making NM drop the part R_j. This can be justified as follows. Consider any solution of Φ. For any L_i in $\mathcal{L}^+$, if $O \cap L_i \neq \emptyset$, then $M \cap p(L_i) = \emptyset$. If $O \cap L_i = \emptyset$ for all L_i in $\mathcal{L}^+$, then the vertices of O must be in parts in $\mathcal{L} - \mathcal{L}^+$, so $NM \cap R_j = \emptyset$.

Case 2.2′. $|X_2| > \frac{9n_2}{10}$. This case is symmetric to Case 2.2.

Case 2.3. $|W_1| > \frac{9n_1}{10}, |W_2| > \frac{9n_2}{10}$.

Suppose there is a vertex $v \in W_1$ with $d_1(v) \leq \frac{9n_1}{10}$. Let $\mathcal{L}^-$ be the set of parts in $\mathcal{L}$ that have a false partner $p(L_i)$ in $\mathcal{R}$. Note that each part $L_i \in \mathcal{L} - \mathcal{L}^-$ has a true partner $p(L_i)$ in $\mathcal{L}$. Construct p new instances, corresponding to each of the p parts of $\mathcal{L}$ that v can be placed in, as follows. For each $L_i \in \mathcal{L}^-$, construct Φ_{L_i} from Φ by making $S_2 \cap N(v)$ drop $p(L_i)$. For each $L_i \in \mathcal{L} - \mathcal{L}^-$, construct Φ_{L_i} from Φ by making $S_1 - N(v)$ drop $p(L_i)$. A routine argument shows the set of p new instances is equivalent to Φ.

A symmetrical argument settles the case in which there is a vertex $v \in W_2$ with $d_2(v) \leq \frac{9n_2}{10}$.

Now, we may assume that each $v \in W_i$ has $d_i(v) > \frac{9n_i}{10}$, for $i = 1, 2$. By Lemma 2.3, we can find either the sets (a) or the sets (b) as follows.

(a) Pairwise disjoint vertex subsets O, T, and NT of $S_1 \cup S_2$ such that all the following hold:
   (a) $\overline{O}$ is connected.
   (b) There are all possible edges between O and T.
   (c) Each vertex in NT is nonadjacent to some vertex in O.
   (d) $|T \cap S_1| \geq \frac{n_1}{10}$.
   (e) $|T \cap S_2| \geq \frac{n_2}{10}$.
   (f) Either $|O \cap S_1| + |NT \cap S_1| \geq \frac{n_1}{10}$ or $|O \cap S_2| + |NT \cap S_2| \geq \frac{n_2}{10}$.

(b) Disjoint vertex subsets O^*, T^* of $S_1 \cup S_2$ such that all the following hold:
   (a) Either $O^* \subseteq S_1$ and $|O^*| \geq \frac{n_1}{10}$, or $O^* \subseteq S_2$ and $|O^*| \geq \frac{n_2}{10}$.
   (b) $|T^* \cap S_1| \geq \frac{n_1}{10}$.
   (c) $|T^* \cap S_2| \geq \frac{n_2}{10}$.
   (d) There are all possible edges between O^* and T^*.

Case (a). Consider the case $|O \cap S_1| + |NT \cap S_1| \geq \frac{n_1}{10}$. (The case $|O \cap S_2| + |NT \cap S_2| \geq \frac{n_2}{10}$ is symmetric.) Construct at most $p + q$ new instances from Φ as follows.

For each part A_i in $\mathcal{L} \cup \mathcal{R}$ with a false partner $p(A_i)$ in $\mathcal{L} \cup \mathcal{R}$, create Φ_{A_i} by making $T \cap S_1$ drop the part $p(A_i)$ if $p(A_i) \in \mathcal{L}$, or $T \cap S_2$ drop part $p(A_i)$ if $p(A_i) \in \mathcal{R}$.

Now, the remaining parts of $\mathcal{L} \cup \mathcal{R}$ (if they exist) can be named $U_1, U_2, \ldots, U_l$ such that condition 10 of Procedure 2 is satisfied. We create instances as follows:

1. If $l = 1$, then $M_{U_1,U_1} = 1$. If $|O| > 1$, do not create a new instance. Otherwise, create instance Φ_1 as follows. Let v be the single member of O. If $v \in S_1$ and $U_1 \in \mathcal{L}$, then v gets label U_1 and $NT \cap S_1$ drops part U_1. If $v \in S_2$ and $U_1 \in \mathcal{R}$, then v gets label U_1 and $NT \cap S_1$ drops a part $A_j \in \mathcal{L}$ such that $M_{U_1,A_j} = 1$. If neither of these conditions are satisfied, do not create a new instance.
2. If $l \geq 2$ and $M_{U_i,U_j} = 1$ for all $j \neq i$, define Φ_{U_i} from Φ as follows. First, suppose $U_i \in \mathcal{L}$. If U_i is also in $\mathcal{R}$ or if $O \cap S_2 = \emptyset$, then define Φ_{U_i} by making O get the list U_i and $NT \cap S_1$ drop a part $A_j \in S_1$ such that $M_{U_i,A_j} = 1$; otherwise, create no new instance (Φ would not have a solution in this eventuality). Now, suppose $U_i \in \mathcal{R} - \mathcal{L}$. If $O \cap S_1 = \emptyset$, then make $NT \cap S_1$ drop a part $A_j \in S_1$ such that $M_{U_i,A_j} = 1$; otherwise, make no new instance (Φ would not have a solution in this eventuality).
3. If $l \geq 2$ and $M_{U_i,U_j} \neq 1$ for some i, j, then we must have $\{i, j\} = \{l-1, l\}$, $M_{U_{l-1},U_{l-1}} = M_{U_l,U_1} = 1$. Test whether O has a unique partition into two cliques K_1, K_2 (if this is not the case then we do not create a new instance, see the explanation below). We define two instances Φ_1, Φ_2 as follows. In Φ_1, K_1 gets the list U_{l-1} (it drops all other parts), K_2 gets the list U_l; for each vertex x in NT, x drops the part U_{l-1} if x is nonadjacent to some vertex in K_1, or x drops the part U_l if x is nonadjacent to some vertex in K_2. The instance Φ_2 is defined similarly with K_1 getting list U_l and K_2 getting list U_{l-1}.

We now show that the set of new instances are equivalent to Φ. Suppose there is a solution $A_1, \ldots, A_k$ to Φ. It must be the case that for some i, $O \cap A_i \neq \emptyset$. If there is a part $A_i \in \mathcal{L} \cup \mathcal{R}$ with a false partner $p(A_i) \in \mathcal{L} \cup \mathcal{R}$ such that $O \cap A_i \neq \emptyset$, then $T \cap S_j \cap p(A_i) = \emptyset$, where $j = 1$ if $p(A_i) \in \mathcal{L}$ and $j = 2$ if $p(A_i) \in \mathcal{R}$. This eventuality is covered by Φ_{A_i}.

Now suppose there is no part with a false partner that has nonempty intersection with O. Let $U_1, \ldots, U_l$ be the parts of $\mathcal{L} \cup \mathcal{R}$ with no false partners in $\mathcal{L} \cup \mathcal{R}$ (if they exist). These parts must have the clique structure. If $l = 1$, then we have $M_{U_1,U_1} = 1$ and $O \subseteq U_1$ in the solution. Since $\overline{O}$ is connected, it follows that if $|O| > 1$, there is no solution in this eventuality. Therefore, O has exactly one vertex v and it is in S_1 or S_2. If $v \in S_1$, there is a solution only if $U_1 \in \mathcal{L}$ and v is placed in U_1. Then no vertex of $NT \cap S_1$ can be in U_1. If $v \in S_2$, there is a solution only if $U_1 \in \mathcal{R}$ and v is placed in U_1. Then no vertex of $NT \cap S_1$ can be in a part that is a true partner of U_1. In this case, since $|O \cap S_1| = 0$, we have $|NT \cap S_1| \geq \frac{n_1}{10}$.

We can now assume $l \geq 2$. Consider a U_i that is a true partner of all U_j with j different from i. If $O \cap U_i \neq \emptyset$, then we have $O \subseteq U_i$. If $U_i \in \mathcal{L}$, then for there to be a solution with $O \subseteq U_i$, we must have either $U_i \in \mathcal{R}$ or $O \cap S_2 = \emptyset$ (or both). If $U_i \in \mathcal{R} - \mathcal{L}$, then for there to be a solution with $O \subseteq U_i$, we need $O \cap S_1 = \emptyset$, and in this case we have $|NT \cap S_1| \geq \frac{n_1}{10}$. This eventuality is covered by Φ_{U_i}.

Last, we consider the case $M_{U_{l-1},U_{l-1}} = M_{U_l,U_1} = 1$ (both belong to $\mathcal{L} \cap \mathcal{R}$ by condition 10 of Procedure 2) and every vertex in O belongs to $U_{l-1} \cup U_l$. Since $\overline{O}$ is connected, there is a unique partition of O into two cliques K_1, K_2 (if this is not the case, then this eventuality has no solution and so we do not need to create a new instance). Since every vertex x in NT is nonadjacent to some vertex in O, x must drop part U_{l-1} or U_l; it follows that this eventuality is covered by Φ_1, Φ_2.

Case (b). Consider the case $O^* \subseteq S_1$, $|O^*| \geq \frac{n_1}{10}$. (The case $O^* \subseteq S_2$, $|O^*| \geq \frac{n_2}{10}$ is symmetric.) We construct two new instances from Φ as follows. Choose an $L_i \in \mathcal{L}$ that has a false partner $p(L_i) \in \mathcal{L} \cup \mathcal{R}$ and create Φ_{L_i} by making $T^* \cap S_1$ drop $p(L_i)$ if $p(L_i) \in \mathcal{L}$, or by making $T^* \cap S_2$ drop $p(L_i)$ if $p(L_i) \in \mathcal{R}$. Then create Φ' by making O^* drop L_i. This can be justified as follows. In any solution to Φ, if for some $L_i \in \mathcal{L}$ we have $L_i \cap O^* \neq \emptyset$, then $T^* \cap S_j \cap p(L_i) = \emptyset$, where $j = 1$ if $p(L_i) \in \mathcal{L}$ and $j = 2$ if $p(L_i) \in \mathcal{R}$; otherwise, $O^* \cap L_i = \emptyset$.

It is easily verified that in each instance Γ created, $|S_{\mathcal{L}}(\Gamma)|\,|S_{\mathcal{R}}(\Gamma)| \leq \frac{9}{10}|S_{\mathcal{L}}(\Phi)|\,|S_{\mathcal{R}}(\Phi)|$. If no new instances are produced by the above analysis, then Φ has no solution. This completes the description of Procedure 2. □

Procedure 3. We note that our Procedure 3, in principle, is the same as Procedure 4 in [15].

Input: Instance Φ of the list M-partition problem with set $\mathcal{Z}$ of parts $A_1, \ldots, A_k$, and two sets $\mathcal{L}$ and $\mathcal{R}$, which are subsets of $\mathcal{Z}$, such that $S_{\mathcal{L}} \neq \emptyset, S_{\mathcal{R}} \neq \emptyset$, $\mathcal{L} \not\subseteq \mathcal{R}$, $\mathcal{R} \not\subseteq \mathcal{L}$, and we can write $\mathcal{L} = \{L_1, L_2\}$ and $\mathcal{R} = \{R_1, \ldots, R_q\}$ ($q \geq 2$) so that L_1 has a false partner in $\mathcal{R}$ and L_2 has a true partner in $\mathcal{R}$.

Output: The set of instances $\{\Phi_1, \Phi_2\}$ that is equivalent to Φ, and such that $|S_{\mathcal{L}}(\Phi_i)|\,|S_{\mathcal{R}}(\Phi_i)| \leq \frac{9}{10}|S_{\mathcal{L}}(\Phi)|\,|S_{\mathcal{R}}(\Phi)|$, or a proof that Φ has no solution.

Note 3. Given an instance Φ on a graph G with n vertices that satisfies the conditions of Procedure 3, recursively applying Procedure 3 produces a polynomial number of instances Φ' for which $S_{\mathcal{L}}(\Phi') = \emptyset$ or $S_{\mathcal{R}}(\Phi') = \emptyset$, and the set of instances Φ' is equivalent to Φ. It is easy to see that the number of instances Φ' is at most $(2)^{\log_{\frac{10}{9}} n^2} = n^{2 \log_{\frac{10}{9}} 2}$. □

Details of Procedure 3. Write $S_1 = S_{\mathcal{L}}, S_2 = S_{\mathcal{R}}$. Let $n_1 = |S_1|$ and $n_2 = |S_2|$. For a vertex $v \in S_1 \cup S_2$, let $d_i(v) = |N(v) \cap S_i|$, $i = 1, 2$. Let $p(L_i) \in \mathcal{R}$ be the partner of L_i, $i = 1, 2$.

Case 1. There is a vertex v in S_1 with $\frac{n_2}{10} \leq d_2(v) \leq \frac{9n_2}{10}$.

Construct two instances from Φ corresponding to v being placed in L_i, $i = 1$, 2. One instance is constructed by making $S_2 \cap N(v)$ drop the part $p(L_1)$ and another is constructed by making $S_2 - N(v)$ drop the part $p(L_2)$. It is a routine matter to verify that the set of new instances is equivalent to Φ.

Case 2. Every vertex in $S_{\mathcal{L}}$ satisfies $d_2(v) < \frac{n_2}{10}$ or $d_2(v) > \frac{9n_2}{10}$.

Define two sets as follows:

$$X_1 = \left\{ v \in S_1 | d_2(v) < \frac{n_2}{10} \right\},\ W_1 = \left\{ v \in S_1 | d_2(v) > \frac{9n_2}{10} \right\}.$$

There are two cases to consider.

Case 2.1. $|X_1| \geq \frac{n_1}{2}$.

Find the sets O, M, and NM as defined in Lemma 2.2.

Suppose first that $|O| \geq \frac{n_1}{10}$ and $|M| > \frac{n_2}{2}$. Replace Φ with two new instances Φ_1, Φ_2 constructed as follows. Φ_1 is obtained from Φ by making M drop the part $p(L_2)$; Φ_2 is obtained from Φ by making O drop the part L_2. This can be justified as follows. Consider any solution of Φ. If $O \cap L_2 \neq \emptyset$, then no vertex in M can be in part $p(L_2)$; otherwise, no vertex of O is in L_2. Thus, the two new instances Φ_1, Φ_2 cover all the eventualities.

Now, we may assume that $\frac{2n_2}{5} \leq |M| \leq \frac{n_2}{2}$ and $|NM| \geq \frac{n_2}{2}$. Replace Φ with two new instances Φ_1, Φ_2 constructed as follows. Φ_1 is obtained from Φ by making

M drop the part $p(L_2)$ and Φ_2 is obtained from Φ by making NM drop the part $p(L_1)$. This can be justified as follows. Consider any solution of Φ. If $O \cap L_2 \neq \emptyset$, then no vertex in M can be in part $p(L_2)$. Otherwise, no vertex of O is in part L_2; hence, every vertex of O is placed in L_1. Since every vertex in NM has a neighbor in O, this implies $NM \cap p(L_1) = \emptyset$. Thus, the two new instances Φ_1, Φ_2 cover all the eventualities.

Case 2.2. $|X_1| < \frac{n_1}{2}$; hence, $|W_1| \geq \frac{n_1}{2}$.

Observe that in this situation, with respect to the adjacencies in the complement of the graph under consideration, we have $|X_1| \geq \frac{n_1}{2}$. Therefore, we can construct a set of two instances equivalent to Φ in this case by using the logic for Case 2.1 in the complement of the given graph using $\overline{M}$ and by simply reversing the roles played by L_1 and L_2.

Finally, it can be easily verified that for each instance Γ created, $|S_{\mathcal{L}}(\Gamma)|\,|S_{\mathcal{R}}(\Gamma)| \leq \frac{9}{10}|S_{\mathcal{L}}(\Phi)|\,|S_{\mathcal{R}}(\Phi)|$. If no new instances are produced by the above analysis, then Φ has no solution. This completes the description of Procedure 3. □

4. The main theorem.

In this section we focus on the main result of the paper which concerns all list M-partition problems where M is a symmetric 4×4 matrix over $\{0, 1, *\}$. In the following we will refer to the four parts of the partition as A, B, C, and D. Recall from section 1 that the *stubborn problem* is the list M-partition problem where $M_{A,A} = 0$, $M_{B,B} = 0$, $M_{D,D} = 1$, $M_{A,C} = M_{C,A} = 0$, and all other entries are asterisks (see Figure 1.1). The stubborn problem has been shown to be solvable in quasi-polynomial time in [22]; hence, it is unlikely to be NP-complete.

THEOREM 4.1. *Suppose M with dimension* 4 *is neither the matrix for the stubborn problem nor its complement. Then the list M-partition problem is solvable in polynomial time or NP-complete. In particular, the list M-partition problem is solvable in polynomial time, except when M contains the matrix for* 3*-colorability, stable cutset, or their complements, or M is the matrix for stable cutset pair, $2K_2$, or their complements, in which cases the problem is NP-complete.*

In proving Theorem 4.1 we employ the tools and procedures described in the previous sections. Given an instance $\mathcal{I}$ of the list M-partition problem, Procedures 1, 2, and 3 are recursively applied to create a polynomial number of new instances $\mathcal{I}_i$ that together are equivalent to the given instance. The resulting instances $\mathcal{I}_i$ are each such that there is a list $\mathcal{L}$ for which the set of vertices $S_{\mathcal{L}}(\mathcal{I}_i)$ with list $\mathcal{L}$ is empty, whereas $S_{\mathcal{L}}(\mathcal{I})$ was not empty. Care must be taken in applying the procedures and tools not to recreate vertices with list $\mathcal{L}$ and thus, reintroduce $S_{\mathcal{L}}$ into subsequent instances of the problem. This can happen as a result of the procedures and tools themselves or the reduction operation that is applied whenever a new instance is created. If any list is (re-)introduced, this list will be a proper subset of a list involved in the operation. This can be easily verified by examining the details of the procedures and the reduction operation.

For simplicity, we write $\mathcal{L}$ (without set brackets) for $\mathrm{S}_{\mathcal{L}}$; for example, $ABC = S_{ABC}$. Theorem 4.1 will be proved via a sequence of lemmata, similar to the treatment in [22].

Proof of Theorem 4.1. If M is a matrix over $\{0, *\}$ or $\{1, *\}$, the result follows from Corollary 1.2 [16, 19, 20, 21]. We can therefore assume that M has at least one 0 and at least one 1. By Theorem 1.3, the only 3-part subproblems that are NP-complete are the stable cutset problem, the 3-colorability problem and their complements, and all others are solvable in polynomial time. By Tool 5, if M contains the matrix for any of these NP-complete subproblems, then the problem is NP-complete. Otherwise, the following lemmata show that the problem can be reduced to a polynomial number of instances that are together equivalent to the given instance, and such that each instance can be solved in polynomial time. The NP-completeness results we employ are well known [14, 27].

The next two lemmata cover the cases when M has an off-diagonal 0 and an off-diagonal 1.

LEMMA 4.2. *Suppose $M_{A,B} = 1$ and $M_{C,D} = 0$. Then the list M-partition problem is solvable in polynomial time or NP-complete.*

Proof. Recall Notes 1, 2, and 3. Given the original instance $\mathcal{I}$, if $ABCD$ is not empty, we recursively apply Procedure 1 with $\mathcal{L} = \{A, B, C, D\}$, $\mathcal{U} = \{A, B\}$, and $\mathcal{F} = \{C, D\}$ to obtain a polynomial number of instances $\mathcal{I}_i$ that together are equivalent to $\mathcal{I}$ such that, for each i, $ABCD(\mathcal{I}_i) = \emptyset$. We now consider the instances $\mathcal{I}_i$.

For each instance $\mathcal{I}_i$, first recursively apply Procedure 2 with $\mathcal{L} = \{A, B, C\}$ and $\mathcal{R} = \{A, B, D\}$, and then (working in $\overline{G}$ using $\overline{M}$) recursively apply Procedure 2 to the resulting instances with $\mathcal{L} = \{A, C, D\}$ and $\mathcal{R} = \{B, C, D\}$ to obtain a polynomial number of instances $\mathcal{J}_j$ that together are equivalent to $\mathcal{I}_i$ and such that, for each j, either $ABC(\mathcal{J}_j) = \emptyset$ or $ABD(\mathcal{J}_j) = \emptyset$, and either $ACD(\mathcal{J}_j) = \emptyset$ or $BCD(\mathcal{J}_j) = \emptyset$.

We now consider the resulting instances $\mathcal{J}_j$. There are four types:

1. $ABC, ACD \neq \emptyset$,
2. $ABC, BCD \neq \emptyset$,
3. $ABD, ACD \neq \emptyset$,
4. $ABD, BCD \neq \emptyset$.

Since the four types are symmetric, we only need to consider instances $\mathcal{J}_j$ of type 1. In this case the possible remaining nonempty sets are ABC, ACD, AB, AC, AD, BC, BD, CD. Recursively, apply Procedure 3 to each $\mathcal{J}_j$ and then to the resulting instances with pairs $\mathcal{L}, \mathcal{R}$, in the following sequence: step (a) $\mathcal{L} = \{B, D\}$ and $\mathcal{R} = \{A, C, D\}$, step (b) $\mathcal{L} = \{B, D\}$ and $\mathcal{R} = \{A, B, C\}$, step (c) $\mathcal{L} = \{A, D\}$ and $\mathcal{R} = \{A, B, C\}$, until one of the two sets involved is empty. This will produce (and will be justified shortly) a polynomial number of instances $\mathcal{K}_k$ that together are equivalent to $\mathcal{J}_j$ and such that each instance $\mathcal{K}_k$ has possible remaining nonempty sets as in one of the following cases:

Case 1. AB, AC, AD, BC, BD, CD,

Case 2. ABC, AB, AC, BC, CD,

Case 3. ABC, ACD, AB, AC, AD, BC, CD.

After step (a), the new instances $\mathcal{K}_k$ either have $BD = \emptyset$ (Case 3) or $ACD = \emptyset$. After step (b), we have either $ABC = \emptyset$ (Case 1) or $BD = \emptyset$ (Case 3 again). In the latter case, we proceed to step (c), after which we have either $AD = \emptyset$ (Case 2) or $ABC = \emptyset$ (Case 1). (Note that if there are vertices with lists of length 3, then the reduction operation may produce a vertex with a list of length 1 or 2 that can be derived from the length 3 list by dropping parts.)

Now we consider the instances $\mathcal{K}_k$.

Case 1. This case can be formulated as a 2-satisfiability problem (2-SAT) and solved in polynomial time (see Tool 1).

Case 2. In the case that $M_{A,C} = 1$ or $M_{B,C} = 1$, we recursively apply Procedure 3 with $\mathcal{L} = \{C, D\}$ and $\mathcal{R} = \{A, B, C\}$. This will create instances in each of which either ABC is empty or CD is empty. In the former case, the problem reduces to 2-SAT. In the latter case, in every instance created there is no vertex with a list containing part D. Thus, the problem is reduced to a 3-part list M-partition problem (3-part problem).

If $M_{A,C} = 0$ ($M_{B,C} = 0$), we recursively apply Procedure 1 with $\mathcal{L} = \{A, B, C\}$, $\mathcal{U} = \{A, B\}$, and $\mathcal{F} = \{A, C\}$ ($\mathcal{L} = \{A, B, C\}$, $\mathcal{U} = \{A, B\}$, and $\mathcal{F} = \{B, C\}$). This will create instances in which ABC is empty, and thus, the problem is reduced to 2-SAT.

Therefore, we can now assume that $M_{A,C} = M_{B,C} = *$.

If $M_{C,C} = 1$, by Tool 3, we create one instance in which no vertex has part C in its list, and at most n instances in each of which no vertex has both C and D in its list. In these cases, the problem is reduced to a 3-part problem.

If $M_{C,C} = 0$, then we recursively apply Procedure 1 with $\mathcal{L} = \{A, B, C\}$, $\mathcal{U} = \{A, B\}$, and $\mathcal{F} = \{C\}$. This will create instances which can be solved using 2-SAT.

Hence, we can now assume that $M_{C,C} = *$.

If $M_{A,A} = 0$ ($M_{B,B} = 0$), by Tool 3, we create one instance in which no vertex has part A (part B) in its list, and at most n instances in each of which no vertex has both A and B in its list. In these cases, the problem is reduced to 2-SAT.

If $M_{A,A} = *$, then as $M_{C,C} = M_{A,C} = *$, the instance can be solved trivially by first placing the vertices whose lists contain A in part A, and then placing any remaining vertices whose lists contain C in part C. Similarly, if $M_{B,B} = *$, the problem can be solved trivially.

So, we can now assume that $M_{A,A} = M_{B,B} = 1$.

If $M_{B,D} = 0$ ($M_{A,D} = 0$), then C dominates B (C dominates A). We can then use Tool 4 to derive an equivalent instance where no vertex has the list ABC, and hence can be solved using 2-SAT.

If $M_{B,D} = 1$ ($M_{A,D} = 1$), by Tool 3, we create one instance in which no vertex has part D in its list, and at most n instances in each of which no vertex has both B and C (both A and C) in its list. In these cases, we either have an instance that is a 3-part problem or can be solved using 2-SAT.

We finally can assume that $M_{A,D} = M_{B,D} = *$. Since B dominates A, we can use Tool 4 to derive an equivalent instance where no vertex has the list ABC, and hence, can be solved using 2-SAT.

Case 3. Suppose we are able to produce an equivalent set of instances in each of which $ACD = \emptyset$, and hence, the possible nonempty sets are ABC, AB, AC, AD, BC, CD. Then, recursively applying Procedure 3 with $\mathcal{L} = \{A, D\}$ and $\mathcal{R} = \{A, B, C\}$ will produce instances each of which either can be solved using 2-SAT or has $AD = \emptyset$, which is settled by Case 2. A similar analysis can be made when an equivalent set of instances can be produced in each of which $ABC = \emptyset$. Suppose $ABC = \emptyset$, recursively applying Procedure 3 with $\mathcal{L} = \{B, C\}$ and $\mathcal{R} = \{A, C, D\}$ will produce instances each of which either can be solved using 2-SAT or has $BC = \emptyset$. The latter case is reduced to Case 2 by working in $\overline{G}$ in place of G and using $\overline{M}$ in place of M. Therefore, we aim to produce equivalent instances in each of which either $ABC = \emptyset$ or $ACD = \emptyset$.

If $M_{A,C} = 0$, then the lists $\mathcal{L} = \{A, B, C\}$ and $\mathcal{R} = \{A, C, D\}$ fail to satisfy the conditions for Procedure 2. However, with respect to $\overline{G}$ and $\overline{M}$, they do satisfy the conditions for Procedure 2. Hence, we recursively apply Procedure 2 with $\mathcal{L}$ and $\mathcal{R}$ in $\overline{G}$ using $\overline{M}$ to create instances in each of which either $ABC = \emptyset$ or $ACD = \emptyset$.

If $M_{A,C} = 1$, then recursively apply Procedure 2 with $\mathcal{L} = \{A, B, C\}$ and $\mathcal{R} = \{A, C, D\}$ to create instances in each of which either $ABC = \emptyset$ or $ACD = \emptyset$.

We can therefore assume that $M_{A,C} = *$.

If $M_{A,A} = 0$ ($M_{B,B} = 0$), using Tool 3, we create one instance in which no vertex has part A (part B) in its list, and at most n instances in each of which no vertex has both A and B in its list; hence, $ABC = \emptyset$.

If $M_{A,A} = 1$, recursively apply Procedure 1 with $\mathcal{L} = \{A, C, D\}$, $\mathcal{U} = \{A\}$, and $\mathcal{F} = \{C, D\}$ to produce instances in each of which $ACD = \emptyset$.

Therefore, we can assume that $M_{A,A} = *$.

If $M_{C,C} = 1$, using Tool 3, we create one instance in which no vertex has part C in its list, and at most n instances in each of which no vertex has both C and D in its list; hence, $ACD = \emptyset$.

If $M_{C,C} = 0$, recursively apply Procedure 1 with $\mathcal{L} = \{A, B, C\}$, $\mathcal{U} = \{A, B\}$, and $\mathcal{F} = \{C\}$ to produce instances in each of which $ABC = \emptyset$.

Therefore, we can assume that $M_{C,C} = *$.

We now have instances in which $M_{A,A} = M_{C,C} = M_{A,C} = *$. Such an instance can be solved trivially by first placing the vertices whose lists contain A in part A, and then placing any remaining vertices whose lists contain C in part C. □

Recall that the *list generalized $\mathcal{P}$ problem* is the list M'-partition problem where M' is obtained from the matrix M for list partition problem $\mathcal{P}$ by changing some asterisks to either 0 or 1.

Corollary 4.3. *Each list generalized skew partition problem is solvable in polynomial time, except when it contains the stable cutset problem or its complement, in which cases the problem is NP-complete.*

Proof. Observe (via Theorem 1.3) that the only possible 3-part subproblems that are NP-complete are the stable cutset problem and its complement, and all others are solvable in polynomial time. By Tool 5, if M contains the matrix for the stable cutset problem or its complement, then the problem is NP-complete. Otherwise, the problem is polynomial-time solvable by Lemma 4.2. □

From here on, we write the proofs in an abbreviated style. When Tool 3 is applied an instance that is a 3-part problem is always created; this will now be assumed and not explicitly stated. The full details can be written in the same manner as the proof of Lemma 4.2.

Lemma 4.4. *Suppose $M_{A,B} = 0$ and $M_{A,D} = 1$. Then the list M-partition problem is solvable in polynomial time or NP-complete.*

Proof. Using Tool 3, either the problem is reduced to a 3-part problem (no A), or no list contains $\{B, D\}$. Assuming the latter, the possible nonempty sets are ABC, ACD, AB, AC, AD, BC, CD. By Lemma 4.2 we may assume $M_{C,D} \neq 1$ and $M_{B,C} \neq 0$.

Suppose $M_{A,C} = 1$. Then no list contains $\{B, C\}$ (using Tool 3). Apply Procedure 3 to the pair AB, ACD. If AB becomes empty, then we have a 3-part problem on $\{A, C, D\}$. Otherwise, the instance can be solved using 2-SAT.

Suppose $M_{A,C} = 0$, and no list contains $\{C, D\}$ (using Tool 3). Apply Procedure 3 to the pair AD, ABC to get either 3-part problems or instances solvable using 2-SAT. Therefore, $M_{A,C} = *$.

If $M_{C,D} = 0$, then no list contains $\{A, C\}$ (using Tool 3) and we get instances solvable using 2-SAT. Therefore, $M_{C,D} = *$.

If $M_{B,C} = 1$, then no list contains $\{A, C\}$ (using Tool 3) and we get instances solvable using 2-SAT. Therefore, $M_{B,C} = *$.

If $M_{C,C} = *$, then C dominates parts A, B, and D, and we get an instance solvable using 2-SAT.

If $M_{C,C} = 0$, apply Procedure 1 to ACD so that ACD becomes empty. Then apply Procedure 3 to the pair AD, ABC. If ABC becomes empty, we get instances solvable using 2-SAT. Otherwise, now apply Procedure 3 to the pair CD, ABC to get 3-part problems or instances solvable using 2-SAT.

Now we have $M_{C,C} = 1$. Apply Procedure 1 to ABC so that ABC becomes empty. Then apply Procedure 3 to the pair AB, ACD. If ACD becomes empty, we get instances solvable using 2-SAT. Otherwise, now apply Procedure 3 to the pair BC, ACD to get 3-part problems or instances solvable using 2-SAT. □

Graphs for which the vertex-set can be partitioned into two stable sets and two cliques are called (2,2)-graphs. Brandstädt [2, 3] introduced this class and gave the first polynomial-time algorithm for recognition. Recognition of (2,2)-graphs is the M-partition problem where $M_{A,A} = 1$, $M_{B,B} = 1$, $M_{C,C} = 0$, and $M_{D,D} = 0$, and all other entries are asterisks. The following result was proved in [22]; we provide a proof using different techniques.

LEMMA 4.5. *All list generalized* (2,2)*-graph recognition problems are solvable in polynomial time.*

Proof. Repeatedly apply Procedure 1 to the following sets to eliminate them, one by one: $ABCD$, ABC, ABD, ACD, BCD. Then use 2-SAT. □

Based on the previous lemmata and Tool 6, we can now assume that 1 occurs only on the diagonal and that the off-diagonal entries are either 0 or $*$. We first consider the case that there are at least two 1's on the diagonal.

LEMMA 4.6. *Suppose there are at least two* 1*'s on the diagonal and all off-diagonal entries are* $*$*. Then the list* M*-partition problem is solvable in polynomial time or NP-complete.*

Proof. If one of the diagonal entries, say $M_{A,A}$, is $*$, A dominates the other parts; hence, the problem can be reduced to a 3-part problem on $\{B, C, D\}$. On the other hand, suppose none of the diagonal entries are $*$. When there are two 1's and two 0's on the diagonal we get a problem solvable in polynomial time (see Lemma 4.5). Otherwise, the problem is NP-complete via the complement of 3-colorability and Tool 5. (This subcase is also covered by Theorem 1.4.) □

We can now assume that there is at least one off-diagonal entry that is 0. The next three lemmata cover the possible position of the off-diagonal 0 with respect to the two or more 1's assumed to be on the diagonal.

LEMMA 4.7. *Suppose all off-diagonal entries are* 0 *or* $*$, $M_{B,B} = M_{D,D} = 1$, *and* $M_{A,C} = 0$. *Then the list* M*-partition problem is solvable in polynomial time or NP-complete.*

Proof. Apply Procedure 1 to eliminate set $ABCD$.

Apply Procedure 1 to eliminate set ABC.

Apply Procedure 1 to eliminate set ACD.

Apply Procedure 2 to the pair ABD, BCD so that one of the sets is eliminated.

Assume $BCD = \emptyset$. (The other case is similar.)

Apply Procedure 3 to the pair BC, ABD so that one of the sets is eliminated.

Apply Procedure 3 to the pair CD, ABD so that one of the sets is eliminated.

We now can assume that the remaining nonempty sets are ABD, AB, AC, AD, BD; otherwise, we can use 2-SAT.

If $M_{A,B} = 0$, then, using Tool 3, no list contains $\{A, B\}$; we can now use 2-SAT.

If $M_{A,D} = 0$, then, using Tool 3, no list contains $\{A, D\}$; we can now use 2-SAT.

Otherwise, the hypothesis of the lemma implies $M_{A,B} = M_{A,D} = *$.

If $M_{A,A} = 0$, then apply Procedure 1 to ABD; we can now use 2-SAT.

If $M_{A,A} = 1$, then, by Tool 3, no list contains $\{A, C\}$; we get a 3-part problem.

Therefore, $M_{A,A} = *$.

If $M_{B,C} = 0$, then A dominates B and no list contains $\{A, B\}$; we can now use 2-SAT.

If $M_{C,D} = 0$, then A dominates D and no list contains $\{A, D\}$; we can now use 2-SAT.

Otherwise, the hypothesis of the lemma implies $M_{B,C} = M_{C,D} = *$.

If $M_{C,C} = 0$, then A dominates C and no list contains $\{A, C\}$; we get a 3-part problem.

If $M_{C,C} = 1$, then, by Tool 3, no list contains $\{A, C\}$; we get a 3-part problem.

Therefore, $M_{C,C} = *$.

Place vertices with list AC in the part A to get a 3-part problem on $\{A, B, D\}$ in which A dominates other parts; we can now use 2-SAT. $\square$

COROLLARY 4.8. *The list* 2*-clique cutset problem is solvable in polynomial time.*

Proof. Lemma 4.7 covers the list 2-clique cutset problem: $M_{B,B} = M_{D,D} = 1$, $M_{A,C} = M_{C,A} = 0$, and all other entries are asterisks. It can be verified (via Theorem 1.3) that every 3-part problem produced in that case in the proof of Lemma 4.7 is solvable in polynomial time. $\square$

COROLLARY 4.9. *Each list generalized* 2*-clique cutset problem is solvable in polynomial time, except when it contains the complement of the* 3*-colorability problem, in which case it is NP-complete.*

Proof. Observe (via Theorem 1.3) that in this case the only possible 3-part subproblem that is NP-complete is the complement of 3-colorability, and all others are solvable in polynomial time. By Tool 5, if M contains the matrix for the complement of 3-colorability, then the problem is NP-complete. Otherwise, the problem is polynomial-time solvable by Lemmata 4.2, 4.4, and 4.7. $\square$

LEMMA 4.10. *Suppose all off-diagonal entries are* 0 *or* $*$*,* $M_{A,A} = M_{B,B} = 1$*, and* $M_{A,C} = 0$*. Then the list* M*-partition problem is solvable in polynomial time or NP-complete.*

Proof. Using Tool 3, no list contains $\{A, C\}$. The possible nonempty sets are ABD, BCD, AB, AD, BC, BD, CD.

By previous lemmata, $M_{C,D} = *$ and $M_{D,D} \neq 1$.

If $M_{D,D} = 0$, then apply Procedure 1 to ABD. Then, apply Procedure 3 to the pair AB, BCD. If BCD is eliminated we can use 2-SAT; otherwise, apply Procedure 1 to AD to get a 3-part problem.

Therefore, $M_{D,D} = *$.

If $M_{A,D} = 0$, then, using Tool 3, no list contains $\{A, D\}$. Apply Procedure 3 to the pair AB, BCD to get a 3-part problem or we can use 2-SAT.

Therefore, $M_{A,D} = *$.

If $M_{B,D} = 0$, then, using Tool 3, no list contains $\{B, D\}$; we can now use 2-SAT.

Therefore, $M_{B,D} = *$.

Now, D dominates the other parts; we get a 3-part problem. $\square$

LEMMA 4.11. *Suppose all off-diagonal entries are* 0 *or* $*$*,* $M_{A,A} = M_{C,C} = 1$*, and* $M_{A,C} = 0$*. Then the list* M*-partition problem is solvable in polynomial time or NP-complete.*

Proof. From previous lemmata, $M_{A,B} = M_{A,D} = M_{B,C} = M_{B,D} = M_{C,D} = *$, $M_{B,B} \neq 1$, and $M_{D,D} \neq 1$. Using Tool 3, no list contains $\{A, C\}$.

If $M_{B,B} = *$, then B dominates other parts; we get a 3-part problem.

If $M_{D,D} = *$, then D dominates other parts; we get a 3-part problem.

Therefore, $M_{B,B} = M_{D,D} = 0$.

Apply Procedure 1 to ABD. Then, apply Procedure 1 to BCD; we can now use 2-SAT. □

In the remaining case M has exactly one 1 and it is on the diagonal; say $M_{D,D} = 1$. Following [22], we define a *separating statement* for $X = A, B$, or C to be "$M_{X,D} = 0$ or $M_{X,X} = 0$." We divide the remaining cases based on the number of separating statements that hold being three, two, or at most one. The following four lemmata cover the cases when three separating statements hold.

LEMMA 4.12. *Suppose the only* 1 *is at* $M_{D,D}$. *If* $M_{A,A} = M_{B,B} = M_{C,C} = 0$, *then the list* M*-partition problem is solvable in polynomial time or NP-complete.*

Proof. If the subproblem on the parts $\{A, B, C\}$ corresponds to 3-colorability, then we have an NP-complete problem by Tool 5. Therefore, without loss of generality, we can assume $M_{A,B} = 0$.

If $M_{B,C} = 0$, then the following reduces the instance to 3-part problems: Apply Procedure 1 to $ABCD$, apply Procedure 1 to ABD, apply Procedure 1 to ACD, apply Procedure 1 to BCD, apply Procedure 1 to AD, apply Procedure 1 to BD, and then apply Procedure 1 to CD.

Therefore, $M_{B,C} = *$. Similarly, $M_{A,C} = *$.

As there is a single 1 in M, each of $M_{A,D}$ and $M_{B,D}$ is constrained to be 0 or $*$. In any such case, A dominates B or B dominates A, and no list contains $\{A, B\}$. Apply Procedure 1 to ACD. Then, apply Procedure 1 to BCD. We can now use 2-SAT. □

LEMMA 4.13. *Suppose the only* 1 *is at* $M_{D,D}$. *If* $M_{B,B} = M_{C,C} = M_{A,D} = 0$, *then the list* M*-partition problem is solvable in polynomial time or NP-complete.*

Proof. Using Tool 3, no list contains $\{A, D\}$. The following will then produce 3-part problems on $\{A, B, C\}$: apply Procedure 1 to BCD, apply Procedure 1 to BD, and then apply Procedure 1 to CD. □

LEMMA 4.14. *Suppose the only* 1 *is at* $M_{D,D}$. *If* $M_{C,C} = M_{A,D} = M_{B,D} = 0$, *then the list* M*-partition problem is solvable in polynomial time or NP-complete.*

Proof. Using Tool 3, no list contains $\{A, D\}$ and also no list contains $\{B, D\}$. Apply Procedure 1 to CD to get 3-part problems on $\{A, B, C\}$. □

LEMMA 4.15. *Suppose the only* 1 *is at* $M_{D,D}$. *If* $M_{A,D} = M_{B,D} = M_{C,D} = 0$, *then the list* M*-partition problem is solvable in polynomial time or NP-complete.*

Proof. Using Tool 3, no list contains $\{A, D\}$, no list contains $\{B, D\}$, and also no list contains $\{C, D\}$. We get 3-part problems on $\{A, B, C\}$. □

The next three lemmata cover the cases when exactly two separating statements hold, say for A and B.

LEMMA 4.16. *Suppose the only* 1 *is at* $M_{D,D}$. *If* $M_{A,D} = M_{B,D} = 0$ *and* $M_{C,C} = M_{C,D} = *$, *then the list* M*-partition problem is solvable in polynomial time or NP-complete.*

Proof. Using Tool 3, no list contains $\{A, D\}$ and also no list contains $\{B, D\}$. C dominates D and no list contains $\{C, D\}$. We get 3-part problems on $\{A, B, C\}$. □

LEMMA 4.17. *Suppose the only* 1 *is at* $M_{D,D}$. *If* $M_{A,A} = M_{B,B} = 0$, $M_{C,C} = M_{C,D} = *$, *and the list* M*-partition problem is different from the stubborn problem, then it is solvable in polynomial time or NP-complete.*

Proof. If $M_{A,C} = M_{B,C} = *$, then C dominates all other parts, so we obtain a 3-part problem. Therefore, without loss of generality, assume $M_{A,C} = 0$.

Suppose $M_{B,C} = 0$. Then we can apply Procedure 1 to eliminate the following sets in sequence: $ABCD, ABD, ACD, BCD, AD, BD$. Let $X = AB$ and Y be the union of sets ABC, AC, BC, and CD (i.e., X is the set of vertices with list $\{A, B\}$ and Y is the set of vertices with any of the possible remaining lists). Suppose we had $u \in X$ and $v \in Y$ such that u and v are adjacent. Since $M_{A,C} = M_{B,C} = 0$, in any solution to the problem, v cannot be placed in part C. Therefore, by making such vertices v drop the part C from their lists (hence, leave the set Y), we get instances where there are no edges between vertices in X and vertices in Y. We can then solve such an instance by placing every vertex in Y in part C and testing whether X induces a bipartite graph. Therefore, we can assume that $M_{B,C} = *$ (and $M_{A,C} = 0$). C dominates A, so no list contains $\{A, C\}$.

Apply Procedure 1 to ABD, then to AD, and then to BD. The possible remaining nonempty sets are BCD, AB, BC, CD.

If $M_{A,B} = 0$, then C dominates B, so no list contains $\{B, C\}$; we can now use 2-SAT.

If $M_{A,D} = 0$, then C dominates D, so no list contains $\{C, D\}$; we can now use 2-SAT.

If $M_{B,D} = 0$, then (using Tool 3) no list contains $\{B, D\}$; we can now use 2-SAT.

Therefore, $M_{A,B} = M_{A,D} = M_{B,D} = *$ and we are left with the stubborn problem. $\square$

We note that the proof of Lemma 4.17 shows that the stubborn problem can be reduced to an equivalent set of instances where for each instance the only possible lists are A, B, C, D, AB, BC, CD, and BCD.

LEMMA 4.18. *Suppose the only* 1 *is at* $M_{D,D}$*. If* $M_{A,D} = M_{B,B} = 0$*,* $M_{A,A} = M_{B,D} = *$*, and* $M_{C,C} = M_{C,D} = *$*, then the list* M*-partition problem is solvable in polynomial time or NP-complete.*

Proof. Using Tool 3, no list contains $\{A, D\}$. Therefore, the possible remaining nonempty sets are ABC, BCD, AB, AC, BC, BD, CD.

If $M_{A,C} = *$ or $M_{A,B} = 0$, then C dominates B, so no list contains $\{B, C\}$, and we can use 2-SAT. Therefore, $M_{A,C} = 0$ and $M_{A,B} = *$.

If $M_{B,C} = *$, then the subproblem on $\{A, B, C\}$ is the stable cutset problem. Therefore, $M_{B,C} = 0$.

Apply Procedure 1 to BCD and then to BD. Now the possible remaining nonempty sets are ABC, AB, AC, BC, CD. Let $X = AB$ and Y be the union of sets ABC, AC, BC, and CD (i.e., X is the set of vertices with list $\{A, B\}$ and Y is the set of vertices with any of the possible remaining lists). Suppose we had $u \in X$ and $v \in Y$ such that u and v are adjacent. Since $M_{A,C} = M_{B,C} = 0$, in any solution to the problem, v cannot be placed in part C. Therefore, by making such vertices v drop the part C from their lists (hence, leave the set Y), we get instances where there are no edges between vertices in X and vertices in Y. We can then solve such an instance by placing every vertex in Y in part C, and placing every vertex in X in part A. $\square$

The only remaining case is when the only 1 is at $M_{D,D}$ and at most one separating statement holds, say, for part A.

LEMMA 4.19. *Suppose the only* 1 *is at* $M_{D,D}$*. If* $M_{B,B} = M_{B,D} = M_{C,C} = M_{C,D} = *$*, then the list* M*-partition problem is solvable in polynomial time or NP-complete.*

Proof. Suppose $M_{B,C} = *$. If $M_{A,B} = *$ ($M_{A,C} = *$), then B (C) dominates all the other parts to yield a 3-part problem. On the other hand, if $M_{A,B} = M_{A,C} = 0$, then rows B and C in M are identical; hence, parts B and C can be identified.

Therefore, we can assume $M_{B,C} = 0$. We divide the cases based on the value of the triple ($M_{A,A}$, $M_{A,B}$, $M_{A,C}$).

Case $(*,*,*)$. If $M_{A,D} = *$, then A dominates all other parts to yield a 3-part problem. Hence, $M_{A,D} = 0$, and by Tool 3, no list contains $\{A, D\}$.

Apply Procedure 1 to BCD. Apply Procedure 3 to the pair BD, CD. Solve the problem as follows: Place vertices with lists $\{A, B, C\}$, $\{A, B\}$, or $\{A, C\}$ in part A. Then, if BD is empty, place vertices with lists $\{B, C\}$ or $\{C, D\}$ in part C, and if CD is empty, place vertices with lists $\{B, C\}$ or $\{B, D\}$ in part B.

Case $(*, 0, *)$. If $M_{A,D} = *$, then rows A and C in M are identical; hence, parts A, C can be identified. Therefore, we can assume $M_{A,D} = 0$. Then, by Tool 3, no list contains $\{A, D\}$. Also, C dominates A and no list contains $\{A, C\}$. Apply Procedure 1 to BCD and then use 2-SAT.

Case $(0, 0, 0)$, $(0, 0, *)$. C dominates A, so no list contains $\{A, C\}$. Apply Procedure 1 to ABD, then to BCD, and use 2-SAT.

Case $(0, *, *)$. This case contains the stable cutset problem; hence, by Tool 5, it is NP-complete.

Case $(*, 0, 0)$. Apply Procedure 1 to the following sets one by one: $ABCD$, ABD, ACD, and BCD. If we had $u \in AB$ (BC, AC) and $v \in ABC$ such that u and v are adjacent, then v must drop C (A, B, respectively) and leave ABC. Therefore, there are no edges between ABC and any of AB, BC, or AC. Now, apply Procedure 3 to AD, CD, then to AD, BD, and finally to CD, BD. We can now assume that exactly one of AD, BD, CD is nonempty.

Suppose AD is nonempty. If $M_{A,D} = 0$, then use Tool 3 to eliminate the set AD and obtain a 3-part problem. If $M_{A,D} = *$, then place vertices with list ABC in part A and solve using 2-SAT. If BD is nonempty, then place vertices with list $\{A, B, C\}$ in part B and solve using 2-SAT. If CD is nonempty, then place vertices with list $\{A, B, C\}$ in part C and solve using 2-SAT.

Case $(0, *, 0)$, $(*, *, 0)$. B dominates A, so no list contains $\{A, B\}$. Apply Procedure 1 to ACD, then to BCD, and solve using 2-SAT. □

Thus, Theorem 4.1 is proved via the following cases.

1. M is a matrix over $\{0, *\}$ or $\{1, *\}$: Corollary 1.2.
2. M has at least one 0 and at least one 1:
   - 2.1 M has an off-diagonal 0 and an off-diagonal 1: Lemmata 4.2 and 4.4.
   - 2.2 M has 1 (resp., 0) only on the diagonal and all off-diagonal entries are either 0 or $*$ (resp., 1 or $*$):
     - 2.2.1 M has at least two 1's (resp., 0's) on the diagonal:
       - 2.2.1.1 all off-diagonal entries are $*$: Lemma 4.6.
       - 2.2.1.2 at least one off-diagonal entry is 0 (resp., 1): Lemmata 4.7, 4.10, and 4.11.
     - 2.2.2 M has exactly one 1 (resp., 0) on the diagonal: Lemmata 4.12 through 4.19. □

Note added in proof. In recent related work the list partition problem on some special classes of graphs [18, 23] and some specific graph partition problems with all parts nonempty [11, 12, 13] have been studied.

REFERENCES

[1] B. Aspvall, F. Plass, and R. E. Tarjan, *A linear time algorithm for testing the truth of certain quantified boolean formulas*, Inform. Process. Lett., 8 (1979), pp. 121–123.

[2] A. Brandstädt, *Partitions of graphs into one or two independent sets and cliques*, Discrete Math., 152 (1996), pp. 47–54; Corrigendum, Discrete Math., 186 (1998), p. 295.

[3] A. Brandstädt, V. B. Le, and T. Szymczak, *The complexity of some problems related to graph 3-colorability*, Discrete Appl. Math., 89 (1998), pp. 59–73.

[4] K. Cameron, E. M. Eschen, C. T. Hoàng, and R. Sritharan, *The list partition problem for graphs*, in Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2004, ACM, New York, pp. 391–399.

[5] M. Chudnovsky, N. Robertson, P. Seymour, and R. Thomas, *The strong perfect graph theorem*, Ann. of Math. (2), 164 (2006), pp. 51–229.

[6] V. Chvátal, *Star-cutsets and perfect graphs*, J. Combin. Theory Ser. B, 39 (1985), pp. 189–199.

[7] V. Chvátal, *Website on Perfect Graph Problems*, http://athos.rutgers.edu/∼chvatal/perfect/problems.html.

[8] M. Conforti, G. Cornuéjols, A. Kapoor, and K. Vusković, *Even-hole-free graphs.* I. *Decomposition theorem*, J. Graph Theory, 39 (2002), pp. 6–49.

[9] M. Conforti, G. Cornuéjols, A. Kapoor, and K. Vusković, *Even-hole-free graphs.* II. *Recognition algorithm*, J. Graph Theory, 40 (2002), pp. 238–266.

[10] A. Cournier and M. Habib, *A new linear algorithm for modular decomposition*, in Trees in Algebra and Programming—CAAP' 94, Lecture Notes in Comput. Sci. 787, Springer, Berlin, 1994, pp. 68–84.

[11] S. Dantas, C. M. H. de Figueiredo, S. Gravier, and S. Klein, *Finding H-partitions efficiently*, Theor. Inform. Appl., 39 (2005), pp. 133–144.

[12] S. Dantas, C. M. H. de Figueiredo, S. Gravier, and S. Klein, *Extended skew partition problem*, Discrete Math., 306 (2006), pp. 2438–2449.

[13] S. Dantas, C. M. H. de Figueiredo, S. Klein, S. Gravier, and B. A. Reed, *Stable skew partition problem*, Discrete Appl. Math., 143 (2004), pp. 17–22.

[14] C. M. H. de Figueiredo and S. Klein, *The NP-completeness of multipartite cutset testing*, Congr. Numer., 119 (1996), pp. 217–222.

[15] C. M. H. de Figueiredo, S. Klein, Y. Kohayakawa, and B. A. Reed, *Finding skew partitions efficiently*, J. Algorithms, 37 (2000), pp. 505–521.

[16] T. Feder and P. Hell, *List homomorphisms to reflexive graphs*, J. Combin. Theory Ser. B, 72 (1998), pp. 236–250.

[17] T. Feder and P. Hell, *Full constraint satisfaction problems*, SIAM J. Comput., 36 (2006), pp. 230–246.

[18] T. Feder, P. Hell, and W. Hochstattler, *Generalized colourings (matrix partitions) of cographs*, in Graph Theory in Paris, A. Bondy et al., eds., Trends Math., Birkhäuser Verlag, Basel, 2007, pp. 149–167.

[19] T. Feder, P. Hell, and J. Huang, *List homomorphisms and circular arc graphs*, Combinatorica, 19 (1999), pp. 487–505.

[20] T. Feder, P. Hell, and J. Huang, *Bi-arc graphs and the complexity of list homomorphisms*, J. Graph Theory, 42 (2003), pp. 61–80.

[21] T. Feder, P. Hell, and J. Huang, *private communicaton.*

[22] T. Feder, P. Hell, S. Klein, and R. Motwani, *List partitions*, SIAM J. Discrete Math., 16 (2003), pp. 449–478.

[23] T. Feder, P. Hell, S. Klein, L. T. Nogueira, and F. Protti, *List matrix partitions of chordal graphs*, Theoret. Comput. Sci., 349 (2005), pp. 52–66.

[24] T. Feder, P. Hell, D. Král, and J. Sgall, *Two algorithms for general list matrix partitions*, in Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2005, ACM, New York, pp. 870–876.

[25] T. Feder, P. Hell, and K. M. Tucker-Nally, *Digraph matrix partitions and trigraph homomorphisms*, Discrete Appl. Math., 154 (2006), pp. 2458–2469.

[26] T. Feder and M. Y. Vardi, *The computational structure of monotone monadic SNP and constraint satisfaction: A study through Datalog and group theory*, SIAM J. Comput., 28 (1998), pp. 57–104.

[27] M. R. Garey and D. S. Johnson, *Computers and Intractability*, W. H. Freeman and Company, San Francisco, 1979.

[28] M. C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[29] R. Hayward and B. A. Reed, *Forbidding holes and antiholes*, in Perfect Graphs, Wiley-Intersci. Ser. Discrete Math. Optim., J. L. Ramirez Alfonsin and B. A. Reed, eds., John Wiley & Sons, Chichester, 2001, pp. 113–137.

[30] P. Hell and J. Nešetřil, *On the complexity of H-coloring*, J. Combin. Theory Ser. B, 48 (1990), pp. 92–110.

[31] Website of the American Institute of Mathematics, *The Perfect Graph Conjecture Workshop*, Oct. 30–Nov. 3, 2002, http://www.aimath.org/WWN/perfectgraph/.

[32] R. M. McConnell and J. P. Spinrad, *Linear time modular decomposition and efficient transitive orientation of comparability graphs*, in Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms (Arlington, VA), 1994, ACM, New York, 1994, pp. 536–545.

[33] R. E. Tarjan, *Decomposition by clique separators*, Discrete Math., 55 (1985), pp. 221–232.

[34] A. Tucker, *Coloring graphs with stable cutsets*, J. Combin. Theory Ser. B, 34 (1983), pp. 258–267.

[35] S. Whitesides, *An algorithm for finding clique cutsets*, Inform. Process. Lett., 12 (1981), pp. 31–32.

[36] S. Whitesides, *A method for solving certain graph recognition and optimization problems, with applications to perfect graphs*, Ann. Discrete Math., 21 (1984), pp. 281–297.

SIAM J. Discrete Math.
Vol. 21, No. 4, pp. 930–935

LOWER BOUNDS ON LOCALITY SENSITIVE HASHING*

RAJEEV MOTWANI†, ASSAF NAOR‡, AND RINA PANIGRAHY†

Abstract. Given a metric space (X, d_X), $c \geq 1$, $r > 0$, and $p, q \in [0, 1]$, a distribution over mappings $\mathscr{H} : X \to \mathbb{N}$ is called a (r, cr, p, q)-sensitive hash family if any two points in X at distance at most r are mapped by $\mathscr{H}$ to the same value with probability at least p, and any two points at distance greater than cr are mapped by $\mathscr{H}$ to the same value with probability at most q. This notion was introduced by Indyk and Motwani in 1998 as the basis for an efficient approximate nearest neighbor search algorithm and has since been used extensively for this purpose. The performance of these algorithms is governed by the parameter $\rho = \frac{\log(1/p)}{\log(1/q)}$, and constructing hash families with small ρ automatically yields improved nearest neighbor algorithms. Here we show that for $X = \ell_1$ it is impossible to achieve $\rho \leq \frac{1}{2c}$. This almost matches the construction of Indyk and Motwani which achieves $\rho \leq \frac{1}{c}$.

Key words. locality sensitive hashing, lower bounds, nearest neighbor search

AMS subject classifications. 68Q17, 68U05

DOI. 10.1137/050646858

1. Introduction.

In this note we study the complexity of finding the nearest neighbor of a query point in certain high dimensional spaces using *locality sensitive hashing* (LSH). The nearest neighbor problem is formulated as follows: Given a database of n points in a metric space, preprocess it so that given a new query point it is possible to quickly find the point closest to it in the data set. This fundamental problem arises in numerous applications, including data mining, information retrieval, and image search, where distinctive features of the objects are represented as points in $\mathbb{R}^d$. There is a vast amount of literature on this topic, which we shall not attempt to discuss here. We refer the interested reader to the papers [6, 5, 4, 8], especially to the references therein, for background on the nearest neighbor problem.

While the exact nearest neighbor problem seems to suffer from the "curse of dimensionality," many efficient techniques have been devised for finding an approximate solution whose distance from the query point is at most c times its distance from the nearest neighbor. One of the most versatile and efficient methods for approximate nearest neighbor search is based on LSH, as introduced by Indyk and Motwani in 1998 [6]. This method has been refined and improved in several papers—the most recent algorithm can be found in [4]. We also refer the reader to the LSH website, where more information on this algorithm can be found, including its implementation and code—all of which can be found at http://web.mit.edu/andoni/www/LSH/index.html. The LSH approach to the approximate nearest neighbor problem is based on the following concept.

DEFINITION 1.1. *Let (X, d_X) be a metric space, $r, R > 0$, and $p, q \in [0, 1]$. A distribution over mappings $\mathscr{H} : X \to \mathbb{N}$ is called a (r, R, p, q)-sensitive hash family if for any $x, y \in X$,*

- $d_X(x, y) \leq r \implies \Pr_{\mathscr{H}}[\mathscr{H}(x) = \mathscr{H}(y)] \geq p,$

[*] Received by the editors December 5, 2005; accepted for publication (in revised form) April 23, 2007; published electronically December 7, 2007.
http://www.siam.org/journals/sidma/21-4/64685.html

[†] Stanford University, Stanford, CA 94305 (rajeev@cs.stanford.edu, rinap@cs.stanford.edu).

[‡] Microsoft Research. Current address: New York University, Courant Institute of Mathematical Sciences, New York, NY 10012 (naor@cims.nyu.edu).

- $d_X(x,y) > R \implies \Pr_{\mathscr{H}}[\mathscr{H}(x) = \mathscr{H}(y)] \le q$.

Given $c \ge 1$ and $q \in (0,1)$, we define $\rho_X(c,q)$ to be the smallest constant $\rho > 0$ such that for every $r > 0$ there exists $p \in (0,1)$ and a (r, cr, p, q)-sensitive hash family $\mathscr{H} : X \to \mathbb{N}$ with $\frac{\log(1/p)}{\log(1/q)} \le \rho$. In other words

(1.1)
$$\rho_X(c,q) = \sup_{r>0} \inf \left\{ \frac{\log(1/p)}{\log(1/q)} : \exists (r,cr,p,q)\text{-sensitive hash family } \mathscr{H} : X \to \mathbb{N} \right\}.$$

Of particular interest is the case $X = \ell_s^d$, for some $s > 0$ and $d \in \mathbb{N}$. Here, and in what follows, ℓ_s^d denotes the space $\mathbb{R}^d$ equipped with the ℓ_s norm $\|(x_1,\dots,x_d)\|_s = (|x_1|^s + \cdots + |x_d|^s)^{1/s}$ (this is only a quasi-norm when $0 < s < 1$). In this case we define

$$\rho_s(c) = \sup_{0<q<1} \limsup_{d\to\infty} \rho_{\ell_s^d}(c,q).$$

The importance of these parameters stems from the following application to approximate nearest neighbor search. It will be convenient to discuss it in the framework of the following decision version of the c-approximate nearest neighbor problem: Given a query point, find any element of the data set which is at distance at most cr from it, provided that there is a data point at distance at most r from the query point. This decision version is known as the (r,cr)-near neighbor problem. It is well known that the reduction to the decision version adds only a logarithmic factor in the time and space complexity [6, 5]. The following theorem was proved in [6]; the exact formulation presented here is taken from [4].

THEOREM 1.2. *Let (X, d_X) be a metric on a subset of $\mathbb{R}^d$. Suppose that (X, d_X) admits a (r, cr, p, q)-sensitive hash family $\mathscr{H}$, and write $\rho = \frac{\log(1/p)}{\log(1/q)}$. Then for any $n \ge \frac{1}{q}$ there exists a randomized algorithm for (r,cr)- near neighbor on n-point subsets of X which uses $O\left(dn + n^{1+\rho}\right)$ space, with query time dominated by $O\left(n^\rho\right)$ distance computations and $O\left(n^\rho \log_{1/q} n\right)$ evaluations of hash functions from $\mathscr{H}$.*

Thus, obtaining bounds on $\rho_X(c)$ is of great algorithmic interest. It is proved in [6] that $\rho_1(c) \le 1/c$, and for small values of c, namely $c \in [1,10]$, is was shown in [4] that this inequality is strict. We refer to [4] for numerical data on the best known estimates for $\rho_1(c)$ for small c. For $s = 2$ a recent result of Andoni and Indyk [1] shows that $\rho_2(c) \le 1/c^2$, and for general $s \in (0,2]$ the best known bounds [4] are $\rho_s(c) \le \max\{1/c, 1/c^s\}$.

The main purpose of this note is to obtain lower bounds on $\rho_1(c)$ and $\rho_2(c)$, which nearly match the bounds obtained from the constructions in [6, 4, 1]. Our main result is as follows.

THEOREM 1.3. *For every $c, s \ge 1$,*

$$\rho_s(c) \ge \frac{e^{\frac{1}{c^s}} - 1}{e^{\frac{1}{c^s}} + 1} \ge \frac{e-1}{e+1} \cdot \frac{1}{c^s} \ge \frac{0.462}{c^s}. \tag{1.2}$$

The second to last inequality in (1.2) follows from concavity of the function $t \mapsto \frac{e^t-1}{e^t+1}$ on $[0,\infty)$. Observe also that as $c \to \infty$, $\frac{e^{1/c}-1}{e^{1/c}+1} \sim \frac{1}{2c}$. It would be very interesting to determine $\limsup_{c\to\infty} c \cdot \rho_1(c)$ exactly—due to Theorem 1.3 and the results of [6], we currently know that this number is in the interval $[1/2, 1]$.

2. Proof of Theorem 1.3. The basic idea in the proof of Theorem 1.3 is simple. Choose a random point $x \in \{0,1\}^d$ and consider the random subset A of the cube $\{0,1\}^d$ consisting of points u for which $\mathscr{H}(u) = \mathscr{H}(x)$. The second condition in Definition 1.1 forces A to be small in expectation. But, when A is small we can bound from above the probability that after r steps, the random walk starting at a random point in A will end up in A. We obtain this upper bound using a Fourier analytic argument, and in combination with the first condition in Definition 1.1 we deduce the desired bound on $\rho_1(c)$.

Theorem 1.3 follows from the following result.

Proposition 2.1. *Let $\mathscr{H}$ be a (r, R, p, q)-sensitive hash family on the Hamming cube $(\{0,1\}^d, \|\cdot\|_1)$. Assume that r is an odd integer and that $R < \frac{d}{2}$. Then*

$$p \le \left(q + e^{-\frac{1}{d}\left(\frac{d}{2}-R\right)^2}\right)^{\frac{e^{2r/d}-1}{e^{2r/d}+1}}.$$

Choosing $R \approx \frac{d}{2} - \sqrt{d \log d}$ and $r \approx R/c$ in Proposition 2.1, and letting $d \to \infty$, yields Theorem 1.3 in the case $s = 1$. The case of general $s \ge 1$ follows from the fact that for $x, y \in \{0,1\}^d$, $\|x-y\|_s = \|x-y\|_1^{1/s}$.

Remark 2.1. Proposition 2.1 implies nontrivial lower bounds on $\frac{\log(1/p)}{\log(1/q)}$ for any (r, cr, p, q)-sensitive hash family on $(\{0,1\}^d, \|\cdot\|_1)$ even if q is allowed to depend on d. Observe that with the definition given in (1.1), Theorem 1.3 implies such a lower bound only for constant q. But, Proposition 2.1 is much stronger and implies a bound which asymptotically coincides with the lower bound in 1.3 for every $q \ge 2^{-o(d)}$.

The proof of Proposition 2.1 will be broken into a few lemmas. We start by bounding the expected size of the inverse images of $\mathscr{H}$, i.e., the expected number of points that are mapped by $\mathscr{H}$ to a fixed value.

Lemma 2.2. *Let $\mathscr{H}$ be a (r, R, p, q)-sensitive hash family on the Hamming cube $(\{0,1\}^d, \|\cdot\|_1)$, and fix $x \in \{0,1\}^d$. Then*

$$\mathbb{E}\left|\mathscr{H}^{-1}\left(\mathscr{H}(x)\right)\right| \le \sum_{k=0}^{\lfloor R\rfloor} \binom{d}{k} + q \cdot \sum_{k=\lfloor R\rfloor+1}^{d} \binom{d}{k}.$$

Proof. We simply write

$$\begin{aligned}\mathbb{E}\left|\mathscr{H}^{-1}\left(\mathscr{H}(x)\right)\right| &= \sum_{u\in\{0,1\}^d} \Pr[\mathscr{H}(u) = \mathscr{H}(x)]\\ &\le \left|\{u \in \{0,1\}^d : \|u-x\|_1 \le R\}\right| + q \cdot \left|\{u \in \{0,1\}^d : \|u-x\|_1 > R\}\right|\\ &= \sum_{k=0}^{\lfloor R\rfloor} \binom{d}{k} + q \cdot \sum_{k=\lfloor R\rfloor+1}^{d} \binom{d}{k}. \qquad \square\end{aligned}$$

Corollary 2.3. *Assume that $R < \frac{d}{2}$. Then, using the notation of Lemma 2.2, we have that*

$$\mathbb{E}\left|\mathscr{H}^{-1}\left(\mathscr{H}(x)\right)\right| \le 2^d \left(q + e^{-\frac{1}{d}\left(\frac{d}{2}-R\right)^2}\right).$$

Proof. This follows from Lemma 2.2 and the standard estimate $\sum_{k\le\frac{d}{2}-a} \binom{d}{k} \le 2^d \cdot e^{-\frac{a^2}{d}}$. □

Our next lemma bounds the probability that a random walk of a given length starting at a uniformly chosen point in a set $B \subseteq \{0,1\}^d$ will land in B. The proof of this lemma is a simple application of the Bonami–Beckner hypercontractive estimate on $\{0,1\}^d$ (see below). The connection between hypercontractivity and isoperimetric estimates in the spirit of Lemma 2.4 is classical, but we did not find the statement that we need in the literature. More recently, in the discrete setting, i.e., in the case of the hypercube $\{0,1\}^d$, similar arguments have been used extensively in theoretical computer science. In particular, Theorem 3.6 in [7] contains a reverse estimate to that of Lemma 2.4, in the case of the lazy random walk.

LEMMA 2.4 (random walk lemma). *Let r be an odd integer. Given $\emptyset \neq B \subseteq \{0,1\}^d$, consider the random variable $Q_B \in \{0,1\}^d$ defined as follows: Choose a point $z \in B$ uniformly at random, and perform r steps of the standard random walk on the Hamming cube starting from z (i.e., in each step we pass to one of the d neighbors of the vertex that we are currently occupying with probability $1/d$). The point obtained in this way will be denoted Q_B. Then*

$$\Pr[Q_B \in B] \le \left(\frac{|B|}{2^d}\right)^{\frac{e^{2r/d}-1}{e^{2r/d}+1}}.$$

Proof. We begin by recalling some background and notation on Fourier analysis on the Hamming cube. Given $S \subseteq \{1,\dots d\}$, the Walsh function $W_S : \{0,1\}^d \to \{-1,1\}$ is defined by

$$W_S(u) = (-1)^{\sum_{j\in S} u_j}.$$

For $f : \{0,1\}^d \to \mathbb{R}$ we set

$$\widehat{f}(S) = \frac{1}{2^d} \sum_{u\in\{0,1\}^d} f(u)W_S(u),$$

so that f can be decomposed as follows:

$$f = \sum_{S\subseteq\{1,\dots,d\}} \widehat{f}(S)W_S.$$

For every $f, g : \{0,1\}^d \to \mathbb{R}$ we write

$$\langle f, g\rangle = \frac{1}{2^d} \sum_{u\in\{0,1\}^d} f(u)g(u).$$

By Parseval's identity,

$$\langle f, g\rangle = \sum_{S\subseteq\{1,\dots,d\}} \widehat{f}(S)\widehat{g}(S).$$

For $\epsilon \in [0,1]$ the Bonami–Beckner operator T_ϵ is defined as

$$T_\epsilon f = \sum_{S\subseteq\{1,\dots,d\}} \epsilon^{|S|}\widehat{f}(S)W_S.$$

The Bonami–Beckner inequality [3, 2] states that for every $f : \{0,1\}^d \to \mathbb{R}$,

$$\sum_{S\subseteq\{1,\ldots,d\}} \epsilon^{2|S|}\widehat{f}(S)^2 = \|T_\epsilon f\|_2^2 = \frac{1}{2^d}\sum_{u\in\{0,1\}^d} (T_\epsilon f(u))^2 \le \|f\|_{1+\epsilon^2}^2$$
$$= \left(\frac{1}{2^d}\sum_{u\in\{0,1\}^d} f(u)^{1+\epsilon^2}\right)^{\frac{2}{1+\epsilon^2}}.$$

Specializing to the indicator of $B \subseteq \{0,1\}^d$ we get that

$$\sum_{S\subseteq\{1,\ldots,d\}} \epsilon^{2|S|}\widehat{\mathbf{1}_B}(S)^2 \le \left(\frac{|B|}{2^d}\right)^{\frac{2}{1+\epsilon^2}}. \tag{2.1}$$

Now, let P be the transition matrix of the standard random walk on $\{0,1\}^d$, i.e., $P_{uv} = 1/d$ if u and v differ in exactly one coordinate; $P_{uv} = 0$ otherwise. By a direct computation we have that for every $S \subseteq \{1,\ldots,d\}$,

$$PW_S = \left(1 - \frac{2|S|}{d}\right) W_S,$$

i.e., W_S is an eigenvector of P with eigenvalue $1 - \frac{2|S|}{d}$. The probability that the random walk starting form a random point in B ends up in B after r steps equals

$$\begin{aligned}\Pr[Q_B \in B] &= \frac{1}{|B|}\sum_{a,b\in B} (P^r)_{ab}\\ &= \frac{2^d}{|B|}\langle P^r\mathbf{1}_B, \mathbf{1}_B\rangle\\ &= \frac{2^d}{|B|}\sum_{S\subseteq\{1,\ldots,d\}} \widehat{\mathbf{1}_B}(S)^2\left(1-\frac{2|S|}{d}\right)^r\\ &\le \frac{2^d}{|B|}\sum_{\substack{S\subseteq\{1,\ldots,d\}\\ |S|\le d/2}} \widehat{\mathbf{1}_B}(S)^2\left(1-\frac{2|S|}{d}\right)^r,\end{aligned}$$

where we used the fact that r is odd (i.e., we dropped negative terms).

Thus, using (2.1) we see that

$$\Pr[Q_B \in B] \le \frac{2^d}{|B|}\sum_{S\subseteq\{1,\ldots,d\}} \widehat{\mathbf{1}_B}(S)^2 \cdot e^{-2r|S|/d} \le \frac{2^d}{|B|}\cdot\left(\frac{|B|}{2^d}\right)^{\frac{2}{1+e^{-2r/d}}}$$
$$= \left(\frac{|B|}{2^d}\right)^{\frac{1-e^{-2r/d}}{1+e^{-2r/d}}}. \qquad \square$$

Proof of Proposition 2.1. Assume that r is an odd integer and $R < \frac{d}{2}$. For $x \in \{0,1\}^d$ let $W_r(x) \in \{0,1\}^d$ be the random point obtained by performing a random walk for r steps starting at x. Since $\|x - W_r(x)\|_1 \le r$ we know that

$\Pr\left[\mathscr{H}\left(W_r(x)\right) = \mathscr{H}(x)\right] \geq p$. Taking expectation with respect to the uniform probability measure on $\{0,1\}^d$ we deduce that

$$\begin{aligned}
p &\leq \mathbb{E}_{x\in\{0,1\}^n} \Pr\left[\mathscr{H}\left(W_r(x)\right) = \mathscr{H}(x)\right] \\
&= \mathbb{E}_{\mathscr{H}} \Pr\left[x \in \{0,1\}^n :\ W_r(x) \in \mathscr{H}^{-1}\left(\mathscr{H}(x)\right)\right] \\
&= \mathbb{E}_{\mathscr{H}} \sum_{k\in\mathbb{N}} \Pr\left[x \in \{0,1\}^n :\ W_r(x) \in \mathscr{H}^{-1}\left(\mathscr{H}(x)\right)\ \wedge\ \mathscr{H}(x) = k\right] \\
&= \mathbb{E}_{\mathscr{H}} \sum_{k\in\mathbb{N}} \frac{\left|\mathscr{H}^{-1}(k)\right|}{2^d} \Pr\left[Q_{\mathscr{H}^{-1}(k)} \in \mathscr{H}^{-1}(k)\right] \\
&\leq \mathbb{E}_{\mathscr{H}} \sum_{k\in\mathbb{N}} \frac{\left|\mathscr{H}^{-1}(k)\right|}{2^d} \cdot \left(\frac{\left|\mathscr{H}^{-1}(k)\right|}{2^d}\right)^{\frac{e^{2r/d}-1}{e^{2r/d}+1}} \tag{2.2} \\
&= \mathbb{E}_{\mathscr{H}}\mathbb{E}_{x\in\{0,1\}^d} \left(\frac{\left|\mathscr{H}^{-1}(\mathscr{H}(x))\right|}{2^d}\right)^{\frac{e^{2r/d}-1}{e^{2r/d}+1}} \\
&\leq \mathbb{E}_{x\in\{0,1\}^d} \left(\frac{\mathbb{E}_{\mathscr{H}}\left|\mathscr{H}^{-1}(\mathscr{H}(x))\right|}{2^d}\right)^{\frac{e^{2r/d}-1}{e^{2r/d}+1}} \tag{2.3} \\
&\leq \left(q + e^{-\frac{1}{d}\left(\frac{d}{2}-R\right)^2}\right)^{\frac{e^{2r/d}-1}{e^{2r/d}+1}}, \tag{2.4}
\end{aligned}$$

where in (2.2) we used Lemma 2.4, in (2.3) we used Jensen's inequality, and in (2.4) we used Corollary 2.3. □

Acknowledgments. We are grateful to Piotr Indyk and Jirka Matoušek for helpful suggestions.

REFERENCES

[1] A. Andoni and P. Indyk, *Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions*, in Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06), Berkeley, CA, IEEE, Washington, DC, pp. 459–468.

[2] W. Beckner, *Inequalities in Fourier analysis*, Ann. of Math. (2), 102 (1975), pp. 159–182.

[3] A. Bonami, *Étude des coefficients de Fourier des fonctions de $L^p(G)$*, Ann. Inst. Fourier (Grenoble), 20 (1971), pp. 335–402.

[4] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, *Locality-sensitive hashing scheme based on p-stable distributions*, in SCG '04: Proceedings of the Twentieth Annual Symposium on Computational Geometry (Brooklyn, NY, 2004), ACM, New York, 2004, pp. 253–262.

[5] S. Har-Peled, *A replacement for Voronoi diagrams of near linear size*, in 42nd IEEE Symposium on Foundations of Computer Science (Las Vegas, NV, 2001), IEEE Computer Soc., Los Alamitos, CA, 2001, pp. 94–103.

[6] P. Indyk and R. Motwani, *Approximate nearest neighbors: Towards removing the curse of dimensionality*, in STOC '98: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing (Dallas, TX, 1998), ACM, New York, 1999, pp. 604–613.

[7] E. Mossel, R. O'Donnell, O. Regev, J. Steif, and B. Sudakov, *Non-interactive correlation distillation, inhomogeneous Markov chains, and the reverse Bonami-Beckner inequality*, Israel J. Math., 154 (2006), pp. 299–336.

[8] R. Panigrahy, *Entropy based nearest neighbor search in high dimensions*, in SODA '06: Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm (Miami, FL, 2006), ACM, New York, 2006, SIAM, Philadelphia, pp. 1186–1195.

SIAM J. Discrete Math.
Vol. 21, No. 4, pp. 936–946

A WORST-CASE ANALYSIS OF THE SEQUENTIAL METHOD TO LIST THE MINIMAL HITTING SETS OF A HYPERGRAPH*

KEN TAKATA†

Abstract. It is open whether the minimal hitting sets of a hypergraph can be listed in time polynomial in the input and output size. We show that a well-known sequential approach described by Berge and studied since the 1950s is not polynomial in the above sense, even if we allow an optimal ordering of the edges. This answers a question posed by H. Hirsh. The proof uses hypergraphs based on read-once formulas. We also offer a generalization of this sequential approach.

Key words. hypergraph, minimal hitting set, listing algorithms, Sequential Method, read-once Boolean functions

AMS subject classifications. 68Q25, 68R05, 05C65, 05C69

DOI. 10.1137/060653032

1. Introduction.

Listing the minimal hitting sets of a hypergraph is a central problem with many applications in combinatorics, logic, learning theory, and knowledge discovery [1, 8, 10]. In this paper we discuss the complexity of a certain class of algorithms used to list minimal hitting sets.

A hypergraph $\mathcal{H}$ is a set of subsets of a finite set. A hitting set is a set of vertices which intersects every subset in $\mathcal{H}$. Using Berge's notation [4], the family of minimal hitting sets of $\mathcal{H}$ is also called the *transversal hypergraph* of the hypergraph and abbreviated as $\mathrm{Tr}(\mathcal{H})$.

Thus, given a hypergraph $\mathcal{H}$, one can ask how difficult it is to list the sets in $\mathrm{Tr}(\mathcal{H})$. The question whether $\mathrm{Tr}(\mathcal{H})$ can be listed in time polynomial in the input and output size or under other criteria for efficient listing [15, 22, 34, 35] is a longstanding open problem for enumeration algorithms. Furthermore, listing $\mathrm{Tr}(\mathcal{H})$ is equivalent to a number of other problems discussed in Eiter and Gottlob [10]. Other recent work shows how listing $\mathrm{Tr}(\mathcal{H})$ is related to problems in data mining (see [5, 7, 8, 14, 16, 19, 27, 33]). In logic, listing $\mathrm{Tr}(\mathcal{H})$ is equivalent to the monotone CNF/DNF dualization problem (see [11, 12]), which is how Fredman and Khachiyan [13] discuss the problem. We mention their algorithm as well as others below.

Paull and Unger [32] have shown that if $\mathcal{H}$ is a graph, then we can list the minimal hitting sets in time polynomial in the input and output size. (Note that in their original formulation, they listed the maximal independent sets. This is equivalent to listing the minimal hitting sets since the complement of a minimal hitting set is a maximal independent set.) Subsequent improvements have come from Tsukiyama et al. [35], who used backtracking algorithms that built a backtracking tree in which the nodes on level i of the tree are the maximal independent sets of the graph restricted to vertices $\{1, \ldots, i\}$. Using this technique, they were able to list the objects using

*Received by the editors February 26, 2006; accepted for publication (in revised form) May 18, 2007; published electronically December 7, 2007. A preliminary version of this paper appeared as "On the Sequential Method for listing minimal hitting sets" in *Proceedings of the Workshop on Discrete Mathematics and Data Mining, Second SIAM International Conference on Data Mining*, Arlington, VA, 2002, pp. 109–120. This research was supported in part by the National Science Foundation under grants CCR-9800070 and CCR-0100336.

http://www.siam.org/journals/sidma/21-4/65303.html

†Department of Mathematics, Hamline University, 1536 Hewitt Avenue, Box 180, St. Paul, MN 55104 (ktakata01@hamline.edu).

polynomial space and with polynomial delay (i.e., a delay of at most $O(n^k)$ between outputting one object and the next one, where n is the size of the input). Johnson, Yannakakis, and Papadimitriou [22] then expanded on these techniques to list the maximal independent sets in lexicographic order. However, as Johnson et al. note in their paper, the same backtracking technique cannot be used to list the maximal independent sets of a hypergraph with polynomial delay unless P=NP.

Most recently, Eiter, Gottlob, and Makino [12] have extended Johnson et al.'s techniques to develop efficient listing algorithms for certain special cases of hypergraphs. Efficient algorithms for the minimal hitting set problem also exist for several other special cases [6, 10, 28].

To list $\mathrm{Tr}(\mathcal{H})$, where $\mathcal{H}$ is an arbitrary hypergraph, there is a simple sequential approach which accepts for an input an ordering of the hypergraph's edges into a sequence and then progressively finds the minimal hitting sets of the first i edges. While the approach, which we will call the Sequential Method, is well known and appears often in the literature (see, for example, [4, 10, 25, 26, 28, 30, 31]), there is little theoretical information about its behavior. (See section 2.3 for more details.)

In comparison, Fredman and Khachiyan [13] developed an algorithm which has a proven time bound. Their algorithm can list the minimal hitting sets of any hypergraph with quasi-polynomial time, where this is quasi-polynomial in the input *and* output size. More specifically, the running time is quasi-polynomial in the input size and the number of objects output *thus far*. Note also that the size of the output could be exponentially larger than the size of the input. (Consider $\frac{n}{2}$ vertex-disjoint, graph-type edges. Such a hypergraph would have $2^{n/2}$ minimal hitting sets. Thus, an algorithm whose delay is quasi-polynomial in the input and output sizes may not work in time polynomial in the input size alone.) More precisely, if V is the number of edges of the hypergraph plus the number of minimal hitting sets output thus far, then the time to list another minimal hitting set is $O(V^{o(\log V)})$.

In addition, it has been shown (see [7, 16, 27]) that many other seemingly more general problems are no more difficult than listing $\mathrm{Tr}(\mathcal{H})$. This, along with the fact that there have been no improvements on Fredman's and Khachiyan's quasi-polynomial algorithm [13], suggests that the Sequential Method is unlikely to list $\mathrm{Tr}(\mathcal{H})$ in time polynomial in the input and output size. Answering a question posed by Hirsh [21], we show that this is the case in a rather strong sense, namely, that the Sequential Method is inefficient no matter how the edges are ordered. The proof is based on showing that a specific infinite family of hypergraphs also studied in [13] has a certain combinatorial property (see Corollary 2). We also propose a generalization of the Sequential Method for which no such negative results are known.

The paper is organized as follows. In section 2, we develop lemmas and constructions which will be useful in describing the Sequential Method and in constructing an infinite family of hypergraphs $\mathcal{H}_i$ ($i \in \mathbb{N}$) (section 3) for which the Sequential Method produces a superpolynomial blowup for any ordering of the edges. We propose a generalization of the Sequential Method (section 4) and show that it lists $\mathrm{Tr}(\mathcal{H}_i)$ efficiently. We conclude (section 5) with some further comments and open problems.

2. Preliminaries. In this section, we provide formal definitions of some basic terms used in this paper. Then we describe two binary operations $\vee$ and $\wedge$ on any hypergraphs $\mathcal{A}$ and $\mathcal{B}$, producing two new hypergraphs $\mathcal{A} \vee \mathcal{B}$ and $\mathcal{A} \wedge \mathcal{B}$. These constructions are used to build an infinite family of hypergraphs $\mathcal{H}_i$ ($i \in \mathbb{N}$) for which the Sequential Method requires a running time which is superpolynomial in the input and output size.

The lemmas about $\mathrm{Tr}(\mathcal{A}\vee\mathcal{B})$ and $\mathrm{Tr}(\mathcal{A}\wedge\mathcal{B})$ will be useful in describing the Sequential Method and the generalized method in section 4. Furthermore, a corollary about $\mathrm{Tr}(\mathcal{A}\vee\mathcal{B})$ and $\mathrm{Tr}(\mathcal{A}\wedge\mathcal{B})$ in the specific case when $\mathcal{A}$ and $\mathcal{B}$ are vertex-disjoint will be useful in calculating the running time of the Sequential Method and its generalization when they list $\mathrm{Tr}(\mathcal{H}_i)$.

2.1. Definitions. A hypergraph $\mathcal{H}$ is a pair $(V,\mathcal{E})$ in which V is a finite set of vertices $\{1,\ldots,n\}$ and $\mathcal{E}$ is a set of subsets $\{e_1,\ldots,e_m\}$ from V. These subsets are called *edges* of $\mathcal{H}$. (For notational convenience, we will usually refer to a hypergraph by its edge set whenever we can do so without ambiguity, and we will use lowercase letters for edges. In addition, unless otherwise stated, $|V| = n$ and $|\mathcal{E}| = m$. The space needed to represent $\mathcal{H}$ as a sequence of incidence vectors is $O(|V|\cdot|\mathcal{E}|)$.) The *minimal* edges in $\mathcal{H}$ are those that do not properly contain other edges in $\mathcal{H}$, and we use the notation $\min(\mathcal{H})$ to designate the hypergraph made of these minimal edges. A hypergraph $\mathcal{H}$ is called *simple* iff no edge contains another (i.e., $\mathcal{H} = \min(\mathcal{H})$).

A set of vertices s is a *hitting set* of $\mathcal{H}$ if $s\cap e \neq \emptyset$ for all edges $e \in \mathcal{H}$. The family of all hitting sets of $\mathcal{H}$ is written as $\mathcal{S}(\mathcal{H})$. Thus, the minimal hitting sets of $\mathcal{H}$ are $\min(\mathcal{S}(\mathcal{H}))$. Using Berge's terminology, we write this as $\mathrm{Tr}(\mathcal{H})$, the transversal hypergraph of $\mathcal{H}$. In addition, since $\mathrm{Tr}(\mathcal{H}) = \mathrm{Tr}(\min(\mathcal{H}))$, it may be assumed for the listing problem that all hypergraphs are simple.

2.2. Constructions. Let $\mathcal{A}$ and $\mathcal{B}$ be any two hypergraphs. Using $\mathcal{A}$ and $\mathcal{B}$ and two binary operators $\vee$ and $\wedge$, we can construct two new hypergraphs.

- $\mathcal{A}\vee\mathcal{B}$ is the hypergraph such that $V_{\mathcal{A}\vee\mathcal{B}} = V_{\mathcal{A}}\cup V_{\mathcal{B}}$ and $\mathcal{E}_{\mathcal{A}\vee\mathcal{B}} = \mathcal{E}_{\mathcal{A}}\cup\mathcal{E}_{\mathcal{B}}$.
- $\mathcal{A}\wedge\mathcal{B}$ is the hypergraph such that $V_{\mathcal{A}\wedge\mathcal{B}} = V_{\mathcal{A}}\cup V_{\mathcal{B}}$ and $\mathcal{E}_{\mathcal{A}\wedge\mathcal{B}} = \{a\cup b : a\in\mathcal{E}_{\mathcal{A}},\, b\in\mathcal{E}_{\mathcal{B}}\}$.

The following standard lemma shows how the minimal hitting sets of $\mathcal{A}\vee\mathcal{B}$ and $\mathcal{A}\wedge\mathcal{B}$ are related to the families of minimal hitting sets of $\mathcal{A}$ and $\mathcal{B}$.

LEMMA 1. *For any two hypergraphs, $\mathcal{A}$ and $\mathcal{B}$,*
1. $\mathrm{Tr}(\mathcal{A}\vee\mathcal{B}) = \min(\mathrm{Tr}(\mathcal{A})\wedge\mathrm{Tr}(\mathcal{B}))$*, and*
2. $\mathrm{Tr}(\mathcal{A}\wedge\mathcal{B}) = \min(\mathrm{Tr}(\mathcal{A})\vee\mathrm{Tr}(\mathcal{B}))$.

A proof, albeit with slightly different notation, can be found in Berge's description of the Sequential Method [4, pp. 52–53].

In the special case when $\mathcal{A}$ and $\mathcal{B}$ are vertex-disjoint, $\mathrm{Tr}\,\mathcal{A}\vee\mathrm{Tr}\,\mathcal{B} = \min(\mathrm{Tr}(\mathcal{A})\vee\mathrm{Tr}(\mathcal{B}))$ and $\mathrm{Tr}(\mathcal{A})\wedge\mathrm{Tr}(\mathcal{B}) = \min(\mathrm{Tr}(\mathcal{A})\wedge\mathrm{Tr}(\mathcal{B}))$. To see this, observe that every set in $\mathrm{Tr}(\mathcal{A})$ is incomparable with every set in $\mathrm{Tr}(\mathcal{B})$. Also note that any two sets in $\mathrm{Tr}(\mathcal{A})\wedge\mathrm{Tr}(\mathcal{B})$ are incomparable. The equalities mentioned above yield the following corollary.

COROLLARY 1. *If $\mathcal{A}$ and $\mathcal{B}$ are vertex-disjoint, then*
1. $\mathrm{Tr}(\mathcal{A}\vee\mathcal{B}) = \mathrm{Tr}(\mathcal{A})\wedge\mathrm{Tr}(\mathcal{B})$*, and* $|\mathrm{Tr}(\mathcal{A}\vee\mathcal{B})| = |\mathrm{Tr}(\mathcal{A})|\cdot|\mathrm{Tr}(\mathcal{B})|$*, and*
2. $\mathrm{Tr}(\mathcal{A}\wedge\mathcal{B}) = \mathrm{Tr}(\mathcal{A})\vee\mathrm{Tr}(\mathcal{B})$*, and* $|\mathrm{Tr}(\mathcal{A}\wedge\mathcal{B})| = |\mathrm{Tr}(\mathcal{A})| + |\mathrm{Tr}(\mathcal{B})|$.

2.3. The Sequential Method to list $\mathrm{Tr}(\mathcal{H})$. The Sequential Method takes as input a hypergraph with its edges ordered into some sequence $(e_1,\ldots,e_m)$, where m is the number of edges in the hypergraph. Applying Lemma 1 to $\mathcal{A} = \{e_1,\ldots,e_i\}$ and $\mathcal{B} = \{e_{i+1}\}$, we get that for any i $(1\leq i\leq m-1)$ it holds that

$$\mathrm{Tr}(\{e_1,\ldots,e_{i+1}\}) = \min(\mathrm{Tr}(\{e_1,\ldots,e_i\})\wedge\mathrm{Tr}(\{e_{i+1}\})).$$

This gives an iterative procedure to find the minimal hitting sets of the subhypergraph $\{e_1,\ldots,e_i\}$, where $1\leq i\leq m$. Note that for any edge $e_i\in\mathcal{H}$ the minimal hitting sets of e_i are the singleton vertices in the edge, i.e., for all $e_i\in\mathcal{H}$, $\mathrm{Tr}(\{e_i\}) = \{\{v\} : v\in e_i\}$.

Thus, it is a simple matter both to initialize the algorithm and to compute $\text{Tr}(\{e_{i+1}\})$ during each iteration.

ALGORITHM 1.

```
Sequential($(e_1, \ldots, e_m)$) {
  for $i = 1$ to $m - 1$ do {
    $\text{Tr}(\{e_1, \ldots, e_{i+1}\}) = \min(\text{Tr}(\{e_1, \ldots, e_i\}) \wedge \text{Tr}(\{e_{i+1}\}))$
  }
}
```

Remark. The number of minimal hitting sets which the algorithm must compute for the ith iteration is $|\text{Tr}(\{e_1, \ldots, e_{i+1}\})|$. Thus, given an initial ordering such as $(e_1, \ldots, e_m)$, the running time is at least $\Omega(\max_i |\text{Tr}(\{e_1, \ldots, e_i\})|)$, where $1 \le i \le m$. Thus, we may compute a lower bound on the running time by determining the number of minimal hitting sets produced in any intermediate stage.

Note that the Sequential Method is a *class* of algorithms rather than a single algorithm since it does not specify *how* one should initially order the edges. There are pathological orderings which create listing times exponential in the input and output size [10, 28]. As an example, order the edges of the complete graph K_n so that the first $\frac{n}{2}$ edges form a perfect matching. Although K_n has only $\binom{n}{2}$ edges and n minimal hitting sets, $2^{n/2}$ minimal hitting sets will be created when processing the first $\frac{n}{2}$ edges of this subgraph.

There have been a number of proposed implementations of the method such as

- ordering the edges according to *lexicographical* order in which lexicographical order is defined by representing the edges $e, f \in \mathcal{H}$ as $\{0,1\}$ vectors of length n where $e <_{lex} f$ iff e as a number in base two is less than f as a number in base two,
- ordering the edges so that, if possible, they always form a connected subhypergraph,
- using the greedy method to select the next edge so that the next edge minimizes the number of minimal hitting sets.

While these modifications work well for many hypergraphs, there have been no upper bounds proven for the running time in general. Thus, Hirsh [21] has asked what the worst-case running time of the Sequential Method is using an optimal edge ordering in terms of the input and output size of the problem. In other words, the question is whether there is a family of hypergraphs $\mathcal{H}_i$ ($i \in \mathbb{N}$) such that for every polynomial p, if i is sufficiently large, then for *any* ordering $(e_1, \ldots, e_j, \ldots, e_m)$ of the edges of $\mathcal{H}_i$ there is a j such that

$$|V_{\mathcal{H}_i}| \cdot |\text{Tr}(\{e_1, \ldots, e_j\})| > p\,(|V_{\mathcal{H}_i}| \cdot |\mathcal{H}_i| + |V_{\mathcal{H}_i}| \cdot |\text{Tr}(\mathcal{H}_i)|).$$

If, say, $|\mathcal{H}_i| + |\text{Tr}(\mathcal{H}_i)| = \Omega(|V_{\mathcal{H}_i}|)$, then this is equivalent to

$$|\text{Tr}(\{e_1, \ldots, e_j\})| > p\,(|\mathcal{H}_i| + |\text{Tr}(\mathcal{H}_i)|)$$

holding for every p and for every i sufficiently large, and this is the form we are going to use. In what follows we answer this question in the affirmative.

3. Lower bounds for running time.

In this section, we prove the main result by describing a family of hypergraphs and then applying it to show that the Sequential Method is inefficient.

3.1. A provably superpolynomial family of hypergraphs. Using the constructions in section 2, we define the infinite family of hypergraphs $\mathcal{H}_i$ ($i \in \mathbb{N}$) recursively as follows:

- $\mathcal{H}_0$ is the hypergraph consisting of one edge, which is a single vertex. (Thus, $V_{\mathcal{H}_0} = \{1\}$ and $\mathcal{E}_{\mathcal{H}_0} = \{\{1\}\}$.)
- $\mathcal{H}_i = (\mathcal{A} \vee \mathcal{B}) \wedge (\mathcal{C} \vee \mathcal{D})$, where $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$ are all vertex-disjoint copies of $\mathcal{H}_{i-1}$.

As an example, $\mathcal{H}_1$ is a cycle on four vertices with four edges, a C_4. (We will return often to this particular example.) Note that since $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$ are all vertex-disjoint, $|V_{\mathcal{H}_i}| = 4|V_{\mathcal{H}_{i-1}}|$. Furthermore, using the hypergraph constructions involving $\vee$ and $\wedge$, $|\mathcal{H}_i| = (|\mathcal{A}|+|\mathcal{B}|)(|\mathcal{C}|+|\mathcal{D}|)$, which is $(2|\mathcal{H}_{i-1}|)^2$. By applying Corollary 1, we know that $|\mathrm{Tr}(\mathcal{H}_i)| = |\mathrm{Tr}(\mathcal{A})||\mathrm{Tr}(\mathcal{B})| + |\mathrm{Tr}(\mathcal{C})||\mathrm{Tr}(\mathcal{D})|$, which is $2|\mathrm{Tr}(\mathcal{H}_{i-1})|^2$.

We can solve these recurrences by applying the fact that $\mathcal{H}_1 = C_4$, which means that $|V_{\mathcal{H}_1}| = 4$, $|\mathcal{H}_1| = 4$, and $|\mathrm{Tr}(\mathcal{H}_1)| = 2$. From this, it follows that

$$|V_{\mathcal{H}_i}| = 4^i, \quad |\mathcal{H}_i| = 2^{2(2^i-1)}, \quad |\mathrm{Tr}(\mathcal{H}_i)| = 2^{2^i-1}.$$

This family of hypergraphs $\mathcal{H}_i$ ($i \in \mathbb{N}$) can also be viewed in terms of monotone Boolean functions. The edges of $\mathcal{H}_i$ correspond to the minimal truth assignments of a certain function, f_i, which can be represented by a read-once Boolean circuit with a very specific structure, and in fact, this paraphrase is what has suggested the terminology ($\vee$ and $\wedge$) and the constructions used in section 2. As a Boolean circuit, the function can be described as a complete binary tree with $\wedge$ at the root (level 1) and all other nodes on odd levels except the leaf level, $\vee$ for all the nodes on even levels, and variables at the leaf level. In addition, since the function is read-once, all the variables are distinct. In the literature, this circuit is sometimes called an alternating tree. This function appears in Fredman and Khachiyan [13] and in Gurvich and Khachiyan [18] in a slightly different form. The alternating tree used by Fredman, Gurvich, and Khachiyan starts with a $\vee$ in the root node and then alternates $\wedge$ and $\vee$ on each level. Because the alternating tree in [13, 18] is slightly different, their constructions for f_i and thus $\mathcal{H}_i$ are not exactly the same. The analogous equations in their paper reflect this difference. For more on read-once functions and their combinatorial properties, see [2, 17, 23, 29].

The family of functions f_i ($i \in \mathbb{N}$) has a recursive definition which is similar to our definition for $\mathcal{H}_i$. f_i is a monotone Boolean function that takes as arguments the variables in X_i, which we also define recursively below.

- $X_0 = \{x_1\}$, and $f_0(X_0) = x_1$.
- Let A, B, C, and D be variable-disjoint copies of $f_{i-1}(X_{i-1})$. Then X_i is the list of variables in A followed by those in B, C, and D, and $f_i(X_i) = (A \vee B) \wedge (C \vee D)$.

Thus, $f_1(x_1, x_2, x_3, x_4) = (x_1 \vee x_2) \wedge (x_3 \vee x_4)$. The minimal truth assignments of f_1 form the edges of a C_4. The edges of this C_4 in lexicographical order are $(e_1, e_2, e_3, e_4) = (\{1,3\}, \{2,3\}, \{1,4\}, \{2,4\})$.

3.2. The Sequential Method always lists $\mathrm{Tr}(\mathcal{H}_i)$ inefficiently. We now proceed to the following lemma, which will be key in showing that the Sequential Method runs slowly on the family of hypergraphs defined above no matter how the edges are ordered. The basic idea is this: using the Sequential Method, we will always have one last remaining edge to process. What we will show is that no matter what this last edge is, the number of hitting sets just before we process that edge will be superpolynomial in the final number of minimal hitting sets plus the original number of edges in the hypergraph for $\mathcal{H}_i$ ($i \in \mathbb{N}$). In other words, the addition of this last edge causes a superpolynomial decrease in the number of minimal hitting sets.

LEMMA 2. *Let $\mathcal{H} = (\mathcal{A} \vee \mathcal{B}) \wedge (\mathcal{C} \vee \mathcal{D})$, where $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$ are vertex-disjoint hypergraphs, and without loss of generality, let us assume that the edge e which we delete from $\mathcal{H}$ is $(b \cup d)$, where $b \in \mathcal{E}_{\mathcal{B}}$ and $d \in \mathcal{E}_{\mathcal{D}}$. Then $|\mathrm{Tr}(\mathcal{H}\backslash\{e\})\backslash\mathrm{Tr}(\mathcal{H})| \geq |\mathrm{Tr}(\mathcal{A})| \cdot |\mathrm{Tr}(\mathcal{B}\backslash\{b\})\backslash\mathrm{Tr}(\mathcal{B})| \cdot |\mathrm{Tr}(\mathcal{C})| \cdot |\mathrm{Tr}(\mathcal{D}\backslash\{d\})\backslash\mathrm{Tr}(\mathcal{D})|$.*

Proof. Let $s = h_{\mathcal{A}} \cup h_{\mathcal{B}^-} \cup h_{\mathcal{C}} \cup h_{\mathcal{D}^-}$, where $h_{\mathcal{A}} \in \mathrm{Tr}(\mathcal{A})$, $h_{\mathcal{B}^-} \in \mathrm{Tr}(\mathcal{B}\backslash\{b\})\backslash\mathrm{Tr}(\mathcal{B})$, $h_{\mathcal{C}} \in \mathrm{Tr}(\mathcal{C})$, $h_{\mathcal{D}^-} \in \mathrm{Tr}(\mathcal{D}\backslash\{d\})\backslash\mathrm{Tr}(\mathcal{D})$. Note that $h_{\mathcal{B}^-} \in \mathrm{Tr}(\mathcal{B}\backslash\{b\})\backslash\mathrm{Tr}(\mathcal{B})$ implies that $h_{\mathcal{B}^-}$ hits every set in $\mathcal{B}$ different from B, but it does not hit B.

We claim that $s \in \mathrm{Tr}(\mathcal{H}\backslash\{e\})\backslash\mathrm{Tr}(\mathcal{H})$. First, observe that $s \notin \mathrm{Tr}(\mathcal{H})$ because there is an edge $(b \cup d) \in \mathcal{H}$ which is not hit. Now we show that $s \in \mathrm{Tr}(\mathcal{H}\backslash\{e\})$. It hits any edge in $\mathcal{H}\backslash\{e\}$ because it hits every edge in $\mathcal{A}$, $(\mathcal{B}\backslash\{b\})$, $\mathcal{C}$, and $(\mathcal{D}\backslash\{d\})$ and because of the way we construct $\mathcal{H}$ from $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$. To show that s is minimal, consider any proper subset s'. There are four different cases to consider, but the argument is basically the same in each case.

Assume that $s' \cap h_{\mathcal{A}} \subsetneq h_{\mathcal{A}}$ (the case of $h_{\mathcal{C}}$ is analogous). Then there exists some edge $a \in \mathcal{A}$ which s' fails to hit. And thus, s' fails to hit the edge $(a \cup d) \in \mathcal{H}\backslash\{e\}$.

If $s' \cap h_{\mathcal{B}^-} \subsetneq h_{\mathcal{B}^-}$, there are at least two edges in $\mathcal{B}$ which s' fails to hit: b and some other edge b'. Thus, s' fails to hit the edge $(b' \cup d) \in \mathcal{H}\backslash\{e\}$. The same argument holds for $h_{\mathcal{D}^-}$. □

Using this lemma and the recurrences for $|V_{\mathcal{H}_i}|$, $|\mathcal{H}_i|$, and $|\mathrm{Tr}(\mathcal{H}_i)|$, we have the materials necessary for proving our main result.

THEOREM 1. *The running time of the Sequential Method on the family of hypergraphs $\mathcal{H}_i$ ($i \in \mathbb{N}$) is superpolynomial in the size of the input and output no matter how the edges are ordered.*

Proof. Let

$$a_i = \min_{e \in \mathcal{H}_i} |\mathrm{Tr}(\mathcal{H}_i\backslash\{e\})\backslash\mathrm{Tr}(\mathcal{H}_i)|.$$

Lemma 2 yields the recurrence

$$a_i \geq |\mathrm{Tr}(\mathcal{H}_{i-1})|^2 \cdot a_{i-1}^2.$$

To start this recurrence, use the fact that $|\mathrm{Tr}(\mathcal{H}_i)| = 2^{2^i-1}$ and note that since $\mathcal{H}_1$ is a C_4 and $\mathcal{H}_1\backslash\{e\}$ is always a P_4, $|\mathrm{Tr}(\mathcal{H}_1\backslash\{e\})| = 3$ for any $e \in \mathcal{H}$, and thus $a_1 = 1$. By induction on i, we can show that for $i \geq 2$, it holds that

$$|\mathrm{Tr}(\mathcal{H}_i\backslash\{e\})\backslash\mathrm{Tr}(\mathcal{H}_i)| \geq 2^{(i-2)2^i+2}.$$

To complete the theorem, we consider an arbitrary ordering of the edges of $\mathcal{H}_i$ and compare M and m, where M represents the size of the largest intermediate result produced by the Sequential Method, and m is the size of the problem's input and output, that is, $m = O(|V_{\mathcal{H}_i}|(|\mathcal{H}_i| + |\mathrm{Tr}(\mathcal{H}_i)|))$. The running time of the Sequential Method will be at least $M \geq |\mathrm{Tr}(\mathcal{H}_i\backslash\{e\})| \geq |\mathrm{Tr}(\mathcal{H}_i\backslash\{e\})\backslash\mathrm{Tr}(\mathcal{H}_i)|$, where $e \in \mathcal{H}_i$. The recurrences for $|\mathcal{H}_i|$, $|\mathrm{Tr}(\mathcal{H}_i)|$, and $|\mathrm{Tr}(\mathcal{H}_i\backslash\{e\})\backslash\mathrm{Tr}(\mathcal{H}_i)|$ show that $M = m^{\Omega(\log\log m)}$. □

This result can also be stated in combinatorial terms without respect to any particular class of algorithms. We state this result in the following corollary, which may be of interest in itself.

COROLLARY 2. $|\mathrm{Tr}(\mathcal{H}_i\backslash\{e\})|$ *is superpolynomial in* $|\mathcal{H}_i|+|\mathrm{Tr}(\mathcal{H}_i)|$ *for any* $e \in \mathcal{H}_i$.

Furthermore, since listing $\mathrm{Tr}(\mathcal{H})$ is equivalent to the monotone CNF/DNF dualization problem, we can also paraphrase the main result by saying that there exist monotone CNFs that, if we multiply them out in any order, will produce an intermediate result which is superpolynomial in the sum of the sizes of the input and output formulas. These are the CNF representations of f_i $(i \in \mathbb{N})$.

4. Generalizing the Sequential Method.

In this section, we propose a generalization of the Sequential Method that can list $\mathrm{Tr}(\mathcal{H}_i)$ efficiently. We will first explain how one can generalize the Sequential Method, then describe this generalization in terms of binary tree representations, and, finally, show that there exist binary trees with a property (see Property 1) that we can use to list $\mathrm{Tr}(\mathcal{H}_i)$ in time polynomial in the input and output.

To see how one can generalize the Sequential Method, note how it uses a special case of the first half of Lemma 1. For the Sequential Method, $\mathcal{A} = \{e_1, \ldots, e_i\}$ and $\mathcal{B} = \{e_{i+1}\}$. However, the lemma is valid for partitions of the edges in $\mathcal{H}$ into other subhypergraphs $\mathcal{A}$ and $\mathcal{B}$. We can represent such a partition by a binary tree with $\mathcal{H}$ as the root and $\mathcal{A}$ and $\mathcal{B}$ as the left and right children, and overall, we can represent a generalization of the Sequential Method with a binary *tree representation*, which we define below.

DEFINITION 1. *A tree representation of a hypergraph (abbreviated as $\mathcal{T}_{\mathcal{H}}$) is a binary tree in which the leaves, $l_1, \ldots, l_m$, are labeled with the edges of the hypergraph, $e_1, \ldots, e_m$, and each internal node is labeled with the family of those edges that are in descendants of that node.*

Using such a tree representation, one can calculate $\mathrm{Tr}(\mathcal{H})$, first by calculating the minimal hitting sets of the leaves and then by applying Lemma 1 to calculate the minimal hitting sets of the internal nodes. We give a backtracking algorithm below that does this by performing a depth-first traversal of $\mathcal{T}_{\mathcal{H}}$. The initial call is `Evaluate(root)`, where `root` is the root of $\mathcal{T}_{\mathcal{H}}$.

ALGORITHM 2.

```
Evaluate(node) {
  if (node has children) {
    $\mathcal{L}$ = Evaluate(left_child);
    $\mathcal{R}$ = Evaluate(right_child);
    return node = $\min(\mathcal{L} \wedge \mathcal{R})$;
  }
  else  { // the node is a leaf
    return node = $\{\{v\} : v \in e_i\}$; // i.e., the min. hit. sets in the leaf
  }
}
```

Note that as the algorithm proceeds on $\mathcal{T}_{\mathcal{H}}$, it relabels each leaf e_i with $\mathrm{Tr}(\{e_i\})$ and also relabels each internal node as $\mathrm{Tr}(\mathcal{S})$, where $\mathcal{S}$ consists of those edges that were in descendants of that internal node. When the algorithm terminates, the root will contain $\mathrm{Tr}(\mathcal{H})$. For ease of discussion, we will refer to each node in the tree by the minimal hitting sets of its family of edges.

Now let us consider $\mathcal{H}_1$ and a specific example of a binary tree $\mathcal{T}_{\mathcal{H}_1}$.

Example. Consider $\mathcal{H}_1$ with $(e_1, e_2, e_3, e_4) = (\{1,3\}, \{2,3\}, \{1,4\}, \{2,4\})$. Now consider $(e_1e_2)(e_3e_4)$, which corresponds to evaluating $\mathrm{Tr}(\mathcal{H}_1)$ using the binary tree in which the root is $\mathrm{Tr}(\mathcal{H}_1)$, the root's left and right children are $\mathrm{Tr}(\{e_1, e_2\})$ and $\mathrm{Tr}(\{e_3, e_4\})$, and the leaves are $\mathrm{Tr}(\{e_1\})$, $\mathrm{Tr}(\{e_2\})$, $\mathrm{Tr}(\{e_3\})$, $\mathrm{Tr}(\{e_4\})$.

Note that this tree representation cannot be expressed by any sequential ordering since it does not leave one last edge to be processed (i.e., in the last step $|\mathcal{B}| > 1$). Thus, tree representations have the ability to process the edges of $\mathcal{H}$ in a way that no sequential ordering can. Furthermore, observe how (in comparison to any sequential ordering) we can use this binary tree $\mathcal{T}_{\mathcal{H}_1}$ to avoid blowups in the intermediate steps in which $|\mathrm{Tr}(\mathcal{A})|$ or $|\mathrm{Tr}(\mathcal{B})|$ exceed $|\mathrm{Tr}(\mathcal{H})|$.

If we can avoid such blowups, then, as Theorem 2 shows, we can list $\mathrm{Tr}(\mathcal{H}_i)$ efficiently. To avoid any blowup, what we will want is a binary tree representation $\mathcal{T}_{\mathcal{H}}$ with the following property.

PROPERTY 1. *Every internal node of the binary tree contains at least as many minimal hitting sets as either of its children. That is to say, if* $\mathrm{Tr}(\mathcal{A} \vee \mathcal{B})$ *is an arbitrary internal node of the tree with children* $\mathrm{Tr}(\mathcal{A})$ *and* $\mathrm{Tr}(\mathcal{B})$*, then*

$$|\mathrm{Tr}(\mathcal{A} \vee \mathcal{B})| \geq \max(|\mathrm{Tr}(\mathcal{A})|, |\mathrm{Tr}(\mathcal{B})|).$$

To see a specific example, consider the binary tree for the example discussed above. (This example will be used later as the basis step for an inductive argument.)

Example (continued). Consider the tree's root $\mathrm{Tr}(\mathcal{H}_1)$. Since $\mathcal{H}_1 = C_4$, $|\mathrm{Tr}(\mathcal{H}_1)| = 2$. Furthermore, the root's left and the right children are $\mathrm{Tr}(\{e_1, e_2\})$ and $\mathrm{Tr}(\{e_3, e_4\})$. Both $\{e_1, e_2\}$ and $\{e_3, e_4\}$ are paths on three vertices P_3 and thus have two minimal hitting sets. Last, the leaves of the tree are $\mathrm{Tr}(\{e_j\})$ $(1 \leq j \leq 4)$. Since every edge in $\mathcal{H}_1$ is a 2-set, $|\mathrm{Tr}(\{e_j\})| = 2$. Thus, this tree representation $\mathcal{T}_{\mathcal{H}_1}$ has Property 1.

To see how one can construct a tree representation $\mathcal{T}_{\mathcal{H}_i}$ with Property 1 for all $\mathcal{H}_i$ $(i \in \mathbb{N})$, we note how $\mathcal{H}_i$ is built from copies of $\mathcal{H}_{i-1}$ (i.e., $\mathcal{H}_i = (\mathcal{A} \vee \mathcal{B}) \wedge (\mathcal{C} \vee \mathcal{D})$, where $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$ are all vertex disjoint copies of $\mathcal{H}_{i-1}$), and we prove a more general claim which shows how one can use binary tree $\mathcal{T}_{\mathcal{H}_{i-1}}$ with Property 1 to construct $\mathcal{T}_{\mathcal{H}_i}$ with Property 1.

LEMMA 3. *Let* $\mathcal{A}$ *and* $\mathcal{B}$ *be vertex-disjoint hypergraphs which have binary tree representations* $\mathcal{T}_{\mathcal{A}}$ *and* $\mathcal{T}_{\mathcal{B}}$ *with Property* 1. *Then one can construct binary trees* $\mathcal{T}_{\mathcal{A} \vee \mathcal{B}}$ *and* $\mathcal{T}_{\mathcal{A} \wedge \mathcal{B}}$ *for hypergraphs* $\mathcal{A} \vee \mathcal{B}$ *and* $\mathcal{A} \wedge \mathcal{B}$ *which also have Property* 1.

Proof. For the proof of the lemma, we provide constructions and then use Corollary 1 from section 2 to prove the claim that every internal node of $\mathcal{T}_{\mathcal{A} \vee \mathcal{B}}$ and $\mathcal{T}_{\mathcal{A} \wedge \mathcal{B}}$ has Property 1. The constructions will rely on Lemma 1.

$\mathcal{T}_{\mathcal{A} \vee \mathcal{B}}$. Construct a tree by defining a root node with $\mathcal{T}_{\mathcal{A}}$ and $\mathcal{T}_{\mathcal{B}}$ as left and right subtrees. To see that the tree satisfies Property 1, the only node we need to check is the root. The children of the root are $\mathrm{Tr}(\mathcal{A})$ and $\mathrm{Tr}(\mathcal{B})$, and the root is $\mathrm{Tr}(\mathcal{A} \vee \mathcal{B})$. Using Corollary 1, we know that $|\mathrm{Tr}(\mathcal{A} \vee \mathcal{B})| = |\mathrm{Tr}(\mathcal{A})| \cdot |\mathrm{Tr}(\mathcal{B})|$. Thus, the root contains at least as many minimal hitting sets as either of its children.

$\mathcal{T}_{\mathcal{A} \wedge \mathcal{B}}$. Construct $\mathcal{T}_{\mathcal{A} \wedge \mathcal{B}}$ first by replacing every internal node $\mathrm{Tr}(\mathcal{S})$ in $\mathcal{T}_{\mathcal{B}}$ with $\mathrm{Tr}(\mathcal{A} \wedge \mathcal{S})$ and then by replacing every leaf node $\mathrm{Tr}(\{b_j\})$ (where $b_j \in \mathcal{B}$) with the root of a subtree $\mathcal{T}_{\mathcal{A} \wedge \{b_j\}}$, which is formed by replacing every node $\mathrm{Tr}(\mathcal{Z})$ in $\mathcal{T}_{\mathcal{A}}$ with the node $\mathrm{Tr}(\mathcal{Z} \wedge \{b_j\})$. Since the leaf nodes of $\mathcal{T}_{\mathcal{A}}$ are $\mathrm{Tr}(\mathcal{Z})$, where $\mathcal{Z}$ is a single edge $a_k \in \mathcal{A}$, the leaf nodes of this new tree will be $\mathrm{Tr}(\{a_k\} \wedge \{b_j\})$, and the new tree will have $|\mathcal{A}||\mathcal{B}|$ leaves. (Note that the new tree will have a shape which can be visualized by taking a $\mathcal{T}_{\mathcal{B}}$ and replacing every leaf node by a subtree in the shape of a $\mathcal{T}_{\mathcal{A}}$.)

To see that this construction is the tree representation for $\mathcal{A} \wedge \mathcal{B}$, we show how the contents of any internal node may be computed from its left and right child. Given an internal node $\mathrm{Tr}(\mathcal{S})$ in $\mathcal{T}_{\mathcal{B}}$ with children $\mathrm{Tr}(\mathcal{L})$ and $\mathrm{Tr}(\mathcal{R})$ (i.e., $\mathcal{L} \vee \mathcal{R} = \mathcal{S} \subseteq \mathcal{B}$), we replace these three nodes in $\mathcal{T}_{\mathcal{B}}$ with $\mathrm{Tr}(\mathcal{A} \wedge \mathcal{S})$, $\mathrm{Tr}(\mathcal{A} \wedge \mathcal{L})$, and $\mathrm{Tr}(\mathcal{A} \wedge \mathcal{R})$, respectively. Since $(\mathcal{A} \wedge \mathcal{L}) \vee (\mathcal{A} \wedge \mathcal{R}) = \mathcal{A} \wedge \mathcal{S} \subseteq \mathcal{A} \wedge \mathcal{B}$, we can calculate $\mathrm{Tr}(\mathcal{A} \wedge \mathcal{S})$ by applying Lemma 1 to that node's left and right children.

Furthermore, note that since $\mathcal{T}_{\mathcal{B}}$ has Property 1, we know that $|\mathrm{Tr}(\mathcal{S})| \geq \max(|\mathrm{Tr}(\mathcal{L})|, |\mathrm{Tr}(\mathcal{R})|)$. Because $\mathcal{L}, \mathcal{R}, \mathcal{S}$ are vertex-disjoint from $\mathcal{A}$, we can apply

Corollary 1 to $\mathrm{Tr}(\mathcal{A} \wedge \mathcal{S})$, $\mathrm{Tr}(\mathcal{A} \wedge \mathcal{L})$, and $\mathrm{Tr}(\mathcal{A} \wedge \mathcal{R})$ and show that

$$\underbrace{|\mathrm{Tr}(\mathcal{A} \wedge \mathcal{S})|}_{|\mathrm{Tr}(\mathcal{A})|+|\mathrm{Tr}(\mathcal{S})|} \geq \max(\underbrace{|\mathrm{Tr}(\mathcal{A} \wedge \mathcal{L})|}_{|\mathrm{Tr}(\mathcal{A})|+|\mathrm{Tr}(\mathcal{L})|}, \underbrace{|\mathrm{Tr}(\mathcal{A} \wedge \mathcal{R})|}_{|\mathrm{Tr}(\mathcal{A})|+|\mathrm{Tr}(\mathcal{R})|}).$$

To complete the proof that the rest of the nodes in $\mathcal{T}_{\mathcal{A}\wedge\mathcal{B}}$ have the desired properties, we look at the subtrees rooted at $\mathrm{Tr}(\mathcal{A} \wedge \{b_j\})$ and apply very similar arguments to nodes $\mathrm{Tr}(\mathcal{Z} \wedge \{b_j\})$, where $\mathcal{Z} \subseteq \mathcal{A}$ and $b_j \in \mathcal{B}$. First, note how they are modified from nodes in $\mathcal{T}_\mathcal{A}$ by replacing the node $\mathrm{Tr}(\mathcal{Z})$ with $\mathrm{Tr}(\mathcal{Z} \wedge \{b_j\})$ (where $\mathcal{Z} = \mathcal{L} \vee \mathcal{R}$ and $\mathcal{L}, \mathcal{R}, \mathcal{Z} \subseteq \mathcal{A}$) and then apply Corollary 1. □

Using the above property and lemma, we may now state the following result about the generalization of the Sequential Method. (Note that Theorem 2 applies only to the infinite family of hypergraphs $\mathcal{H}_i$ ($i \in \mathbb{N}$). A worst-case analysis of this generalization remains an open problem.)

Theorem 2. *There exists a tree representation $\mathcal{T}_{\mathcal{H}_i}$ such that, using this representation, a generalization of the Sequential Method can list $\mathrm{Tr}(\mathcal{H}_i)$ in time $O(|V_{\mathcal{H}_i}| \cdot |\mathcal{H}_i| \cdot |\mathrm{Tr}(\mathcal{H}_i)|^4)$.*

Proof. First, we show the existence of a tree for $\mathcal{H}_i$ having Property 1 by induction on i. The basis step for $\mathcal{H}_i$ is described in the example above. The induction step follows by applying Lemma 3.

Second, we demonstrate the running time. To see how the stated running time follows from the existence of such a tree, note how Algorithm 2 proceeds by performing a depth-first traversal of the tree. The time it takes to determine the contents of any internal node $\mathrm{Tr}(\mathcal{A} \vee \mathcal{B})$ is the time to compute $\mathrm{Tr}(\mathcal{A} \vee \mathcal{B})$ from $\mathrm{Tr}(\mathcal{A})$ and $\mathrm{Tr}(\mathcal{B})$. Lemma 1 shows how this can be done by finding minimal sets in $\mathrm{Tr}(\mathcal{A}) \wedge \mathrm{Tr}(\mathcal{B})$. Using Property 1, we know that $|\mathrm{Tr}(\mathcal{A})|$, $|\mathrm{Tr}(\mathcal{B})|$, and $|\mathrm{Tr}(\mathcal{A}\vee\mathcal{B})|$ are at most $|\mathrm{Tr}(\mathcal{H})|$. Therefore, $|\mathrm{Tr}(\mathcal{A}) \wedge \mathrm{Tr}(\mathcal{B})| \leq |\mathrm{Tr}(\mathcal{H})|^2$. One can compute $\min(\mathrm{Tr}(\mathcal{A}) \wedge \mathrm{Tr}(\mathcal{B}))$ by comparing every two sets and keeping only the minimal ones. This can be done in $O\big(|V_{\mathcal{H}_i}| \cdot \binom{|\mathrm{Tr}(\mathcal{H}_i)|^2}{2}\big)$ time for every node. Since $\mathcal{T}_\mathcal{H}$ is a binary tree with $|\mathcal{H}|$ leaves and, thus, $|\mathcal{H}| - 1$ internal nodes, the time it will take to perform a depth-first traversal of $\mathcal{T}_\mathcal{H}$ will be $O(|V_{\mathcal{H}_i}| \cdot |\mathcal{H}_i| \cdot |\mathrm{Tr}(\mathcal{H}_i)|^4)$. □

5. Other remarks and further work.

While the question which Hirsh originally posed has been resolved, its analogue for the generalized version still remains unanswered. Does there exist a hypergraph whose minimal hitting sets cannot be listed efficiently regardless of what tree representation we use to process the edges? That is to say, while the generalized Sequential Method is able to process $\mathcal{H}_i$ efficiently, does it fail on some other hypergraph?

It would also be interesting to improve the $m^{\Omega(\log\log m)}$ lower bound of Theorem 1 for the running time of the Sequential Method with the optimal ordering of the edges. In particular, as this lower bound is smaller than the upper bound of the Fredman–Khachiyan algorithm, it could still be the case that the edges can always be ordered (perhaps even by an efficient algorithm) in such a way that using this ordering the Sequential Method outperforms the Fredman–Khachiyan algorithm. Experiments show that for the lexicographic ordering of the edges of $\mathcal{H}_i$, the largest blowup occurs before the penultimate step. Thus it may be possible to get some improvement from considering $\mathcal{H}_i$ in more detail. This would also involve the question of whether the lexicographic order is the best one for $\mathcal{H}_i$. In addition, given the relevance of listing $\mathrm{Tr}(\mathcal{H})$ to data mining, it would be useful to investigate implementations related to the Sequential Method that have worked well on practically relevant data. Recent results by Hagen [20] provide theoretical bounds for three such algorithms [3, 9, 24].

Another area to study is the probabilistic analysis of the Sequential Method. Experiments suggest that using lexicographical ordering performs well on random hypergraphs. We also try to develop heuristics for the application of the generalization of the Sequential Method.

Acknowledgments. I would like to thank Prof. H. Hirsh for suggesting the problem discussed in this paper and also the two anonymous referees as well as Prof. Gy. Turán for useful suggestions and remarks.

REFERENCES

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I Verkamo, *Fast discovery of association rules*, in Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, 1996, pp. 307–328.

[2] D. Angluin, L. Hellerstein, and M. Karpinski, *Learning read-once formulas with queries*, J. Assoc. Comput. Mach., 40 (1993), pp. 185–210.

[3] J. Bailey, T. Manoukian, and K. Ramanohanarao, *A fast algorithm for computing hypergraph transversals and its application in mining emerging patterns*, in Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), IEEE Computer Society Press, Los Alamitos, CA, pp. 485–488.

[4] C. Berge, *Hypergraphs*, North–Holland Math. Library 455, North–Holland, Amsterdam, 1989.

[5] J. C. Bioch and T. Ibaraki, *Complexity of identification and dualization of positive Boolean functions*, Inform. and Comput., 123 (1995), pp. 50–63.

[6] E. Boros, K. Elbassioni, V. Gurvich, and L. Khachiyan, *An efficient incremental algorithm for generating all maximal independent sets in hypergraphs of bounded dimension*, Parallel Process. Lett., 10 (2000), pp. 253–266.

[7] E. Boros, V. Gurvich, L. Khachiyan, and K. Makino, *On the complexity of generating maximal frequent and minimal infrequent sets*, in STACS 2002, Lecture Notes in Comput. Sci. 2285, Springer, Berlin, 2002, pp. 133–141.

[8] E. Boros, V. Gurvich, L. Khachiyan, and K. Makino, *Dual-bounded generating problems: Partial and multiple transversals of a hypergraph*, SIAM J. Comput., 30 (2001), pp. 2036–2050.

[9] G. Dong and J. Li, *Mining border descriptions of emerging patterns from dataset pairs*, Knowledge and Information Systems, 8 (2005), pp. 178–202.

[10] T. Eiter and G. Gottlob, *Identifying the minimal transversals of a hypergraph and related problems*, SIAM J. Comput., 24 (1995), pp. 1278–1304.

[11] T. Eiter and G. Gottlob, *Hypergraph transversal computation and related problems in logic and AI*, in Proceedings of the 8th European Conference on Logics in Artificial Intelligence (JELIA 2002), Lecture Notes in Comput. Sci. 2424, Springer, Berlin, 2002, pp. 549–564.

[12] T. Eiter, G. Gottlob, and K. Makino, *New results on monotone dualization and generating hypergraph transversals*, SIAM J. Comput., 32 (2003), pp. 514–537.

[13] M. L. Fredman and L. Khachiyan, *On the complexity of dualization of monotone disjunctive normal forms*, J. Algorithms, 21 (1996), pp. 618–628.

[14] Z. Füredi, R. H. Sloan, K. Takata, and Gy. Turán, *On set systems with a threshold property*, Discrete Math., 306 (2006), pp. 3097–3111.

[15] L. A. Goldberg, *Efficient Algorithms for Listing Combinatorial Structures*, Cambridge University Press, Cambridge, UK, 1993.

[16] D. Gunopulos, R. Khardon, H. Mannila, and H. Toivonen, *Data mining, hypergraph transversals, and machine learning*, in Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, ACM Press, New York, 1997, pp. 209–216.

[17] V. Gurvich, *On repetition-free Boolean functions*, Uspehi Mat. Nauk, 32 (1977), pp. 183–184.

[18] V. Gurvich and L. Khachiyan, *On the frequency of the most frequently occurring variable in dual monotone DNFs*, Discrete Math., 169 (1997), pp. 245–248.

[19] V. Gurvich and L. Khachiyan, *On generating the irredundant conjunctive and disjunctive normal forms of monotone Boolean functions*, Discrete Appl. Math., 96/97 (1999), pp. 363–373.

[20] M. Hagen, *Lower bounds for three algorithms for the transversal hypergraph generation*, in Proceedings of the 33rd International Workshop on Graph-Theoretic Concepts in Computer

Science (WG 2007), Lecture Notes in Comput. Sci. 4769, Springer, Berlin, 2007, pp. 316–327.

[21] H. Hirsh, *private communication*, 2000.

[22] D. Johnson, M. Yannakakis, and C. H. Papadimitriou, *On generating all maximal independent sets*, Inform. Process. Lett., 27 (1988), pp. 119–123.

[23] M. Karchmer, N. Linial, I. Newman, M. Saks, and A. Widgerson, *Combinatorial characterization of read-once formulae*, Discrete Math., 114 (1993), pp. 275–282.

[24] D. J. Kavvadias and E. C. Stavropoulos, *An efficient algorithm for the transversal hypergraph generation*, J. Graph Algorithms Appl., 9 (2005), pp. 239–264.

[25] E. L. Lawler, *Covering problems: Duality relations and a new method of solution*, SIAM J. Appl. Math., 14 (1966), pp. 1115–1132.

[26] K. Maghout, *Sur la détermination des nombres de stabilité et du nombre chromatique d'un graphe*, C. R. Acad. Sci. Paris, 248 (1959), pp. 3522–3523.

[27] K. Makino and T. Ibaraki, *Inner-core and outer-core functions of partially defined Boolean functions*, Discrete Appl. Math., 96/97 (1999), pp. 443–460.

[28] N. Mishra and L. Pitt, *Generating all maximal independent sets of bounded-degree hypergraphs*, in Proceedings of the 10th Annual Conference on Computational Learning Theory, ACM Press, New York, 1997, pp. 211–217.

[29] D. Mundici, *Functions computed by monotone Boolean formulas with no repeated variables*, Theoret. Comput. Sci., 66 (1989), pp. 113–114.

[30] R. J. Nelson, *Simplest normal truth functions*, J. Symb. Logic, 20 (1955), pp. 105–108.

[31] R. J. Nelson, *Weak simplest normal truth functions*, J. Symb. Logic, 20 (1955), pp. 232–234.

[32] M. C. Paull and S. H. Unger, *Minimizing the number of states in incompletely specified sequential switching functions*, IRE Trans. Electr. Comput., 8 (1959), pp. 356–367.

[33] R. H. Sloan, K. Takata, and Gy. Turán, *On frequent sets of Boolean matrices*, Ann. Math. Artificial Intelligence, 24 (1998), pp. 193–209.

[34] K. Takata, *Listing Algorithms*, Ph.D. thesis, University of Illinois at Chicago, Chicago, IL, 2004.

[35] S. Tsukiyama, M. Ide, H. Ariyoshi, and I. Shirakawa, *A new algorithm for generating all the maximal independent sets*, SIAM J. Comput., 6 (1977), pp. 505–517.

SIAM J. Discrete Math.
Vol. 21, No. 4, pp. 947–958

ON THE NUMBER OF FIXED PAIRS IN A RANDOM INSTANCE OF THE STABLE MARRIAGE PROBLEM*

B. PITTEL[†], L. SHEPP[‡], AND E. VEKLEROV[§]

Abstract. Consider a group of n men and n women, each ranking the members of the opposite sex as a potential marriage partner. A matching (marriage) of men and women is called stable if there is no pair (man, woman) who are not matched but prefer each other to their partners in the matching. It is known that, for every instance of the rankings, there is at least one stable matching and that there are instances with exponentially many stable matchings. Assume that the instance is chosen uniformly at random among all $(n!)^{2n}$ possibilities. In this case the likely number of stable matchings is known to be $n^{1/2-o(1)}$, with high probability, and of order $n \ln n$, with probability 0.84 at least. In this paper we show that the average number of fixed pairs (man, woman), i.e., pairs common to all stable matchings, is asymptotic to $\ln^2 n$. More generally, the average number of women (men) with k stable husbands (wives) is asymptotic to $(\ln^{k+1} n)/(k-1)!$.

Key words. stable matching, proposal algorithm, random instance, likely behavior, asymptotics

AMS subject classifications. 05A05, 05A16, 60C05

DOI. 10.1137/070696155

1. Introduction and main result.

In a group of n men and n women, each person ranks the members of the opposite sex as a potential marriage partner. A matching (marriage) $\mathcal{M}$ between the set of men and the set of women is called stable if there is no pair (man, woman) who are not matched in $\mathcal{M}$ but prefer each other to their partners in $\mathcal{M}$. Back in 1962, Gale and Shapley [2] showed that, for any given rankings instance, at least one stable matching exists always. They proved it by developing a "proposal" algorithm which always finds a special stable matching $\mathcal{M}_1$. In this algorithm men propose to women in rounds, with each woman resolving "collisions" in favor of a currently best suitor and rejected men each proposing next to the best woman among those who haven't rejected him earlier. Furthermore, they proved that in any other stable matching $\mathcal{M}$, if any exists, no man (woman) has a better wife (worse husband) than in $\mathcal{M}_1$. In 1971–1972, McVitie and Wilson [8, 11] introduced a sequential version of the Gale–Shapley algorithm, in which proposals are made one at a time. This algorithm delivers the same matching $\mathcal{M}_1$, and each man makes the same proposals as in the proposals-by-rounds version, regardless of the order in which the "free" men propose; cf. Gusfield and Irving [3].

By establishing a connection with a *coupon collector* problem, Wilson proved that the expected running time (total number of proposals) is at most

$$n \sum_{j=1}^{n} \frac{1}{j} \sim n \ln n,$$

if the ranking instance is chosen uniformly at random among all $(n!)^{2n}$ instances. In 1976, Knuth [4] undertook a systematic study of the stable matchings. He found a

[*] Received by the editors July 3, 2007; accepted for publication (in revised form) September 8, 2007; published electronically December 7, 2007.
http://www.siam.org/journals/sidma/21-4/69615.html

[†] Department of Mathematics, Ohio State University, Columbus, OH 43210 (bgp@math.ohio-state.edu). Research for this author was supported in part by NSF grant DMS-0406024.

[‡] Department of Statistics, Rutgers University, Piscataway, NJ 08854 (shepp@stat.rutgers.edu).

[§] Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720 (eveklerov@lbl.gov).

lower bound for the expected running time which matched Wilson's upper bound and gave an example of the rankings with $2^{n/2}$ stable matchings. Years later, Gusfield and Irving [3] proved that the largest number of stable matchings grows as 2^n at least and constructed an n-size instance (n being a power of 2), for which they expected an even larger number of solutions. Knuth [5] used the recurrence from [3] to show that, indeed, that number is of order 2.28^n. He also found an integral formula for the probability that a given matching is stable, suggesting a problem of estimating this integral asymptotically [4]. Pittel [9] proved that the integral is asymptotic to $(n \ln n)/(en!)$, so that the expected number of solutions is about $e^{-1} n \ln n$. Knuth, Motwani, and Pittel [6] found an extension of the McVitie–Wilson algorithm that delivers all of the stable husbands of any given woman and used this algorithm to show that, with high probability (whp), the number of those stable husbands is between $(1/2 - \varepsilon) \ln n$ and $(1 + \varepsilon) \ln n$. Thus a *likely* number of stable matchings was shown to grow with n as $\ln n$ at least. Later Pittel [10] proved that, in fact, this number grows much faster, at least as $n^{1/2-o(1)}$. Recently Lennon and Pittel [7] extended the techniques in [9, 10], based on Knuth-type integral formulas, to show that the likely number of solutions is of order $n \ln n$, with probability 0.84 at least. It is quite plausible that this probability approaches 1; i.e., the likely values of the number of solutions are on the order of its expected value.

Thus, typically the random rankings instance has plenty of stable matchings. Surprisingly, whp, for every stable matching, the product of the total wives' rank and the total husbands' rank is relatively asymptotic to the same quantity n^3. Thus, switching from one stable matching to another, what one side collectively gains is what another side collectively loses.

In light of many stable matchings, what is then a likely number of the *fixed* pairs $(m,\, w)$, i.e., the pairs common to all stable matchings? It was proved in [10] that whp the number of women whose husband in Gale–Shapley's matching $\mathcal{M}_1$ is their best choice is close to $\ln n$. Clearly each of these women is matched with the same man in every stable matching, implying that whp the number of fixed pairs grows as $\ln n$ at least.

Our goal in this paper is to demonstrate that the expected number of the fixed pairs is asymptotic to $\ln^2 n$ and that, more generally, the expected number of women (men) with k stable husbands (wives) is asymptotic to $(\ln^{k+1} n)/(k-1)!$, $k \geq 1$. We do this by using the probabilistic analysis of the extended proposal algorithm and some estimates in [10].

We conjecture that, in fact, the distribution of the number of women (men) with k stable husbands (wives) is concentrated around $(\ln^{k+1} n)/(k-1)!$. A proof of this conjecture, if based on the second moment method, would almost certainly depend on the availability of a sequential proposal algorithm which determines all stable husbands of *two* women, and no such algorithm comes to mind at this moment. We hasten to add that there is known an algorithm by Irving and Gusfield [3], that determines all stable pairs in $O(n^2)$ steps, but a sharp probabilistic analysis of this algorithm is not in sight.

2. Proofs. The argument is based on properties of a proposal algorithm that, for a given rankings instance, determines *all* stable husbands of a particular woman w_1 [6].

(a) The first phase is the McVitie–Wilson proposal algorithm [8], which delivers the stable matching $\mathcal{M}_1$, simultaneously male-optimal and female-pessimal. (Each man's wife (woman's husband) in $\mathcal{M}_1$ is his best (her worst) partner he (she) can have in any stable matching.)

Here is how it works. The *arbitrarily ordered* men propose in turn to women. Each man always proposes to a woman of his first choice among the women to whom he has not proposed so far. The chosen woman accepts his proposal temporarily if she does not hold a proposal already; if she does, then the collision is resolved in favor of a man whom she likes better. Whoever is rejected goes back to the queue of free men, and, when his turn comes, proposes to his next best choice. Once all of the women have been proposed to, we have a complete matching, which turns out to be $\mathcal{M}_1$, regardless of a queue discipline. Bearing in mind how the collisions are resolved, and knowing that $\mathcal{M}_1$ is female-pessimal, we see that in this phase no woman has been proposed to by a stable husband other than her worst stable husband in $\mathcal{M}_1$. However, for every woman, he is still best among the men who have proposed to her so far.

Let us look at this phase from a woman w_1's point of view. Suppose that by some moment she has been proposed to by men $m_1, \dots, m_k$, in that chronological order. Let $M = \big\{m_j : j \in [k]\big\}$. By the rule, she retains the best man, m_i say, and each of the men from the set $M \setminus \{m_i\}$ is matched in $\mathcal{M}_1$ with a woman below w_1 on his preference list. w_1 will receive no further proposals, and consequently w_1 is matched with m_i in $\mathcal{M}_1$, iff every other man's partner in $\mathcal{M}_1$ is higher than w_1 on his preference list. Thus, the chronological ordering of the set M is immaterial for whether or not w_1 will receive other proposals.

(b) The second phase begins with woman w_1 rejecting her stable husband in $\mathcal{M}_1$. He is forced to propose to a woman of his next best choice, thus triggering a sequence of collisions always resolved in favor of a better suitor. If w_1 receives a proposal before one of the men runs out of his choices, we have a matching. w_1 rejects her latest proposer, his next proposal leads to another sequence of collisions, and so on. By rejecting "blindly" all of the proposals she receives, w_1 forces a sequence of collisions to run until one of the men is rejected by a woman of his last (nth) choice, at which moment the process stops. It was proven in [6] that all of the stable husbands of w_1, besides her worst stable husband in $\mathcal{M}_1$, are among the men who proposed to her in the second phase. Specifically, a man who has proposed to w_1 in the second phase is her stable husband, and the resulting matching is stable, iff she prefers him to all of the previous proposers and thus to all of the stable husbands determined so far. So her best stable husband is the last proposer whom she likes more than all of the previous proposers. A key feature of this phase is that w_1 can postpone determination of who those stable husbands are until the proposal process terminates. In particular, w_1's stable husband from $\mathcal{M}_1$ is her only stable husband iff in phase 2 she does not receive any better proposal.

We will refer to this algorithm as the extended proposal algorithm, EPA for brevity.

Suppose that the rankings instance is chosen uniformly at random among all $(n!)^{2n}$ possible instances. Let $X_n = X_n(w_1)$ denote the random number of stable husbands of w_1. (By symmetry, $X_n(w_1)$ equals $X_n(w)$ in distribution, for any other woman w.) A probabilistic study of the EPA showed that

$$P\big\{(1/2 - \varepsilon)\ln n \le X_n \le (1+\varepsilon)\ln n\big\} \to 1, \quad n \to \infty,$$

for every fixed $\varepsilon > 0$ [6]. A follow-up analysis in [10] of this algorithm produced a sharper result: X_n is asymptotically $\mathcal{N}(\ln n, \ln n)$, i.e., normal, with mean and variance $\ln n$. As an afterthought it was demonstrated there (section 6, note 2) that Y_n, the total number of women whose husband in $\mathcal{M}_1$ has rank 1, is also $\mathcal{N}(\ln n, \ln n)$, asymptotically.

Let Z_n denote the total number of women each having exactly one stable husband. Clearly $Z_n \geq Y_n$, and a question raised in [10] was how large, typically, is $Z_n - Y_n$.

THEOREM 2.1.

$$\lim_{n\to\infty} \frac{E[Z_n]}{\ln^2 n} = 1. \tag{2.1}$$

More generally, denoting by $Z_{n,m}$ the total number of women, each having exactly m stable husbands ($Z_{n,1} = Z_n$),

$$\lim_{n\to\infty} \frac{E[Z_{n,m}]}{\ln^{m+1} n} = \frac{1}{(m-1)!}, \quad m \geq 1. \tag{2.2}$$

Note. We conjecture that, in fact, the distribution of $Z_{n,m}$ is concentrated around $E[Z_{n,m}]$. A proof of this conjecture based on the second moment method would almost certainly depend on the availability of a sequential proposal algorithm which determines all stable husbands of *two* women, and no such an extension of the EPA comes to mind at this moment. We hasten to add that there is known an algorithm [3] that determines all stable pairs in $O(n^2)$ steps, but a sharp probabilistic analysis of this algorithm is not in sight.

Proof. Step 1. Adopting a "principle of deferred decisions" [6], we postulate that a man, whose turn it is to propose, proposes to a woman chosen uniformly at random from among women he has not yet proposed to, independently of the past choices by other players. Dually, if a chosen woman has already been proposed to by some k men, then she ranks the current proposer, relative to those k men, as being jth best with probability $1/(k+1)$, $j = 1, \ldots, k+1$, independently of the past choices by other players. In particular, if at some time t during phase 1 a certain woman, w_1 say, has received her vth proposal, her current suitor is rejected in favor of the vth proposer with probability $1/v$. Also, the total number of proposers put on temporary hold by this woman up to and including time t is distributed as the number of left-record values in the uniformly random permutation of the set $[v] = \{1, \ldots, v\}$.

Introduce $V_{n,1}$ and V_n, the total number of proposals to w_1 in phase 1 and in both phases, respectively. From the key feautures of the EPA, namely irrelevance of chronology of proposals to w_1 in phase 1 and blind rejection by w_1 of all proposals in phase 2, it follows that, conditioned on $(V_{n,1}, V_n)$, the ranking, by w_1, of V_n proposals is the uniformly random permutation of $[V_n]$.

In particular, the conditional probability that w_1 has a single stable husband, i.e., $X_n = 1$, is the probability that the entry 1 of the uniformly random permutation of $[V_n]$ is among its first $V_{n,1}$ entries, so that

$$P\{X_n = 1 \,|\, (V_{n,1}, V_n)\} = \frac{V_{n,1}}{V_n} \Longrightarrow P\{X_n = 1\} = E\left[\frac{V_{n,1}}{V_n}\right]. \tag{2.3}$$

So our task is to estimate sharply the last expectation, since

$$E[Z_n] = nP\{X_n = 1\}.$$

To this end, introduce U_n, the rank of the best stable husband of w_1. By symmetry, U_n coincides in distribution with the rank of the best wife of a man, and this rank equals the number of proposals made by this man in phase 1. Since proposals made are proposals received, we see that

$$nE[U_n] = nE[V_{n,1}] \Longrightarrow E[U_n] = E[V_{n,1}]. \tag{2.4}$$

Furthermore, $U_n - 1$ is the number of men out of $n - V_n$ nonproposers to w_1, who would have been ranked by w_1 higher than any proposer. That is, *conditioned on* $(V_{n,1}, V_n)$, U_n has the same distribution as 1+ the occupancy number of a fixed box in the uniformly random allocation scheme with $V_n + 1$ boxes and $n - V_n$ *indistinguishable* balls, i.e.,

$$\begin{aligned} P\big\{U_n > u \,|\, (V_{n,1}, V_n)\big\} &= \frac{|\{\mathbf{x} \ge \mathbf{0} \,:\, x_1 + \cdots + x_{V_n+1} = n - V_n,\, x_1 \ge u\}|}{|\{\mathbf{x} \ge \mathbf{0} \,:\, x_1 + \cdots + x_{V_n+1} = n - V_n\}|} \\ &= \frac{\binom{n-u}{V_n}}{\binom{n}{V_n}} \ge 1 - \frac{uv}{n-v} \quad (0 \le u \le n - V_n). \end{aligned} \tag{2.5}$$

In particular,

$$E\big[U_n \,|\, (V_{n,1}, V_n)\big] = 1 + \frac{n - V_n}{V_n + 1} = \frac{n+1}{V_n + 1},$$

so that

$$\frac{1}{V_n + 1} = \frac{1}{n+1} E[U_n \,|\, (V_{n,1}, V_n)]. \tag{2.6}$$

Since V_n is likely to be large, (2.6) is a key tool for estimating $E[V_{n,1}/V_n]$.

Step 2. It was proven in [10] (estimates (7.5), (7.7)) that

$$E[V_{n,1}] = \ln n - O(n^{-1} \ln^3 n), \tag{2.7}$$

so, by (2.4),

$$E[U_n] = \ln n - O\big(n^{-1} \ln^3 n\big). \tag{2.8}$$

The bounds (2.5) and (2.8) lead (see the proof of Theorem 2.1 [10]) to

$$P\left\{V_n \ge \frac{n}{\ln^7 n}\right\} \ge 1 - \ln^{-3} n. \tag{2.9}$$

Furthermore, by Theorem 6.1 (estimates (6.3)–(6.4) with $a = 11$) in [10],

$$P\{U_n \le 13 \ln^2 n\} \ge 1 - n^{-3}. \tag{2.10}$$

While (2.8)–(2.10) is all we will need to know about U_n, V_n, it is critically important that the distribution of $V_{n,1}$ is sharply concentrated around $\ln n$. (Not that it matters, but we suspect that there is no such concentration for U_n.)

To prove concentration of $V_{n,1}$, we use the ingenious reduction of the phase 1 proposal algorithm, applied to the uniformly random instance, to the coupon collector problem by Wilson [11]. He suggested that each man always proposed to a woman chosen uniformly at random from among all women, with the natural proviso that every proposal to a woman to whom the man had proposed before would be rejected. This relaxed version will deliver the exact same stable matching $\mathcal{M}_1$, and the total number of proposals in it is distributed as N_n, the number of coupon draws in the n-coupon collector problem. Let $\tilde{V}_{n,1}$ denote the total number of proposals received by woman w_1; this number counts both initial (nonredundant) proposals and repeated (redundant) proposals to w_1 by men already rejected by her. Clearly the number of nonredundant proposals has the same distribution as $V_{n,1}$, so we have coupled $V_{n,1}$ and $\tilde{V}_{n,1}$ in such a way that $V_{n,1} \le \tilde{V}_{n,1}$. Now

$$N_n = \sum_{j=0}^{n-1} G_j,$$

where $G_0, \dots, G_{n-1}$ are independent geometrics, with success probability $1, 1-1/n, \dots, 1-(n-1)/n$, respectively. So

$$E[\tilde{V}_{n,1}] = \frac{E[N_n]}{n} = \sum_{j=1}^{n} \frac{1}{j} = \ln n + O(1),$$

and by combining this with (2.7) we obtain

$$E\big[\tilde{V}_{n,1} - V_{n,1}\big] = O\big(n^{-1} \ln^3 n\big). \tag{2.11}$$

Let us show that N_n is concentrated around $n \ln n$. First, given $\varepsilon \in (0,1)$, we have

$$\big\{N_n \le [(1-\varepsilon)n \ln n]\big\} = \bigcap_{j=1}^{n} \{O_j > 0\},$$

where $O_1, \dots, O_n$ are the occupancy numbers in the uniformly random allocation of $[(1-\varepsilon)n \ln n]$ *distinguishable* balls among n boxes. Then, since the events $\{O_j > 0\}$, $j \in [n]$, are negatively correlated,

$$\begin{aligned}
P\{N_n \le [(1-\varepsilon)n\ln n]\} &\le P^n\{O_1 > 0\} \\
&= \left(1 - \left(1 - \frac{1}{n}\right)^{[(1-\varepsilon)n\ln n]}\right)^n \le \exp\left(-n\left(1-\frac{1}{n}\right)^{(1-\varepsilon)n\ln n}\right) \\
&= \exp\left(-n\exp\left(-(1-\varepsilon)\ln n + O(n^{-1}\ln^2 n)\right)\right) \\
&= \exp\big(-n^{\varepsilon}(1 + O(n^{-1}\ln^2 n))\big) \\
&\le \exp\big(-0.5 n^{\varepsilon}\big),
\end{aligned}$$

uniformly for ε. Picking $\varepsilon = \ln^{-2/3} n$, we have then

$$P\{N_n \le [n\ln n - n\ln^{1/3} n]\} \le \exp\big(-0.5\exp(\ln^{1/3} n)\big).$$

Second,

$$\begin{aligned}
P\{N_n \ge [n\ln n + n\ln^{1/3} n]\} &= P\left\{\bigcap_{j=1}^{n}\{O_j = 0\}\right\} \\
&\le nP\{O_1 = 0\},
\end{aligned} \tag{2.12}$$

where $O_1, \dots, O_n$ are the occupancy numbers for the random allocation of $[n(\ln n + n\ln^{1/3} n)]$ balls. An easy computation shows then that

$$P\{N_n \ge [n\ln n + n\ln^{1/3} n]\} \le \exp\big(-0.5\ln^{1/3} n\big). \tag{2.13}$$

Thus

$$\begin{aligned}
&P(A_n) \ge 1 - 2\exp\big(-0.5\ln^{1/3} n\big), \\
&\quad A_n := \{[n\ln n - n\ln^{1/3} n] \le N_n \le [n\ln n + n\ln^{1/3} n]\}.
\end{aligned} \tag{2.14}$$

Further, in the event A_n, $\tilde{V}_{n,1}$ is sandwiched, stochastically, between two binomials,

$$B_1 := \mathrm{Bin}\big([n \ln n - n \ln^{1/3} n], p = 1/n\big), \quad B_2 := \mathrm{Bin}\big([n \ln n + n \ln^{1/3} n], p = 1/n\big).$$

Observe that

$$E[B_i] = \ln n + O(\ln^{1/3} n), \quad \mathrm{Var}[B_i] = \ln n + O(\ln^{1/3} n),$$

so that the standard deviation of B_i ($\sim \ln^{1/2} n$) far exceeds $\big|E[B_i] - \ln n\big|$. (That's the reason for our choice of ε !) Using the standard, Chernoff-type, bound for the tail probabilities of a binomial random variable, we get that, for some absolute constant $c > 0$,

$$\text{(2.15)} \qquad P\{|B_i - \ln n| \geq \ln^{2/3} n\} \leq \exp\left(-c\,\frac{(\ln^{2/3} n)^2}{\ln n}\right) = \exp\big(-c \ln^{1/3} n\big).$$

Combining (2.14) and (2.15), we conclude that

$$P\{|\tilde{V}_{n,1} - \ln n| \geq \ln^{2/3} n\} \leq \exp\big(-c^* \ln^{1/3} n\big)$$

for some absolute constant $c^* > 0$. Hence (see (2.11)),

$$\text{(2.16)} \qquad P\{|V_{n,1} - \ln n| \geq \ln^{2/3} n\} \leq \exp\big(-c_0 \ln^{1/3} n\big), \quad c_0 < c^*;$$

that is, $V_{n,1}$ is concentrated around $\ln n$ with probability $1 - \exp\big(-c^* \ln^{1/3} n\big)$, at least. In addition, with a considerably higher probability, $V_{n,1}$ is of order $\ln n$, at most. Indeed, analogously to (2.12)–(2.13),

$$P\{N_n < [3n \ln n]\} \geq 1 - \frac{1}{n^2},$$

and, using Chernoff's inequality,

$$P\{\mathrm{Bin}(m,p) \geq k\} \leq \exp\left[m\left(\frac{k}{m}\ln\frac{p}{k/m} + \left(1 - \frac{k}{m}\right)\ln\frac{1-p}{1-k/m}\right)\right] \quad (k \geq pm),$$

we also have

$$P\{\mathrm{Bin}([3n \ln n], p = 1/n) < 8 \ln n\} \geq 1 - \frac{1}{n^2}.$$

Therefore

$$P\{\tilde{V}_{n,1} < 8 \ln n\} \geq 1 - \frac{1}{n^2},$$

whence, as $V_{n,1} \leq \tilde{V}_{n,1}$,

$$\text{(2.17)} \qquad P\{V_{n,1} < 8 \ln n\} \geq 1 - \frac{1}{n^2},$$

as well.

Step 3. We are ready now for the asymptotic evaluation of $E[V_{n,1}/V_n]$. First of all, using the notation

$$E[X; A] := E[X \mathbf{1}_A],$$

we write

$$E\left[\frac{V_{n,1}}{V_n}\right] = E\left[\frac{V_{n,1}}{V_n}\,;\, V_n \ge \frac{n}{\ln^7 n}\right] + E\left[\frac{V_{n,1}}{V_n}\,;\, V_n < \frac{n}{\ln^7 n}\right] \tag{2.18}$$
$$= E_1^{(0)} + E_2^{(0)}.$$

Here, by (2.6), (2.9), (2.10), and (2.17),

$$\begin{aligned} E_2^{(0)} &\le 2E\left[\frac{V_{n,1}}{V_n+1}\,;\, V_n < \frac{n}{\ln^7 n}\right] = \frac{2}{n+1}E\left[V_{n,1}U_n\,;\, V_n < \frac{n}{\ln^7 n}\right] \\ &\le \frac{2n^2}{n+1}P\big\{(V_{n,1} \ge 8\ln n) \cup (U_n \ge 13\ln^2 n)\big\} + \frac{208\ln^3 n}{n+1}P\left\{V_n < \frac{n}{\ln^7 n}\right\} \\ &\le \frac{4n^2}{n+1}\cdot n^{-2} + \frac{208}{n+1}, \end{aligned}$$

so that

$$E_2^{(0)} = O(n^{-1}). \tag{2.19}$$

Next, within the factor $1 + O(n^{-1}\ln^7 n)$ coming from (2.6) and

$$V_n = (V_n+1)\big(1 + O(n^{-1}\ln^7 n)\big) \quad \text{on} \quad \left\{V_n \ge \frac{n}{\ln^7 n}\right\},$$

we have

$$\begin{aligned} E_1^{(0)} &= \frac{1}{n}E\left[U_nV_{n,1}\,;\, V_n \ge \frac{n}{\ln^7 n}\right] \\ &= \frac{1}{n}E\left[U_nV_{n,1}\,;\, \left(V_n \ge \frac{n}{\ln^7 n}\right) \cap \big((V_{n,1} < 8\ln n) \cap (U_n \le 13\ln^2 n)\big)\right] \\ &\quad + \frac{1}{n}E\left[U_nV_{n,1}\,;\, \left(V_n \ge \frac{n}{\ln^7 n}\right) \cap \big((V_{n,1} \ge 8\ln n) \cup (U_n \ge 13\ln^2 n)\big)\right] \\ &= E_1^{(1)} + E_2^{(1)}. \end{aligned} \tag{2.20}$$

Here

$$E_2^{(1)} \le nP\big\{(V_{n,1} \ge 8\ln n) \cup (U_n \ge 13\ln^2 n)\big\} = O(n^{-1}), \tag{2.21}$$

and

$$\begin{aligned} E_1^{(1)} &= \frac{1}{n}E\big[U_nV_{n,1}\,;\, (V_{n,1} < 8\ln n) \cap (U_n \le 13\ln^2 n)\big] \\ &\quad - O\big(n^{-1}\ln^3 nP\{V_n < n/\ln^7 n\}\big) \\ &= E_1^{(2)} - O(n^{-1}), \end{aligned} \tag{2.22}$$

the remainder estimate being based on (2.9). Further, by (2.16),

$$\begin{aligned} E_1^{(2)} &= \frac{1}{n}E\big[U_nV_{n,1}\,;\, (V_{n,1} < 8\ln n) \cap (U_n \le 13\ln^2 n) \cap (|V_{n,1} - \ln n| < \ln^{2/3} n)\big] \\ &\quad - O\big(n^{-1}\ln^3 n\,\exp(-c_0\ln^{1/3} n)\big) \\ &= E_1^{(3)} - o(n^{-1}), \end{aligned} \tag{2.23}$$

and

$$\begin{aligned}
E_1^{(3)} &= \frac{\ln n}{n}\, E\big[U_n\,;\,(V_{n,1} < 8\ln n) \cap (U_n \le 13\ln^2 n) \cap (|V_{n,1} - \ln n| < \ln^{2/3} n)\big]\\
&\quad + O\big(n^{-1} E[U_n] \ln^{2/3} n\big)\\
&= \frac{\ln n}{n}\, E\big[U_n\,;\,(V_{n,1} < 8\ln n) \cap (U_n \le 13\ln^2 n)\big]\\
&\quad + O\big(n^{-1}\ln^2 n\, P\{|V_{n,1} - \ln n| > \ln^{2/3} n\}\big) + O\big(n^{-1}\ln^{5/3} n\big)\\
&= E_1^{(4)} + O\big(n^{-1}\ln^{5/3} n\big),
\end{aligned} \tag{2.24}$$

the remainders estimates being based on (2.8) and (2.16). Finally, by (2.10), (2.17), and (2.8),

$$E_1^{(4)} = \frac{\ln n}{n}\, E[U_n] - O(n^{-2}\ln n) \sim \frac{\ln^2 n}{n}. \tag{2.25}$$

By combining the relations (2.18)–(2.25), we arrive at

$$P\{X_n = 1\} = E\left[\frac{V_{n,1}}{V_n}\right] \sim \frac{\ln^2 n}{n},$$

which proves (2.1).

Step 4. The proof of (2.2) runs parallel to the argument above, so we will concentrate on new technicalities. Let $m \ge 2$. Observe that

$$P\{X_n = m \,|\, (V_{n,1}, V_n)\} = P\{\omega^v \in \mathcal{A}^{v_1,v}\}|_{v_1 = V_{n,1},\, v = V_n}\,; \tag{2.26}$$

here ω^v is the uniformly random permutation of $[v]$, and $\mathcal{A}^{v_1,v}$ is the set of all permutations $\omega = (\omega_1, \dots, \omega_v)$ of $[v]$ such that the subsequence of $(\omega_{v_1+1}, \dots, \omega_v)$ consisting of ω_j's below $\min\big\{\omega_j \,:\, j \in [v_1]\big\}$ has $m-1$ left-record values. Clearly

$$P\{\omega^v \in \mathcal{A}^{v_1,v}\} = 0 \quad \forall\, v - v_1 < m - 1.$$

Let $v - v_1 \ge m - 1$, so $v \ge m$, in particular. Interpreting the entries of ω^v as the absolute ranks of the v independent $(0,1)$-uniforms, we have then

$$P\{\omega^v \in \mathcal{A}^{v_1,v}\} = \sum_{j \ge m-1} P_j(m-1)\, v_1 \binom{v - v_1}{j} \int_0^1 x^j (1-x)^{v-j-1}\, dx,$$

where $P_j(\mu)$ is the probability that ω^j, the uniformly random permutation of $[j]$, has μ left-record values. By performing the integration, we obtain

$$P\{\omega^v \in \mathcal{A}^{v_1,v}\} = \frac{v_1}{v} \sum_{j \ge m-1} P_j(m-1)\, \frac{(v - v_1)_j}{(v-1)_j}, \tag{2.27}$$

$$\le \frac{v_1}{v} \sum_{j=m-1}^{v} P_j(m-1). \tag{2.28}$$

Notice that the number of left-record values of ω^j is distributed as $\sum_{r=1}^{j} I_r$, where

$I_1, \dots, I_j \in \{0,1\}$ are independent, and $P\{I_r = 1\} = 1/r$. So

$$\begin{aligned} P_j(\mu) &= \sum_{\substack{S\subseteq\{2,\dots,j\}\\ |S|=\mu-1}} \prod_{r\in S} \frac{1}{r} \prod_{\rho\in\{2,\dots,j\}\setminus S} \left(1-\frac{1}{\rho}\right) \\ &\le \frac{\prod_{\rho\in[j]}(1-1/\rho)}{\prod_{r=2}^{\mu}(1-1/r)} \cdot \sum_{\substack{S\subseteq[j]\\ |S|=\mu-1}} \prod_{s\in S}\frac{1}{s} \\ &\le \frac{\mu}{j} \sum_{1\le r_1,\dots,r_{\mu-1}\le j} \prod_{i=1}^{\mu-1} \frac{1}{r_i} = \frac{\mu}{j}\left(\sum_{r=1}^{j}\frac{1}{r}\right)^{\mu-1} \\ &\le \frac{\mu}{j}(\ln j + 1)^{\mu-1}. \end{aligned} \tag{2.29}$$

Then, by (2.28),

$$\begin{aligned} P\{\omega^v \in \mathcal{A}^{v_1,v}\} &\le m\frac{v_1}{v}(\ln v + 1)^{m-2}\sum_{j=1}^{v}\frac{1}{j} \\ &\le c(m)\frac{v_1}{v}\ln^{m-1} v \le c(m)\frac{v_1}{v}(\ln n)^{m-1} \end{aligned}$$

for some constant $c(m) > 0$. This *upper* bound differs from the exact formula

$$P\{\omega^v \in \mathcal{A}^{v_1,v}\} = \frac{v_1}{v}, \quad m = 1,$$

only by the deterministic factor $c(m)(\ln n)^{m-1}$. So, by doing the counterparts of the estimates (2.18)–(2.23), we obtain

$$\begin{aligned} P\{X_n = m\} &= E\left[P\{\omega^v \in \mathcal{A}^{v_1,v}\}|_{\substack{v_1=V_{n,1}\\ v=V_n}}\right] \\ &= E\big[P\{\omega^v \in \mathcal{A}^{v_1,v}\}|_{\substack{v_1=V_{n,1}\\ v=V_n}}\,; (V_n \ge n\ln^{-7} n) \cap (V_{n,1} < 8\ln n) \\ &\qquad \cap(|V_{n,1} - \ln n| < \ln^{2/3} n) \cap (U_n \le 13\ln^2 n)\big] + o\big(n^{-1}\ln^{m+1} n\big) \\ &= \hat{E} + o\big(n^{-1}\ln^{m+1} n\big). \end{aligned} \tag{2.30}$$

The next step is to replace $P\{\omega^v \in \mathcal{A}^{v_1,v}\}$ with a simpler asymptotic approximation, assuming that v_1 and v meet the constraints imposed on $V_{n,1}$ and V_n in the expectation $\hat{E}$. Let $j(v) = [v/\ln n]$. By (2.29),

$$\begin{aligned} \frac{v_1}{v}\sum_{j=j(v)}^{v-v_1} P_j(m-1)\frac{(v-v_1)_j}{(v-1)_j} &\le c_1(m)\frac{v_1}{v}(\ln n)^{m-2}\sum_{j\ge j(v)}\frac{1}{j}\left(\frac{v-v_1}{v-1}\right)^j \\ &\le c_1(m)\frac{v_1}{v}(\ln n)^{m-2}\frac{1}{j(v)}\left(\frac{v-v_1}{v-1}\right)^{j(v)}\frac{v-1}{v_1-1} \\ &\le c_1(m)\frac{v_1}{v}(\ln n)^{m-2}\cdot\frac{\exp\left(-j(v)\frac{v_1-1}{v-1}\right)}{j(v)\frac{v_1-1}{v-1}} \\ &\le c_2(m)\frac{v_1}{v}(\ln n)^{m-2}. \end{aligned}$$

This computation shows also that

$$\frac{v_1}{v}\sum_{j=j(v)}^{\infty} P_j(m-1)\left(\frac{v-v_1}{v}\right)^j = O\left(\frac{v_1}{v}(\ln n)^{m-2}\right).$$

Furthermore, for $j \le j(v)$,

$$\begin{aligned}\frac{(v-v_1)_j}{(v-1)_j} &= \prod_{i=v-v_1+1}^{v-1}\left(1-\frac{j}{i}\right)\\ &= \exp\left[-j\sum_{i=v-v_1+1}^{v-1}\frac{1}{i}+O\left(j^2\sum_{i=v-v_1+1}^{v-1}\frac{1}{i^2}\right)\right]\\ &= \exp\left(-j\sum_{i=v-v_1+1}^{v-1}\frac{1}{i}+O\big(j^2(v)v_1/v^2\big)\right)\\ &= \exp\left(-j\sum_{i=v-v_1+1}^{v-1}\frac{1}{i}+O\big(\ln^{-1}n\big)\right)\\ &= \left(\frac{v-v_1}{v}\right)^j\big(1+O(\ln^{-1}n)\big).\end{aligned}$$

Therefore, "swapping the tails,"

$$\begin{aligned}\frac{v_1}{v}\sum_{j=m-1}^{v-v_1}P_j(m-1)\frac{(v-v_1)_j}{(v-1)_j} &= \big(1+O(\ln^{-1}n)\big)\frac{v_1}{v}\sum_{j=m-1}^{\infty}P_j(m-1)\left(\frac{v-v_1}{v}\right)^j\\ &\quad+O\left(\frac{v_1}{v}(\ln n)^{m-2}\right).\end{aligned}$$

The punchline is that, for $\mu \ge 1$,

$$\sum_{j=\mu}^{\infty}P_j(\mu)x^j = \frac{1}{\mu!}\ln^{\mu}\frac{1}{1-x}, \quad |x|<1.$$

This identity can be proved either from scratch by using the representation of the number of record values in the random permutation ω^j as the sum $I_1+\cdots+I_j$ of independent indicators or by observing that $P_j(\mu)=C(j,\mu)/j!$, where $C(j,\mu)$ is the Stirling number of the first kind, and using the classic exponential identity

$$\sum_{j,\mu}\frac{C(j,\mu)}{j!}x^j y^{\mu} = \exp\left(y\sum_{j=1}^{\infty}\frac{x^j}{j}\right) = \sum_{\mu\ge 0}\frac{y^{\mu}}{\mu!}\ln^{\mu}\frac{1}{1-x}, \quad |x|<1;$$

see Comtet [1, section 5.5], for example. Thus

$$\begin{aligned}P\{\omega^v \in \mathcal{A}^{v_1,v}\} &= \frac{v_1}{v}\sum_{j=m-1}^{v-v_1}P_j(m-1)\frac{(v-v_1)_j}{(v-1)_j}\\ &= \big(1+O(\ln^{-1}n)\big)\frac{v_1}{v}\,\frac{\ln^{m-1}\frac{v}{v_1}}{(m-1)!}+O\left(\frac{v_1}{v}(\ln n)^{m-2}\right)\\ &= \frac{v_1}{v}\,\frac{\ln^{m-1}\frac{v}{v_1}}{(m-1)!}+O\left(\frac{v_1}{v}(\ln n)^{m-2}\right)\\ &= \frac{v_1}{v}\,\frac{\ln^{m-1}n}{(m-1)!}+O\left(\frac{v_1}{v}(\ln n)^{m-2}\ln\ln n\right).\end{aligned}$$

By using this asymptotic formula $P\{\omega^v \in \mathcal{A}^{v_1,v}\}$ in the formula for $\hat{E}$ in (2.30) and arguing as in Step 3, we obtain

$$\begin{aligned} P\{X_n = m\} &= E\left[P\{\omega^v \in \mathcal{A}^{v_1,v}\}|_{\substack{v_1=V_{n,1}\\ v=V_n}} \right] \\ &\sim \frac{\ln^{m+1} n}{n(m-1)!}, \end{aligned}$$

which proves (2.2). □

A final remark. So the expected number of fixed pairs grows as $\ln^2 n$. Let us define a closed clique as a set of some ν men and ν women, $\nu < n$, such that in every stable matching these men and women are matched with each other only. If there are at least two stable matchings, the men and the women in fixed pairs are one such closed clique. It would be very interesting to determine a likely size of the largest closed clique with at most $n/2$ men and $n/2$ women. How does it compare with $\ln^2 n$? The question seems to be quite hard. It is not even clear how to estimate the probability that some 2 men and 2 women form a fixed clique. (Our result means that a man and a woman form a closed clique (fixed pair) with probability asymptotic to $\frac{\ln^2 n}{n^2}$.)

Acknowledgement. We are grateful to a referee for his/her interest in this work and thoughtful comments.

REFERENCES

[1] L. Comtet, *Advanced Combinatorics*, Dordrecht-Holland, Boston, 1974.

[2] D. Gale and L. S. Shapley, *College admissions and the stability of marriage*, Amer. Math. Monthly, 69 (1962), pp. 9–15.

[3] D. Gusfield and R. W. Irving, *The Stable Marriage Problem (Structure and Algorithms)*, MIT Press, Cambridge, MA, 1989.

[4] D. E. Knuth, *Marriages Stables et Leurs Relations avec d' Autres Problèmes Combinatoires*, les Presses de l' Université de Montréaul, 1976.

[5] D. E. Knuth, *private communication*, 1988.

[6] D. E. Knuth, R. Motwani, and B. Pittel, *Stable husbands*, Random Structures and Algorithms, 1 (1990), pp. 1–14.

[7] C. Lennon and B. Pittel, *On a likely number of solutions for a stable marriage problem*, Combin. Probab. Comput., submitted.

[8] D. G. McVitie and L. B. Wilson, *The stable marriage problem*, Comm. ACM, 14 (1971), pp. 486–490.

[9] B. Pittel, *The average number of stable matchings*, SIAM J. Discrete Math., 2 (1989), pp. 530–549.

[10] B. Pittel, *On likely solutions of a stable marriage problem* Ann. Appl. Probab., 2 (1992), pp. 358–401.

[11] L. B. Wilson, *An analysis of the stable assignment problem*, BIT, 12 (1972), pp. 569–575.

SIAM J. DISCRETE MATH.
Vol. 21, No. 4, pp. 959–979

CONSTANT WEIGHT CONFLICT-AVOIDING CODES[*]

KOJI MOMIHARA[†], MEINARD MÜLLER[‡], JUNYA SATOH[†], AND MASAKAZU JIMBO[†]

Abstract. A conflict-avoiding code (CAC) C of length n with weight k is a family of binary sequences of length n and weight k satisfying $\sum_{0\le t\le n-1} x_{it}x_{j,t+s} \le \lambda$ for any distinct codewords $x_i = (x_{i0}, x_{i1}, \dots, x_{i,n-1})$ and $x_j = (x_{j0}, x_{j1}, \dots, x_{j,n-1})$ in C and for any integer s, where the subscripts are taken modulo n. A CAC with maximum code size for given n and k is said to be optimal. A CAC has been studied for sending messages correctly through a multiple-access channel. The use of an optimal CAC enables the largest possible number of potential users to transmit information efficiently and reliably. In this paper, the case $\lambda = 1$ is treated, and various direct and recursive constructions of optimal CACs for weight $k = 4$ and 5 are obtained by providing constructions of CACs for general weight k. In particular, the maximum code size of CACs satisfying certain sufficient conditions is determined through number theoretical and combinatorial approaches.

Key words. conflict-avoiding codes, cyclotomic cosets, Kronecker density, recursive constructions

AMS subject classifications. 94B25, 94C30, 11R45

DOI. 10.1137/06067852X

1. Introduction.

Several authors in [14, 15, 16, 5, 13, 19, 10, 12, 11] have investigated protocol sequences for a multiple-access channel without feedback. In such a multiple-access channel model, the time axis is partitioned into slots whose duration corresponds to the transmission time for one packet and all users are supposed to have slot synchronization, but no other synchronization is assumed. If more than one user is sending packets in a particular slot simultaneously, then there is a conflict and the channel output in that slot is the unreadable collision symbol, called an *erasure*.

If the binary protocol sequence $x_i = (x_{i0}, x_{i1}, \dots, x_{i,n-1})$ has Hamming weight k, then user i sends k packets in each frame of n slots, where his or her protocol sequence appears. When a user i is sending a message by using a protocol sequence x_i, a different message from the other user may be sent by a protocol sequence x_j or its cyclic shift since only slot synchronization is assumed. The set $C = \{x_1, x_2, \dots, x_N\}$ of N binary sequences is called an (N, u, n, σ) *protocol sequence set* if any $x_i \in C$ is of length n and has the property that at least σ successful packet transmissions in a frame are guaranteed for each active user, provided that at most u users out of N potential users are active. On the assumption that the number of collisions of any two distinct sequences is at most λ, in order to guarantee each user that at least σ packets in a frame survive from collision, the weight k of the (N, u, n, σ) protocol sequence set satisfies $k \ge \sigma + \lambda(u - 1)$. If there is at least one packet that survives from collision, then there may be a chance to use an inner code for erasure correction. Let ℓ be the bit length of each slot. An $(n' = k\ell, k' = \sigma\ell, d' = k\ell - \sigma\ell + 1)$ shortened

[*] Received by the editors December 24, 2006; accepted for publication (in revised form) April 25, 2007; published electronically December 12, 2007.
http://www.siam.org/journals/sidma/21-4/67852.html

[†] Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, 464-8601, Japan (momihara@math.cm.is.nagoya-u.ac.jp, jsatoh@math.cm.is.nagoya-u.ac.jp, jimbo@is.nagoya-u.ac.jp). The last author was partially supported by JSPS Scientific Research (B)18340024 and by Monbu Kagakusyo Exploratory Research 17650080.

[‡] Department of Computer Science, University of Bonn, Römerstr. 164, 53117, Bonn, Germany (meinard@cs.uni-bonn.de). This author was partially supported by FY2005 JSPS Postdoctoral Fellowship for Foreign Researchers.

Reed–Solomon (RS) code over GF(q) can be used as a code for each user to decode his or her σ survives packets into k transmitted packets, since a $(k\ell, \sigma\ell, k\ell - \sigma\ell + 1)$ shortened RS code can correct up to $k\ell - \sigma\ell$ position erasures where the user's packets suffer from collision and then the $k - \sigma$ erased packets are recovered by the σ alive packets at the receiver. In order to use an inner code, every protocol sequence of C should have constant weight k. Such an (N, u, n, σ) protocol sequence set is also called a *conflict-avoiding code* (CAC) *of length n with weight k*. In this paper, it is not our objective to discuss inner codes for erasure correction but rather to provide an upper bound on the code size N of a CAC for given n and k in the case of $\lambda = 1$ and $k = \sigma + u - 1$, and to construct "optimal" CACs attaining the bound.

Let $\mathcal{P}(n, k)$ denote the set of all k-subsets of $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$, the residue ring of rational integers modulo n. Hereafter we denote each coset $[i]$, $0 \le i \le n-1$, in $\mathbb{Z}_n$ by i for simplicity. Each element $x \in \mathcal{P}(n, k)$ can be identified with a binary sequence in $\{0, 1\}^n$ of Hamming weight k representing the indices of the nonzero positions. Given a k-subset $x \in \mathcal{P}(n, k)$, we define the *difference set* of x by

$$\Delta(x) = \big\{ j - i \,|\, i, j \in x, i \ne j \big\}.$$

Note that $\Delta(x)$ contains at most $k(k-1)$ differences. Furthermore, $i \in \Delta(x)$ implies $(n - i) \in \Delta(x)$, i.e., $\Delta(x)$ is symmetric with respect to $n/2$. In a set notation, for the case of $\lambda = 1$ a CAC of length n with weight k is a subset $C \subset \mathcal{P}(n, k)$ satisfying the condition

$$\Delta(x) \cap \Delta(y) = \emptyset \text{ for any } x, y \in C \text{ with } x \ne y. \tag{1.1}$$

Each element $x \in C$ is called a *codeword* of *length n* and *weight k*. Without loss of generality, we can assume that any codeword contains the element 0. For given n and k, let CAC(n, k) denote the class of all conflict-avoiding codes of length n with weight k. The maximum size of some code in CAC(n, k) is denoted by M(n, k), i.e.,

$$\mathrm{M}(n, k) = \max\{|C| \,|\, C \in \mathrm{CAC}(n, k)\}.$$

A code $C \in$ CAC(n, k) is said to be *optimal* if $|C| = \mathrm{M}(n, k)$. The advantage of using an optimal CAC is that it enables the largest number of potential users to transmit packets efficiently and reliably in such a multiple-access channel model.

In the case of weight $k = 3$, Levenshtein and Tonchev [12] showed a construction of optimal CACs of length n for every $n \equiv 2 \pmod 4$ and for sufficiently large odd prime n. Levenshtein [10] extended the result to the case of sufficiently large odd integer n. Jimbo et al. [8] obtained a construction of an optimal CAC in the case when $n = 4m$ and $m \equiv 2 \pmod 4$ for $k = 3$. In the case of general k, Levenshtein [11] gave an infinite series of CACs with $n = p^r$, $k = (p+1)/2$, and code size $|C| = (n-1)/(2(k-1))$ for any prime $p \ge 3$ and integer $r \ge 2$.

In the remainder of this paper, we will describe various direct and recursive constructions to obtain optimal CACs. In section 2, we show upper bounds on the code size of CACs for weight $k = 4$ and 5. In sections 3 and 5, we give some direct constructions of CACs for general weight k and obtain new optimal CACs for $k = 4$ and 5. In particular, we obtain constructions of optimal codes:

(i) $C \in$ CAC$(n = p, 4)$ with $|C| = \frac{n-1}{6}$ for infinitely many primes $p \equiv 1 \pmod 6$.
(ii) $C \in$ CAC$(n = 3p, 4)$ with $|C| = \frac{n-3}{6}$ for all primes $p \equiv 7 \pmod 8$.
(iii) $C \in$ CAC$(n = 4p, 4)$ with $|C| = \frac{n+2}{6}$ for infinitely many primes $p \equiv 13 \pmod{24}$.

(iv) $C \in \mathrm{CAC}(n = 2p, 5)$ with $|C| = \frac{n-2}{8}$ for all primes $p \equiv 5 \pmod{24}$.

(v) $C \in \mathrm{CAC}(n = 4p, 5)$ with $|C| = \frac{n-4}{8}$ for all primes $p \equiv 11 \pmod{12}$.

Furthermore, we investigate the Kronecker density of those primes by using the Chevotarëv's density theorem. In section 4, we see the asymptotic behavior of maximum code size of CACs of length $n \equiv 0 \pmod 3$ and weight $k = 4$ by using the Euler's φ-function. Moreover, a recursive construction is given in section 6.

2. Equidifference CACs and upper bounds on code size. In order to find codes of large size, the condition given in (1.1) suggests using many codewords that possess a difference set of small size. This motivates the following definition. A codeword $x \in \mathcal{P}(n, k)$ is called *equidifference with generator* $i \in \mathbb{Z}_n \setminus \{0\}$ if it is of the form

$$x = x_i = \{0, i, 2i, \ldots, (k-1)i\}. \tag{2.1}$$

Note that the assumption that $x = x_i$ is k-subset implies the condition $ji \not\equiv 0 \pmod n$ holds for every j, $1 \leq j \leq k-1$. Furthermore, for an equidifference codeword x_i one has $\Delta(x_i) = \{\pm ji \,|\, 1 \leq j \leq k-1\}$ and $|\Delta(x_i)| \leq 2(k-1)$. A codeword with $|\Delta(x)| < 2(k-1)$ is said to be *exceptional.* It should be noted that there may exist exceptional codewords which are not equidifference. A code $C \in \mathrm{CAC}(n, k)$ is said to be *equidifference* if it consists entirely of equidifference codewords. The set of generators of such a code will be denoted by $\Gamma(C)$. Furthermore, the subclass consisting of equidifference codes in $\mathrm{CAC}(n, k)$ will be denoted by $\mathrm{CAC}^{\mathrm{e}}(n, k)$ and the maximum size of some equidifference CACs by $\mathrm{M}^{\mathrm{e}}(n, k)$. Obviously, one has $\mathrm{M}^{\mathrm{e}}(n, k) \leq \mathrm{M}(n, k)$.

Now we consider the case of equidifference conflict-avoiding codes of weight $k = 4$. The equidifference codewords of weight $k = 4$ are of the form $x_i = \{0, i, 2i, 3i\}$ for $i \in \mathbb{Z}_n \setminus \{0, n/2, n/3, 2n/3\}$, where the notation $\mathbb{Z}_n \setminus \{0, n/2, n/3, 2n/3\}$ means that $n/2$, $n/3$, and $2n/3$ are removed from $\mathbb{Z}_n$ only when $2|n$, $3|n$, or $3|2n$. It is not hard to see that for a general codeword x of weight four one has $3 \leq |\Delta(x)| \leq 12$.Furthermore, for an exceptional codeword x one can tediously check that

$$|\Delta(x)| = \begin{cases} 3 & \text{iff} \quad x = \{0, n/4, n/2, 3n/4\}, \\ 4 & \text{iff} \quad x = \{0, n/5, 2n/5, 3n/5\}, \\ 5 & \text{iff} \quad x = \{0, d, n/2, n/2 + d\} \text{ or } x = \{0, d, n/2, n - d\} \\ & \qquad \text{for any } d \in \mathbb{Z}_n \setminus \{0, n/4, n/2, 3n/4\}. \end{cases} \tag{2.2}$$

Note that for given $C \in \mathrm{CAC}(n, 4)$, the difference sets $\Delta(x)$ for $x \in C$ are pairwise disjoint subsets of $\mathbb{Z}_n$. From this fact, one obtains the following upper bound on code size.

Lemma 2.1. *Let $n = 2^r 5^s m$, where m is not divisible by 2 and 5. Then it holds that*

$$\mathrm{M}(n, 4) \leq \begin{cases} \lfloor n/6 \rfloor & \text{if } r = 1,\ s = 0, \\ \lfloor (n+1)/6 \rfloor & \text{if } r = 0,\ s \geq 1, \\ \lfloor (n+2)/6 \rfloor & \text{if } r \geq 2,\ s = 0, \text{ or } r = 1,\ s \geq 1, \\ \lfloor (n+4)/6 \rfloor & \text{if } r \geq 2,\ s \geq 1, \\ \lfloor (n-1)/6 \rfloor & \text{if } r = s = 0. \end{cases}$$

For example, if $r \geq 2$ and $s \geq 1$, since some $C \in \mathrm{CAC}(n, 4)$ can contain the two exceptional codewords $x_{n/4} = \{0, n/4, n/2, 3n/4\}$ and $x_{n/5} = \{0, n/5, 2n/5, 3n/5\}$, then we have $\mathrm{M}(n,4) \le \lfloor (n-1-|\Delta(x_{n/4})|-|\Delta(x_{n/5})|)/6\rfloor + 2 = \lfloor (n+4)/6\rfloor$. The other cases are checked similarly.

Example 2.2. For $n = 21$ one has $\mathrm{M}(n,4) \le 3$ by Lemma 2.1. The difference sets of the equidifference codewords x_1, x_4, and x_5 are given by $\Delta(x_1) = \{1,2,3,20,19,18\}$, $\Delta(x_4) = \{4,8,12,17,13,9\}$, and $\Delta(x_5) = \{5,10,15,16,11,6\}$, respectively. Thus, we have $\{x_1,x_4,x_5\} \in \mathrm{CAC}^{\mathrm{e}}(21,4)$ and $\mathrm{M}(n,4) = \mathrm{M}^{\mathrm{e}}(n,4) = 3$.

In the case of $k = 5$, for exceptional codewords we have

$$|\Delta(x)| = \begin{cases} 4 \;\text{ iff }\; x = \{0, n/5, 2n/5, 3n/5, 4n/5\}, \\ 5 \;\text{ iff }\; x = \{0, n/6, n/3, n/2, 2n/3\}, \\ 6 \;\text{ iff }\; x = \{0, n/7, 2n/7, 3n/7, 4n/7\},\ x = \{0, n/7, 2n/7, 3n/7, 5n/7\}, \\ \qquad \text{or } x = \{0, n/7, 2n/7, 4n/7, 5n/7\}, \\ 7 \;\text{ iff }\; x = \{0, n/8, n/4, 3n/8, n/2\},\ x = \{0, n/8, 2n/8, 3n/8, 5n/8\}, \\ \qquad x = \{0, n/8, n/4, n/2, 5n/8\},\ x = \{0, n/8, n/4, n/2, 3n/4\}, \\ \qquad \text{or } x = \{0, n/8, 3n/8, n/2, 3n/4\}, \end{cases}$$

and obtain the following upper bound on code size, similar to the case $k = 4$.

LEMMA 2.3. *Let $n = 2^r3^s5^t7^um$, where m is not divisible by $2, 3, 5$, and 7. Then it holds that*

$$\mathrm{M}(n,5) \le \begin{cases} \lfloor n/8\rfloor & \textit{if } r \ge 3, s = t = u = 0, \\ \lfloor (n+1)/8\rfloor & \textit{if } s \ge 0, u \ge 1, r = t = 0, \textit{ or } 1 \le r \le 2, u \ge 1, s = t = 0, \\ \lfloor (n+2)/8\rfloor & \textit{if } r, s \ge 1, t = u = 0, \textit{ or } r \ge 3, u = 1, s = t = 0, \\ \lfloor (n+3)/8\rfloor & \textit{if } s \ge 0, t \ge 1, r = u = 0, \textit{ or } 1 \le r \le 2, t \ge 1, s = u = 0, \\ \lfloor (n+4)/8\rfloor & \textit{if } r, s \ge 1, t = 0, u \ge 1, \textit{ or } r \ge 3, t = 1, s = u = 0, \\ \lfloor (n+5)/8\rfloor & \textit{if } t, u \ge 1, r = 0, s \ge 0, \textit{ or } t, u \ge 1, 1 \le r \le 2, s = 0, \\ \lfloor (n+6)/8\rfloor & \textit{if } r, s, t \ge 1, u = 0, \textit{ or } t, u \ge 1, r \ge 3, s = 0, \\ \lfloor (n+8)/8\rfloor & \textit{if } r, s, t, u \ge 1, \\ \lfloor (n-1)/8\rfloor & \textit{if } s \ge 0, r = t = u = 0, \textit{ or } 0 \le r \le 2, s = t = u = 0. \end{cases}$$

Our aim is to give an explicit construction of codes $C \in \mathrm{CAC}^{\mathrm{e}}(n,k)$ for certain parameters n such that $|C|$ attains the upper bound given in Lemmas 2.1 and 2.3 implying $\mathrm{M}(n,k) = \mathrm{M}^{\mathrm{e}}(n,k) = |C|$. However, note that the upper bounds on $\mathrm{M}(n,k)$ for general weight k are not known.

3. Direct constructions of CACs from finite fields. In the rest of this paper, we use the following notation. Given a prime p, a primitive element $\alpha \in \mathbb{Z}_p$ and some divisor $e|(p-1)$, let $\gamma = \alpha^e$ and denote the multiplicative subgroup with generator γ by $\langle\gamma\rangle$. The cosets $H_j^e(p) = \alpha^j\langle\gamma\rangle$, $0 \le j < e$, are called the *cyclotomic cosets* of $\mathbb{Z}_p$ of index e, denoted by $\mathcal{H}^e(p)$. Given a list $(i_1, i_2, \ldots, i_e)$ of elements in $\mathbb{Z}_p^{\times}$, if each coset $H_j^e(p)$, $0 \le j < e$, contains exactly one element of the list, then we say that the list forms a *system of distinct representative* of $\mathcal{H}^e(p)$, denoted by $\mathrm{SDR}(\mathcal{H}^e(p))$. Let ζ_e be a primitive eth root of unity. We denote the eth power residue symbol in $\mathbb{Q}(\zeta_e)$ by $(\frac{\mathfrak{a}}{\mathfrak{p}})_e$, where $\mathfrak{p}$ is a prime ideal in $\mathbb{Q}(\zeta_e)$ lying over (p) and $\mathfrak{a}$ is an ideal in $\mathbb{Q}(\zeta_e)$ prime to $\mathfrak{p}$. (See [7, 6] for the definition and basic properties.) Furthermore, if the integer ring of $\mathbb{Q}(\zeta_e)$ is a principal ideal ring, we may denote an ideal $\mathfrak{p}$ in $\mathbb{Q}(\zeta_e)$ by an algebraic number π generating $\mathfrak{p}$.

We consider a code $C \in \mathrm{CAC}^{\mathrm{e}}(n,k)$ of the form $C = \{x_{i_1}, x_{i_2}, \dots, x_{i_m}\}$ with m equidifference codewords x_{i_j}. To ease the notation, we will use the concept of difference lists as defined, e.g., in [1, 20]. In this notation, the union of all differences $\Delta(x_{i_j})$ can be written as the following product of lists:

$$\Delta(C) = \bigcup_{j=1}^{m} \Delta(x_{i_j}) = (i_1, i_2, \dots, i_m) \cdot (1, 2, \dots, k-1, -1, -2, \dots, -(k-1)).$$

If $n = p = 2(k-1)m+1$ is a prime, we have $|\Delta(x_{i_j})| = 2(k-1)$ and the list $\Delta(C)$ must cover each element of $\mathbb{Z}_p^\times$ exactly once in order that $|C| = m$. The following theorem gives a sufficient condition to construct such equidifference CACs, and the technique is similar to that of construction for difference families; see, e.g., [2, 3, 9, 20].

THEOREM 3.1. *Let $n = p = 2(k-1)m + 1$ be a prime such that $(1, 2, \dots, k-1)$ forms an $\mathrm{SDR}(\mathcal{H}^{k-1}(p))$. Then there exists a code $C \in \mathrm{CAC}^{\mathrm{e}}(n = p, k)$ with $|C| = \mathrm{M}^{\mathrm{e}}(n,k) = m$.*

Proof. Let p satisfy the condition of the theorem, and let $\gamma = \alpha^{k-1}$ for a primitive element $\alpha \in \mathbb{Z}_p$. Since $(p-1)/2 = (k-1)m$ is a multiple of $k-1$, we have $-1 = \alpha^{(k-1)m} = \gamma^m \in H_0^{k-1}(p)$. Let $\Gamma(C) = \{1, \gamma, \dots, \gamma^{m-1}\}$; then the list of all differences of C is given by

$$\begin{aligned}
\Delta(C) &= (1, \gamma, \dots, \gamma^{m-1}) \cdot (1, 2, \dots, k-1, -1, -2, \dots, -(k-1)) \\
&= (1, \gamma, \dots, \gamma^{m-1}) \cdot (1, \gamma^m) \cdot (1, 2, \dots, k-1) \\
&= (1, \gamma, \dots, \gamma^{2m-1}) \cdot (1, 2, \dots, k-1) \\
&= H_0^{k-1}(p) \cdot (1, 2, \dots, k-1) \\
&= \mathbb{Z}_p^\times.
\end{aligned}$$

In other words, all elements of $\mathbb{Z}_p^\times$ appear exactly once as differences in $\Delta(C)$, which proves the theorem. □

Note that equidifference CACs of $k = 4$ and 5 constructed by this theorem are optimal by Lemmas 2.1 and 2.3 since $n = p$ is a prime.

Example 3.2. Let $p = 37$ and $k = 4$; then $\alpha = 2 \in \mathbb{Z}_{37}$ is a primitive element. Since $1 \in H^3(p)$, $2 = \alpha \in H_1^3(p)$, and $3 = \alpha^{26} \in H_2^3(p)$, $(1, 2, 3)$ forms an $\mathrm{SDR}(\mathcal{H}^3(p))$, and p satisfies the condition of Theorem 3.1. Let $\gamma = \alpha^3 = 8$; then $(1, \gamma, \dots, \gamma^5) = (1, 8, 27, 31, 26, 23)$ defines a list of generators for an optimal code $C \in \mathrm{CAC}^{\mathrm{e}}(37, 4)$ with $|C| = \mathrm{M}(37, 4) = 6$.

Now the statements of Theorem 3.1 can be expressed in another way by the following lemma.

LEMMA 3.3. *Let p be a rational prime and e a rational integer prime to p. Then* (i) *a list $(i_1, i_2, \dots, i_e)$ of $\mathbb{Z}_p^\times$ forms an $\mathrm{SDR}(\mathcal{H}^e(p))$ iff* (ii) *$\left(\frac{i_j}{\mathfrak{p}}\right)_e$, $1 \le j \le e$, are distinct from each other, where $\mathfrak{p}$ is a prime ideal in $\mathbb{Q}(\zeta_e)$ lying over (p).*

Proof. Let i be a rational integer prime to p and α a primitive element of $\mathbb{Z}_p$. We define $x_i, y_i \in \mathbb{Z}$ by

$$i \equiv \alpha^{x_i} \pmod{p} \text{ and } \left(\frac{i}{\mathfrak{p}}\right)_e = \zeta_e^{y_i}.$$

We note that x_i and y_i are uniquely determined by i modulo $p-1$ and e, respectively. By the definition of the eth power residue symbol, that is, $\left(\frac{i}{\mathfrak{p}}\right)_e \equiv i^{\frac{N\mathfrak{p}-1}{e}} \pmod{\mathfrak{p}}$, we have

$$\zeta_e^{y_i} \equiv i^{\frac{N_{\mathfrak{p}}-1}{e}} \equiv \alpha^{x_i \frac{N_{\mathfrak{p}}-1}{e}} \mod \mathfrak{p}),$$

where $N_{\mathfrak{p}}$ is the norm of $\mathfrak{p}$. In particular, if $i = \alpha$, then $x_\alpha \equiv 1 \pmod{p-1}$ and $\zeta_e^{y_\alpha} \equiv \alpha^{\frac{N_{\mathfrak{p}}-1}{e}} \pmod{\mathfrak{p}}$. Hence we have

$$\zeta_e^{y_i} \equiv \zeta_e^{y_\alpha x_i} \pmod{\mathfrak{p}}. \tag{3.1}$$

Since p and e are relatively prime, the equality holds for the congruence (3.1). Hence we have $y_i \equiv y_\alpha x_i \pmod{e}$.

((i)⇒(ii)) If $(i_1, i_2, \ldots, i_e)$ forms an SDR($\mathcal{H}^e(p)$), then x_{i_j}, $1 \le j \le e$, are distinct from each other modulo e and

$$y_{i_j} \equiv y_\alpha x_{i_j} \pmod{e} \tag{3.2}$$

holds by the above argument. Furthermore, we have $N_{\mathfrak{p}} = p$ and $\zeta_e^{y_\alpha} \equiv \alpha^{\frac{p-1}{e}}(\mathfrak{p})$ since obviously $p \equiv 1 \pmod{e}$ by the definition of $\mathcal{H}^e(p)$. This implies $(y_\alpha, e) = 1$. Therefore, y_{i_j}'s for $1 \le j \le e$ are distinct from each other modulo e, i.e., $\left(\frac{i_j}{\mathfrak{p}}\right)_e$'s for $1 \le j \le e$ are distinct from each other.

((ii)⇒(i)) If $\left(\frac{i_j}{\mathfrak{p}}\right)_e$'s for $1 \le j \le e$ are distinct from each other, then we have $(y_\alpha, e) = 1$ since (3.2) holds for any i_j , $1 \le j \le e$, and y_{i_j} are distinct from each other modulo e. This implies that x_{i_j}'s for $1 \le j \le e$ are distinct from each other modulo e, i.e., $(i_1, i_2, \ldots, i_e)$ forms an SDR($\mathcal{H}^e(p)$). □

In the case of $k = 4$, we can obtain the following by using Lemma 3.3.

COROLLARY 3.4. *Let $p = 6m + 1$ be a prime and $\pi = a + b\zeta_3 \in \mathbb{Z}[\zeta_3]$ be a prime element such that $p = \pi\bar{\pi}$ satisfying*

$$\begin{cases} a \equiv 2 \pmod{6} \\ b \equiv 3 \pmod{18} \end{cases} \text{or} \begin{cases} a \equiv 5 \pmod{6}, \\ b \equiv 15 \pmod{18}, \end{cases}$$

where $\bar{\pi}$ means the complex conjugate of π. Then there exists an optimal code $C \in \mathrm{CAC^e}(n = p, 4)$ with $|C| = \mathrm{M}(n, 4) = \mathrm{M^e}(n, 4) = m$.

Proof. By Lemma 3.3, it is enough to show that $\left(\frac{i}{\pi}\right)_3$'s for $1 \le i \le 3$ are distinct from each other iff π satisfies the condition of the corollary. Without loss of generality, we can assume that $a \equiv 2 \pmod{3}$ and $b \equiv 0 \pmod{3}$ for a prime element $\pi = a + b\zeta_3 \in \mathbb{Z}[\zeta_3]$ satisfying $p = \pi\bar{\pi}$. It is obvious that $\left(\frac{1}{\pi}\right)_3 = 1$. By the cubic reciprocity law, we have

$$\left(\frac{2}{\pi}\right)_3 \equiv \begin{cases} 1 & \text{if } (a, b) \equiv (1, 0) \pmod{(2, 2)}, \\ \zeta_3 & \text{if } (a, b) \equiv (0, 1) \pmod{(2, 2)}, \\ \zeta_3^2 & \text{if } (a, b) \equiv (1, 1) \pmod{(2, 2)}, \end{cases}$$

since 2 is also a prime element of $\mathbb{Z}[\zeta_3]$ (see [7, 6]). Also we have

$$\left(\frac{3}{\pi}\right)_3 \equiv \zeta_3^{\frac{ab}{3}} \pmod{\pi}$$

by using $3 = -\zeta_3^2(1 - \zeta_3^2)^2$ and the supplement law of cubic reciprocity. Then one can readily check that $\left(\frac{2}{\pi}\right)_3 = \zeta_3$ and $\left(\frac{3}{\pi}\right)_3 = \zeta_3^2$ iff $(a, b) \equiv (2, 3) \pmod{(6, 18)}$, and $\left(\frac{2}{\pi}\right)_3 = \zeta_3^2$ and $\left(\frac{3}{\pi}\right)_3 = \zeta_3$ iff $(a, b) \equiv (5, 15) \pmod{(6, 18)}$. □

Let K be a Galois extension of an algebraic number field F and C the conjugate class of $\sigma \in G = \mathrm{Gal}(K/F)$, i.e., $C = \{\gamma^{-1}\sigma\gamma \mid \gamma \in G\}$. We define a set M_σ of prime ideals in F for a fixed $\sigma \in G$ as follows:

$$M_\sigma = \{\mathfrak{P} \cap F \mid \mathfrak{P} \text{ is a prime ideal in } K \text{ such that } \sigma_{\mathfrak{P}} = \sigma\},$$

where $\sigma_{\mathfrak{P}}$ is a Frobenius substitution with respect to $\mathfrak{P}$ in K/F. If K/F is an abelian extension, then $\sigma_{\mathfrak{P}}$ depends on only the prime ideal $\mathfrak{p}$ of F lying under $\mathfrak{P}$. So $\sigma_{\mathfrak{P}}$ may be denoted by the Artin symbol $\left(\frac{K/F}{\mathfrak{p}}\right)$.

By utilizing the following proposition, we can show that primes satisfying the condition of Corollary 3.4 exist infinitely many. The proposition is well known as Chebotarëv's density theorem [17].

PROPOSITION 3.5. *The Kronecker density $\delta(M_\sigma)$ of M_σ is equal to $\frac{|C|}{|G|}$, i.e.,*

$$\delta(M_\sigma) = \lim_{s\to 1+0} \sum_{\mathfrak{p}\in M_\sigma} \frac{1}{(N_{\mathfrak{p}})^s} \Big/ \log \frac{1}{s-1} = \frac{|C|}{|G|}.$$

If the extension K/F is also abelian, then there exists infinitely many prime ideals $\mathfrak{p}$ in F such that $\left(\frac{K/F}{\mathfrak{p}}\right) = \sigma$ for each $\sigma \in \mathrm{Gal}(K/F)$, and the density of the set of all those prime ideals is equal to $\frac{1}{[K:F]}$, where $\left(\frac{K/F}{\mathfrak{p}}\right)$ is the Artin symbol.

Then the following corollary can be shown by noting that

$$(3.3) \qquad \left(\frac{\alpha}{\mathfrak{p}}\right)_k = 1 \Longleftrightarrow \left(\frac{\mathbb{Q}(\zeta_k, \sqrt[k]{\alpha})/\mathbb{Q}(\zeta_k)}{\mathfrak{p}}\right) = 1.$$

COROLLARY 3.6. *The Kronecker density of the set of all primes satisfying the condition of Corollary* 3.4 *is equal to $\frac{1}{9} = 0.11\ldots$, and there exist infinitely many of those primes.*

Proof. By Lemma 3.3, $(1, 2, 3)$ forms an $\mathrm{SDR}(\mathcal{H}^3(p))$ iff

$$(3.4) \qquad \left(\frac{6}{\pi}\right)_3 = 1 \text{ and } \left(\frac{2}{\pi}\right)_3 \neq 1,$$

where $\mathfrak{p} = (\pi)$ is a prime ideal in $\mathbb{Q}(\zeta_3)$ lying over (p). Let $\mathfrak{P}$ be a prime ideal in $\mathbb{Q}(\zeta_3, \sqrt[3]{6})$ lying over (p) and let

$$\sigma = \left(\frac{\mathbb{Q}(\zeta_3, \sqrt[3]{2}, \sqrt[3]{6})/\mathbb{Q}(\zeta_3, \sqrt[3]{6})}{\mathfrak{P}}\right).$$

Then, by (3.3), a necessary and sufficient condition for $\left(\frac{2}{\pi}\right)_3 \neq 1$ under $\left(\frac{6}{\pi}\right)_3 = 1$ is $\sigma \neq 1$. Now we apply Proposition 3.5 to the case of $K = \mathbb{Q}(\zeta_3, \sqrt[3]{2}, \sqrt[3]{6})$, $F = \mathbb{Q}(\zeta_3, \sqrt[3]{6})$, and $\sigma = 1$. Since $\mathrm{Gal}(\mathbb{Q}(\zeta_3, \sqrt[3]{2}, \sqrt[3]{6})/\mathbb{Q}(\zeta_3, \sqrt[3]{6}))$ is cyclic, the density of $\{\mathfrak{P}\}$ in $\mathbb{Q}(\zeta_3, \sqrt[3]{6})$ for $\sigma \neq 1$ is equal to

$$1 - \delta(M_1) = 1 - \frac{1}{[\mathbb{Q}(\zeta_3, \sqrt[3]{2}, \sqrt[3]{6}) : \mathbb{Q}(\zeta_3, \sqrt[3]{6})]} = \frac{2}{3}.$$

Furthermore, by a similar discussion, the density of $\{\mathfrak{p}\}$ in $\mathbb{Q}(\zeta_3)$ is equal to

$$\frac{2}{3} \frac{1}{[\mathbb{Q}(\zeta_3, \sqrt[3]{6}) : \mathbb{Q}(\zeta_3)]} = \frac{2}{9}.$$

It follows that the Kronecker density of the set of all those primes is equal to

$$\frac{2}{9}\frac{1}{[\mathbb{Q}(\zeta_3):\mathbb{Q}]}=\frac{1}{9}. \qquad \square$$

In fact, by computer search, the frequency ratio of those primes in the first $1,000$ primes is equal to $\frac{110}{1000} \fallingdotseq \frac{1}{9}$. Table 7.1 in section 7 shows such 110 primes.

Note that when $k = 5$ and $p = 8m + 1$ is a prime, $(1, 2, 3, 4)$ does not form an $\mathrm{SDR}(\mathcal{H}^4(p))$. The fact is easily seen. Since 2 is a square in $\mathbb{Z}_p$, we can obtain $2 \in H_0^4(p) \cup H_2^4(p)$. This implies that $4 = 2^2 \in H_0^4(p)$. Since we also have $1 \in H_0^4(p)$, $(1, 2, 3, 4)$ cannot form an $\mathrm{SDR}(\mathcal{H}^4(p))$.

Now, we give a further direct construction.

THEOREM 3.7. *Let $e \geq 1$ and $s > 1$ be integers, and let $p = 2em + 1$ be a prime such that each of $(i - es, i - (e - 1)s, \ldots, i + (e - 1)s)$, $1 \leq i \leq s - 1$, and $(\pm s, \pm 2s, \ldots, \pm es)$ forms an $\mathrm{SDR}(\mathcal{H}^{2e}(p))$. Then there exists a code $C \in \mathrm{CAC^e}(n = sp, k = es + 1)$ with $|C| = \mathrm{M^e}(n, k) = m$, which satisfies $\mathbb{Z}_n \setminus \Delta(C) = p\mathbb{Z}_n$.*

Proof. For e, s, and p satisfying the condition of the theorem, let $\gamma = \alpha^{2e}$ for a primitive element $\alpha \in \mathbb{Z}_p$. Note that we can assume $p > s$. In fact, if $p \leq s$, then we have

$$0 \in \bigcup_{1\leq i\leq s-1} \{i - es, i - (e - 1)s, \ldots, i + (e - 1)s\},$$

i.e., there is a coset of $\mathcal{H}^{2e}(p)$ which contains the element 0. Therefore, since p is a prime and $p > s$, $\mathbb{Z}_s \times \mathbb{Z}_p$ can be identified with $\mathbb{Z}_{sp}$. Let $\Gamma(C) = \{1\} \times H_0^{2e}(p)$ over $\mathbb{Z}_s \times \mathbb{Z}_p$. The differences arose from each codeword $x_{(1,\gamma^j)}$ with generator $(1, \gamma^j)$ are

$$\Delta^i(x_{(1,\gamma^j)}) = \begin{cases} \{0\} \times \gamma^j \cdot \{\pm s, \pm 2s, \ldots, \pm es\}, & i = 0, \\ \{i\} \times \gamma^j \cdot \{i - es, i - (e - 1)s, \ldots, i + (e - 1)s\}, & 1 \leq i \leq s - 1, \end{cases}$$

where $\Delta^i(x_{(1,\gamma^j)})$ is the set of differences of the form $(i, -)$ that arose from $x_{(1,\gamma^j)}$. Then the list of all differences of C is given by

$$\begin{aligned} \Delta(C) &= (\{1\} \times H_0^{2e}(p)) \cdot (1, 2, \ldots, k - 1, -1, -2, \ldots, -(k - 1)) \\ &= \left(\bigcup_{1\leq i\leq s-1} (\{i\} \times (H_0^{2e}(p)) \cdot (i - es, i - (e - 1)s, \ldots, i + (e - 1)s)) \right) \\ &\quad \cup \Big(\{0\} \times (H_0^{2e}(p)) \cdot (\pm s, \pm 2s, \ldots, \pm es) \Big) \\ &= \bigcup_{0\leq i\leq s-1} (\{i\} \times \mathbb{Z}_p^\times) = \mathbb{Z}_s \times (\mathbb{Z}_p^\times). \end{aligned}$$

In other words, all elements of $\mathbb{Z}_s \times (\mathbb{Z}_p^\times)$ appear exactly once as differences in $\Delta(C)$, and $(\mathbb{Z}_s \times \mathbb{Z}_p) \setminus \Delta(C) = \mathbb{Z}_s \times \{0\} \simeq p\mathbb{Z}_{sp}$ holds, which proves the theorem. $\square$

Note that equidifference CACs of $k = 4$ and 5 constructed by this theorem are optimal by Lemmas 2.1 and 2.3. In the case of $k = 4$ and 5, the statements of Theorem 3.7 can be expressed in another way by using quadratic residue. In the case of $e = 1$ and $s = 3$, we obtain the following infinite series of an optimal CAC for $k = 4$.

COROLLARY 3.8. *Let $p = 2m + 1$ be a prime such that $p \equiv 7 \pmod 8$. Then there exists an optimal code $C \in \mathrm{CAC}^{\mathrm{e}}(n = 3p, 4)$ with $|C| = \mathrm{M}(n, 4) = \mathrm{M}^{\mathrm{e}}(n, 4) = m$, which satisfies $\mathbb{Z}_n \setminus \Delta(C) = p\mathbb{Z}_n$.*

Proof. By Theorem 3.7 and Lemma 3.3, it is enough to show that $\left(\frac{i}{p}\right)_2$ are distinct for each of $i \in \{1, -1\}$ and $i \in \{1, -2\}$ iff $p \equiv 7 \pmod 8$. Obviously $\left(\frac{1}{p}\right)_2 = 1$, and $\left(\frac{-1}{p}\right)_2 = -1$ iff $p \equiv 3 \pmod 4$. By the supplement of quadratic reciprocity,

$$\left(\frac{2}{p}\right)_2 = \begin{cases} 1 & \text{iff } p \equiv 1, 7 \pmod 8, \\ -1 & \text{iff } p \equiv 3, 5 \pmod 8, \end{cases} \tag{3.5}$$

holds. Hence, $\left(\frac{-1}{p}\right)_2 = \left(\frac{-2}{p}\right)_2 = -1$ iff $p \equiv 7 \pmod 8$. Thus each of $\{1, -1\}$ and $\{1, -2\}$ forms an $\mathrm{SDR}(\mathcal{H}^2(p))$ iff $p \equiv 7 \pmod 8$. □

Example 3.9. Let $p = 7$, $s = 3$, and $e = 1$. Note that 3 is the primitive elements of $\mathbb{Z}_7$ and $H_0^2(7) = \{1, 2, 4\}$. Then, $((1, 1), (1, 2), (1, 4))$ over $\mathbb{Z}_3 \times \mathbb{Z}_7$ (or $(1, 16, 4)$ over $\mathbb{Z}_{21}$) defines a list of generators for an optimal code $C \in \mathrm{CAC}^{\mathrm{e}}(21, 4)$ with $|C| = \mathrm{M}(21, 4) = 3$.

In the case of $e = 2$ and $s = 2$, we can obtain an infinite series of optimal CACs for $k = 5$ as follows.

COROLLARY 3.10. *Let $p = 4m + 1$ be a prime such that $p \equiv 5 \pmod{24}$. Then there exists an optimal code $C \in \mathrm{CAC}^{\mathrm{e}}(n = 2p, 5)$ with $|C| = \mathrm{M}(n, 5) = \mathrm{M}^{\mathrm{e}}(n, 5) = m$, which satisfies $\mathbb{Z}_n \setminus \Delta(C) = p\mathbb{Z}_n$.*

Proof. By Theorem 3.7, we show that each of $\{2, 4, -2, -4\}$ and $\{1, 3, -1, -3\}$ forms an $\mathrm{SDR}(\mathcal{H}^4(p))$ iff $p \equiv 5 \pmod{24}$. Since $-1 \in H_2^4(p)$ iff $p \equiv 5 \pmod 8$, it is enough to show that each of $\{2, 4\}$ and $\{1, 3\}$ forms an $\mathrm{SDR}(\mathcal{H}^2(p))$, i.e, $\left(\frac{i}{p}\right)_2$ are distinct for each of $i \in \{2, 4\}$ and $i \in \{1, 3\}$ iff $p \equiv 5 \pmod{24}$. Obviously $\left(\frac{1}{p}\right)_2 = \left(\frac{4}{p}\right)_2 = 1$ holds. By (3.5), $\left(\frac{2}{p}\right)_2 = -1$ iff $p \equiv 3, 5 \pmod 8$. Furthermore, by quadratic reciprocity, we have

$$\left(\frac{3}{p}\right)_2 = \begin{cases} 1 & \text{iff } p \equiv 1, 11 \pmod{12}, \\ -1 & \text{iff } p \equiv 5, 7 \pmod{12}. \end{cases} \tag{3.6}$$

Hence each of $\{2, 4\}$ and $\{1, 3\}$ forms an $\mathrm{SDR}(\mathcal{H}^2(p))$ iff $p \equiv 5 \pmod{24}$. □

In the case of $e = 1$ and $s = 4$, we obtain the following result.

COROLLARY 3.11. *Let $p = 2m + 1$ be a prime such that $p \equiv 11 \pmod{12}$. Then there exists an optimal $C \in \mathrm{CAC}^{\mathrm{e}}(n = 4p, 5)$ with $|C| = \mathrm{M}(n, 5) = \mathrm{M}^{\mathrm{e}}(n, 5) = m$, which satisfies $\mathbb{Z}_n \setminus \Delta(C) = p\mathbb{Z}_n$.*

Proof. By Theorem 3.7 and Lemma 3.3, it is enough to show that $\left(\frac{i}{p}\right)_2$ are distinct for each of $i \in \{1, -1\}$ and $i \in \{1, -3\}$ iff $p \equiv 11 \pmod{12}$. Obviously $\left(\frac{1}{p}\right)_2 = 1$, and $\left(\frac{-1}{p}\right)_2 = -1$ iff $p \equiv 3 \pmod 4$. Furthermore, $\left(\frac{3}{p}\right)_2 = 1$ iff $p \equiv 1, 11 \pmod{12}$ by (3.6). Thus each of $\{1, -1\}$ and $\{1, -3\}$ forms an $\mathrm{SDR}(\mathcal{H}^2(p))$ iff $p \equiv 11 \pmod{12}$. □

In the case of $k \geq 6$, we can obtain some infinite series of CACs by similar investigations; however, we cannot judge whether the resultant CACs are optimal or not.

Remark 3.12. We denote the set $\Delta_2(x) = \Delta(x) \cap \{1, 2, \ldots, \lceil n/2 \rceil\}$ for any codeword x of a code $C \in \mathrm{CAC}^{\mathrm{e}}(n, k)$. We note that in the case $n = 6m + 1$ and for an optimal code $C \in \mathrm{CAC}^{\mathrm{e}}(n, 4)$ with $|C| = m$, the sets $\Delta_2(x)$ for $x \in C$, form a partition of the set $\{1, 2, \ldots, 3m\} \subset \mathbb{Z}_n$. Then, it follows that the triples $\Delta_2(x)$ for $x \in C$, are a solution to the first Heffter difference problem [4]. In the case $n = 6m + 3$ and for an optimal equidifference code $C \in \mathrm{CAC^e}(n, 4)$ with $|C| = m$, the sets $\Delta_2(x)$ for $x \in C$, form a partition of the set $\{1, 2, \ldots, 3m\} \setminus \{2m+1\}$. Again, it follows that the triples $\Delta_2(x)$ for $x \in C$, are a solution to the second Heffter difference problem [4]. The notions of Heffter difference problems were introduced for generating Steiner triple systems.

4. Asymptotic behavior of maximum code size. In the beginning of this section, we introduce a general problem. For a given integer v and a collection $\mathcal{A}$ of unordered pairs of $\mathbb{Z}_v \setminus \{0\}$, if there exists an h-subset S_v, $h \le v$, of $\mathbb{Z}_v \setminus \{0\}$ such that

$$|S_v| = |xS_v| = |yS_v|, \ xS_v \cap yS_v = \emptyset, \text{ and } 0 \notin xS_v \cup yS_v$$

for every $\{x, y\} \in \mathcal{A}$, then S_v is called a *halving set* of size h for $\mathcal{A}$.

The following is a natural generalization of the case $(e, s) = (1, k-1)$ of Theorem 3.7.

LEMMA 4.1. *Let v be an integer such that $(v, k-1) = 1$ and let*

$$\mathcal{A}_k = \{\{k-1, -(k-1)\}\} \cup \{\{i, -(k-1-i)\} \,|\, 1 \le i \le k-2\}.$$

If there exists a halving set S_v of size h for $\mathcal{A}_k$, then there exists a code $C \in \mathrm{CAC^e}(n = (k-1)v, k)$ with $|C| = h$.

Proof. Since $(v, k-1) = 1$, $\mathbb{Z}_{k-1} \times \mathbb{Z}_v$ can be identified with $\mathbb{Z}_{(k-1)v}$. Let $\Gamma(C) = \{1\} \times S_v$ over $\mathbb{Z}_{k-1} \times \mathbb{Z}_v$; then it is easy to follow that $\Gamma(C)$ defines a list of generators of a code $C \in \mathrm{CAC^e}((k-1)v, k)$ with $|C| = h$, similar to the case $(e, s) = (1, k-1)$ of Theorem 3.7. □

We are interested in the asymptotic behavior of the size of a halving set. We use a graph theoretical approach to see the maximum size of a halving set for $\mathcal{A}_4$, in other words, the maximum size of codewords of $\mathrm{CAC^e}(n = 3v, 4)$ constructed in Lemma 4.1. The result in the following theorem implies $\mathrm{M}(n = 3v, 4) \sim \frac{n}{6}$ for sufficiently large v such that $(v, i) = 1$ for $i = 2$ and 3. Note that the condition $(v, i) = 1$ for $i = 2$ and 3 implies that $|S_v| = |xS_v| = |yS_v|$ for $\{x, y\} \in \mathcal{A}_4$. The similar technique in the proof was used by Levenshtein in [10].

Let v be an odd integer such that $(v, 3) = 1$. A graph $G(v)$ has a vertex set $V = \mathbb{Z}_v$ and an edge set E, where $\{a, b\} \in E$ when $a \equiv -2b \pmod{v}$, $b \equiv -2a \pmod{v}$, or $a \equiv -b \pmod{v}$. Then the degree of each vertex of $G(v)$ is exactly three and the connected component containing $a \in V$ of $G(v)$ is either $G_a^1(v) = (V_a^1, E_a^1)$ or $G_a^2(v) = (V_a^2, E_a^2)$ of Figure 4.1. Let

$$r_a(v) = \min\{r > 0 \,|\, ((-2)^r - 1)a \equiv 0 \ \text{ or } \ ((-2)^r + 1)a \equiv 0 \pmod{v}\}$$

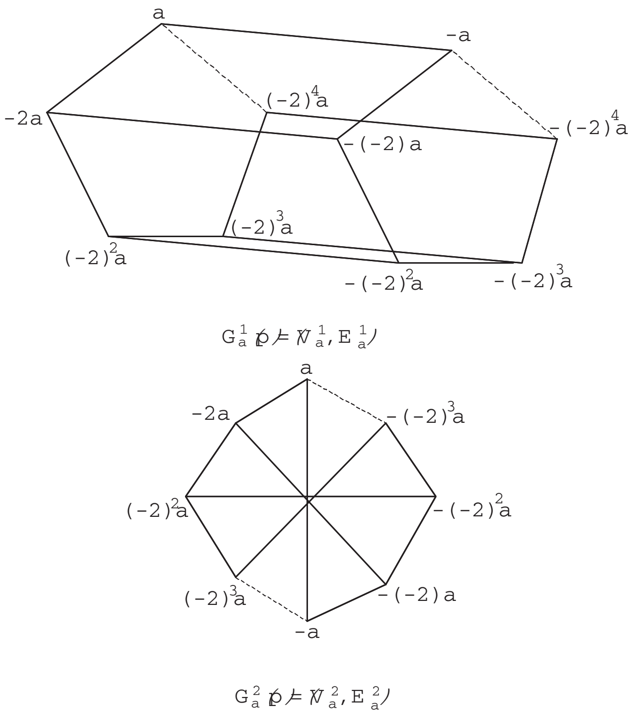
and

$$r(v) = \min\{r > 0 \,|\, (-2)^r - 1 \equiv 0 \ \text{ or } \ (-2)^r + 1 \equiv 0 \pmod{v}\}.$$

THEOREM 4.2. *Let v be an odd integer such that $(v, 3) = 1$. Then*

$$\mathrm{M}(3v, 4) \ge \mathrm{M^e}(3v, 4) \ge \frac{v}{2} + O\left(\frac{v}{\log_2 v}\right).$$

Proof. By the definition of $G(v)$, the maximum size of a halving set for $\mathcal{A}_4$ equals the maximum size of an independent set of $G(v)$. Now we construct a halving set S_v by choosing an independent set with maximum size from each connected component



$G_a^1(v) = (V_a^1, E_a^1)$

$G_a^2(v) = (V_a^2, E_a^2)$

FIG. 4.1. *The connected component of $G(v) = (V, E)$ containing a vertex $a \in V$ is either $G_a^1(v) = (V_a^1, E_a^1)$ or $G_a^2(v) = (V_a^2, E_a^2)$.*

of $G(v)$. Note that, by the definition of $r_a(v)$, $r_a(v) = |V_a^1|/2$, i.e., $(-2)^{r_a(v)}a' \equiv a' \pmod{v}$ for a, $a' \in V_a^1$, and $r_a(v) = |V_a^2|/2$, i.e., $(-2)^{r_a(v)}a' \equiv -a' \pmod{v}$ for $a, a' \in V_a^2$. If $|V_a^1| \equiv 0 \pmod 4$, then we can choose $|V_a^1|/2 = r_a(v)$ vertices as an independent set of $G_a^1(v)$, e.g., $\{a, 2a, 2^2a, \dots, 2^{r_a(v)-1}a\}$; otherwise, we can choose $|V_a^1|/2 - 1 = r_a(v) - 1$ vertices, e.g., $\{a, 2a, 2^2a, \dots, 2^{r_a(v)-2}a\}$. Furthermore, if $|V_a^2| \equiv 0 \pmod 4$, then we can choose $|V_a^2|/2 - 1 = r_a(v) - 1$ vertices as an independent set of $G_a^2(v)$, e.g., $\{a, 2a, 2^2a, \dots, 2^{r_a(v)-2}a\}$; otherwise, we can choose $|V_a^2|/2 = r_a(v)$ vertices, e.g., $\{a, 2a, 2^2a, \dots, 2^{r_a(v)-1}a\}$. In other words, we can choose $r_a(v)$ vertices as an independent set of $G_a^1(v)$ or $G_a^2(v)$ depending on whether $r_a(v)$ is even or odd iff $(2^{r_a(v)} - 1)a \equiv 0 \pmod 4$, and we can choose $r_a(v) - 1$ vertices as an independent set of $G_a^1(v)$ or $G_a^2(v)$ depending on whether $r_a(v)$ is odd or even iff $(2^{r_a(v)} + 1)a \equiv 0 \pmod{v}$.

Here, let $V(b) = \{a \in V \mid (a, v) = b\}$ for each divisor b, $1 \le b \le (v-1)/2$, of v and let $d = v/b$. For each integer h, $1 \le h < d/2$, such that $(h, d) = 1$, hb and $(d-h)b$ belong to $V(b)$. Therefore, $|V(b)| = \varphi(d)$, where $\varphi(d)$ is Euler's φ-function. By the definition of $r(d)$, all vertices of $V(b)$ are partitioned into some connected components with same size $2r(d)$. Hence, we have

$$\begin{aligned}|S_v| &= \sum_{1<d|v;\, 2^{r(d)}-1\equiv 0 \bmod d} \frac{\varphi(d)}{2r(d)} \cdot r(d) + \sum_{1<d|v;\, 2^{r(d)}+1\equiv 0 \bmod d} \frac{\varphi(d)}{2r(d)} \cdot (r(d) - 1)\\ &= \frac{1}{2}\sum_{1<d|v} \varphi(d) - \sum_{1<d|v;\, 2^{r(d)}+1\equiv 0 \bmod d} \frac{\varphi(d)}{2r(d)} \ge \frac{v-1}{2} - \sum_{1<d|v} \frac{\varphi(d)}{2r(d)}.\end{aligned}$$

By applying Levenshtein's results in [10], that is,

$$\sum_{1<d|v} \frac{\varphi(d)}{2r(d)} < \frac{2v}{\log_2 v} + v^{1/2}v^{\Theta(1)},$$

we obtain the desired assertion. □

We give a sufficient condition to construct optimal CACs of length $n = 3p$ and weight $k = 4$.

Corollary 4.3. *Let* $p = 2m+1$ (> 3) *be a prime such that* $p \equiv 3, 5 \pmod 8$ *and* 2 *is a primitive element of* $\mathbb{Z}_p$*. Then there exists an optimal code* $C \in \mathrm{CAC}^{\mathrm{e}}(n = 3p, 4)$ *with* $|C| = \mathrm{M}^{\mathrm{e}}(n, 4) = \mathrm{M}(n, 4) = m - 1$.

Remark 4.4. We prove the corollary by noting the following. Let n be an integer and x a codeword of a code $C \in \mathrm{CAC}(n, 4)$.

(i) In the case of $|\Delta(x)| = 6$, it can be tediously checked that $x = \{0, \frac{n}{7}, \frac{2n}{7}, \frac{4n}{7}\}$, $x = \{0, \frac{n}{7}, \frac{2n}{7}, \frac{5n}{7}\}$, or x is an equidifference codeword. In the first and second cases, x can be replaced by the equidifference codeword $x_{n/7}$ since $\Delta(x_{n/7}) = \Delta(x)$.

(ii) The number $|\Delta(x)|$ is odd iff $2|n$ and $n/2 \in \Delta(x)$. In particular, if $2|n$, the code C can contain only one of the codewords such that $|\Delta(x)|$ is odd.

Proof. Assume that p satisfies the condition of the corollary. Then $p(= v)$ also satisfies the condition of Theorem 4.2, and note that $2 \in H_1^2(p)$ iff $p \equiv 3, 5 \pmod 8$. Since 2 is a primitive element of $\mathbb{Z}_p$, $2^{\frac{p-1}{2}} + 1 \equiv 0 \pmod 8$ and $2^i + 1 \not\equiv 0 \pmod p$ for any i, $1 \leq i \leq \frac{p-1}{2} - 1$. Here, by Theorem 4.2, there exists a halving set S_p of maximum size

$$|S_p| = \frac{p-1}{2} - \sum_{1<d|p;\, 2^{r(d)}+1\equiv 0 \bmod d} \frac{\varphi(d)}{2r(d)} = \frac{p-1}{2} - \frac{\varphi(p)}{2r(p)} = m - 1. \tag{4.1}$$

This follows that there exists a code $C \in \mathrm{CAC}^{\mathrm{e}}(3p, 4)$ with $|C| = m - 1$, and hence it is enough to show that $\mathrm{M}(3p, 4) = m - 1$. By Lemma 2.1 $\mathrm{M}(3p, 4) \leq m$ holds, and suppose that there is a code $C^* \in \mathrm{CAC}(3p, 4)$ with $|C^*| = m$. Then, by the fact that n is odd and by (2.2) and Remark 4.4, for $p > 5$, C^* cannot contain any codeword x with $\Delta(x) = 3, 5$, or 7. Note that for $p = 5$ if a code $C \in \mathrm{CAC}(15, 4)$ contains a codeword x with $\Delta(x) = \{3, 6, 9, 12\}$, the code C cannot contain any other codeword since $\Delta(y) \cap \Delta(x) \neq \emptyset$ for any $y \in \mathcal{P}(15, 4)$, which implies that C^* cannot contain x with $\Delta(x) = \{3, 6, 9, 12\}$. It immediately follows that there are two possible cases.

(i) The code C^* contains m equidifference codewords with $|\Delta(x)| = 6$.

(ii) The code C^* contains one nonequidifference codeword x with $|\Delta(x)| = 8$ and $m - 1$ equidifference codewords with $|\Delta(x)| = 6$.

Now we derive a contradiction for each case.

Case (i). Each of the m equidifference codewords is generated by either of the generators $(0, a_0)$, $(1, a_1)$, or $(2, a_2)$ for $a_0, a_1, a_2 \in \mathbb{Z}_p^\times$ since $\mathbb{Z}_{3p} \simeq \mathbb{Z}_3 \times \mathbb{Z}_p$. Note that we can replace an equidifference codeword $x_{(2,a_2)}$ by $x_{(1,-a_2)}$ since $\Delta(x_{(2,a_2)}) = \Delta(x_{(1,-a_2)})$. If C^* has $\ell > 0$ codewords with generator $(0, a_0)$'s for some $a_0 \in \mathbb{Z}_p^\times$, then C^* can contain at most $(2m - 6\ell)/2$ equidifference codewords with generator $(1, a_1)$'s for some $a_1 \in \mathbb{Z}_p^\times$ since $|\Delta(x_{(0,a_0)}) \cap (\{0\} \times \mathbb{Z}_p^\times)| = 6$ and $|\Delta(x_{(1,a_1)}) \cap (\{0\} \times \mathbb{Z}_p^\times)| = 2$. Then the total number of codewords is at most $\ell + (2m - 6\ell)/2 = m - 2\ell < m - 1$, which contradicts the assumption. Hence, we can assume that C^* contains m equidifference codewords with generator $(1, a_1)$'s, $a_1 \in \mathbb{Z}_p^\times$, i.e., the maximum size of halving sets for $\mathcal{A}_4$ is equal to m, which also contradicts (4.1).

Case (ii). Similar to Case (i), we can assume that C^* contains $m-1$ equidifference codewords with generator $(1,a_1)$'s, $a_1 \in \mathbb{Z}_p^\times$. Let E be the set of such $m-1$ equidifference codewords contained in C^*, and let $A = ((\mathbb{Z}_3 \times \mathbb{Z}_p) \setminus \{(0,0)\}) \setminus \Delta(E)$. Then there exists an element $a \in \mathbb{Z}_p^\times$ such that $(1,a),(1,-a) \notin \Gamma(E)$ since $|E| = m-1$. In particular, we can assume $(1,a) \notin \Delta(E)$. In fact, if $(1,a),(1,-a) \in \Delta(E)$, there must be two generators $(1,b),(1,c) \in \Gamma(E)$ such that $(1,a) \equiv (1,-2b) \pmod{(3,p)}$ and $(1,-a) \equiv (1,-2c) \pmod{(3,p)}$, and then $b \equiv -c \pmod{p}$ holds, which implies that $\Delta(x_{(1,b)}) \cap \Delta(x_{(1,c)})) \neq \emptyset$. For such $a \in \mathbb{Z}_p^\times$, the set $\Delta(E)$ does not contain $(0,3a)$ since $(0,3a) \in \Delta(E)$ iff $(1,a) \in \Gamma(E)$ or $(1,-a) \in \Gamma(E)$. Furthermore, $(1,-2a) \in \Gamma(E)$ since $(1,(-2)^{-1}a),(1,a),(1,-a) \notin \Gamma(E)$ and $|E| = m-1$, which implies that $(1,2a) \notin \Gamma(E)$. Then, by using $(1,-a) \notin \Gamma(E)$ again, we have $(1,2a) \notin \Delta(E)$. Therefore, by $t \notin \Delta(E)$ iff $-t \notin \Delta(E)$ for any $t \in \mathbb{Z}_3 \times \mathbb{Z}_p$, we have

$$A = \{(1,0),(2,0),(0,3a),(0,-3a),(1,a),(2,-a),(1,2a),(2,-2a)\} \tag{4.2}$$

and x must cover the eight elements of A as differences. In order that $A = \Delta(x)$, for every $y \in A$ there must be at least one element of A, say $y' \in A$, such that $y + y' \in A$. However, by the fact that p is a prime and $a \in \mathbb{Z}_p^\times$, it is easily checked that such y' does not exist in A for any $y \in A \setminus \{(1,0),(2,0)\}$. Hence A cannot be the set of differences of x. □

Small primes $p < 1000$ satisfying the condition of Corollary 4.3 are listed below:

$$\begin{aligned} p \;=\;& 5, 11, 13, 19, 29, 37, 53, 59, 61, 67, 83, 101, 107, 131, 139, 149, 163, 173, 179, 181,\\ & 197, 211, 227, 269, 293, 317, 347, 349, 373, 379, 389, 419, 421, 443, 461, 467, 491,\\ & 509, 523, 541, 547, 557, 563, 587, 613, 619, 653, 659, 661, 677, 701, 709, 757, 773,\\ & 787, 797, 821, 827, 829, 853, 859, 877, 883, 907, 941, 947. \end{aligned}$$

5. Constructions of optimal CACs of length $n = 4p$ with weight $k = 4$.

In this section, we obtain some sufficient conditions in order to obtain optimal CACs of length $n = 4p$ and weight $k = 4$. The following construction is an application of halving sets. Let $\mathcal{A}'_k = \mathcal{A}_k \setminus \{\{k-1, -(k-1)\}\}$, where $\mathcal{A}_k$ was defined in section 4.

THEOREM 5.1. *Let v be an integer such that $(v,k) = 1$. If there exist a code $C \in \mathrm{CAC^e}(v,k)$ with $m_1 = |C|$ and a halving set S_v of size m_2 for $\mathcal{A}'_{k+1}$, then there exists a code $C' \in \mathrm{CAC^e}(n = kv, k)$ with $|C'| = m_1 + m_2 + 1$.*

Proof. Let v, k, C, and S_v satisfy the condition of the theorem. Since $(v,k) = 1$, $\mathbb{Z}_k \times \mathbb{Z}_v$ can be identified with $\mathbb{Z}_{kv}$. Let

$$\Gamma(C') = (\{0\} \times \Gamma(C)) \cup (\{1\} \times S_v) \cup \{(1,0)\}$$

over $\mathbb{Z}_k \times \mathbb{Z}_v$. Note that the differences arose from each codeword $x_{(1,a)}$ with generator $(1,a)$ for $a \in S_v$ are

$$\Delta^i(x_{(1,a)}) = \{(i, ia), (i, -(k-i)a)\}$$

for each i, $1 \le i \le k-1$. We show that $\Gamma(C')$ forms a set of generators of $C' \in \mathrm{CAC^e}(kv,k)$ with $|C'| = m_1 + m_2 + 1$. Obviously, $|\Gamma(C')| = m_1 + m_2 + 1$ holds. By the assumption $C \in \mathrm{CAC^e}(v,k)$ and the definition of S_v, we have

$$(\{0\} \times \Gamma(C)) \cdot (1, 2, \ldots, k-1, -1, -2, \ldots, -(k-1)) \subseteq \{0\} \times (\mathbb{Z}_v \setminus \{0\})$$

and

$$\begin{aligned}&(\{1\}\times S_v)\cdot(1,2,\dots,k-1,-1,-2,\dots,-(k-1))\\ &=\bigcup_{1\le i\le k-1}(\{i\}\times(S_v)\cdot(i,-(k-i)))\\ &\subseteq\bigcup_{1\le i\le k-1}(\{i\}\times(\mathbb{Z}_v\setminus\{0\}))=(\mathbb{Z}_k\setminus\{0\})\times(\mathbb{Z}_v\setminus\{0\}).\end{aligned}$$

Finally, the differences of the form $(\ell,0)$ occur only in the codeword $x_{(1,0)}$ with generator $(1,0)$. Thus, all elements of $(\mathbb{Z}_{k-1}\times\mathbb{Z}_v)\setminus\{(0,0)\}$ appear at most once as differences in $\Delta(C)$. □

Now, we give two sufficient conditions to construct optimal CACs of length $n=4p$ and weight $k=4$. We use the quartic residue to give the first sufficient condition. The following lemma is a preparation for the first assertion.

LEMMA 5.2. *Let p be a rational prime and ρ a prime element of $\mathbb{Z}[\zeta_4]$ lying over (p). Then $-1,-3\in H_2^4(p)$ iff $\left(\frac{-1}{\rho}\right)_4\equiv-1$ and $\left(\frac{-3}{\rho}\right)_4\equiv-1$.*

Proof. By the definition of the quartic residue symbol, we have

$$\left(\frac{-1}{\rho}\right)_4\equiv(-1)^{\frac{N_\rho-1}{4}}=\begin{cases}(-1)^{\frac{p-1}{4}} & \text{iff } p\equiv 1 \pmod 4,\\ (-1)^{\frac{p^2-1}{4}} & \text{iff } p\equiv 3 \pmod 4.\end{cases}$$

Then one can check that $\left(\frac{-1}{\rho}\right)_4\equiv-1$ iff $p\equiv5 \pmod 8$. On the other hand, it is obvious that $-1\in H_2^4(p)$ iff $p\equiv 5 \pmod 8$. We define $i\in\mathbb{Z}$ by $-3\equiv\alpha^i \pmod p$, where α is a primitive element of $\mathbb{Z}_p$. Then we have

$$\begin{aligned}\left(\frac{-3}{\rho}\right)_4&\equiv(-3)^{\frac{N_\rho-1}{4}}\equiv\alpha^{i\frac{N_\rho-1}{4}}=\alpha^{\frac{i}{2}\cdot\frac{p-1}{2}}\equiv(-1)^{\frac{i}{2}}\\ &\equiv\begin{cases}1 & \text{iff } i\equiv 0 \pmod 4,\\ \zeta_4 & \text{iff } i\equiv 1 \pmod 4,\\ -1 & \text{iff } i\equiv 2 \pmod 4,\\ -\zeta_4 & \text{iff } i\equiv 3 \pmod 4\end{cases}\end{aligned}$$

by using $N_\rho=p$ and $\alpha^{\frac{p-1}{2}}\equiv-1 \pmod\rho$. Hence $\left(\frac{-1}{\rho}\right)_4\equiv-1$ and $\left(\frac{-3}{\rho}\right)_4\equiv-1$ iff $-1,-3\in H_2^4(p)$. □

Now, we give the first sufficient condition.

COROLLARY 5.3. *Let $p=24m+13$ be a prime satisfying the condition of Corollary 3.4 and $\rho=a+b\zeta_4\in\mathbb{Z}[\zeta_4]$ a prime element such that $p=\rho\bar{\rho}$ satisfying*

$$\begin{cases}a\equiv 3 \pmod{12}\\ b\equiv 2 \pmod{12}\end{cases} \quad or \quad \begin{cases}a\equiv 3 \pmod{12},\\ b\equiv 10 \pmod{12}.\end{cases}$$

Then there exists an optimal code $C\in\mathrm{CAC}^{\mathrm{e}}(n=4p,4)$ with $|C|=\mathrm{M}^{\mathrm{e}}(n,4)=\mathrm{M}(n,4)=16m+9$.

Proof. By Lemma 5.2, $\left(\frac{-1}{\rho}\right)_4\equiv-1$ iff $p\equiv5 \pmod 8$. Without loss of generality, we can assume that $a\equiv 3 \pmod 4$ and $b\equiv 2 \pmod 4$ for a prime element $\rho=$

$a + b\zeta_4 \in \mathbb{Z}[\zeta_4]$ which satisfies $p = \rho\bar{\rho}$. By quartic reciprocity,

$$\left(\frac{-3}{\rho}\right)_4 \equiv \begin{cases} 1 & \text{if } (a,b) \equiv (\pm 1, 0) \pmod{(3,3)}, \\ \zeta_4 & \text{if } (a,b) \equiv (\pm 1, \mp 1) \pmod{(3,3)}. \\ -1 & \text{if } (a,b) \equiv (0, \pm 1) \pmod{(3,3)}, \\ -\zeta_4 & \text{if } (a,b) \equiv (\pm 1, \pm 1) \pmod{(3,3)}. \end{cases}$$

Hence, $-1, -3 \in H_2^4(p)$ iff $(a,b) \equiv (3,2)$ or $(3,10) \pmod{(12,12)}$. Then $S_p = H_0^4(p) \cup H_1^4(p)$ defines a halving set of size $12m+6$ for $\mathcal{A}_5'$. In fact, since $-S_p = -H_0^4(p) \cup -H_1^4(p) = H_2^4(p) \cup H_3^4(p)$ and $-3S_p = -3H_0^4(p) \cup -3H_1^4(p) = H_2^4(p) \cup H_3^4(p)$ hold, we have $S_p \cap -S_p = \emptyset$ and $S_p \cap -3S_p = \emptyset$. By combining Corollary 3.4 and Theorem 5.1, we have a code $C \in \mathrm{CAC^e}(4p, 4)$ with $|C| = 16m + 9$. By Lemma 2.1, we also have $\mathrm{M}(4p, 4) = 16m + 9$, i.e., the resultant CAC is optimal. □

By utilizing Proposition 3.5, we can show that the primes satisfying the conditions of Corollary 5.3 exist infinitely many as follows.

Corollary 5.4. *The Kronecker density of the set of all primes satisfying the condition of Corollary* 5.3 *is equal to* $\frac{1}{2^3 \cdot 3^2} = 0.0138\cdots$, *and there exist infinitely many of those primes.*

Proof. By Lemmas 3.3 and 5.2, $(1,2,3)$ forms an $\mathrm{SDR}(\mathcal{H}^3(p))$ and $-1, -3 \in H_2^4(p)$ iff

$$\left(\frac{6}{\pi}\right)_3 = 1, \quad \left(\frac{3}{\rho}\right)_4 = 1, \tag{5.1}$$

$$\left(\frac{2}{\pi}\right)_3 \neq 1, \text{ and } \left(\frac{-1}{\rho}\right)_4 = -1, \tag{5.2}$$

where $\mathfrak{p} = (\pi)$ is a prime ideal in $\mathbb{Q}(\zeta_3)$ lying over (p) and $\mathfrak{r} = (\rho)$ is a prime ideal in $\mathbb{Q}(\zeta_4)$ lying over (p). Let $\mathfrak{P}$ be a prime ideal in $\mathbb{Q}(\zeta_4, \sqrt[3]{6}, \sqrt[4]{3})$ lying over (p) and

$$\sigma = \left(\frac{\mathbb{Q}(\zeta_8, \sqrt[3]{6}, \sqrt[3]{2}, \sqrt[4]{3})/\mathbb{Q}(\zeta_4, \sqrt[3]{6}, \sqrt[4]{3})}{P}\right).$$

Note that $\zeta_3 \in \mathbb{Q}(\zeta_4, \sqrt[4]{3})$. Then a necessary and sufficient condition such that (5.2) holds under (5.1) is

$$\sigma(\zeta_8) \neq \zeta_8 \text{ and } \sigma(\sqrt[3]{2}) \neq \sqrt[3]{2}. \tag{5.3}$$

Hence the density of prime ideals $\mathfrak{P}$ satisfying (5.3) in $\mathbb{Q}(\zeta_4, \sqrt[3]{6}, \sqrt[4]{3})$ is equal to $\frac{1}{3}$ and the density of rational primes p satisfying the condition of the corollary is equal to $\frac{1}{2^3 \cdot 3^2}$. □

By our computer search, the frequency ratio of those primes in the first 3,000,000 primes is equal to $\frac{41684}{3{,}000{,}000} \fallingdotseq \frac{1}{2^3 \cdot 3^2}$.

Next, we give the second sufficient condition.

Corollary 5.5. *Let* $p = 12m + 7$ *be a prime satisfying the condition of Corollary* 3.4 *such that* 3 *is a primitive element of* $\mathbb{Z}_p$. *Then there exists an optimal code* $C \in \mathrm{CAC}(n = 4p, 4)$ *with* $|C| = \mathrm{M^e}(n, 4) = \mathrm{M}(n, 4) = 8m + 4$.

Proof. Let p satisfy the condition of the corollary, and let $\alpha = 3 \in \mathbb{Z}_p$ be a primitive element. Then we can take a halving set of size $6m + 2$ for $\mathcal{A}_5'$, e.g., $S_p = \{\alpha^0, \alpha^1, \ldots, \alpha^{6m+1}\}$. In fact, since

$$-S_p = \alpha^{\frac{p-1}{2}} \cdot S_p = \{\alpha^{6m+3}, \alpha^{6m+4}, \ldots, \alpha^{12m+4}\}$$

and

$$-3S_p = \alpha^{\frac{p+1}{2}} \cdot S_p = \{\alpha^{6m+4}, \alpha^{6m+5}, \ldots, \alpha^{12m+5}\},$$

we have $S_p \cap -S_p = \emptyset$ and $S_p \cap -3S_p = \emptyset$. Furthermore, since $-3 \in H_0^2(p)$ and $\frac{p-1}{2}$ is odd, the maximum size of halving sets for $\mathcal{A}'_5$ is exactly $6m+2$. Now, by combining Corollary 3.4 and Theorem 5.1, we obtain a code $C \in \mathrm{CAC}^{\mathrm{e}}(n = 4p, 4)$ with $|C| = \mathrm{M}^{\mathrm{e}}(n,4) = 8m+4$. Note that $p = 12m+7$ is a sufficient condition for $3 \in H_1^2(p)$. Hence, it is enough to show that $\mathrm{M}(n,4) = 8m+4$. By Lemma 2.1 $\mathrm{M}(n,4) \leq 8m+5$ holds, and suppose that there exists a code C^* with $|C^*| = 8m+5$. Then, by (2.2) and Remark 4.4, C^* can contain only one codeword x with $|\Delta(x)| = 3, 5$, and 7. On the other hand, since $n-1 = 6(8m+4)+3$ and $|C^*| = 8m+5$, C^* must contain a codeword with $|\Delta(x)| = 3$, which implies that the cases $|\Delta(x)| = 5$ and 7 are impossible. Therefore, we can assume that C^* contains the exceptional codeword $x_{(1,0)}$ and $8m+4$ equidifference codewords, which have generators of the form $(0,a_0)$, $(1,a_1)$, $(2,a_2)$, or $(3,a_3)$ for some $a_0,a_1,a_2,a_3 \in \mathbb{Z}_p^\times$, since $\mathbb{Z}_{4p} \simeq \mathbb{Z}_4 \times \mathbb{Z}_p$. Here, an equidifference codeword $x_{(3,a_3)}$ can be replaced by $x_{(1,-a_3)}$ since $\Delta(x_{(3,a_3)}) = \Delta(x_{(1,-a_3)})$. If C^* has $\ell > 0$ codewords with generator $(2,a_2)$'s for $a_2 \in \mathbb{Z}_p^\times$, then C^* contains at most $(12m+6-4\ell)/2$ equidifference codewords with generator $(1,a_1)$'s for $a_1 \in \mathbb{Z}_p^\times$, since $|\Delta(x_{(2,a_2)}) \cap (\{2\} \times \mathbb{Z}_p^\times)| = 4$ and $|\Delta(x_{(1,a_1)}) \cap (\{2\} \times \mathbb{Z}_p^\times)| = 2$. Furthermore, C^* contains at most $(12m+6-2\ell)/6$ equidifference codewords with generator $(0,a_0)$'s for $a_0 \in \mathbb{Z}_p^\times$, since $|\Delta(x_{(2,a_2)}) \cap (\{0\} \times \mathbb{Z}_p^\times)| = 2$ and $|\Delta(x_{(0,a_0)}) \cap (\{0\} \times \mathbb{Z}_p^\times)| = 6$. Then we have

$$|C^*| \leq \ell + (12m+6-4\ell)/2 + (12m+6-2\ell)/6 + 1 < 8m+5,$$

which contradicts the assumption $|C^*| = 8m+5$. Therefore, C^* contains no equidifference codewords with generator $(2,a_2)$'s for $a_2 \in \mathbb{Z}_p^\times$. Moreover, since the maximum number of codewords with generator $(0,a_0)$'s for $a_0 \in \mathbb{Z}_p^\times$, is equal to $2m+1$ by Corollary 3.4, in order that $|C^*| = 8m+5$, then the maximum number of codewords with generator $(1,a_1)$'s for $a_1 \in \mathbb{Z}_p^\times$, must be equal to $6m+3$; i.e., the maximum size of halving sets for $\mathcal{A}'_5$ must be equal to $6m+3$. However, this also contradicts our first argument. Thus $\mathrm{M}(n,4) = 8m+4$ holds. □

Small primes p satisfying the condition of Corollaries 5.3 and 5.5 are listed in Table 7.1.

Example 5.6. Let $p = 7$ and $k = 4$, then $4p = 28$. Note that 3 is a primitive element of $\mathbb{Z}_7$ and $S_p = \{1,3\}$ is a halving set of size 2 for $\mathcal{A}'_5$. Let $C \in \mathrm{CAC}^{\mathrm{e}}(7,4)$ which has one generator 1. Then, $((0,1),(1,1),(1,3),(1,0))$ over $\mathbb{Z}_4 \times \mathbb{Z}_7$ (or $(1,8,17,21)$ over $\mathbb{Z}_{28}$) defines a list of generators for a code $C' \in \mathrm{CAC}^{\mathrm{e}}(28,4)$ with $|C'| = \mathrm{M}(28,4) = 4$.

6. A recursive construction of equidifference CACs. In this section, we give a recursive construction of equidifference CACs.

THEOREM 6.1. *Let $k \geq 3$, and let n_1, n_2, and s be positive integers satisfying $s \mid n_1$ and $(n_2, \ell) = 1$ for each ℓ, $1 \leq \ell \leq k-1$. Let $C_1 \in \mathrm{CAC}^{\mathrm{e}}(n_1,k)$ with $t_1 = |C_1|$ nonexceptional codewords satisfying*

$$\mathbb{Z}_{n_1} \setminus \Delta(C_1) \supseteq (n_1/s)\ \ \mathbb{Z}_{n_1}, \tag{6.1}$$

and let $C_2 \in \mathrm{CAC}^{\mathrm{e}}(sn_2,k)$ with $t_2 = |C_2|$ codewords. Then there exists a code $C \in \mathrm{CAC}^{\mathrm{e}}(n_1 n_2, k)$ with $t = |C| = n_2 t_1 + t_2$.

Proof. Let

$$\Gamma_1 = \{i + jn_1 \mid i \in \Gamma(C_1), 0 \leq j \leq n_2 - 1\} \text{ and } \Gamma_2 = \{j(n_1/s) \mid j \in \Gamma(C_2)\},$$

where each element is reduced modulo n_1n_2. Then $\Gamma(C) = \Gamma_1 \cup \Gamma_2$ defines a code C consisting of $n_2t_1 + t_2$ equidifference codewords. We prove that the difference sets of any two codewords of C are disjoint. By (6.1) and the definition of Γ_1, it holds that

$$\mathbb{Z}_{n_1n_2} \setminus \bigcup_{\ell\in\pm\{1,2,\ldots,k-1\}} \ell \cdot \Gamma_1 \supseteq (n_1/s)\ \mathbb{Z}_{n_1n_2}.$$

Furthermore, since every element of Γ_2 is a multiple of (n_1/s), it is obvious that

$$\bigcup_{\ell\in\pm\{1,2,\ldots,k-1\}} \ell \cdot \Gamma_2 \subseteq (n_1/s)\ \mathbb{Z}_{n_1n_2}$$

holds. These imply that

$$\left(\bigcup_{\ell\in\pm\{1,2,\ldots,k-1\}} \ell \cdot \Gamma_1\right) \cap \left(\bigcup_{\ell\in\pm\{1,2,\ldots,k-1\}} \ell \cdot \Gamma_2\right) = \emptyset.$$

Now we show that the difference sets of any two codewords with generators from Γ_1 are disjoint. Assume $\ell(i+jn_1) \equiv \ell'(i'+j'n_1) \pmod{n_1n_2}$ for some $\ell, \ell' \in \pm\{1, 2, \ldots, k-1\}$ and two generators $i+jn_1$ and $i'+j'n_1$ from Γ_1. Then we need to show that $i = i'$ and $j = j'$. By the above assumption, since $(\ell i-\ell' i')+(\ell j-\ell' j')n_1 \equiv 0 \pmod{n_1n_2}$, we then have $\ell i \equiv \ell' i' \pmod{n_1}$. By the definition of C_1, $\ell i \neq 0, \ell' i' \neq 0$, and $i = i'$ hold. Furthermore, since C_1 has no exceptional codewords, we also have $\ell = \ell'$ and $(\ell j - \ell j')n_1 \equiv 0 \pmod{n_1n_2}$, i.e., $\ell(j - j') \equiv 0 \pmod{n_2}$. Then $(n_2, \ell) = 1$ implies $j = j'$. Similarly, the difference sets of any two codewords with generators from Γ_2 are disjoint, since $C_2 \in \mathrm{CAC^e}(sn_2, k)$. □

COROLLARY 6.2. $\mathrm{M}(35, 4) = 6$, $\mathrm{M}(77, 4) = 12$, *and* $\mathrm{M}(91, 4) = 14$.

Proof. For $n_1 = 7$ we have an equidifference CAC with $\Gamma(C_1) = \{1\}$ consisting of one nonexceptional codeword x_1. For $n_2 = 5$, $\Gamma(C_2) = \{1\}$ defines an equidifference CAC with an exceptional codeword. By Theorem 6.1, we obtain a code C for $n = 35$ with $|C| = 6$. Then Lemma 2.1 implies $\mathrm{M}(35, 4) = \mathrm{M^e}(35, 4) = 6$.

Similarly, for $n_2 = 11$ and $\Gamma(C_2) = \{1\}$, we obtain an optimal code C for $n = 77$ with $|C| = \mathrm{M}(77, 4) = \mathrm{M^e}(77, 4) = 12$.

For $n_2 = 13$, we easily see that $\mathrm{M}(n_2, 4) = 1$. By using $\Gamma(C_2) = \{1\}$, we obtain a code C for $n = 91$ with $|C| = 14$. Note that $\mathrm{M}(91, 4) \leq 15$ by Lemma 2.1. Suppose that there exists a code C' with $|C'| = 15$. Since there are no exceptional codewords for $n = 91$, C' must be an equidifference code. Now, for an equidifference codeword x_i of C', the difference set $\Delta(x_i)$ intersects with $7\mathbb{Z}_{91} \setminus \{0\} \simeq \mathbb{Z}_{13}^{\times}$ iff $i \in 7\mathbb{Z}_{91} \setminus \{0\}$. In other words, $7\mathbb{Z}_{91} \setminus \{0\}$ is covered by differences iff $\mathrm{M^e}(13, 4) = 2$, which contradicts $\mathrm{M^e}(13, 4) = 1$ (see Table 7.2). Hence it follows that $\mathrm{M}(91, 4) = 14 = |C|$. □

When p is an odd prime, 1 is a generator of a code $C \in \mathrm{CAC^e}(p, (p + 1)/2)$. By applying Theorem 6.1 to C recursively, we obtain the following corollary.

COROLLARY 6.3 (Levenshtein [11]). *Let p be an odd prime and r be a positive integer. Then there exists an optimal code $C \in \mathrm{CAC^e}(n, k)$ with parameters $n = p^r$, $k = \frac{p+1}{2}$, and $|C| = \frac{n-1}{2(k-1)}$.*

Furthermore, some infinite series of optimal CACs are obtained.

COROLLARY 6.4. *Let $p_1, p_2, \ldots, p_r$ be primes such that $p_i \equiv 1 \pmod 6$ satisfying the condition of Corollary 3.4 for each i, $1 \leq i \leq r$. Then there exists an optimal code $C \in \mathrm{CAC^e}(n = \prod_{1\leq i\leq r} p_i, 4)$ with $|C| = \frac{n-1}{6}$.*

Proof. Let $C_i \in \mathrm{CAC}^{\mathrm{e}}(p_i, 4)$ be an optimal code constructed in Corollary 3.4 for each i, $1 \leq i \leq r$. We have only to check the number of codewords for the code given by the recursive construction in Theorem 6.1. Each code C_i has $m_i = \frac{p_i - 1}{6}$ codewords, which attains the upper limit of Lemma 2.1. By applying the recursive construction to C_1 and C_2, we have an equidifference code of length $p_1 p_2 = 6(6m_1m_2 + m_1 + m_2) + 1$ with $6m_1m_2 + m_1 + m_2$ codewords, which also attains the upper limit of Lemma 2.1. By continuing this process, we have the desired optimal code $C \in \mathrm{CAC}^{\mathrm{e}}(n = \prod_{1 \leq i \leq r} p_i, 4)$ with $|C| = \frac{n-1}{6}$. □

In the following corollaries, it is enough to check the case of $r = 2$ since the similar process can be applied recursively.

Corollary 6.5. *Let $p_1, p_2, \ldots, p_r$ be primes such that $p_i \equiv 7 \pmod 8$ for each i, $1 \leq i \leq r$. Then there exists an optimal code $C \in \mathrm{CAC}^{\mathrm{e}}(n = 3\prod_{1 \leq i \leq r} p_i, 4)$ with $|C| = \frac{n-3}{6}$.*

Proof. Let $C_i \in \mathrm{CAC}^{\mathrm{e}}(3p_i, 4)$ be an optimal code constructed in Corollary 3.8 for each i, $1 \leq i \leq r$. Each code C_i has $m_i = \frac{3p_i - 3}{6}$ codewords, which attains the upper limit of Lemma 2.1. The composed code of C_1 and C_2 is an equidifference code of length $3p_1p_2 = 3(2(2m_1m_2 + m_1 + m_2) + 1)$ with $2m_1m_2 + m_1 + m_2$ codewords, which also attains the upper limit of Lemma 2.1. □

Corollary 6.6. *Let $p_1, p_2, \ldots, p_r$ be primes such that $p_i \equiv 13 \pmod{24}$ satisfying the condition of Corollary 5.3 for each i, $1 \leq i \leq r$. Then there exists an optimal code $C \in \mathrm{CAC}^{\mathrm{e}}(n = 4\prod_{1 \leq i \leq r} p_i, 4)$ with $|C| = \frac{n+2}{6}$.*

Proof. Let $C_i \in \mathrm{CAC}^{\mathrm{e}}(4p_i, 4)$ be an optimal code constructed in Corollary 5.3 for each i, $1 \leq i \leq r$. Each code C_i has $m_i = \frac{4p_i + 2}{6}$ codewords, which attains the upper limit of Lemma 2.1. Here, we can assume $m_i = 2\ell_i + 1$ for some $\ell \in \mathbb{N}$ since $p_i \equiv 1 \pmod 3$. Let C_1' be a code derived by deleting an exceptional codeword with generator p_1 from C_1. By composing C_1' and C_2, we have an equidifference code of length $4p_1p_2 = 4(3(3\ell_1\ell_2 + \ell_1 + \ell_2) + 1)$ with $2(3\ell_1\ell_2 + \ell_1 + \ell_2) + 1$ codewords, which also attains the upper limit of Lemma 2.1. □

Corollary 6.7. *Let $p_1, p_2, \ldots, p_r$ be primes such that $p_i \equiv 5 \pmod{24}$ for each i, $1 \leq i \leq r$. Then there exists an optimal code $C \in \mathrm{CAC}^{\mathrm{e}}(n = 2\prod_{1 \leq i \leq r} p_i, 5)$ with $|C| = \frac{n-2}{8}$.*

Proof. Let $C_i \in \mathrm{CAC}^{\mathrm{e}}(2p_i, 5)$ be an optimal code constructed in Corollary 3.10 for each i, $1 \leq i \leq r$. Each code C_i has $m_i = \frac{2p_i - 2}{8}$ codewords, which attains the upper limit of Lemma 2.3. By composing C_1 and C_2, we have an equidifference code of length $2p_1p_2 = 2(4(4m_1m_2 + m_1 + m_2) + 1)$ with $4m_1m_2 + m_1 + m_2$ codewords, which also attains the upper limit of Lemma 2.3. □

Corollary 6.8. *Let $p_1, p_2, \ldots, p_r$ be primes such that $p_i \equiv 11 \pmod{12}$ for each i, $1 \leq i \leq r$. Then there exists an optimal code $C \in \mathrm{CAC}^{\mathrm{e}}(n = 4\prod_{1 \leq i \leq r} p_i, 5)$ with $|C| = \frac{n-4}{8}$.*

Proof. Let $C_i \in \mathrm{CAC}^{\mathrm{e}}(4p_i, 5)$ be an optimal code constructed in Corollary 3.11 for each i, $1 \leq i \leq r$. Each code C_i has $m_i = \frac{4p_i - 4}{8}$ codewords, which attains the upper limit of Lemma 2.3. By composing C_1 and C_2, we have an equidifference code of length $4p_1p_2 = 4(2(2m_1m_2 + m_1 + m_2) + 1)$ with $2m_1m_2 + m_1 + m_2$ codewords, which also attains the upper limit of Lemma 2.3. □

7. Tables. In this section, we give some tables for the existence of equidifference CACs of small code length. Table 7.1 shows the first 110 primes satisfying the condition of Corollary 3.4. Table 7.2 shows the maximum size $\mathrm{M}^{\mathrm{e}}(n, 4)$ of an equidifference CAC for each n, $4 \leq n \leq 100$, and its corresponding generators. We also refer to [18] for more tables of CACs for weight $k = 3$, 4, and 5.

Table 7.1

The first 110 primes $p = 6m + 1$ satisfying the condition of Corollary 3.4. $\alpha \in \mathbb{Z}_p$ denotes a primitive element and $\gamma = \alpha^3$. The code $C \in \mathrm{CAC}(n = p, 4)$ defined by the list of generators $(1, \gamma, \ldots, \gamma^{m-1})$ is optimal. The column c_1 (or c_2) indicates the length $n = 4p$ if p satisfies the condition of Corollary 5.3 (or Corollary 5.5, respectively).

p	m	α	γ	c_1	c_2
7	1	3	6	-	28
37	6	2	8	-	-
139	23	2	8	-	556
163	27	2	8	-	652
181	30	2	8	724	-
241	40	7	102	-	-
313	52	10	61	-	-
337	56	10	326	-	-
349	58	2	8	-	-
379	63	2	8	-	1516
409	68	21	263	-	-
421	70	2	8	1684	-
541	90	2	8	2164	-
571	95	3	27	-	2284
607	101	3	27	-	2428
631	105	3	27	-	2524
751	125	3	27	-	3004
859	143	2	8	-	3436
877	146	2	8	-	-
937	156	5	125	-	-
1033	172	5	125	-	-
1087	181	3	27	-	4348
1123	187	2	8	-	4492
1171	195	2	8	-	-
1291	215	2	8	-	5164
1297	216	10	1000	-	-
1447	241	3	27	-	5788
1453	242	2	8	5812	-
1483	247	2	8	-	5932
1693	282	2	8	-	-
1741	290	2	8	-	-
1747	291	2	8	-	6988
2011	335	3	27	-	8044
2161	360	23	1362	-	-
2239	373	3	27	-	8956
2311	385	3	27	-	9244
2371	395	2	8	-	9484
2473	412	5	125	-	-
2539	423	2	8	-	10156
2647	441	3	27	-	10588
2677	446	2	8	10708	-
2707	451	2	8	-	10828
2719	453	3	27	-	10876
2857	476	11	1331	-	-
3169	528	7	343	-	-
3361	560	22	565	-	-
3433	572	5	125	-	-
3511	585	7	343	-	-
3547	591	2	8	-	14188
3559	593	3	27	-	14236
3571	595	2	8	-	14284
3613	602	2	8	-	-
3637	606	2	8	14548	-
3727	621	3	27	-	14908
3877	646	2	8	-	-
3919	653	3	27	-	15676
3931	655	2	8	-	15724
4003	667	2	8	-	16012
4021	670	2	8	16084	-
4111	685	12	1728	-	-
4201	700	11	1331	-	-
4219	703	2	8	-	16876
4261	710	2	8	-	-
4297	716	5	125	-	-
4357	726	2	8	-	-
4363	727	2	8	-	17452
4441	740	21	379	-	-
4507	751	2	8	-	18028
4561	760	11	1331	-	-
4603	767	2	8	-	18412
4801	800	7	343	-	-
4831	805	3	27	-	19324
4861	810	11	1331	19444	-
4903	817	3	27	-	19612
4987	831	2	8	-	19948
4999	833	3	27	-	19996
5023	837	3	27	-	20092
5107	851	2	8	-	-
5119	853	3	27	-	20476
5431	905	3	27	-	21724
5479	913	3	27	-	21916
5563	927	2	8	-	22252
5683	947	2	8	-	22732
5689	948	11	1331	-	-
5743	957	10	1000	-	-
5749	958	2	8	22996	-
5827	971	2	8	-	23308
5857	976	7	343	-	-
5869	978	2	8	23476	-
5881	980	316	386	-	-
5923	987	2	8	-	23692
6073	1012	10	1000	-	-
6343	1057	3	27	-	25372
6379	1063	2	8	-	25516
6397	1066	2	8	-	-
6469	1078	2	8	25876	-
6571	1095	3	27	-	26284
6577	1096	5	125	-	-
6733	1122	2	8	26932	-
6781	1130	2	8	27124	-
6823	1137	3	27	-	27292
6907	1151	2	8	-	27628
6949	1158	2	8	27796	-
7129	1188	7	343	-	-
7159	1193	3	27	-	28636
7237	1206	2	8	28948	-
7243	1207	2	8	-	28972
7759	1293	3	27	-	31036
7789	1298	2	8	-	-
7879	1313	3	27	-	31516

Table 7.2

This table shows for each n, $4 \leq n \leq 100$, with $n = 6m + c$, $0 \leq c \leq 5$, the maximum size $t = \mathrm{M}^{\mathrm{e}}(n, 4)$ of an equidifference CAC. $\Gamma(C)$ is the set of generators of such a maximum equidifference code C (the lexicographical smallest with respect to the generators).

n	m	c	t	$\Gamma(C)$
4	0	4	**1**	1
5	0	5	**1**	1
6	1	0	**1**	1
7	1	1	**1**	1
8	1	2	**1**	1
9	1	3	**1**	1
10	1	4	**1**	1
11	1	5	**1**	1
12	2	0	**1**	1
13	2	1	**1**	1
14	2	2	**1**	1
15	2	3	**1**	1
16	2	4	**2**	1, 4
17	2	5	**2**	1, 4
18	3	0	**2**	1, 4
19	3	1	**2**	1, 4
20	3	2	**3**	1, 4, 5
21	3	3	**3**	1, 4, 5
22	3	4	**2**	1, 4
23	3	5	**2**	1, 4
24	4	0	**3**	1, 4, 5
25	4	1	**3**	1, 4, 5
26	4	2	**3**	1, 4, 5
27	4	3	**3**	1, 4, 7
28	4	4	**4**	1, 4, 5, 7
29	4	5	**3**	1, 4, 5
30	5	0	**4**	1, 4, 5, 7
31	5	1	**3**	1, 4, 5
32	5	2	**4**	1, 4, 5, 7
33	5	3	**4**	1, 4, 5, 13
34	5	4	**4**	1, 4, 5, 7
35	5	5	**6**	1, 5, 6, 7, 8, 13
36	6	0	**5**	1, 4, 7, 9, 10
37	6	1	**6**	1, 6, 8, 10, 11, 14
38	6	2	**5**	1, 4, 5, 7, 9
39	6	3	**5**	1, 4, 5, 7, 11
40	6	4	**6**	1, 4, 5, 7, 9, 17
41	6	5	**5**	1, 4, 10, 16, 18
42	7	0	**5**	1, 4, 5, 7, 11
43	7	1	**6**	1, 5, 6, 7, 8, 13
44	7	2	**7**	1, 4, 5, 7, 9, 11, 19
45	7	3	**6**	1, 4, 5, 7, 9, 13
46	7	4	**6**	1, 4, 5, 7, 9, 11
47	7	5	**6**	1, 4, 11, 19, 20, 21
48	8	0	**6**	1, 5, 6, 7, 8, 13
49	8	1	**8**	1, 6, 7, 8, 13, 15, 20, 22
50	8	2	**7**	1, 4, 5, 7, 9, 13, 22
51	8	3	**6**	1, 4, 5, 7, 9, 19
52	8	4	**8**	1, 4, 5, 7, 9, 11, 13, 23
53	8	5	**7**	1, 6, 7, 8, 10, 19, 22
54	9	0	**7**	1, 4, 7, 9, 10, 13, 16
55	9	1	**7**	1, 4, 5, 7, 9, 11, 13
56	9	2	**8**	1, 4, 5, 7, 9, 11, 13, 25
57	9	3	**8**	1, 4, 5, 7, 11, 13, 16, 17
58	9	4	**8**	1, 4, 5, 9, 11, 13, 14, 17
59	9	5	**8**	1, 4, 10, 14, 18, 22, 24, 25
60	10	0	**7**	1, 4, 5, 7, 11, 17, 18
61	10	1	**8**	1, 5, 6, 7, 8, 11, 19, 26
62	10	2	**8**	1, 4, 7, 9, 10, 11, 13, 19
63	10	3	**8**	1, 4, 5, 7, 9, 11, 13, 19
64	10	4	**9**	1, 4, 5, 7, 9, 11, 13, 16, 29
65	10	5	**8**	1, 4, 5, 7, 13, 16, 18, 28
66	11	0	**9**	1, 4, 5, 7, 11, 13, 16, 19, 30
67	11	1	**9**	1, 4, 5, 11, 14, 16, 18, 29, 30
68	11	2	**10**	1, 4, 5, 7, 9, 11, 13, 16, 17, 31
69	11	3	**11**	1, 4, 5, 11, 13, 14, 16, 17, 20, 25, 31
70	11	4	**8**	1, 4, 5, 7, 9, 11, 13, 17
71	11	5	**10**	1, 5, 6, 8, 13, 14, 17, 25, 30, 31
72	12	0	**9**	1, 4, 5, 9, 11, 14, 16, 17, 26
73	12	1	**8**	1, 4, 5, 7, 9, 11, 13, 16
74	12	2	**9**	1, 4, 5, 7, 9, 11, 13, 19, 34
75	12	3	**10**	1, 4, 5, 7, 11, 13, 16, 17, 19, 23
76	12	4	**11**	1, 4, 5, 7, 9, 11, 13, 16, 17, 19, 35
77	12	5	**12**	1, 6, 7, 8, 11, 13, 15, 20, 27, 29, 34, 36
78	13	0	**10**	1, 4, 5, 7, 11, 13, 16, 17, 18, 29
79	13	1	**10**	1, 4, 5, 7, 11, 16, 17, 18, 20, 35
80	13	2	**12**	1, 4, 5, 7, 9, 11, 13, 16, 17, 19, 20, 37
81	13	3	**10**	1, 4, 5, 7, 9, 11, 13, 17, 19, 25
82	13	4	**10**	1, 4, 7, 9, 10, 11, 13, 16, 19, 29
83	13	5	**11**	1, 4, 7, 15, 18, 20, 22, 24, 26, 32, 37
84	14	0	**10**	1, 4, 5, 7, 11, 13, 16, 19, 20, 25
85	14	1	**13**	1, 4, 5, 14, 17, 18, 20, 22, 23, 24, 26, 32, 38
86	14	2	**11**	1, 4, 7, 9, 13, 15, 22, 23, 25, 35, 38
87	14	3	**13**	1, 4, 5, 7, 11, 16, 17, 19, 20, 23, 26, 31, 37
88	14	4	**13**	1, 4, 7, 9, 11, 15, 16, 17, 23, 25, 31, 39, 41
89	14	5	**12**	1, 4, 5, 9, 14, 20, 22, 24, 26, 32, 34, 35
90	15	0	**11**	1, 4, 5, 7, 9, 13, 16, 19, 20, 22, 28
91	15	1	**14**	1, 6, 7, 8, 13, 15, 20, 22, 27, 29, 34, 36, 41, 43
92	15	2	**14**	1, 4, 7, 9, 11, 13, 15, 16, 19, 23, 25, 29, 41, 43
93	15	3	**15**	1, 4, 5, 7, 13, 16, 17, 19, 20, 22, 23, 25, 28, 29, 41
94	15	4	**12**	1, 4, 5, 9, 13, 16, 17, 19, 21, 22, 29, 35
95	15	5	**13**	1, 5, 6, 7, 8, 19, 20, 23, 33, 39, 41, 42, 43
96	16	0	**11**	1, 4, 5, 7, 11, 13, 16, 17, 18, 23, 29
97	16	1	**12**	1, 4, 5, 7, 9, 16, 17, 20, 22, 26, 28, 36
98	16	2	**13**	1, 4, 7, 9, 15, 17, 19, 22, 25, 26, 29, 31, 37
99	16	3	**12**	1, 4, 5, 7, 9, 11, 13, 16, 19, 23, 25, 31
100	16	4	**14**	1, 4, 5, 7, 9, 11, 13, 16, 17, 19, 20, 23, 25, 47

Acknowledgments. The authors would like to thank the anonymous referees for their helpful comments and valuable suggestions.

REFERENCES

[1] T. Beth, D. Jungnickel, and H. Lenz, *Design Theory*, Cambridge University Press, Cambridge, UK, 1999.

[2] M. Buratti, *On simple radical difference families*, J. Combin. Des., 3 (1995), pp. 161–168.

[3] M. Buratti, *Cyclic designs with block size* 4 *and related optimal optical orthogonal codes*, Des. Codes Cryptogr., 26 (2002), pp. 111–125.

[4] C. J. Colbourn, J. H. Dinitz, K. Chen, and L. Zhu, *The CRC Handbook of Combinatorial Designs*, CRC Press, Boca Raton, FL, 1996.

[5] L. Györfi and I. Vajda, *Constructions of protocol sequences for multiple access collision channel without feedback*, IEEE Trans. Inform. Theory, 39 (1993), pp. 1762–1765.

[6] H. Hasse, *Mathmatische Abhandlungen Band* 1, Walter de Gruyter, Berlin, 1975.

[7] K. Ireland and M. Rosen, *A Classical Introduction to Modern Number Theory*, Springer-Verlag, New York, 1982.

[8] M. Jimbo, M. Mishima, S. Janiszewski, A. Y. Teymorian, and V. Tonchev, *On conflict-avoiding codes of length $n = 4m$ for three active users*, IEEE Trans. Inform. Theory, 53 (2007), pp. 2732–2742.

[9] C. Lam and Y. Miao, *$(C_k \oplus G, k, \lambda)$ difference families*, Des. Codes Cryptogr., 24 (2001), pp. 291–304.

[10] V. I. Levenshtein, *Conflict-avoiding Codes for Three Active Users and Cyclic Triple Systems*, preprint.

[11] V. I. Levenshtein, *Conflict-avoiding codes for many active users*, in Problems of Theoretic Cybernetics, Abstracts of the 14th International Conference, Penza, Publishing House of the Mechanical Mathematics Department of Moscow State University, 2005, Lomonosov, Russia, p. 86 (in Russian).

[12] V. I. Levenshtein and V. D. Tonchev, *Conflict-avoiding codes and cyclic triple systems*, in Proceedings of the 2005 IEEE International Symposium on Information Theory, Adelaide, Australia, 2005, pp. 535–537.

[13] V. I. Levenshtein and A. J. H. Vinck, *Perfect (d, k)-codes capable of correcting single peak-shifts*, IEEE Trans. Inform. Theory, 39 (1993), pp. 656–662.

[14] J. L. Massey and P. Mathys, *The collision channel without feedback*, IEEE Trans. Inform. Theory, 31 (1985), pp. 192–204.

[15] P. Mathys, *A class of codes for T active users out of N multiple-access communication system*, IEEE Trans. Inform. Theory, 36 (1990), pp. 1206–1219.

[16] Q. A. Nguyen, L. Györfi, and J. L. Massey, *Constructions of binary constant weight cyclic codes and cyclically permutable codes*, IEEE Trans. Inform. Theory, 38 (1992), pp. 940–949.

[17] P. Ribenboim, *Classical Theory of Algebraic Numbers*, Springer-Verlag, New York, 2001.

[18] V. D. Tonchev, *Tables of Conflict-Avoiding Codes*, available online at http://www.math.mtu.edu/~tonchev/CAC.html (2005).

[19] B. S. Tsybakov and A. R. Rubinov, *Some constructions of conflict-avoiding codes*, Probl. Inf. Transm., 38 (2002), pp. 268–279.

[20] R. M. Wilson, *Cyclotomy and difference families in elementary abelian groups*, J. Number Theory, 4 (1972), pp. 17–47.

SIAM J. DISCRETE MATH.
Vol. 21, No. 4, pp. 980–986

RAMSEY NUMBERS AND THE SIZE OF GRAPHS[*]

BENNY SUDAKOV[†]

Abstract. For two graphs H and G, the Ramsey number $r(H, G)$ is the smallest positive integer n such that every red-blue edge coloring of the complete graph K_n on n vertices contains either a red copy of H or a blue copy of G. Motivated by questions posed by Erdős and Harary, in this note we study how the Ramsey number $r(K_s, G)$ depends on the size of the graph G. For $s \geq 3$, we prove that for every G with m edges, $r(K_s, G) \geq c(m/\log m)^{(s+1)/(s+3)}$ for some positive constant c depending only on s. This lower bound improves an earlier result of Erdős, Faudree, Rousseau, and Schelp, and it is tight up to a polylogarithmic factor when $s = 3$. We also study the maximum value of $r(K_s, G)$ as a function of m.

Key words. Ramsey numbers, probabilistic methods

AMS subject classifications. 05C55, 05D40, 05D10, 05C80

DOI. 10.1137/060667360

1. Introduction.

For two graphs H and G, the *Ramsey number* $r(H, G)$ is the smallest positive integer n such that any red-blue coloring of the edges of the complete graph K_n on n vertices contains either a red copy of H or a blue copy of G. If $H = G$, we usually denote $r(G, G)$ by $r(G)$. The problem of determining or accurately estimating Ramsey numbers is one of the central problems in modern combinatorics, and it has received considerable attention; see, e.g., [10], [5]. In most cases the Ramsey number is estimated in terms of the order (number of vertices) of the graph. However, in the early 1980's Erdős and Harary asked about the relation between $r(H, G)$ and the sizes (number of edges) of the graphs H and G.

The first partial answers to this general problem were obtained by Erdős et al. [8]. They determined up to a constant factor the minimum value of $r(G)$ for all graphs of size m and showed that the order of magnitude of this minimum is $\Theta(m/\log m)$. They also proved that for fixed $s \geq 3$ there exist constants c_1, c_2 such that

$$c_1 m^{\frac{s}{s+2}} < \min_{e(G)=m} r(K_s, G) < c_2\, m^{\frac{s-1}{s}}.$$

This estimate is not sharp, and there is a large gap between the upper and lower bounds even when the complete graph is a triangle ($s = 3$). In this case Erdős [6] conjectured that the upper bound $O(m^{2/3})$ is closer to the truth. Our first result improves the bounds from [8] and confirms this conjecture.

THEOREM 1.1. *Let $s \geq 3$ and let G be a graph with m edges. Then there exists a constant c depending only on s such that*

$$r(K_s, G) \geq c\big(m/\log m\big)^{\frac{s+1}{s+3}}.$$

On the other hand, there exists a graph G of size m such that $r(K_s, G) \leq O(m^{\frac{s-1}{s}}/\log^{\frac{s-2}{s}} m)$.

[*] Received by the editors August 11, 2006; accepted for publication (in revised form) May 23, 2007; published electronically December 12, 2007. This research was supported in part by NSF CAREER award DMS-0546523, NSF grant DMS-0355497, USA-Israeli BSF grant, an Alfred P. Sloan fellowship, and the State of New Jersey.

http://www.siam.org/journals/sidma/21-4/66736.html

[†] Department of Mathematics, Princeton University, Princeton, NJ 08544, and Institute for Advanced Study, Princeton, NJ 08540 (bsudakov@math.princeton.edu).

In particular, when $s = 3$ this determines the minimum value of $r(K_3, G)$ for all graphs G of size m up to a polylogarithmic factor and shows that

$$\Omega\left(\frac{m^{2/3}}{\log^{2/3} m}\right) < \min_{e(G)=m} r(K_3, G) < O\left(\frac{m^{2/3}}{\log^{1/3} m}\right). \tag{1}$$

Another specific question which is part of the general problem mentioned in the first paragraph is how to bound the maximum value of $r(H, G)$ when the graphs H and G have a given size. One of the basic results in Ramsey theory is the fact that for the complete graph G with m edges, $r(G) = 2^{\Theta(\sqrt{m})}$. A conjecture of Erdős [7] (see also [5]) asserts that there is an absolute constant c such that for any graph G with m edges, $r(G) \leq 2^{c\sqrt{m}}$. This conjecture is still open. For bipartite graphs it was proved by Alon, Krivelevich, and Sudakov [3]. They also show that for general graphs G with m edges, $r(G) \leq 2^{c\sqrt{m}\log m}$ for some absolute positive constant c. For the first off-diagonal case, Harary conjectured and Sidorenko [15] proved that $r(K_3, G) \leq 2m + 1$ for any graph G of size m without isolated vertices. This inequality is best possible, since $r(K_3, G) = 2m + 1$ for any tree with m edges. Thus for $s = 3$ the graphs which maximize $r(K_3, G)$ are very sparse. However, for $s > 3$ Erdős conjectured that exactly the opposite is true and that to maximize $r(K_s, G)$ over all graphs with m edges one should make G as nearly complete as possible. Motivated by this question we obtain the following general upper bound on $r(K_s, G)$ for graphs G of size m.

THEOREM 1.2. *Let $s \geq 3$ and let G be a graph with m edges and without isolated vertices. Then there exists a constant c depending only on s such that*

$$r(K_s, G) \leq c\, m^{\frac{s-1}{2}} / \log^{\frac{s-3}{2}} m\,.$$

When G is a clique with m edges it is known by the result of Ajtai, Komlós, and Szemerédi [1] that $r(K_s, G) \leq O(m^{\frac{s-1}{2}} / \log^{s-2} m)$. Hence our estimate has, up to a polylogarithmic factor, similar order of magnitude to the best known upper bound for off-diagonal Ramsey numbers of cliques.

The rest of this short note is organized as follows. In the next section we present proofs of our main results. The final section contains some concluding remarks and open problems. Throughout the paper we make no attempts to optimize various absolute constants. To simplify the presentation, we often omit floor and ceiling signs whenever these are not crucial. All logarithms are in the natural base e.

2. Proofs. To prove Theorem 1.1 we use an approach developed by Krivelevich [13], which is based on probabilistic arguments together with large deviation inequalities. The first inequality we need is a standard bound of Chernoff (see Appendix A in [2]) which states that if X is a binomially distributed random variable with parameters m and p, then for every $a > 0$

$$\mathbb{P}[X - pm < -a] \leq e^{-\frac{a^2}{2pm}}.$$

Another large deviation bound, which we use in the proof, was obtained by Erdős and Tetali [9] (see also Chapter 8.4 in [2]).

Let Ω be a finite set (in our instance it is the set of edges of a complete graph) and let R be a random subset of Ω such that $\mathbb{P}[\omega \in R] = p_\omega$ independently for all $\omega \in \Omega$. Let $C_i, i \in I$, be subsets of Ω, where I is some finite index set. For every C_i we define A_i to be the event that $C_i \subseteq R$. Let X_i be the indicator random variable of event A_i and let $X = \sum_{i \in I} X_i$ be the number of $C_i \subseteq R$. Finally, let X_0 be the maximum number of pairwise disjoint subsets C_i which belong to R. Obviously, $X_0 \leq X$. Let μ be the expectation of X; then Erdős and Tetali gave the following bound on the possible size of X_0:

$$\mathbb{P}[X_0 \geq k] \leq \frac{\mu^k}{k!} \leq \left(\frac{e\,\mu}{k}\right)^k .$$

Proof of Theorem 1.1. Let $n = \frac{1}{3s^3}\big(m/\log m\big)^{\frac{s+1}{s+3}}$ and consider coloring the edges of the complete graph K_n such that each edge is colored randomly and independently red with probability $p = \frac{1}{3s}n^{-\frac{2}{s+1}}$ and blue with probability $1-p$. Let $G_1, \ldots, G_t$ be all subgraphs of K_n which are isomorphic to G. The number of such subgraphs t is clearly bounded by the number of injective functions from $V(G)$ to K_n, which in turn is at most the number of permutations on n elements. Thus $t \leq n!$. For every subgraph G_i, let X_i be the random variable that counts the number of red edges in G_i. By definition, X_i is binomially distributed with parameters m and p. Hence, by Chernoff's inequality

$$\mathbb{P}[X_i < mp/2] = \mathbb{P}[X_i - mp < -mp/2] \leq e^{-mp/8}.$$

Also for every subgraph G_i define Y_i to be the number of red cliques of order s which share at least one edge with G_i. Since G_i has m edges, the number of s-cliques sharing at least one edge with G_i is bounded by mn^{s-2}. The probability that an s-clique is red is clearly $p^{\binom{s}{2}}$. Therefore

$$\mathbb{E}[Y_i] \leq mn^{s-2}p^{\binom{s}{2}} = mpn^{s-2}p^{\frac{(s+1)(s-2)}{2}} = \left(\frac{1}{3s}\right)^{\frac{(s+1)(s-2)}{2}} mp \leq \frac{mp}{9s^2} .$$

Let Y_i' be the maximum number of edge disjoint red s-cliques which share at least one edge with G_i. Then by the Erdős–Tetali inequality we have that

$$\mathbb{P}\big[Y_i' \geq mp/s^2\big] \leq \left(\frac{e\,\mathbb{E}[Y_i]}{mp/s^2}\right)^{mp/s^2} \leq \left(\frac{e}{9}\right)^{mp/s^2} \leq e^{-mp/s^2} .$$

By definition of n and p we have that $mp = (3s^3n)^{\frac{s+3}{s+1}}p\log m \geq s^2 n\log n > 8n\log n$. Therefore the probability that for some index i either $X_i < mp/2$ or $Y_i' \geq mp/s^2$ is bounded by $t\big(e^{-mp/8} + e^{-mp/s^2}\big) \leq 2n!\,n^{-n} = o(1)$. In particular there exists a red-blue edge coloring of K_n such that for every $1 \leq i \leq t$, subgraph G_i contains at least $mp/2$ red edges and there are at most mp/s^2 edge disjoint red s-cliques each sharing at least one edge with G_i.

Fix such a coloring and let Γ be the subgraph of red edges in it. Also let $\mathcal{C}$ be the maximum (under inclusion) collection of edge disjoint cliques of order s in Γ. Recolor edges in all cliques from $\mathcal{C}$ by blue and denote the remaining red graph by Γ'. Note that by recoloring we removed from Γ the maximum collection of edge disjoint s-cliques. Thus Γ' contains no clique of order s. On the other hand, in every subgraph G_i we changed the color of at most $\binom{s}{2}mp/s^2 < mp/2$ red edges. Since G_i originally had at least $mp/2$ red edges, we obtain that every subgraph of K_n isomorphic to G still has at least one red edge. This implies that new coloring contains no blue copy of G and no red copy of K_s and completes the proof of the first statement of the theorem.

To prove the second part, let G be the union of $\frac{2m}{k(k-1)}$ vertex disjoint cliques of order $k = m^{1/s}(\log m)^{\frac{s-2}{s}}$. By definition, the number of edges in G is at least m. To estimate the Ramsey number $r(K_s, G)$ we use the result of Ajtai, Komlós, and Szemerédi [1] (see also Theorem 12.17 in [4]) which bounds off-diagonal Ramsey numbers. They prove that there exists a constant c such that $r(K_s, K_k) \leq c\frac{k^{s-1}}{\log^{s-2} k}$. Let

$$n = c\frac{k^{s-1}}{\log^{s-2} k} + \frac{2m}{k-1} = O\left(\frac{m^{\frac{s-1}{s}}}{\log^{\frac{s-2}{s}} m}\right)$$

and consider any red-blue edge coloring of the complete graph K_n. We can assume that there is no red s-clique, or else we are done. Then, since $n \geq r(K_s, K_k)$, we can find a blue clique of order k. Delete it from the graph and continue this process. Note that as long as we deleted less than $\frac{2m}{k(k-1)}$ cliques of order k the remaining number of vertices is still larger than $r(K_s, K_k)$ and we can find a new blue k-clique. In the end we will find at least $\frac{2m}{k(k-1)}$ blue cliques of size k, i.e., a copy of G. This implies that $r(K_s, G) \leq O(m^{\frac{s-1}{s}} / \log^{\frac{s-2}{s}} m)$ and completes the proof. □

Proof of Theorem 1.2. We prove the theorem by induction on s. Consider the case $s = 3$. Although one can use results from [8] and [15] to show that $r(K_3, G) \leq O(m)$, we include here the simple proof that $r(K_3, G) \leq 3m$ for the sake of completeness. Clearly, we can assume that G is connected, since $r(K_3, G_1 \cup G_2) \leq r(K_3, G_1) + r(K_3, G_2)$. Hence the number of vertices of G is at most $m + 1$. Let $n = 3m$ and suppose that the edges of K_n are red-blue colored with no red triangle. Pick the vertex with maximum red degree in this coloring and let $X, |X| = t$, be the set of its red neighbors. Note that all the edges inside X are blue, since there is no red triangle. Partition the vertices of G into two sets $V(G) = V' \cup V''$, where V' consists of the t vertices with the highest degree. Since the sum of the degrees in G is $2m$, we have that all the vertices in V'' have degree at most $2m/(t + 1)$. Now we will find the blue copy of G as follows. Embed the vertices of V' into X arbitrarily, and then embed the vertices of V'' one by one. Given a vertex $v \in V''$, let Y be the set of vertices of K_n where we already embedded neighbors of v. Since the maximum red degree in the coloring is t and $|Y| \leq d(v) \leq 2m/(t + 1)$, we have that K_n contains at least $3m - t|Y| \geq m + 1 - |Y|$ vertices which are adjacent to all vertices in Y by blue edges. As the total number of vertices of G is at most $m + 1$, one such vertex is still unoccupied and can be used to embed v. Continuing this process we find a blue copy of G.

Now suppose $s > 3$ and by induction we have that $r(K_{s-1}, G) \leq c_1\, m^{\frac{s-2}{2}} / \log^{\frac{s-4}{2}} m$. Let $n = c_2 m^{\frac{s-1}{2}} / \log^{\frac{s-3}{2}} m$, where c_2 is a sufficiently large constant which depends on c_1 and which we fix later. Consider a red-blue coloring of the complete graph K_n such that there is no red copy of K_s. If there is a vertex which has at least $d = c_1\, m^{\frac{s-2}{2}} / \log^{\frac{s-4}{2}} m$ red neighbors, then this set cannot contain a red copy of K_{s-1}. Therefore by the induction hypothesis it will contain a blue copy of G, and we are done. Thus we can assume that the maximum degree in the red subgraph of K_n is at most d. Set $k = \sqrt{m \log m}$. It is easy to check that, by definition, $n = \Omega(\frac{k^{s-1}}{\log^{s-2} k})$. Therefore, by choosing c_2 large enough and using the result of Ajtai, Komlós, and Szemerédi [1] on Ramsey numbers, we get that $n \geq r(K_s, K_k)$. Hence there exists a set X of k vertices which spans only blue edges. Again partition the vertices of G into two sets $V(G) = V' \cup V''$, where V' consists of the k vertices with the highest degree. Since the sum of the degrees in G is $2m$, we have that all the vertices in V'' have degree at most $2m/(k+1)$. Embed the vertices of V' into X arbitrarily, and then embed the vertices of V'' one by one as follows. Given a vertex $v \in V''$, let Y be the set of vertices of K_n where we already embedded neighbors of v. Since the maximum red degree in the coloring is d and $|Y| \leq d(v) \leq 2m/(k+1)$, by choosing sufficiently large c_2, we have that there are at least

$$n - d|Y| \geq n - \frac{2md}{k+1} > \frac{c_2 m^{\frac{s-1}{2}}}{\log^{\frac{s-3}{2}} m} - \frac{2m}{\sqrt{m \log m}} \left(\frac{c_1\, m^{\frac{s-2}{2}}}{\log^{\frac{s-4}{2}} m} \right) > 2m$$

vertices in K_n which are adjacent to all vertices in Y by blue edges. Note that the total number of vertices of G is at most $2m$, as it has no isolated vertices. Therefore there exists an unoccupied vertex of K_n which is connected to all vertices in Y by blue edges. This vertex can be used to embed v. In the end of this procedure we obtain a blue copy of G. This completes the proof of the theorem. □

3. Concluding remarks. Let H be a graph with $v_H \geq 3$ vertices and e_H edges. The *density* $\rho(H)$ of H is defined as $\rho(H) = \frac{e_H - 1}{v_H - 2}$. Also define

$$\rho^*(H) = \max_{H' \subseteq H} \rho(H').$$

For example, for the complete graph of order s we have $\rho^*(K_s) = \frac{s+1}{2}$. The arguments in the proof of Theorem 1.1 can be used to obtain the following more general result. Since the proof of this statement does not require new ideas and contains somewhat tedious computations, we omit it here.

THEOREM 3.1. *Let H be a fixed graph. Then there exists a constant c depending only on H such that for every graph G with m edges,*

$$r(H, G) \geq c\big(m/\log m\big)^{\frac{\rho^*}{1+\rho^*}}.$$

In addition to the triangle, this result is nearly tight when H is the complete bipartite graph $K_{p,q}$ with $q \gg p$. Indeed it is easy to check from the definition that if p is fixed and $q \to \infty$, then $\rho^*(K_{p,q}) \to p$. Therefore for every p and $\epsilon > 0$ there exists q such that $\frac{\rho^*(K_{p,q})}{1+\rho^*(K_{p,q})} > \frac{p}{1+p} - \epsilon$. Thus, from Theorem 3.1 we have that $r(K_{p,q}, G) \geq \Omega(m^{\frac{p}{1+p} - \epsilon})$ for every G with m edges. On the other hand, from the result of Kövari, Sós, and Turán [12] that $K_{p,q}$-free graphs on n vertices can have at most $O(n^{2-1/p})$ edges, it follows that such a graph has an independent set of size $\Omega\big(n^{1/p}\big)$. This implies that $r(K_{p,q}, K_k) \leq O(k^p)$ (see also [3, 14] for a slightly better estimate). Using this bound together with the argument from the proof of the second part of Theorem 1.1, we can show that if G is the disjoint union of $\Theta\big(m^{\frac{p-1}{p+1}}\big)$ cliques of order $m^{1/(p+1)}$, then $r(K_{p,q}, G) \leq O(m^{\frac{p}{1+p}})$.

For $s = 3$ the lower bound in Theorem 1.1 is tight up to a multiplicative factor of $\log^{1/3} m$. It would be very interesting to close this gap. We think that our upper bound in (1) is closer to the truth and there exists an absolute constant c such that $r(K_3, G) \geq cm^{2/3}/\log^{1/3} m$ for every graph G of size m. To prove this one might try to use an approach based on the semirandom method which was developed by Kim [11] to determine the asymptotic behavior of Ramsey numbers $r(K_3, K_k)$.

It would be interesting to extend an upper bound in Theorem 1.2 to the general case when H and G are arbitrary graphs with sizes t and m and with no isolated vertices. We conjecture that if t is fixed and m is sufficiently large, then

$$r(H,G) \leq m^{O(\sqrt{t})}.$$

This estimate, if true, is tight up to a constant in the exponent, since the known lower bounds (see [16, 13]) on off-diagonal Ramsey numbers imply that $r(H,G) \geq m^{\Omega(\sqrt{t})}$ when H and G are complete graphs with t and m edges, respectively. In [3] it was proved that if H is a graph with chromatic number ℓ and maximum degree $d \geq \ell$, then for all sufficiently large m, $r(H, K_{2m}) \leq m^{\ell d}$. Using this estimate it is easy to obtain the following partial result, which shows that our conjecture holds if the chromatic number of H is a fixed constant.

PROPOSITION 3.2. *Let H and G be two graphs with no isolated vertices such that the size of G is m, the size of H is t, and H has chromatic number $\ell \geq 2$. Then there exists a constant c depending only on ℓ such that for sufficiently large m, $r(H,G) \leq m^{c\sqrt{t}}$.*

Sketch of proof. We use induction on t. Let v be the vertex of maximum degree in H. Since G has m edges, it has at most $2m$ vertices. Therefore, if the maximum degree of H is at most $2\sqrt{t}$, it follows from the above cited estimate in [3] that $r(H,G) \leq m^{2\sqrt{t}\ell}$. Otherwise, the degree of v in H is larger than $2\sqrt{t}$. Delete it and denote $H_1 = H \setminus \{v\}$. This graph has $t_1 \leq t - 2\sqrt{t}$ edges and $\sqrt{t_1} \leq \sqrt{t} - 1$.

Let $n = m^{2\sqrt{t}\ell}$ and consider red-blue coloring of the edges of the complete graph K_n. Since G has at most $2m$ vertices, we can assume that there is no red K_{2m} in this coloring. Therefore, by Turán's theorem, there is a vertex x in K_n, whose blue degree is at least $n/2m \gg m^{(2\sqrt{t}-2)\ell} \geq m^{2\sqrt{t_1}\ell}$. Let U be the set of blue neighbors of x. Clearly, this set contains no blue copy of H_1. Now we can use induction to conclude that it contains a red copy of G. □

REFERENCES

[1] M. Ajtai, J. Komlós, and E. Szemerédi, *A note on Ramsey numbers*, J. Combin. Theory Ser. A, 29 (1980), pp. 354–360.

[2] N. Alon and J. Spencer, *The Probabilistic Method*, 2nd ed., Wiley, New York, 2000.

[3] N. Alon, M. Krivelevich, and B. Sudakov, *Turán numbers of bipartite graphs and related Ramsey-type questions*, Combin. Probab. Comput., 12 (2003), pp. 477–494.

[4] B. Bollobás, *Random Graphs*, 2nd ed., Cambridge Stud. Adv. Math. 73, Cambridge University Press, Cambridge, UK, 2001.

[5] F. Chung and R. Graham, *Erdős on Graphs. His Legacy of Unsolved Problems*, A K Peters, Ltd., Wellesley, MA, 1998.

[6] P. Erdős, *Solved and unsolved problems in combinatorics and combinatorial number theory*, Congr. Numer., 32 (1981), 49–62.

[7] P. Erdős, *On some problems in graph theory, combinatorial analysis and combinatorial number theory*, in Graph Theory and Combinatorics (Cambridge, 1983), Academic Press, London, 1984, pp. 1–17.

[8] P. Erdős, R. Faudree, C. Rousseau, and R. Schelp, *A Ramsey problem of Harary on graphs with prescribed size*, Discrete Math., 67 (1987), pp. 227–233.

[9] P. Erdős and P. Tetali, *Representation of integers as the sum of k terms*, Random Structures Algorithms, 1 (1990), pp. 245–261.

[10] R. Graham, B. Rothschild, and J. Spencer, *Ramsey Theory*, 2nd ed., Wiley, New York, 1990.

[11] J. H. Kim, *The Ramsey number $R(3,t)$ has order of magnitude $t^2/\log t$*, Random Structures Algorithms, 7 (1995), pp. 173–207.

[12] T. Kövari, V. T. Sós, and P. Turán, *On a problem of K. Zarankiewicz*, Colloquium Math., 3 (1954), pp. 50–57.

[13] M. Krivelevich, *Bounding Ramsey numbers through large deviation inequalities*, Random Structures Algorithms, 7 (1995), pp. 145–155.

[14] Y. Li and W. Zang, *Ramsey numbers involving large dense graphs and bipartite Turán numbers*, J. Combin. Theory Ser. B, 87 (2003), pp. 280–288.

[15] A. Sidorenko, *The Ramsey number of an n-edge graph versus triangle is at most* $2n + 1$, J. Combin. Theory Ser. B, 58 (1993), pp. 185–196.

[16] J. Spencer, *Asymptotic lower bounds for Ramsey numbers*, Discrete Math., 20 (1977), pp. 69–76.

SIAM J. Discrete Math.
Vol. 21, No. 4, pp. 987–998

ON K-TERM DNF WITH THE LARGEST NUMBER OF PRIME IMPLICANTS[*]

ROBERT H. SLOAN[†], BALÁZS SZÖRÉNYI[‡], AND GYÖRGY TURÁN[§]

Abstract. It is known that a k-term DNF can have at most $2^k - 1$ prime implicants and that this bound is sharp. We determine all k-term DNF having the maximal number of prime implicants. It is shown that a DNF is maximal if and only if it corresponds to a nonrepeating decision tree with literals assigned to the leaves in a certain way. We also mention some related results and open problems.

Key words. disjoint DNF, disjunctive normal form, prime implicant

AMS subject classification. 68R05

DOI. 10.1137/050632026

1. Introduction.

Prime implicants of a Boolean function, or, in other words, maximal subcubes of a subset of the n-dimensional hypercube, form a basic concept for the theory of Boolean functions and their applications. Concerning the maximal number of prime implicants, it is known that an n-variable Boolean function can have at most $O(\frac{3^n}{\sqrt{n}})$ prime implicants, and there are n-variable Boolean functions with $\Omega(\frac{3^n}{n})$ prime implicants (see, e.g., [4]).

Another case considered is the maximal number of prime implicants of Boolean functions represented by disjunctive normal forms (DNF) with a bounded number of terms. The result that a k-term DNF can have at most $2^k - 1$ prime implicants was discovered independently by Chandra and Markowsky [4], Levin [17], and McMullen and Shearer [19]. For a recent application in computational learning theory, see Hellerstein and Raghavan [9]. It was shown by Laborde [16], Levin [17], and McMullen and Shearer [19] that the bound is sharp, i.e., there are k-term DNF with $2^k - 1$ prime implicants (Chandra and Markowsky gave an example with more than $2^{k/2}$ prime implicants). In view of these results, we call a DNF *maximal* if it has k terms and $2^k - 1$ prime implicants for some k.

In this paper we complete the results of [4, 16, 17, 19] by determining all the maximal DNF. In order to formulate the description, let us introduce the following definition.

By a tree we mean a rooted binary tree such that for every inner node, the edge leading to its left (resp., right) child is labeled 0 (resp., 1). For a given $k \geq 2$ and $r \geq 0$, let us consider the pairwise distinct variables $x_1, \ldots, x_{k-1}, y_1, \ldots, y_k$, and $z_1, \ldots, z_r$. For each of the y and z variables, pick an orientation, i.e., form the literals $y_i^{\epsilon_i}$ ($i = 1, \ldots, k$) and $z_j^{\delta_j}$ ($j = 1, \ldots, r$), where for ϵ_i and δ_j the value 1 (resp., 0)

[*] Received by the editors May 20, 2005; accepted for publication (in revised form) April 10, 2007; published electronically January 9, 2008. This material is based upon work supported by the National Science Foundation under grants CCR-0100036 and CCF-0431059.

http://www.siam.org/journals/sidma/21-4/63202.html

[†] University of Illinois at Chicago, Chicago, IL 60607 (sloan@uic.edu, http://www.cs.uic.edu/∼sloan).

[‡] Hungarian Academy of Sciences and University of Szeged, Research Group on Artificial Intelligence, Szeged, Hungary (szorenyi@inf.u-szeged.hu).

[§] University of Illinois at Chicago, Chicago, IL 60607 and Hungarian Academy of Sciences and University of Szeged, Research Group on Artificial Intelligence, Szeged, Hungary (gyt@uic.edu).

corresponds, as usual, to an unnegated (resp., negated) variable. A *nonrepeating, unate-leaf decision tree (NUD) T* over these variables and literals is constructed by taking a tree with $k-1$ inner nodes (and thus with k leaves), assigning to each inner node a distinct x variable, assigning to each leaf a distinct y literal from those formed above, and, in addition, assigning to each leaf an arbitrary subset of the z literals formed above. The set of leaves of T is denoted by L. If we want to mention the number of x variables and y literals used in the construction, then we refer to T as a k-NUD (the value r is irrelevant). Figure 1 gives an example of a 5-NUD (the labeling of the edges is omitted for simplicity).

A k-NUD represents a k-term DNF, determined as follows. For a leaf $\ell \in L$, let the term t_ℓ be the conjunction of the x literals along the path leading to ℓ (where traversing an edge labeled 1 corresponds to an unnegated literal, and traversing an edge labeled 0 corresponds to a negated literal) and of the y and z literals assigned to ℓ. The k-term DNF represented by the k-NUD T is

$$\varphi_T = \bigvee_{\ell \in L} t_\ell.$$

For example, the 5-term DNF represented by the 5-NUD of Figure 1 is

$$\overline{x_1}\,\overline{x_2}\,\overline{x_4}\,y_1\,z_1 \;\vee\; \overline{x_1}\,\overline{x_2}\,x_4\,\overline{y_2}\,\overline{z_2}\,z_3 \;\vee\; \overline{x_1}\,x_2\,y_3\,z_1 \;\vee\; x_1\,\overline{x_3}\,\overline{y_4}\,z_1\,z_4 \;\vee x_1\,x_3\,y_5\,\overline{z_2}.$$

The Boolean function represented by φ_T can also be thought of in the following way: Given a truth assignment a to all the variables, use the values of the x variables to determine a path from the root to a leaf. The function value is 1 if a makes all the y and z literals assigned to this leaf true, and it is 0 otherwise. It is clear from the definition that the input vectors accepted at a leaf ℓ are precisely those vectors which satisfy the term t_ℓ. The function φ_T is a generalized addressing function or multiplexer [20, 25]. If a DNF φ comes from a NUD T, then T can be reconstructed from φ. The y and z literals are those which are unate in φ, i.e., their negation does not occur in φ, while the x variables are those which occur both negated and unnegated. Among the x variables, the one labeling the root is the only one which occurs in every term (either unnegated or negated). The left child is the only x variable which occurs in every term containing the negation of the root variable, etc. In view of this correspondence, with some abuse of terminology, we can talk about a DNF being a NUD, rather than corresponding to a NUD. The maximal DNF of [16, 19] (resp., [17]) corresponds to a tree which is a single path (resp., a complete binary tree), without any z literals. A NUD generalizes these examples by allowing for a binary arbitrary tree and for the additional z literals. Now we can formulate the description of maximal DNF.

THEOREM 1. *A DNF is maximal if and only if it corresponds to a NUD.*

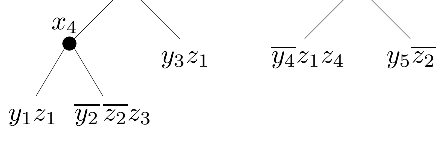


FIG. 1. *A nonrepeating, unate-leaf decision tree (NUD).*

A closely related class of DNF *tautologies* is obtained if we consider trees with the same kind of inner nodes, but without any literals assigned to the leaves. In the case of the example of Figure 1, the corresponding DNF tautology is

$$\overline{x_1}\,\overline{x_2}\,\overline{x_4} \lor \overline{x_1}\,\overline{x_2}\,x_4 \lor \overline{x_1}\,x_2 \lor x_1\,\overline{x_3} \lor x_1\,x_3\,.$$

Let us refer to this class of tautologies as *nonrepeating decision tree* tautologies, or *ND*'s. The main step in the proof of Theorem 1, the ND lemma (Lemma 11) is to show that for every DNF tautology the following two properties are equivalent: (a) any two of its terms have exactly one conflicting pair of literals (in other words, the terms are pairwise neighboring), (b) it is an ND. Lemma 11 was proven recently, independently from our work, by Kullmann [14, 15]. Kullmann's proof uses the concept of Hermitian defect and other concepts from linear algebra. (The Hermitian rank of a symmetric matrix is the maximum of the number of positive and the number of negative eigenvalues of the matrix (Gregory, Watts, and Shader [7]), and the Hermitian defect is the difference of the order of the matrix and its Hermitian rank [14, 15].) Kullmann also uses the characterization of ND's as strongly minimal tautologies with the additional property that the number of terms is one more than the number of variables (Aharoni and Linial [1], Davydov, Davydova, and Kieine Böning [5], Kullmann [13]), proved using Hall's theorem or resolution techniques. (A tautology is strongly minimal if deleting any term, or adding any literal to a term results in a nontautology.) Our proof is an elementary combinatorial argument.

We note that ND's come up in other contexts as well, e.g., in connection with the complexity of analytic tableaux (Urquhart [24], referring to an earlier unpublished work of Cook, and Arai, Pitassi, and Urquhart [2]). Another related topic is the decision tree complexity of tautologies (Lovász et al. [18]).

The characterization of ND's as pairwise neighboring DNF tautologies is a direct consequence of the following *splitting lemma* (Lemma 10): if the n-dimensional hypercube is partitioned into subcubes of pairwise distance 1 then there is a split of the whole cube into two half cubes such that every cube of the partition is contained in one of the two halves. We also consider the question of what can be said about cube partitions without the distance assumption. The goodness of a split into two half cubes can be measured by the fraction of the total volume of subcubes contained in one of the two halves (thus in the distance 1 case one always has a split of measure 1). This measures the fraction of points for which flipping the component corresponding to the two half cubes gives a point in a different subcube of the partition. Thus the goodness of the split measures the influence of the variable corresponding to the half cubes, on the partition (for other notions of influence, see, e.g., Hammer, Kogan, and Rothblum [8] and Kahn, Kalai, and Linial [11]). We give general lower and upper bounds for the best achievable split. The upper bound uses a result of Savický and Sgall [21] on DNF tautologies with bounded occurrences of the variables.

Recent related work on the combinatorial aspects of the satisfiability problem (see Kullmann [15] for a recent survey) makes use of the connection with partitioning complete graphs into complete bipartite graphs (bicliques). This connection, and in particular, the Graham–Pollak theorem [6] is used by Laborde [16] to show that a maximal k-term DNF contains at least $2\,k-1$ variables. (This result, in turn, follows immediately from Theorem 1 without using the Graham–Pollak theorem.) We give an application of the splitting lemma (Lemma 10) to show that the family of recursive partitions into complete bipartite graphs has an extremal property among all partitions into complete bipartite graphs.

The paper is organized as follows. After some preliminaries in section 2, the results of [4, 16, 17, 19] are presented in section 3. The proof of Theorem 1 is given in section 4. Section 5 contains the bounds for the general splitting problem. The connection to partitions of complete graphs into complete bipartite graphs is discussed briefly in section 6. Section 7 contains some further open problems on the number of prime implicants.

2. Preliminaries. A literal is a variable or a negated variable, a term is a conjunction (or a set) of literals, and a disjunctive normal form (DNF) is a disjunction of terms. The empty conjunction (resp., disjunction) is identically true (resp., false). It is assumed that terms do not contain both a variable and its negation. The size of a term t, denoted by $|t|$, is the number of its literals. The number of *conflicts* between two terms is the number of variables occurring unnegated in one term and negated in the other. A DNF is *disjoint* if any two of its terms have at least one conflict. We write $\psi \leq \varphi$ if every truth assignment satisfying ψ also satisfies φ, and $\psi < \varphi$ if, in addition, there is a truth assignment a with $\psi(a) = 0$ and $\varphi(a) = 1$. The set of vectors in $\{0,1\}^n$ satisfying φ are denoted by $T(\varphi)$. If t is a term, then $T(t)$ is a subcube (or simply cube) in $\{0,1\}^n$, with $|T(t)| = 2^{n-|t|}$. With an abuse of notation, we usually write cube t instead of cube $T(t)$. (This is an example of switching freely between syntactic and semantic views of the same object, which occurs frequently in the paper and is, in general, useful in the study of Boolean functions.) For a literal z, the z half cube of $\{0,1\}^n$ is the $(n-1)$-dimensional subcube formed by the vectors for which z is true.

A term t is an *implicant* of a DNF $\varphi = t_1 \vee \cdots \vee t_k$ if $t \leq \varphi$. In this case we also say that φ is a *cover* of t, as the union of the cubes $T(t_i)$ covers the cube $T(t)$. Note that the variables occurring in t and φ may differ. It may be assumed w.l.o.g. that by a truth assignment we mean an assignment of truth values to every variable occurring in t or φ. The term t is a *prime implicant* of φ, if t is an implicant of φ, but every term obtained by deleting a literal from t is not an implicant of φ. The DNF φ is a *minimal cover* of the term t, if φ is a cover of t (i.e., t is an implicant of φ), but every DNF obtained from φ by deleting a term is not a cover of t.

Let t be a term and $\varphi = t_1 \vee \cdots \vee t_k$ be a DNF. Every term t_i of φ can be uniquely written in the form

$$t_i = t_i' \wedge t_i'', \tag{1}$$

where t_i' contains all the literals from t_i which also occur in t, and t_i'' contains the remaining literals of t_i.

Given a DNF φ, let $Var(\varphi)$ (resp., $Lit(\varphi)$) denote the set of variables (resp., literals) occurring in any term of φ, and let

$$UL(\varphi) = \{z \in Lit(\varphi) : \bar{z} \notin Lit(\varphi)\} \tag{2}$$

be the set of *unate* literals in φ, i.e., the set of those literals occurring in φ, for which their negation does not occur in φ.

For $a \in \{0,1\}^n$, the vector $a^{(\ell)}$ is the vector obtained from a by flipping its component corresponding to the literal ℓ, e.g., for variables x_1, x_2, x_3, x_4 one has $1010^{(x_2)} = 1110$ and also $1010^{(\bar{x}_2)} = 1110$. Given $a, b \in \{0,1\}^n$, the term corresponding to the smallest subcube containing both a and b is obtained by including every literal corresponding to components where a and b agree. For example, the smallest subcube containing both 1010 and 1100 is $x_1\bar{x}_4$. The Hamming distance $d(a,b)$ of $a,b \in \{0,1\}^n$ is the number of components where a and b differ. The graph of the n-dimensional cube has $\{0,1\}^n$ as vertices, and edges (a,b) for every a,b of Hamming distance 1. The distance of two subcubes C_1 and C_2 is $\min\{d(a,b) : a \in C_1, b \in C_2\}$. Note that the distance of $T(t_1)$ and $T(t_2)$ is equal to the number of conflicts between the terms t_1 and t_2. A partition of the cube into subcubes can also be viewed as a disjoint DNF tautology. A partition of a cube into subcubes is *pairwise neighboring*, if any two subcubes in the partition have distance 1. A set of terms forms a pairwise neighboring partition if the corresponding set of cubes forms a pairwise neighboring partition.

3. Prime implicants and k-term DNF.

In this section we describe the results of [4, 16, 17, 19] on prime implicants of k-term DNF. We give a complete presentation in order to make the paper self-contained, to clarify what are the consequences of the separate assumptions of being an implicant, a prime implicant (resp., a minimal cover), and to give an explicit formulation of results implicit in [16]. We use the notation introduced in (1) and (2).

PROPOSITION 2. *A term t is an implicant of a DNF φ if and only if $\bigvee_{i=1}^k t_i'' = 1$.*

Proof. For the $\Leftarrow$ direction, let a be a truth assignment such that $t(a) = 1$. Then $t_i'(a) = 1$ for every i and $t_i''(a) = 1$ for some i, so $t_i(a) = 1$ for some i, and thus $\varphi(a) = 1$.

For the $\Rightarrow$ direction assume $\bigvee_{i=1}^k t_i'' < 1$, i.e., $(\bigvee_{i=1}^k t_i'')(a) = 0$ for some a. The literals occurring in $\bigvee_{i=1}^k t_i''$ do not occur in t, but it may be the case that the negation of such a literal occurs in t. Let b be the truth assignment obtained from a by setting all the literals of t to 1. Then every literal in $\bigvee_{i=1}^k t_i''$ is either unchanged, or is changed to 0, thus $(\bigvee_{i=1}^k t_i'')(b) = 0$, and so $\varphi(b) = 0$. But $t(b) = 1$, contradicting the fact that t is an implicant of φ. ◻

PROPOSITION 3. *If t is a prime implicant of φ, then*

(a) $t = \bigwedge_{i=1}^k t_i'$,

(b) *every literal of t occurs in φ.*

Proof. For a), it follows from the definition that $t \leq \bigwedge_{i=1}^k t_i'$. Assume that a variable x in t does not occur in any t_i. Then x does not occur in φ at all, though $\bar{x}$ may occur in some t_i''. But then t is an implicant of the disjunction of those terms in φ which do not contain $\bar{x}$, and so by deleting x from t we still get an implicant of φ. Part b) follows trivially from a). ◻

PROPOSITION 4. *If φ is a minimal cover of t, then*

(a) $Lit(t) \cap Lit(\varphi) = UL(\varphi)$,

(b) $\bigvee_{i=1}^k t_i''$ *is a minimal cover of* 1.

Proof. For the $\subseteq$ part of (a), note that if t contains a nonunate literal z of φ, then terms containing $\bar{z}$ can be deleted from φ and we still get a cover of t, contradicting the minimality of φ. For the $\supseteq$ part of (a), assume that a unate literal z is not contained in t. Then $\bar{z}\,t$ is also an implicant of φ, which is covered by the terms of φ not containing z. As these terms do not contain $\bar{z}$ either, their disjunction covers t as well, again contradicting the minimality of φ. Part (b) follows from Proposition 2. ◻

Putting together Propositions 2, 3, and 4, we get the following theorem.

THEOREM 5. *If t is a prime implicant of φ and φ is a minimal cover of t, then*

(a) *t is the conjunction of the literals in $UL(\varphi)$,*

(b) $\bigvee_{i=1}^k t_i''$ *is a minimal cover of* 1.

THEOREM 6 (see [4, 17, 19]). *Every k-term DNF has at most $2^k - 1$ prime implicants.*

Proof. Let φ be a k-term DNF and t be a prime implicant of φ. Consider a minimal set of terms of φ covering t. Then, by Theorem 5(a), t is uniquely determined by this nonempty set of terms. □

The next result gives important structural information on maximal DNF's.

THEOREM 7 (see [16]). *Let $\varphi = t_1 \vee \cdots \vee t_k$ be a k-term DNF with $2^k - 1$ prime implicants, and let t be the term formed by the literals in $UL(\varphi)$.*

Then

(a) $\bigvee_{i=1}^{k} t_i''$ *is a minimal cover of* 1,

(b) t_i'' *and* t_j'' *conflict in exactly one variable, for every* $1 \le i < j \le k$.

Proof. By Theorems 5 and 6, every nonempty subset of the terms of φ is a minimal covering of some prime implicant of φ. Part (a) follows by applying Theorem 5(b) to all the terms.

Let us consider now $\psi_{i,j} = t_i \vee t_j$. Again, this is a minimal cover of a prime implicant of φ. If t_i and t_j do not conflict in any variable, then, by Theorem 5(a), the corresponding prime implicant is the term formed by all the literals in t_i and t_j. But that term is not a prime implicant. Indeed, it must be the case that $t_i \neq t_j$, and so $t_i \wedge t_j < t_i$ or $t_i \wedge t_j < t_j$. If t_i and t_j conflict in more than one variable, then we get a contradiction to Theorem 5(b), as the disjunction of two terms with at least two conflicts cannot be 1. □

4. Proof of Theorem 1.

In this section we prove Theorem 1: a DNF is maximal if and only if it corresponds to a NUD.

First we consider the ⇐ direction.

LEMMA 8. *Every NUD corresponds to a maximal DNF.*

Proof. Let T be a k-NUD, and let H be a nonempty subset of its leaves. Define the term

$$t_H = UL(\{t_\ell : \ell \in H\}).$$

Let a be a truth assignment satisfying t_H. It follows by induction on the number of inner nodes evaluated that on input a we arrive at a leaf belonging to H, and it follows from the definition of t_H that a satisfies every literal assigned to that leaf. Thus t_H is an implicant of φ_T.

Assume that we delete an x literal, say x_i^ϵ from t_H, to get the term t'. As $x_i^\epsilon \in UL(\{t_\ell : \ell \in H\})$, there is a leaf ℓ_1 belonging to H below the ϵ-child of the inner node x_i, but no leaf below the $(1-\epsilon)$-child of x_i is in H. Let a be the vector satisfying all the literals in t_{ℓ_1} and t_H, with every literal of the form $y_j^{\epsilon_j}$ not occurring in these terms set to 0. Let $b = a^{(x_i)}$. On the input b we arrive at a leaf ℓ_2 below the $(1-\epsilon)$-child of x_i. But the y literal assigned to ℓ_2 is set to 0 in b, and hence $\varphi_T(b) = 0$. On the other hand, b still satisfies t'. Thus t' is not an implicant.

Assume now that we delete a y literal, say $y_j^{\epsilon_j}$, from t_H, to get the term t'. Let ℓ be the leaf containing $y_j^{\epsilon_j}$. It follows from the definition of t_H that $\ell \in H$. Let a be a vector satisfying t_ℓ and t_H, and let $b = a^{(y_j)}$. Then the input b leads to ℓ, but as the literal $y_j^{\epsilon_j}$ has value 0 for vector b, we get $\varphi_T(b) = 0$. On the other hand, b still satisfies t'. Thus t' is not an implicant. The case when we delete a z literal, say $z_j^{\delta_j}$, from t_H is the same, except now there may be several leaves in H containing $z_j^{\delta_j}$. We can choose any such leaf and repeat the same argument as for $y_j^{\epsilon_j}$. It again follows that the term obtained after deleting the literal is not an implicant.

Thus the term t_H is a prime implicant of φ_T. Terms corresponding to different subsets of L are different, as each leaf has its unique y literal. Hence φ_T has at least $2^k - 1$ prime implicants, and so it is maximal by Theorem 6. □

The rest of this section contains the proof of the ⇒ direction of Theorem 1.

LEMMA 9. *Every maximal DNF corresponds to a NUD.*

Proof. Let $\varphi = t_1 \vee \cdots \vee t_k$ be a k-term DNF with $2^k - 1$ prime implicants. Consider the term $t = UL(\varphi)$, and the decomposition $t_i = t_i' \wedge t_i''$ of the terms of φ with respect to t, as in (1). According to Theorem 7, the terms $t_1'', \dots, t_k''$ form a pairwise neighboring partition over the nonunate variables occurring in φ, i.e., over $\{0,1\}^s$, where $s = |Var(\varphi) \backslash UL(\varphi)|$. The following lemma states a basic combinatorial property of pairwise neighboring partitions.

LEMMA 10 (splitting lemma). *If a set of $k \geq 2$ terms forms a pairwise neighboring partition, then there is a variable that occurs (unnegated or negated) in every term.*

Proof. We proceed by induction on the number of variables; the case of one or two variables is trivial. Let $u_1, \dots, u_k$ be terms forming a pairwise neighboring partition of $\{0,1\}^s$.

Consider the ℓ half cube corresponding to an arbitrary literal ℓ. The restriction of $u_1, \dots, u_k$ to the ℓ half cube is formed by deleting terms which contain the literal $\overline{\ell}$. It follows directly from the definitions that the restriction gives a pairwise neighboring partition of the ℓ half cube. If the restriction consists of a single cube, then ℓ is a term of the original partition. In this case every other term of the original partition must contain $\overline{\ell}$ and we are done. Hence in what follows we may assume that the restrictions always contain at least two terms.

Applying the induction hypothesis to the pairwise neighboring partition of the $s-1$ dimensional cube obtained by deleting the component corresponding to ℓ, and deleting the literal ℓ from each of the remaining terms, it follows that there is a variable *Split*(ℓ), different from the variable of ℓ, contained (negated or unnegated) in every term covering a point in the ℓ half cube. As there are $2s$ literals and s variables, there are literals ℓ_1 and ℓ_2 such that *Split*$(\ell_1) =$ *Split*$(\ell_2) = z$ for some variable z.

We claim that z occurs (negated or unnegated) in every term of the partition $u_1, \dots, u_k$. If ℓ_1 is the negation of ℓ_2, then z must occur in every term and we are done; henceforth we can assume that ℓ_1 and ℓ_2 have different variables. Assume now for contradiction that z is not in every term of the partition. Let u be a term of the partition containing neither z nor $\overline{z}$, and let a be a point in u. Then a belongs to neither the ℓ_1 subcube, nor the ℓ_2 subcube.

Consider the points $a^{(\ell_1)}$ and $a^{(\ell_2)}$, covered, respectively, by terms u_{ℓ_1} and u_{ℓ_2} of the partition. Note that u_{ℓ_1} and u_{ℓ_2} are different. Indeed, if $u_{\ell_1} = u_{\ell_2} = u'$, then, as $a^{(\ell_1)}$ and $a^{(\ell_2)}$ differ in both their ℓ_1 and ℓ_2 components, u' contains neither ℓ_1 nor ℓ_2, and hence it covers a as well. This contradicts the definition of a.

The points $a^{(\ell_1)}$ and $a^{(\ell_2)}$ differ only in their ℓ_1 and ℓ_2 components; hence the unique conflict of the terms u_{ℓ_1} and u_{ℓ_2} is either ℓ_1 or ℓ_2. Assume w.l.o.g, that the conflict is ℓ_1. By definition, both u_{ℓ_1} and u_{ℓ_2} contain a z literal. As $a^{(\ell_1)}$ and $a^{(\ell_2)}$ do not conflict on z, both u_{ℓ_1} and u_{ℓ_2} contain the same z literal, say z. Thus so far we have that u_{ℓ_1} contains ℓ_1^ϵ and z, and u_{ℓ_2} contains $\ell_1^{1-\epsilon}$ and z, for some $\epsilon \in \{0,1\}$.

Now consider the point $a^{(\ell_1,z)}$ covered by the term $u_{\ell_1,z}$ of the partition. As $a^{(\ell_1,z)}$ is in the ℓ_1 subcube, it contains a z literal, which must be $\overline{z}$. What is the unique conflict of u (the term covering a) and $u_{\ell_1,z}$? As $a^{(\ell_1,z)}$ and a conflict only on their ℓ_1 and z components, but u contains no z literal, it must be ℓ_1. Thus $u_{\ell_1,z}$ contains ℓ_1^ϵ and $\overline{z}$. But then u_{ℓ_2} and $u_{\ell_1,z}$ conflict in at least two components, a contradiction. □

The splitting lemma is now used to prove the characterization of nonrepeating decision tree tautologies mentioned in the introduction.

LEMMA 11 (see the ND Lemma [14]). *A set of $k \geq 2$ terms forms a pairwise neighboring partition if and only if it is an ND.*

Proof. We begin by applying Lemma 10 to the pairwise neighboring partition to get a variable x_1 occurring in every term. It must be the case that x_1 occurs both unnegated and negated, as otherwise the cubes would not cover the whole cube. If the x_1^ϵ half cube contains just one cube, then we stop at that branch, otherwise we use the lemma again to get a variable which occurs in every subcube of the partition, belonging to the x_1^ϵ half cube, etc. In this way we get a tree, where the inner nodes are labeled with variables and there are k leaves $\ell_1, \ldots, \ell_k$ corresponding to the cubes in the partition. (The tree constructed is (the dual of) a special *search tree* in the sense of [18] for the partition.) The labels of the inner nodes are *different*, as the same label appearing twice would mean that some pair of cubes have distance at least 2. Indeed, if variable x_i occurs twice, then let x_j be the variable labeling the least common ancestor of the two occurrences in the tree. By construction, there are terms containing $\bar{x}_i\,\bar{x}_j$ (resp., $x_i\,x_j$). Thus the partition is an ND. □

Now we can complete the proof of Lemma 9. Lemma 11 gives an ND for the pairwise neighboring terms $t_1'', \ldots, t_k''$. We claim that by adding the literals in t_i' to the leaf ℓ_i, we get a k-NUD for φ. Consider any truth assignment a to the variables in φ. Evaluating the tree on a, we arrive at a leaf corresponding to a term t_i''. As $\varphi(a) = 1$ if and only if $t_i'(a) = 1$, the tree computes φ correctly. By construction, all the literals in the leaves are unate. Thus, in order to verify the NUD-ity of the tree, it only remains to show that for every leaf there is a literal which occurs only in that leaf (that literal will be its y literal). Assume that this is not the case, and every (unate) literal assigned to leaf ℓ_i occurs in some other leaf. Let x_j^ϵ be the last literal on the path leading to ℓ_i. Then $x_j^{1-\epsilon} \in UL(\varphi \setminus \{t_i\})$. We claim that $UL(\varphi \setminus \{t_i\}) \setminus \{x_j^{1-\epsilon}\}$ is an implicant of φ. Let a be a truth assignment satisfying every literal in $UL(\varphi \setminus \{t_i\}) \setminus \{x_j^{1-\epsilon}\}$, and let us evaluate the tree on a. If we arrive at a leaf other than ℓ_i, then $\varphi(a) = 1$ by construction. But $\varphi(a) = 1$ if we arrive at ℓ_i as well, as all unate literals in ℓ_i occur in other leaves, and thus they must be set to 1 in a. Thus $UL(\varphi \setminus \{t_i\})$ is not a prime implicant of φ, contradicting Theorems 5 and 6. □

5. The general splitting problem for cube partitions.

According to the splitting lemma (Lemma 10), for every pairwise neighboring cube partition, the whole cube can be split into two halves in such a way that every cube of the partition is contained in one of the halves. In this section we consider the following question: What can be said without the pairwise neighboring property? Given an arbitrary partition of the whole cube into subcubes and a split into two halves, let us say that a cube in the partition is good if it is contained in either one of the halves. We would like to find a split such that the good cubes contain many points.

Thus we consider the following quantities. Given a cube partition φ over the variables $x_1, \ldots, x_n$ and a variable x_j, let

$$v_{\varphi,j} = \sum \left\{ 2^{-|t|} : t \in \varphi,\ x_j \in t \text{ } or \text{ } \bar{x}_j \in t \right\}$$

be the fraction of the volume of good cubes in φ with respect to the x_j split of the cube, and let

$$\alpha_n = \min_{\varphi} \max_{1\le j\le n} v_{\varphi,j},$$

where φ ranges over all cube partitions, or in other words, over all disjoint DNF tautologies. Note that as φ is a partition it holds that

$$\sum_{t\in\varphi} 2^{-|t|} = 1. \tag{3}$$

Theorem 12. *It holds that*

$$\frac{\log n - \log\log n}{n} \le \alpha_n \le O\left(n^{-\frac{1}{5}}\right).$$

Proof. Let $\varphi = t_1 \vee \cdots \vee t_r$ be a disjoint DNF tautology over the variables $x_1, \ldots, x_n$. If the term t_i contains x_j or $\bar{x}_j$, then t_i contributes $2^{-|t_i|}$ to $v_{\varphi,j}$. Thus

$$\sum_{j=1}^{n} v_{\varphi,j} = \sum_{i=1}^{r} |t_i| \cdot 2^{-|t_i|},$$

and there is a variable x_j with

$$v_{\varphi,j} \ge \frac{1}{n} \sum_{i=1}^{r} |t_i| \cdot 2^{-|t_i|}.$$

Let s denote the size of the shortest term in φ. As every term has size at least s, it follows from (3) that

$$\frac{1}{n} \sum_{i=1}^{r} |t_i| \cdot 2^{-|t_i|} \ge \frac{s}{n} \sum_{i=1}^{r} 2^{-|t_i|} = \frac{s}{n}.$$

On the other hand, for every variable x_j occurring in a shortest term t_i it holds that $v_{\varphi,j} \ge 2^{-s}$. Thus

$$\alpha_n \ge \min\left(\frac{s}{n},\, 2^{-s}\right). \tag{4}$$

The lower bound then follows by taking $s = \log n - \log\log n$, for which the two terms in (4) are close to each other.

The upper bound follows from a construction of Savický and Sgall [21], providing an upper bound on the number of variable occurrences in tautological k-DNF formulas (a problem introduced by Tovey [23] and Kratochvíl, Savický and Tuza [12]). They constructed disjoint DNF tautologies over $n = 4^{\ell}$ variables, having $2^{3^{\ell}}$ terms of size 3^{ℓ}, such that every variable occurs in at most a $(3/4)^{\ell}$ fraction of the terms. The bound then follows by a direct calculation. □

We note that the upper bound of Savický and Sgall [21] has recently been improved almost optimally by Hoory and Szeider [10]. The improved constructions do not appear to improve the upper bound, since the DNF constructed are not disjoint.

In view of Theorems 1 and 12, it may be of interest to consider the quantity α_n^d, which is defined as α_n, except that φ is restricted to cube partitions with pairwise distances bounded by d. In the construction of [21] the maximal distance grows linearly with n.

6. Partitions of complete graphs into complete bipartite graphs. Given a set of pairwise disjoint cubes in $\{0,1\}^n$, corresponding to terms $t_1, \dots, t_r$, one can construct a covering

$$\mathcal{G} = \{G_1, \dots, G_n\}$$

of the r-vertex complete graph K_r by complete bipartite graphs, where G_u has an edge connecting vertices v_i and v_j if terms t_i and t_j conflict in the variable x_u. If the set of cubes is pairwise neighboring, then this covering is a partition, as the complete bipartite graphs are edge disjoint.

Conversely, given a covering $\mathcal{G} = \{G_1, \dots, G_n\}$ of K_r by complete bipartite graphs, we can construct a set of pairwise disjoint cubes $t_1, \dots, t_r$ in $\{0,1\}^n$. For every G_u, fix arbitrarily one of the sides as the left side. The term t_i contains x_u (resp., $\bar{x}_u$), if vertex v_i is contained in the left (resp., right) side of G_u. If $\mathcal{G}$ is a partition, then it follows that the t_i's are pairwise neighboring. The cubes thus constructed do not necessarily form a partition of $\{0,1\}^n$ (an example is given below).

The Graham–Pollak theorem [6] states that every partition of K_r into complete bipartite graphs consists of at least $r-1$ graphs. A large class of such partitions, which can be called *recursive* partitions, is obtained as follows. Take a complete bipartite graph on the whole vertex set. This "takes care" of all edges connecting the two sides. In order to partition the remaining edges (those having both endpoints on the same side), repeat the same construction, i.e., recursively add similar partitions of the complete graphs formed by the two sides of this bipartite graph (see, e.g., [3]).

Consider a partition $\mathcal{G} = \{G_1, \dots, G_n\}$ of K_r into complete bipartite graphs. Let the degree of a vertex v with respect to $\mathcal{G}$, denoted by $d_{\mathcal{G}}(v)$, be the number of G_i's containing v, and let the *volume* $vol(\mathcal{G})$ of the partition be defined as

$$vol(\mathcal{G}) = \sum_v 2^{-d_{\mathcal{G}}(v)}.$$

In view of the translation into a set of pairwise disjoint cubes in $\{0,1\}^n$ described above, $vol(\mathcal{G}) \le 1$ for every $\mathcal{G}$, as $d_{\mathcal{G}}(v_i) = |t_i|$ for every $i = 1, \dots, r$, and $vol(\mathcal{G}) = 1$ if and only if the cubes form a partition of $\{0,1\}^n$. For example, the partition of K_4 into the three complete bipartite graphs $(\{1\},\{3,4\})$, $(\{2\},\{1,4\})$, and $(\{3\},\{2,4\})$ (mentioned in [16]) has volume $\frac{7}{8}$. This partition of K_4 is not recursive. (It was actually this example which suggested Lemma 10.) As a corollary to the splitting lemma (Lemma 10) one gets the following characterization of recursive partitions. This characterization is also a direct consequence of Kullmann's [13, 14, 15] results.

COROLLARY 13. *A partition $\mathcal{G}$ is recursive if and only if $vol(\mathcal{G}) = 1$.*

Proof. The ⇒ direction follows directly by induction on the number of vertices by considering the bipartite graph from $\mathcal{G}$ which contains all the vertices.

For the ⇐ direction, one only has to note that the set of terms $t_1, \dots, t_r$ constructed above is pairwise neighboring, and by the volume condition it is also a partition of the whole cube.

Applying Lemma 10 we get that there is a variable which occurs (unnegated or negated) in every term. This means that the corresponding bipartite graph contains all the r vertices. The remaining partitions of the two sides of this bipartite graph have total volume 2, and thus each side must have volume 1. The statement then follows by induction. □

The corollary shows that among partitions of K_r into complete bipartite graphs, recursive ones have the largest possible volume. Among the partitions of K_r into $r-1$ complete bipartite graphs, which ones have *minimal* volume?

7. Other open problems.

In this paper we have discussed k-term DNF with the largest number of prime implicants. Similar results do not appear to be known for *shortest* prime implicants, i.e., prime implicants containing the smallest possible number of literals. The k-term DNF

$$x_1\bar{x}_2 \vee x_2\bar{x}_3 \vee \cdots \vee x_{k-1}\bar{x}_k \vee x_k\bar{x}_1,$$

which is false for 0^k and 1^k, and true everywhere else, has $k(k-1)$ prime implicants, namely $x_i\bar{x}_j$ for every $i \neq j$. These prime implicants are all shortest prime implicants, as the DNF has no prime implicants consisting of a single literal. How many shortest prime implicants can a k-term DNF have in general?

Another question concerns the maximal number of prime implicants of a Boolean function which is true at a given number of points. As noted by Levin [17], every implicant is determined by the top and bottom of the corresponding subcube, in the componentwise partial ordering of the hypercube (the top and bottom may also be identical). Thus if a function is true at m points, then it has $O(m^2)$ prime implicants. It is also noted in [17] that the n-variable function which is true for vectors of weight between $\frac{n}{3}$ and $\frac{2n}{3}$, has $m^{\log 3 - o(1)}$ prime implicants. (This is the function with the largest known number of prime implicants among n-variable functions.) Thus the maximal number of prime implicants is bounded by two polynomial functions of m, and the question is to get sharper bounds.

Acknowledgments. We are especially grateful to Jiri Sgall for providing an essential simplification of the proof of the splitting lemma (Lemma 10). Also, we would like to thank the referees for their extremely useful comments.

REFERENCES

[1] R. Aharoni and N. Linial, *Minimal non-two-colorable hypergraphs and minimal unsatisfiable formulas*, J. Combin. Theory Ser. A, 43 (1986), pp. 196–204.

[2] N. H. Arai, T. Pitassi, and A. Urquhart, *The complexity of analytic tableaux*, in Proceedings of the 33rd Annual ACM Symposium on Theory Computing (STOC), 2001, ACM, New York, pp. 356–363.

[3] L. Babai and P. Frankl, *Linear Algebra Methods in Combinatorics*, preliminary version 2, available from University of Chicago Computer Science Dept., Chicago, IL, 1992.

[4] A. K. Chandra and G. Markowsky, *On the number of prime implicants*, Discrete Math., 24 (1978), pp. 7–11.

[5] G. Davydov, I. Davydova, and H. Kleine Büning, *An efficient algorithm for the minimal unsatisfiability problem for a subclass of CNF*, Ann. Math. Artificial Intelligence, 23 (1998), pp. 229–245.

[6] R. L. Graham and H. O. Pollak, *On the addressing problem for loop switching*, Bell System Tech. J., 50 (1971), pp. 2495–2519.

[7] D. A. Gregory, V. L. Watts, and B. L. Shader, *Biclique decompositions and Hermitian rank*, Linear Algebra Appl., 292 (1999), pp. 267–280.

[8] P. L. Hammer, A. Kogan, and U. G. Rothblum, *Evaluation, strength and relevance of variables of Boolean functions*, SIAM J. Discrete Math., 13 (2000), pp. 302–312.

[9] L. Hellerstein and V. Raghavan, *Exact learning of DNF formulas using DNF hypotheses*, J. Comput. System Sci., 70 (2005), pp. 435–470.

[10] S. Hoory and S. Szeider, *A note on unsatisfiable k-CNF formulas with few occurrences per variable*, SIAM J. Discrete Math., 20 (2006), pp. 523–528.

[11] J. Kahn, G. Kalai, and N. Linial, *The influence of variables on Boolean functions*, in Proceedings of the 29th Annual Symposium on Foundations of Computer Science (FOCS), White Plains, NY, 1988, pp. 68–80.

[12] J. Kratochvíl, P. Savický, and Zs. Tuza, *One more occurrence of variables makes satisfiability jump from trivial to NP-complete*, SIAM J. Comput., 22 (1993), pp. 203–210.

[13] O. Kullmann, *An application of matroid theory to the SAT problem*, in 15th IEEE Conference on Computational Complexity, Florence, Italy, 2000, pp. 116–124.

[14] O. Kullmann, *The combinatorics of conflicts between clauses*, in Theory and Application of Satisfiability Testing, Lecture Notes in Comput. Sci. 2919, Springer, 2003, pp. 426–440.

[15] O. Kullmann, *On the Conflict Matrix of Clause-Sets*, Technical report CSR 7-2003, University of Wales at Swansea, Swansea, UK, 2003.

[16] J.-M. Laborde, *Sur le cardinal maximum de la base complète d'une fonction booléenne, en fonction du nombre de conjunctions de l'une de ses formes normales*, Discrete Math., 32 (1980), pp. 209–212.

[17] A. A. Levin, *Comparative complexity of disjunctive normal forms*, Metody Diskret. Analiz., 36 (1981), pp. 23–38 (in Russian).

[18] L. Lovász, M. Naor, I. Newman, and A. Wigderson, *Search problems in the decision tree model*, SIAM J. Discrete. Math., 8 (1995), pp. 119–132.

[19] C. McMullen and J. Shearer, *Prime implicants, minimum covers, and the complexity of logic simplification*, IEEE Trans. Comput. C-35 (1986), pp. 761–762.

[20] J. E. Savage, *Models of Computation: Exploring the Power of Computing*, Addison-Wesley, Reading, MA, 1998.

[21] P. Savický and J. Sgall, *DNF tautologies with a limited number of occurrences of every variable*, Theoret. Comput. Sci., 238 (2000), pp. 495–498.

[22] R. H. Sloan, B. Szörényi, and Gy. Turán, *Projective DNF formulae and their revision*, in Learning Theory and Kernel Machines, 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, 2003, Proceedings, in: Lecture Notes in Artificial Intelligence 2777, Springer, Berlin 2003, pp. 625–639.

[23] C. A. Tovey, *A simplified NP-complete satisfiability problem*, Discrete Appl. Math., 8 (1984), pp. 85–89.

[24] A. Urquhart, *The complexity of propositional proofs*, Bull. Symbolic Logic, 1 (1995), pp. 425–467.

[25] I. Wegener, *The Complexity of Boolean Functions*, Wiley–Teubner, Chichester, UK, Stuttgart, Germany, 1987.

SIAM J. DISCRETE MATH.
Vol. 21, No. 4, pp. 999–1018

COLORING BULL-FREE PERFECTLY CONTRACTILE GRAPHS*

BENJAMIN LÉVÊQUE† AND FRÉDÉRIC MAFFRAY‡

Abstract. We consider the class of graphs that contain no bull, no odd hole, and no antihole of length at least five. We present a new algorithm that colors optimally the vertices of every graph in this class. This algorithm is based on the existence in every such graph of an ordering of the vertices with a special property. More generally we prove, using a variant of lexicographic breadth-first search, that in every graph that contains no bull and no hole of length at least five there is a vertex that is not the middle of a chordless path on five vertices. This latter fact also generalizes known results about chordal bipartite graphs, totally balanced matrices, and strongly chordal graphs.

Key words. perfect graph, bull-free graph, coloring, LexBFS

AMS subject classifications. 05C15, 05C17, 05C85

DOI. 10.1137/06065948X

1. Introduction.

The chromatic number of a graph G is the smallest integer $\chi(G)$ for which it is possible to assign one color from the set $\{1, \ldots, \chi(G)\}$ to each vertex so that any two adjacent vertices receive different colors. A graph G is *perfect* if the chromatic number of every induced subgraph H of G is equal to $\omega(H)$, where $\omega(H)$ is the maximum clique size in H. A *hole* is a chordless cycle on at least four vertices. The complement of a hole is called an *antihole.* A hole or an antihole is odd if it has an odd number of vertices. Graphs that do not contain an odd hole or an odd antihole of length at least five are usually called Berge graphs. Berge [2, 3] conjectured that such graphs are perfect, and this famous problem, known as the strong perfect graph conjecture, was solved by Chudnovsky et al. [5]. Earlier, Grötschel, Lovász, and Schrijver [15] gave a polynomial time algorithm that computes the chromatic number of every perfect graph; but this algorithm, based on the ellipsoid method, is considered very impractical, and it is still an open problem to find a purely combinatorial algorithm to color optimally the vertices of all perfect graphs in polynomial time. Here we consider the class of bull-free graphs.

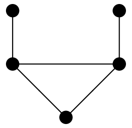


FIG. 1. *The bull.*

A *bull* is a graph with five vertices a, b, c, d, e and edges ab, bc, cd, de, bd; see Figure 1. We will denote such a bull by *a-bcd-e*. In a bull *a-bcd-e*, we call the edge bd the *central* edge and vertices b, d the *ears* of the bull. Chvátal and Sbihi [8] proved that

*Received by the editors May 10, 2006; accepted for publication (in revised form) May 11, 2007; published electronically January 9, 2008. This research was supported by the ADONET network of the European Community, which is a Marie Curie Training Network.

http://www.siam.org/journals/sidma/21-4/65948.html

†Laboratoire G-SCOP, 46 avenue Félix Viallet, 38031 Grenoble Cedex, France (benjamin.leveque@g-scop.inpg.fr).

‡C.N.R.S., Laboratoire G-SCOP, 46 avenue Félix Viallet, 38031 Grenoble Cedex, France (frederic.maffray@g-scop.inpg.fr).

the strong perfect graph conjecture holds for bull-free graphs, that is, every bull-free Berge graph is perfect. Subsequently, the structure of bull-free Berge graphs was also studied by Reed and Sbihi [31]; De Figueiredo, Maffray, and Porto [10, 11]; and Hayward [18]. De Figueiredo and Maffray [9] gave a combinatorial algorithm, based on the results from [8, 10], that optimally colors every bull-free Berge graph G with n vertices and m edges in time $\mathcal{O}(n^5m^3)$.

Let $\mathcal{B}$ be the class of bull-free Berge graphs that contain no antihole of length at least five. We will present an $\mathcal{O}(mn)$ algorithm that computes an optimal coloring for every graph in class $\mathcal{B}$. This algorithm is based on new structural results concerning the graphs in that class. Before doing so, we want to review the known methods that perform such a task, and for this purpose we need to introduce a few more definitions.

A graph G is *weakly chordal* [17] if G contains no hole of length at least five and no antihole of length at least five. A graph G is *transitively orientable* [14, 28] if we can assign one orientation to each of its edges so that for every directed path $u \rightarrow v \rightarrow w$ the arc $u \rightarrow w$ is present in the orientation. A graph G is *perfectly orderable* [6] if it admits an ordering $<$ such that, for every induced subgraph H of G, applying the greedy coloring algorithm on $(H, <)$ produces an optimal coloring (such an ordering is called a perfect ordering). A *homogeneous set* in a graph G is a set $S \subset V(G)$ with $|S| \geq 2$, $S \neq V(G)$, such that every vertex of $V(G) \setminus S$ is adjacent to either all or none of the vertices of S. A *prism* is a graph that consists in two disjoint triangles and three disjoint paths between the two triangles, with no edge between any two of these three paths other than the triangles' edges. A prism is odd if these three paths have odd length. A graph G is an *Artemis* graph [12] if it contains no odd hole, no antihole of length at least five, and no prism. A graph G is a *Grenoble* graph [12] if it contains no odd hole, no antihole of length at least five, and no odd prism. It was proved in [10] that every graph in class $\mathcal{B}$ is "perfectly contractile" in the sense of Bertschi [4]; see section 5. Note that a prism either is the complement of a cycle of length six or contains a bull. Therefore, "bull-free Artemis," "bull-free Grenoble," and "bull-free perfectly contractile" are just different names for class $\mathcal{B}$.

We know of three purely combinatorial methods to color graphs in class $\mathcal{B}$, which we summarize briefly:

• Method 1: Results from [10, 11] say that every graph in class $\mathcal{B}$ either is weakly chordal, or has a homogeneous set, or is transitively orientable. Homogeneous sets can be handled by the so-called *modular decomposition*, which decomposes any graph into $\mathcal{O}(n)$ subgraphs that have no homogeneous sets. Modular decomposition can be performed in time $\mathcal{O}(n + m)$; see, for example, [16]. By [10, 11], for a graph in class $\mathcal{B}$, these indecomposable subgraphs are either weakly chordal or transitively orientable. One can find an optimal coloring for these subgraphs in time $\mathcal{O}(nm)$ for weakly chordal graphs [19] and in time $\mathcal{O}(m)$ for transitively orientable graphs [27]. One can then combine these optimal colorings along the modular decomposition to obtain an optimal coloring of the original graph (details are omitted). Thus we can estimate the complexity of this method at $\mathcal{O}(n^2m)$.

• Method 2: Chvátal [7] conjectured that every graph in class $\mathcal{B}$ is perfectly orderable, and Hayward [18] proved that conjecture, using some results from [10, 11]. We estimate the technique in [18] at $\mathcal{O}(n^5)$ (the exponent 5 is due to the search for an induced P_5 performed in [11]), and so, combining the techniques in [10, 11, 18], and using again a linear-time algorithm for modular decomposition such as [16], one can find a perfect ordering of any graph in class $\mathcal{B}$ in time $\mathcal{O}(n^5(n + m))$. Then applying the greedy coloring on this ordering produces an optimal coloring in time $\mathcal{O}(m)$. Thus the total complexity of this method can be estimated at $\mathcal{O}(n^5(n+m))$.

• Method 3: Since every graph in class $\mathcal{B}$ is an Artemis graph, one can use the algorithm from [25], which colors every Artemis graph in time $\mathcal{O}(n^2m)$.

Our aim here is to present an algorithm that we think is conceptually simpler than all of the above and whose complexity is also lower.

First let us fix some terminology and notation. We say that a vertex *a sees* a vertex b when ab is an edge of the graph, otherwise vertex *a misses b*. The complement of a graph G is denoted by $\overline{G}$. The neighborhood of a vertex v is denoted by $N(v)$. The degree of a vertex v in G is denoted by $d(v)$. A chordless path on k vertices is denoted by P_k. A *house* is a graph with five vertices a, b, c, d, e and edges ac, ce, eb, bd, da, ae; vertex c is called the *top* of the house. Note that a house is the complement of a P_5. We will establish the following result.

THEOREM 1.1. *Every graph in $\mathcal{B}$ has a vertex that is not the top of a house.*

The above theorem implies the following. Let G be any graph in $\mathcal{B}$. So G has a vertex v_1 that is not the top of a house, and for $i = 2, \dots, n$, the subgraph $G \setminus \{v_1, \dots, v_{i-1}\}$ has a vertex v_i that is not the top of a house in this subgraph. We may call the ordering $v_1, \dots, v_n$ of the vertices of G an *NTH elimination ordering.* In section 3 we show how such an ordering can be computed in time $\mathcal{O}(nm)$, using the algorithm described in section 2. After such an ordering is obtained, we run an $\mathcal{O}(nm)$ coloring algorithm called COSINE*, which is a new algorithm based on Hertz's coloring algorithm COSINE [21]. Algorithm COSINE works on a graph whose vertices need not be ordered, while COSINE* uses the NTH elimination ordering. In section 5 we prove the optimallity of this coloring algorithm for every graph in $\mathcal{B}$. In section 6 we present an extension of this algorithm that finds a clique of maximum size in a graph in $\mathcal{B}$. This yields an $\mathcal{O}(nm)$ robust algorithm to color graphs in $\mathcal{B}$.

Let $\mathcal{C}$ be the class of graphs that contain no bull and no hole of length at least five. Clearly $\mathcal{B}$ is strictly contained in $\overline{\mathcal{C}}$, and Theorem 1.1 is an immediate consequence of the following.

THEOREM 1.2. *Every graph in $\mathcal{C}$ has a vertex that is not the middle of a P_5.*

The above theorem will be proved in section 3. Note that this theorem implies the following. Let G be any graph in $\mathcal{C}$. So G has a vertex v_1 that is not the middle of a P_5, and for $i = 2, \dots, n$, the subgraph $G \setminus \{v_1, \dots, v_{i-1}\}$ has a vertex v_i that is not the middle of a P_5 in this subgraph. We may call the ordering $v_1, \dots, v_n$ of the vertices of G an *NMP$_5$ elimination ordering.* The proof of Theorem 1.2 is an $\mathcal{O}(nm)$ algorithm called LEXBFS* that finds such an ordering.

We mention a theoretical consequence of this theorem. Recall that a graph is *chordal bipartite* if it is bipartite and it contains no hole of length at least six. A classical result is the existence in every chordal bipartite graph of a vertex that is not the middle of a P_5. This result is known under several equivalent variants, such as the existence of a *simple* vertex in every *strongly chordal* graph, or the existence of a Γ*-free* ordering in every *totally balanced* matrix [26]. Since every chordal bipartite graph is in class $\mathcal{C}$, our Theorem 1.2 generalizes this result.

2. Algorithm LEXBFS*.

Algorithm LEXBFS* is a particular case of Algorithm LEXBFS (lexicographic breadth-first search). Algorithm LEXBFS, due to Rose, Tarjan, and Lueker [32], explores a graph and numbers its vertices one by one, from n to 1. At the general step, each unnumbered vertex has a label, which is the set of numbers of its already numbered neighbors. A lexicographic order is defined on the labels: label $L(a)$ is strictly greater than label $L(b)$ if there exists an integer i such that $i \in L(a) \setminus L(b)$ and $\forall j > i$, either $j \in L(a) \cap L(b)$ or $j \notin L(a) \cup L(b)$. The next vertex to be numbered is any unnumbered vertex whose label is lexicographically maximal. Ties in LexBFS are broken arbitrarily.

In LexBFS*, we need to break ties according to the following rule. Suppose that at a given step the set A of unnumbered vertices with maximal label satisfies $|A| \geq 2$. Let $L(A)$ be the label of the vertices in A. Let U be the set of unnumbered vertices not in A. For each $u \in U$, set $L'(u) := L(u) \setminus L(A)$, and let the vertices of U be ordered lexicographically according to L'. Then the first (i.e., maximal according to the L' ordering) vertex u of U "votes" by eliminating from A the nonneighbors of u (except if that causes A to become empty; in that case u has no effect); then the second vertex of U votes, etc. The procedure stops when all vertices of U have voted; then ties are broken arbitrarily. Here is a formal description of the algorithm:

```
Algorithm LexBFS*
Input: A graph $G$ with $n$ vertices.
Output: An ordering $\sigma$ on the vertices of $G$.
Initialization: For every vertex $a$ of $G$, set $L(a) := \emptyset$;
General step: For $i = n, \dots, 1$ do:
1. Let $A$ be the set of unnumbered vertices whose label is maximum,
and let $U$ be the other unnumbered vertices.
2. While $U \neq \emptyset$ do:
  2.1. Select a vertex $u \in U$ for which $L(u) \setminus L(A)$ is maximum.
  2.2. Set $U := U \setminus \{u\}$. If $A \cap N(u) \neq \emptyset$, then set $A := A \cap N(u)$.
3. Pick any vertex $a \in A$ and set $\sigma(a) := i$.
4. For each unnumbered neighbor $v$ of $a$, add $i$ to $L(v)$.
```

Complexity analysis. Let us analyze the complexity of Algorithm LexBFS*. Rose, Tarjan, and Lueker [32] showed that Algorithm LexBFS can be implemented in time $\mathcal{O}(n+m)$ as follows, where n is the number of vertices and m the number of edges of the graph in input. Ordering the vertices according to the value of $L(v)$ can be done with the usual techniques, such as bucket sort [1]: For each label ℓ, we maintain the set S_ℓ of the unnumbered vertices v such that $L(v) = \ell$. This set is implemented as a doubly linked list, where each element also points to the head of the list, which is a special cell containing their label. The heads of the nonempty S_ℓ's are themselves put in decreasing lexicographic label order into a doubly linked list M. During the initialization step, all vertices are put into $S_\emptyset$, and $S_\emptyset$ is the only element of M. Thus the initialization takes time $\mathcal{O}(n)$. Set A of step 1 of the algorithm is the first set in M. When a vertex a of A is selected at step 3, it is removed from the data structure, and each neighbor u of a is removed from the set S_ℓ that contains u and added into a (new) set $S_{\ell \cup \{\sigma(a)\}} = S_\ell \cap N(A)$ which is placed just before S_ℓ in M (empty sets are removed from M). This operation of splitting the S_ℓ's takes time $\mathcal{O}(d(a))$. So the total cost of steps 3 and 4 is $\mathcal{O}(n+m)$. This is how LexBFS is implemented in [32].

Unfortunately, breaking the ties in LexBFS* increases the complexity to $\mathcal{O}(nm)$ as we show now. Consider the set U defined on line 1 of the algorithm. Set U is ordered according to $L'(u)$ by using the same data structure as before. This takes time $\mathcal{O}(n+m)$. This ordering procedure is performed only once, at the beginning of step 2. Then, at step 2.1 we take the maximum vertex u in the ordered set U (which takes constant time), and the operations performed in step 2.2 take time $\mathcal{O}(d(u))$. So the total cost of step 2 is $\mathcal{O}(n+m)$. Since this step is performed n times, the total running time of Algorithm LexBFS* is $\mathcal{O}(n(n+m))$.

Actually, we will need to apply Algorithm LexBFS* on the complement $\overline{G}$ of a graph G. Let $\overline{m}$ be the number of edges in $\overline{G}$. Since $\overline{m} = \mathcal{O}(n^2)$, this might lead to a complexity of $\mathcal{O}(n^3)$, but we can avoid this as follows. When applied on $\overline{G}$, splitting the sets S_ℓ take time $\mathcal{O}(\overline{d}(a))$, where $\overline{d}$ is the degree function in $\overline{G}$, but we can do it in time $\mathcal{O}(d(a))$ if, instead of removing each neighbor u of a (in $\overline{G}$) from the set S_ℓ that contains u and adding it into the new set $S_\ell \cap N_{\overline{G}}(A)$, we remove each neighbor u of a (in G) from the set S_ℓ that contains u and add it into a new set $S_\ell \setminus N_G(A)$, which is placed just after S_ℓ in M. The same idea can be used to sort the set U and to update A in time $\mathcal{O}(n+m)$. In conclusion, the total running time of Algorithm LexBFS* applied on the complement $\overline{G}$ of a graph G with n vertices and m edges is $O(nm)$.

Properties of LexBFS. Here are some notation and properties for Algorithm LexBFS. When the algorithm selects a vertex $a \in A$ at step 3 of Algorithm LexBFS, we denote by $L_a(u)$ the current value of the label of any vertex u at this step of the algorithm. We denote by $a < b$ the fact that $\sigma(a) < \sigma(b)$.

Lemma 2.1. *Suppose that $a < u$, $b \leq u$, and $L_u(a) < L_u(b)$. Then $a < b$ and, $\forall v$ such that $v \leq u$, $L_v(a) < L_v(b)$.*

Proof. Suppose $a < u$, $b \leq u$, and $L_u(a) < L_u(b)$. At the step of the algorithm when u is numbered, there exists $i > \sigma(u)$ such that $i \in L_u(b) \setminus L_u(a)$ and $\forall j > i$, either $j \in L_u(a) \cap L_u(b)$ or $j \notin L_u(a) \cup L_u(b)$. After u is numbered, integers that may be added to $L(a)$ and $L(b)$ are smaller than $\sigma(u)$ and therefore strictly smaller than i, so the inequality $L(a) < L(b)$ still holds throughout the rest of the execution of the algorithm. Thus the lemma holds. □

Lemma 2.2. *Suppose that $a < b$ and $L_b(a) \neq L_b(b)$. Then there exists a vertex $> b$ that sees b and misses a. Let $f(b,a)$ be a maximum such vertex. Then we have the following properties:*

- *For every u that sees a and misses b, we have $u < f(b,a)$.*
- *Every u such that $f(b,a) < u$ either sees both a, b or misses both a, b.*

Proof. Suppose $a < b$ and $L_b(a) \neq L_b(b)$. Then $L_b(a) < L_b(b)$ because b is selected before a. Then there exists i such that $i \in L_b(b) \setminus L_b(a)$ and $\forall j > i$, either $j \in L_b(a) \cap L_b(b)$ or $j \notin L_b(a) \cup L_b(b)$. Vertex $f(b,a)$ is the vertex such that $\sigma(f(b,a)) = i$.

Suppose a vertex u sees a, misses b, and $u > f(b,a)$. Let $j = \sigma(u)$. Since u sees a, we have $j \in L_b(a)$. Since u misses b, we have $j \notin L_b(b)$. So $j \in L_b(a) \setminus L_b(b)$, a contradiction to the definition of i.

Let u' be a vertex such that $f(b,a) < u'$. Let $j' = \sigma(u')$. Since $j' = \sigma(u') > \sigma(f(b,a)) = i$, we have $j \in L_b(a) \cap L_b(b)$ or $j \notin L_b(a) \cup L_b(b)$, and so u' either sees both a, b or u' misses both. Thus the lemma holds. □

Lemma 2.3. *Suppose that $a < b < u$, and u sees a and misses b. Let $a_0 = a$, $b_0 = b$, $a_1 = u$, $b_1 = f(b,a)$, and define vertices a_i and b_i, for $i \geq 2$, as follows, as long as possible:*

- *If b_i misses a_i, then let $a_{i+1} = f(a_i, b_{i-1})$.*
- *If a_{i+1} misses b_i, then let $b_{i+1} = f(b_i, a_i)$.*

Let k be the maximum integer such that a_k is defined. Let ℓ be the maximum integer such that b_ℓ is defined, so ℓ is equal to k or $k+1$. Denote by $\mathcal{P}(u,b,a)$ the path a_0-$\cdots$-a_k-b_ℓ-$\cdots$-b_0. If a misses b, then $\mathcal{P}(u,b,a)$ is a chordless path. If a sees b, then $\mathcal{P}(u,b,a)$ is a hole.

Proof. Suppose $\ell = k$ for convenience (the same can be done when $\ell = k+1$). We prove by induction on $j \leq k$ the property that the sequences $(a_i)_{i \leq j}$, $(b_i)_{i \leq j}$ are well defined, $a_0 < b_0 < a_1 < b_1 < \cdots < a_j < b_j$, a_0-$\cdots$-a_j and b_0-$\cdots$-b_j are chordless paths, and there is no edge between the (a_i)'s and the (b_i)'s, except for $a_k b_k$ and possibly a_0b_0.

If $j = 1$, then a_1 sees a_0, misses b_0, and $a_0 < b_0 < a_1$, so $L_{b_0}(a_0) \neq L_{b_0}(b_0)$. So vertex $b_1 = f(b_0, a_0)$ is well defined by Lemma 2.2. Vertex b_1 sees b_0, misses a_0, and $a_1 < b_1$. So the property is true for $j = 1$.

Now suppose that $1 \leq j < k$ and that the property is true for j. Since b_j sees b_{j-1}, misses a_j, and $b_{j-1} < a_j < b_j$, we have $L_{a_j}(b_{j-1}) \neq L_{a_j}(a_j)$. Apply Lemma 2.2 to define $a_{j+1} = f(a_j, b_{j-1})$. Vertex a_{j+1} sees a_j, misses b_{j-1}, and $b_j < a_{j+1}$. Since a_{j+1} misses b_{j-1}, and $a_0 < b_0 < a_1 < b_1 = f(b_0, a_0) \cdots < a_j = f(a_{j-1}, b_{j-2}) < b_j = f(b_{j-1}, a_{j-1})$, it follows that a_{j+1} misses $a_0, \dots, a_{j-1}, b_0, \dots, b_{j-1}$. The same can be done to define b_{j+1}. So the property is true for $j + 1$. Thus the lemma holds. □

Lemma 2.4. *In a graph that contains no hole of length at least five, suppose that $a < b < u$, u sees a, u misses b, and a sees b. Then $f(b, a)$ sees u.*

Proof. Consider the path $\mathcal{P}(u, b, a)$ of Lemma 2.3. Since a sees b, that path is a hole, so it is a hole of length four, so $f(b, a)$ sees u. □

Properties of LexBFS*. Here are some notation and properties for Algorithm LexBFS*. When the algorithm selects a vertex $a \in A$ at step 3 of Algorithm LexBFS*, we put $L'_a(u) = L_a(u) \setminus L_a(a)$ for every (unnumbered) vertex u.

Lemma 2.5. *Suppose that $a < b$, $L_b(a) = L_b(b)$, and $N(a) \neq N(b)$. Then, during the loop of step 2 of algorithm* LexBFS*, *vertex a has been removed from A by a vertex $u = g(b, a)$ that sees b and misses a. We have the following properties:*

- *$u < a$,*
- *$L_b(u) < L_b(b)$,*

if there exists a vertex $v < a$ that sees a, misses b, and $L_b(v) \neq L_b(b)$, then $L'_b(v) \leq L'_b(u)$. If $L'_b(v) \neq L'_b(u)$, then there exists a vertex $> b$ that sees u and misses a, b, v, denote by $x = h(u, v)$ a maximum such vertex. We have the following properties:

- *For all y that sees v and misses a, b, u, we have $y < x$.*
- *For all y such that $x < y$ and y misses a, b, we have y sees u, v or y misses u, v.*

Proof. The definition of u and its properties follows from the definition of the algorithm. Suppose there exists a vertex $v < a$ that sees a, misses b, and $L_b(v) \neq L_b(b)$.

Suppose that $L'_b(v) > L'_b(u)$. Then v should have been selected at step 2.1 before u. Then, at step 2.2, $A \cap N(v)$ should be empty, otherwise b is removed from A and b is not the selected vertex at step 3. Since a is in $N(v)$, it has been previously removed from A by a vertex w with $L'_b(w) \geq L'_b(v)$. Since $L'_b(w) \geq L'_b(v) > L'_b(u)$, so $w \neq u$. This contradicts the definition of $u = g(b, a)$, so $L'_b(v) \leq L'_b(u)$.

If $L'_b(v) \neq L'_b(u)$, then $x = h(u, v)$ is well defined.

Suppose there exists a vertex y that sees v, misses a, b, u, and $x < y$. Then $L'_b(v) < L'_b(u)$ implies that there exists a vertex $> y$ that sees u and misses a, b, v; a contradiction to the definition of x.

Let y' be a vertex such that $x < y'$ and y' misses a, b. By the preceding property, it is not possible that y' sees v and misses u. If y' sees u and misses v, then this is a contradiction to the definition of x. So y sees u, v or y misses u, v. Thus the lemma holds. □

3. Proof of Theorem 1.2.

Recall that $\mathcal{C}$ denotes the class of graphs that contain no bull and no hole of length at least five. In this section we prove that when the input graph is in $\mathcal{C}$, the ordering given by Algorithm LexBFS* is an NMP$_5$ elimination ordering. It may be worth pointing out that this outcome does not hold for

LexBFS. For an example, consider the graph made of a chordless path a-b-c-d-e-f-g plus one vertex h adjacent to a, c, e, g. Then LexBFS can produce the ordering h, a, g, c, e, b, f, d, and d is the middle of the P_5 b-c-d-e-f. It is this example that led us to define the tie-breaking rule of LexBFS*.

Before proving the main result, we need the following lemma.

Lemma 3.1. *In a graph $G \in \mathcal{C}$, let $P = a_0$-a_1-$\cdots$-a_r be a chordless path with $r \geq 4$, and let u be a vertex that sees the two endvertices a_0, a_r of P. Then one of the following holds*:

- *u sees all vertices of P,*
- *r is even, and u sees $a_0, a_2, \ldots, a_r$ and misses $a_1, a_3, \ldots, a_{r-1}$, or*
- *$r = 4$, and u sees a_2 and exactly one of a_1, a_3.*

Consequently, in any case, u sees a_2 and a_{r-2}.

Proof. Denote a *segment* as any subpath of P, of length at least one, whose endvertices see u and interior vertices do not. So P is (edgewise) partitioned into its segments. Since G contains no hole of length at least five, every segment has length one or two. For $\ell = 1, 2$, let s_ℓ be the number of segments of P of length ℓ. So $r = s_1 + 2s_2$. If $s_1 = 0$, then every segment has length two, and we have the second outcome of the lemma. Now let $s_1 > 0$. So u sees two consecutive vertices of P. Suppose that we do not have the first outcome, so u has a nonneighbor in P. Thus, up to symmetry, there is an integer i such that u sees a_i and a_{i+1} and not a_{i+2}. Then $i \leq 1$, for otherwise a_0-ua_ia_{i+1}-a_{i+2} is a bull, and $r \leq i+3$, for otherwise a_r-ua_ia_{i+1}-a_{i+2} is a bull. It follows that $r = 4$ and $i = 1$, and we have the third outcome. Thus the lemma holds. □

Now we prove the following theorem, which implies Theorem 1.2. For any path P, let P^* denote the path formed by the interior vertices of P.

Theorem 3.2. *When the input graph is a graph in $\mathcal{C}$, Algorithm* LexBFS* *produces an NMP$_5$ ordering of the vertices of G.*

Proof of Theorem 3.2. Say that a P_5 a-b-c-d-e in G is *bad* if $c < \min\{a, b, d, e\}$. Say that a bad P_5 a-b-c-d-e is *worse* than another bad P_5 a'-b'-c'-d'-e' if $a \geq a', b \geq b', c \geq c', d \geq d', e \geq e'$, and at least one of these five inequalities is strict. Our aim is to prove that there is no bad P_5, so let us assume the contrary and show that this leads to a contradiction. Let a-b-c-d-e be a worst P_5. Up to symmetry we may assume that $e < a$.

Claim 1. $e < b$.

Proof. Suppose the claim is false, so $c < b < e < a$.

Since a sees b, misses e, and $b < e < a$, we can consider the chordless path $R = \mathcal{P}(a, e, b)$ of Lemma 2.3. If none of c, d has a neighbor in R^*, then $R \cup \{c, d\}$ is a cycle of length at least six, so one of c, d has a neighbor in R^*. Let q be the vertex of R^* closest to a that sees one of c, d. If q misses c, then $R[b, q] \cup \{d, c\}$ is a hole of length ≥ 5. So q sees c. The hole $R[b, q] \cup \{c\}$ must have length < 5, so q sees a and so $q \neq d$.

Since q sees c, misses b, and $c < b < q$, we have $L_b(c) \neq L_b(b)$. Apply Lemma 2.2 to define $r = f(b, c)$. Vertex r sees b, misses c, and $q < r$. Since b sees c, vertex r sees q by Lemma 2.4. Since r sees b, we have $r \neq d$. Since $f(e, b)$ is the neighbor of e on $\mathcal{R}$, it follows that $f(e, b) \leq q < r$, and r sees b so r sees e by Lemma 2.2. If r sees a, then there is a bull e-rab-c, a contradiction, so r misses a. If r misses d, then r, b, c, d, e is a hole, so r sees d. Suppose $\mathcal{R}$ has length > 3, then $f(e, b) < q = f(a, e) < r$, r sees e and misses a, a contradiction. So $\mathcal{R}$ has length 3, q sees e, and $q = f(e, b)$.

Since r sees e, misses a, and $e < a < r$, we have $L_a(e) \neq L_a(a)$. Apply Lemma 2.2

to define $s = f(a, e)$. Vertex s sees a, misses e, and $r < s$. Since s sees a, we have $s \neq d$. Since s misses e and $q = f(e, b) < r = f(b, c) < s$, it follows that s misses b, c by Lemma 2.2. If s sees d, then s, a, b, c, d is a hole, so s misses d. If s sees q, then b-asq-e is a bull, so s misses q. If s sees r, then c-der-s is a bull, so s misses r.

Since s sees a, misses q, and $a < q < s$, we have $L_q(a) \neq L_q(q)$. Apply Lemma 2.2 to define $t = f(q, a)$. Vertex t sees q, misses a, and $s < t$. Since q sees a, vertex t sees s by Lemma 2.4. Since t misses a and $q = f(e, b) < r = f(b, c) < s = f(a, e) < t$, vertex t misses b, c, e by Lemma 2.2. Since t misses c, we have $t \neq d$. If t sees r, then t, r, b, a, s is a hole, so t misses r, but then b-req-t is a bull, a contradiction. Thus the claim holds. □

Now we go on with the proof of the theorem. Since b sees c, misses e, and $c < e < b$, we have $L_e(c) \neq L_e(e)$. Apply Lemma 2.2 to define $p = f(e, c)$. Vertex p sees e, misses c, and $b < p$. Since p sees e and misses c, we have $p \neq a$ and $p \neq d$. If p sees a, then p sees the extremities of the P_5 a, b, c, d, e without seeing c, a contradiction to Lemma 3.1, so p misses a. If p sees b, then p sees d, otherwise p, b, c, d, e is a hole. If p misses b, then p misses d, otherwise the bad P_5 a-b-c-d-p is worse than a-b-c-d-e. So p either sees both b, d or misses both b, d.

Claim 2. $a < b$.

Proof. Suppose the claim is false, so $c < e < b < a$ by Claim 1.

Case 1. *$p < a$ and p sees b, d.* Since a sees b, misses p, and $b < p < a$, we have $L_p(b) \neq L_p(p)$. Apply Lemma 2.2 to define $q = f(p, b)$. Vertex q sees p, misses b, and $a < q$. Since p sees b, vertex q sees a by Lemma 2.4. Since q sees a, we have $q \neq d$. Since $p = f(e, c) < q$, vertex q either sees both e, c or misses both e, c. Suppose q misses e, c. If q sees d, then q, a, b, c, d is a hole, so q misses d. Then c-dep-q is a bull, a contradiction. So q sees e, c. Since q sees c, misses b, and $c < b < q$, we have $L_b(c) \neq L_b(b)$. Apply Lemma 2.2 to define $r = f(b, c)$. Vertex r sees b, misses c, and $q < r$. Since b sees c, vertex r sees q by Lemma 2.4. Since r sees b, misses c, and $p = f(e, c) < q = f(p, b) < r$, it follows that r sees p and misses e. But then c-brp-e is a bull, a contradiction.

Case 2. *$p < a$ and p misses b, d.* Since a sees b, misses p, and $b < p < a$, we can consider the chordless path $R = \mathcal{P}(a, p, b)$ of Lemma 2.3. If none of c, d, e has a neighbor in R^*, then the $R \cup \{c, d, e\}$ is a cycle of length at least 7, so one of c, d, e has a neighbor in R^*. Let q be the vertex of R^* closest to a that sees one of c, d, e. If q misses c, then one of $R[b, q] \cup \{c, d\}$, $R[b, q] \cup \{c, d, e\}$ is a hole of length ≥ 5. So q sees c. Since q sees c and $p = f(e, c) < q$, vertex q sees e. The hole $R[b, q] \cup \{c\}$ must have length < 5, so q sees a and so $q \neq d$. Since q sees c, misses b, and $c < b < q$, we have $L_b(c) \neq L_b(b)$. Apply Lemma 2.2 to define $r = f(b, c)$. Vertex r sees b, misses c, and $q < r$. Since b sees c, vertex r sees q by Lemma 2.4. Since r sees b, misses c, and $p = f(e, c) < f(p, b) \leq q < r$, vertex r sees p and misses e. Suppose $\mathcal{R}$ has length > 3, then $f(p, b) < q = f(a, p) < r$, r sees p, so r sees a and then c-bar-p is a bull, a contradiction. So $\mathcal{R}$ has length three and q sees p. But then q sees the extremities of the P_6 a-b-c-d-e-p without seeing b, a contradiction to Lemma 3.1.

Case 3. *$a < p$ and p sees b, d.* Since p sees b, misses a, and $b < a < p$, we have $L_a(b) \neq L_a(a)$. Apply Lemma 2.2 to define $q = f(a, b)$. Vertex q sees a, misses b, and $p < q$. Since a sees b, vertex q sees p by Lemma 2.4. Since q sees a, we have $q \neq d$. Since $p = f(e, c) < q$, vertex q either sees both e, c or misses both e, c. Suppose q misses e, c. If q sees d, then q, a, b, c, d is a hole, so q misses d. Then c-dep-q is a bull, a contradiction, so q sees c, e. Since q sees c, misses b, and $c < b < q$, we have $L_b(c) \neq L_b(b)$. Apply Lemma 2.2 to define $r = f(b, c)$. Vertex r sees b, misses c, and

$q < r$. Since b sees c, vertex r sees q by Lemma 2.4. Since r sees b, we have $r \neq d$. Since r sees b, misses c, and $p = f(e, c) < q = f(a, b) < r$, vertex r sees a and misses e. If r sees p, then *c-brp-e* is a bull, so r misses p. Since r sees a, misses p, and $a < p < r$, we have $L_p(a) \neq L_p(p)$. Apply Lemma 2.2 to define $s = f(p, a)$. Vertex s sees p, misses a and $r < s$. Since s misses a and $p = f(e, c) < q = f(a, b) < r = f(b, c) < s$, vertex s misses a, b, c, e. Since s misses c, we have $s \neq d$. If s misses d, then *c-dep-s* is a bull, so s sees d. But then *a-b-c-d-s* is a bad P_5 that is worse than *a-b-c-d-e*, a contradiction.

Case 4. *$a < p$ and p misses b, d.* Since p sees e, misses b, and $e < b < p$, we have $L_b(e) \neq L_b(b)$. Apply Lemma 2.2 to define $q = f(b, e)$. Vertex q sees b, misses e, and $p < q$. Since q sees b, we have $q \neq d$. Since q misses e and $p = f(e, c) < q$, vertex q misses c. If q misses d, then *q-b-c-d-e* is worse than *a-b-c-d-e*, so q sees d.

Case 4.1. *q misses a.* Since q sees b, misses a, and $b < a < q$, we have $L_a(b) \neq L_a(a)$. Apply Lemma 2.2 to define $r = f(a, b)$. Vertex r sees a, misses b, and $q < r$. Since a sees b, vertex r sees q by Lemma 2.4. Since r sees a, we have $r \neq d$. Since r misses b and $p = f(e, c) < q = f(b, e) < r$, vertex r misses b, c, e. If r sees d, then a, b, c, d, r is a hole, so r misses d. If r sees p, then a, b, c, d, e, p, r is a hole, so r misses p. Since r sees a, misses p, and $a < p < r$, we can consider the chordless path $R = \mathcal{P}(r, p, a)$ of Lemma 2.3. Every vertex u of R^* misses a and satisfies $p = f(e, c) < q = f(b, e) < r = f(a, b) < u$, so u misses a, b, c, e. If d has no neighbor in R^*, then $R \cup \{b, c, d, e\}$ is a cycle of length at least eight, so d has a neighbor in R^*. Let s be the vertex of R^* closest to a that sees d. Then $R[a, s] \cup \{b, c, d\}$ is a hole of length ≥ 5, a contradiction.

Case 4.2. *q sees a.* If q sees p, then *c-baq-p* is a bull, so q misses p. Since q sees a, misses p, and $a < p < q$, we can consider the chordless path $R = \mathcal{P}(q, p, a)$ of Lemma 2.3. Since $p = f(e, c) < q = f(b, e)$, every vertex of R^* either sees b, c, e or misses b, c, e. Let r be the neighbor of q in R^*. Vertex r misses a, and $f(p, a) \leq r$. If r misses b, c, e, then *c-baq-r* is a bull, so r sees b, c, e. Then *a-bcr-e* is a bull, a contradiction. Thus the claim holds. □

Claims 1 and 2 imply that $c < e < a < b$.

Since p sees e, misses a, and $e < a < p$, we have $L_a(e) \neq L_a(a)$. Apply Lemma 2.2 to define $q = f(a, e)$. Vertex q sees a, misses e, and $p < q$. Since q sees a, we have $q \neq d$. Since d sees e and misses a, it follows that $d < q = f(a, e)$. Since q misses e and $p = f(e, c) < q$, vertex q misses c. If q sees d, then q sees b, otherwise q, a, b, c, d is a hole. If q misses d, then q misses b, otherwise the bad P_5 *q-b-c-d-e* is worse than *a-b-c-d-e*. So q sees b, d or misses b, d.

Claim 3. *The path q-a-b-c-d-e-p is chordless.*

Proof. Suppose that q sees p. Then q sees the extremities of the path *a-b-c-d-e-p* without seeing c, so, by Lemma 3.1, the path is not chordless, so p sees b, d. If q misses b, d, then p sees the extremities of the path *q-a-b-c-d-e* without seeing c, a contradiction to Lemma 3.1, so q sees b, d. But then *c-bqp-e* is a bull, a contradiction. So q misses p.

Since q sees a, misses p, and $a < p < q$, we have $L_p(a) \neq L_p(p)$. Apply Lemma 2.2 to define $r = f(p, a)$. Vertex r sees p, misses a, and $q < r$. Since r misses a and $p = f(e, c) < q = f(a, e) < r$, vertex r misses c, e.

Suppose p sees b, d. If r misses d, then *c-dep-r* is a bull, so r sees d. If r misses b, then the bad P_5 *a-b-c-d-r* is worse than *a-b-c-d-e*, so r sees b. Then *a-brp-e* is a bull, a contradiction. So p misses b, d.

Suppose q sees b, d and r sees q. If r misses b, then *c-baq-r* is a bull, so r sees b.

If r misses d, then the bad P_5 r-b-c-d-e is worse than a-b-c-d-e, so r sees d. Then e-drq-a is a bull, a contradiction.

Suppose q sees b, d and r misses q. Since r sees p, misses q, and $p < q < r$, we have $L_q(p) \neq L_q(q)$. Apply Lemma 2.2 to define $s = f(q, p)$. Vertex s sees q, misses p, and $r < s$. Since s misses p and $p = f(e, c) < q = f(a, e) < r = f(p, a)$, vertex s misses a, c, e. If s misses b, then c-baq-s is a bull, so s sees b. If s misses d, then the bad P_5 s-b-c-d-e is worse than a-b-c-d-e, so s sees d. Then e-dsq-a is a bull, a contradiction. So q misses b, d. Thus the claim holds. □

CLAIM 4. $d < b$.

Proof. Suppose the claim is false, then $c < e < a < b < d$ by Claims 1 and 2.

Case 1. $L_d(b) \neq L_d(d)$. Apply Lemma 2.2 to define $s = f(d, b)$. Vertex s sees d, misses b, and $d < s$. Since s sees d, we have $s \neq p$. Suppose s sees c. If s misses e, then b-csd-e is a bull, so s sees e. If s misses a, then the bad P_5 a-b-c-s-e is worse than a-b-c-d-e, so s sees a. If s misses p, then a-sde-p is a bull, so s sees p. Then b-cds-p is a bull, a contradiction, so s misses c. If s sees a, then a, b, c, d, s is a hole, so s misses a. Then the bad P_5 a-b-c-d-s is worse than a-b-c-d-e, a contradiction.

Case 2. $L_d(b) = L_d(d)$. Since a sees b and misses d, we have $N(b) \neq N(d)$. Apply Lemma 2.5 to define $s = g(d, b)$. Vertex s sees d, misses b, $s < b$, and $L_d(s) < L_d(d)$. Since s misses b, we have $s \neq a, c$. Since q sees a and misses d, we have $L_d(a) \neq L_d(d)$, and since a sees b and misses d, we have $L'_d(a) \leq L'_d(s)$. If s sees q, then s sees the extremities of the P_5 q, a, b, c, d without seeing b, a contradiction to Lemma 3.1, so s misses q. So $L'_d(a) \neq L'_d(s)$. Apply Lemma 2.5 to define $t = h(s, a)$. Vertex t sees s, misses a, b, d, and $q < t$. Since t misses a and $p = f(e, c) < q = f(a, e) < t$, vertex t misses c, e. Since t misses e, we have $s \neq e$. If s sees c, then b-cds-t is a bull, so s misses c. If s sees a, then a, b, c, d, s is a hole, so s misses a. Suppose $s < e$. Since t sees s, misses e, and $s < e < t$, we have $L_e(s) \neq L_e(e)$. Apply Lemma 2.2 to define $u = f(e, s)$. Vertex u sees e, misses s, and $t < u$. Since u sees e and $p = f(e, c) < q = f(a, e) < t < u$, vertex u sees a, c. Since u sees a, misses s, $t = h(s, a) < u$, and $s = g(b, d)$, vertex u sees b, d. But then a-ucd-s is a bull, a contradiction. So $e < s$. Then the bad P_5 a-b-c-d-s is worse than a-b-c-d-e, a contradiction. Thus the claim holds. □

CLAIM 5. $L_b(d) = L_b(b)$.

Proof. Suppose the claim is false, so $L_b(d) \neq L_b(b)$. Apply Lemma 2.2 to define $s = f(b, d)$. Vertex s sees b, misses d, and $b < s$. Since s sees b, we have $s \neq q$. Suppose s sees c. If s misses a, then d-csb-a is a bull, so s sees a. If s misses e, then the bad P_5 a-s-c-d-e is worse than a-b-c-d-e, so s sees e. If s misses q, then e-sba-q is a bull, so s sees q. Then d-cbs-q is a bull, a contradiction, so s misses c. If s sees e, then b, c, d, e, s is a hole, so s misses e. Then s-b-c-d-e is a bad P_5 that is worse than a-b-c-d-e, a contradiction. Thus the claim holds. □

CLAIM 6. $a < d$.

Proof. Suppose the claim is false, then $d < a < b$. By Lemma 2.1, $L_b(d) \leq L_b(a) \leq L_b(b)$, and, by Claim 5, $L_b(d) = L_b(b)$, so $L_b(a) = L_b(b)$. Vertex q sees a, misses b, and $a < b < q$, a contradiction. Thus the claim holds. □

With the preceding claims, we have established that $c < e < a < d < b < p = f(e, c) < q = f(a, e)$, $L_b(d) = L_b(b)$, and q-a-b-c-d-e-p is a chordless path. Define sequences $(a_i), (b_i), (d_i), (e_i)$ as follows:

- $a_0 = a$, $b_0 = b$, $d_0 = d$, $e_0 = e$, $b_1 = q = f(a, e)$, $d_1 = p = f(e, c)$.
- For $i \geq 1$, $a_i = g(b_i, d_i)$, $e_i = g(d_i, b_{i-1})$.
- For $i \geq 2$, $b_i = h(a_{i-1}, e_{i-1})$, $d_i = h(e_{i-1}, a_{i-2})$.

For any $k \geq 1$, let us say that a-b-c-d-e admits an extension of order k, noted $\mathcal{W}_k$, if the sequences $(a_i)_{i<k}$, $(b_i)_{i\leq k}$, $(d_i)_{i\leq k}$, $(e_i)_{i<k}$ are well defined, and have the following property:

- $c < e_0 < a_0 < \cdots < e_{k-1} < a_{k-1} < d_0 < b_0 < \cdots < d_k < b_k$.
- $L_{b_{k-1}}(b_0) = \cdots = L_{b_{k-1}}(b_{k-1}) = L_{b_{k-1}}(d_0) = \cdots = L_{b_{k-1}}(d_{k-1})$.
- b_k-a_{k-1}-b_{k-1}-$\cdots$-b_1-a_0-b_0-c-d_0-e_0-d_1-$\cdots$-d_{k-1}-e_{k-1}-d_k is a chordless path.

Claims 1–6 and the definition of p, q shows that a-b-c-d-e admits an extension of order 1. Let k be the greatest integer such that a-b-c-d-e admits an extension $\mathcal{W}_k$ of order k. We will prove that a-b-c-d-e admits an extension of order $k+1$. Since G is finite, this is a contradiction that will complete the proof that there is no bad P_5.

CLAIM 7. $L_{d_k}(b_{k-1}) = L_{d_k}(d_k)$.

Proof. For suppose that $L_{d_k}(b_{k-1}) \neq L_{d_k}(d_k)$. Since $b_{k-1} < d_k$ we can apply Lemma 2.2 to define $r = f(d_k, b_{k-1})$. Vertex r sees d_k, misses b_{k-1}, and $d_k < r$. Since r sees d_k, we have $r \neq b_k$. Since r misses b_{k-1} and $L_{b_{k-1}}(b_0) = \cdots = L_{b_{k-1}}(b_{k-1}) = L_{b_{k-1}}(d_0) = \cdots = L_{b_{k-1}}(d_{k-1})$, it follows that r misses $b_0, \dots, b_{k-1}, d_0, \dots, d_{k-1}$. Since r misses $b_0, \dots, b_{k-1}, d_0, \dots, d_{k-1}$ and $e_1 = g(d_1, b_0) < a_1 = g(b_1, d_1) < \cdots < a_{k-2} = g(b_{k-2}, d_{k-2}) < e_{k-1} = g(d_{k-1}, b_{k-2}) < d_1 = f(e_0, c) < b_1 = f(a_0, e_0) < d_2 = h(e_1, a_0) < b_2 = h(a_1, e_1) < \cdots < b_{k-1} = h(a_{k-2}, e_{k-2}) < d_k = h(e_{k-1}, a_{k-2}) < r$, it follows that r either sees all of $c, a_0, \dots, a_{k-2}, e_0, \dots, e_{k-1}$ or misses all of them. If r sees them, then d_{k-1}-$e_{k-1}d_kr$-a_{k-2} is a bull, so r misses them. If r sees one of a_{k-1}, b_k, then $\mathcal{W}_k \cup \{r\}$ contains a hole of length at least six, a contradiction, so r misses a_{k-1}, b_k.

Case 1. $r < b_k$. Since b_k sees a_{k-1}, misses b_{k-1}, and $a_{k-1} < b_{k-1} < b_k$, we have $L_{b_{k-1}}(a_{k-1}) \neq L_{b_{k-1}}(b_{k-1})$. Apply Lemma 2.2 to define $s = f(b_{k-1}, a_{k-1})$. Vertex s sees b_{k-1}, misses a_{k-1} and $b_k < s$. Since b_{k-1} sees a_{k-1}, vertex s sees b_k by Lemma 2.4. Since s sees b_{k-1} and $r = f(d_k, b_{k-1}) < b_k < s$, vertex s sees d_k. Since s sees b_k, d_k and misses a_{k-1}, it follows from Lemma 3.1 that r sees all of $b_0, \dots, b_k, d_0, \dots, d_k$ and misses all of c, $a_0, \dots, a_{k-1}, e_0, \dots, e_{k-1}$. If s sees r, then e_{k-1}-d_krs-b_k is a bull, so s misses r.

Since s sees d_k, misses r, and $d_k < r < s$, we have $L_r(d_k) \neq L_r(r)$. Apply Lemma 2.2 to define $t = f(r, d_k)$. Vertex t sees r, misses d_k, and $s < t$. Since r sees d_k, vertex t sees s by Lemma 2.4. Since t misses d_k and $r = f(d_k, b_{k-1}) < s = f(b_{k-1}, a_{k-1}) < t$, vertex t misses a_{k-1}, b_{k-1}. Since t misses b_{k-1} and $L_{b_{k-1}}(b_0) = \cdots = L_{b_{k-1}}(b_{k-1}) = L_{b_{k-1}}(d_0) = \cdots = L_{b_{k-1}}(d_{k-1})$, it follows that t misses all of $b_0, \dots, b_{k-1}, d_0, \dots, d_{k-1}$. Since t misses $a_{k-1}, b_0, \dots, b_{k-1}, d_0, \dots, d_k$, and $e_1 = g(d_1, b_0) < a_1 = g(b_1, d_1) < \cdots < e_{k-1} = g(d_{k-1}, b_{k-2}) < a_{k-1} = g(b_{k-1}, d_{k-1}) < d_1 = f(e_0, c) < b_1 = f(a_0, e_0) < d_2 = h(e_1, a_0) < b_2 = h(a_1, e_1) < \cdots < d_k = h(e_{k-1}, a_{k-2}) < b_k = h(a_{k-1}, e_{k-1}) < t$, it follows that t misses all of $c, a_0, \dots, a_{k-1}, e_0, \dots, e_{k-1}$.

Since t sees r, misses b_k, and $r < b_k < t$, we can consider the chordless path $R = \mathcal{P}(t, b_k, r)$ of Lemma 2.3. Every vertex u of R^* misses r and satisfies $t = f(r, d_k) < u$, so u misses d_k. The cycle $R \cup \mathcal{W}_k$ has length at least ten, so one of $\mathcal{W}_k \setminus \{b_k\}$ has a neighbor in R^*. Let u be the vertex of R^* closest to t that sees one of $\mathcal{W}_k \setminus \{b_k\}$, then $R[u, r] \cup \mathcal{W}_k$ contains a hole of size ≥ 5, a contradiction.

Case 2. $b_k < r$. Since r sees d_k, misses b_k, and $d_k < b_k < r$, we can consider the chordless path $R = \mathcal{P}(r, b_k, d_k)$ of Lemma 2.3. Every vertex u of R^* misses d_k and satisfies $r = f(d_k, b_{k-1}) < u$, so u misses b_{k-1}. Then, since $L_{b_{k-1}}(b_0) = \cdots = L_{b_{k-1}}(b_{k-1}) = L_{b_{k-1}}(d_0) = \cdots = L_{b_{k-1}}(d_{k-1})$, vertex u misses all of $b_0, \dots, b_{k-1}$, $d_0, \dots, d_{k-1}$. Since u misses $b_0, \dots, b_{k-1}, d_0, \dots, d_k$ and $e_1 = g(d_1, b_0) < a_1 =$

$g(b_1, d_1) < \cdots < e_{k-1} = g(d_{k-1}, b_{k-2}) < a_{k-1} = g(b_{k-1}, d_{k-1}) < d_1 = f(e_0, c) < b_1 = f(a_0, e_0) < d_2 = h(e_1, a_0) < b_2 = h(a_1, e_1) < \cdots d_k = h(e_{k-1}, a_{k-2}) < b_k = h(a_{k-1}, e_{k-1}) < u$, vertex u either sees all of $c, a_0, \ldots, a_{k-1}, e_0, \ldots, e_{k-1}$ or misses all of them.

Let t be the neighbor of b_k in R^*, so $t = f(b_k, d_k)$. If t sees $c, a_0, \ldots, a_{k-1}, e_0, \ldots, e_{k-1}$, then b_{k-1}-$a_{k-1}b_kt$-e_{k-1} is a bull. So t misses $c, a_0, \ldots, a_{k-1}$, $e_0, \ldots, e_{k-1}$. If t sees r, then $\mathcal{W}_k \cup \{r, t\}$ is a hole, so t misses r.

Let u be the neighbor of r in R^*, so $u = f(r, b_k)$. Vertex u misses $b_0, \ldots, b_k, d_0, \ldots, d_k$. If u misses $c, a_0, \ldots, a_{k-1}, e_0, \ldots, e_{k-1}$, then $R \cup \mathcal{W}_k$ contains a hole of size ≥ 5, so u sees $c, a_0, \ldots, a_{k-1}, e_0, \ldots, e_{k-1}$.

Since u sees c, misses b_0, and $c < b_0 < u$, we have $L_{b_0}(c) \neq L_{b_0}(b_0)$. Apply Lemma 2.2 to define $v = f(b_0, c)$. Vertex v sees b_0, misses c, and $u < v$. Since b_0 sees c, vertex v sees u by Lemma 2.4. Since v sees b_0 and $L_{b_{k-1}}(b_0) = \cdots = L_{b_{k-1}}(b_{k-1}) = L_{b_{k-1}}(d_0) = \cdots = L_{b_{k-1}}(d_{k-1})$, vertex v misses $b_0, \ldots, b_{k-1}, d_0, \ldots, d_{k-1}$. Since v sees b_{k-1}, misses c, and $d_1 = f(e_0, c) < r = f(d_k, b_{k-1}) < t = f(b_k, d_k) < u = f(r, b_k) < v$, vertex v sees d_k, b_k, r and misses e_0. But then b_0-vru-e_0 is a bull, a contradiction. Thus the claim holds. □

CLAIM 8. $L_{d_k}(b_0) = \cdots = L_{d_k}(b_{k-1}) = L_{d_k}(d_0) = \cdots = L_{d_k}(d_k)$.

Proof. By Claim 7, $L_{d_k}(b_{k-1}) = L_{d_k}(d_k)$, and $L_{b_{k-1}}(b_0) = \cdots = L_{b_{k-1}}(b_{k-1}) = L_{b_{k-1}}(d_0) = \cdots = L_{b_{k-1}}(d_{k-1})$, and $b_{k-1} < d_k$, so $L_{d_k}(b_0) = \cdots = L_{d_k}(b_{k-1}) = L_{d_k}(d_0) = \cdots = L_{d_k}(d_k)$. Thus the claim holds. □

Since a_{k-1} sees b_{k-1} and misses d_k, we have $N(b_{k-1}) \neq N(d_k)$. Apply Lemma 2.5 to define $e_k = g(d_k, b_{k-1})$. Vertex e_k sees d_k, misses b_{k-1}, $e_k < b_{k-1}$, and $L_{d_k}(e_k) < L_{d_k}(d_k) = L_{d_k}(d_0)$. Since $L_{d_k}(e_k) < L_{d_k}(d_0)$, so $e_k < d_0$ by Lemma 2.1. Since e_k sees d_k, so $e_k \notin \mathcal{W}_k \setminus \{e_{k-1}\}$. Since a_{k-1} sees b_{k-1} and misses d_k, we have $L'_{d_k}(a_{k-1}) \leq L'_{d_k}(e_k)$. If e_k sees b_k, then e_k sees the extremities of the chordless path $\mathcal{W}_k$ without seeing b_{k-1}, a contradiction to Lemma 3.1, so e_k misses b_k. So $L'_{d_k}(a_{k-1}) < L'_{d_k}(e_k)$. Apply Lemma 2.5 to define $d_{k+1} = h(e_k, a_{k-1})$. Vertex d_{k+1} sees e_k, misses a_{k-1}, b_{k-1}, d_k, and $b_k < d_{k+1}$.

CLAIM 9. *$\mathcal{W}_k$-e_k-d_{k+1} is a chordless path.*

Proof. Since d_{k+1} misses d_k and $L_{d_k}(b_0) = \cdots = L_{d_k}(b_{k-1}) = L_{d_k}(d_0) = \cdots = L_{d_k}(d_k)$, vertex d_{k+1} misses $b_0, \ldots, b_{k-1}, d_0, \ldots, d_k$. Since d_{k+1} misses a_{k-1}, $b_0, \ldots$, b_{k-1}, $d_0, \ldots, d_k$, and $e_1 = g(d_1, b_0) < a_1 = g(b_1, d_1) < \cdots < e_{k-1} = g(d_{k-1}, b_{k-2}) < a_{k-1} = g(b_{k-1}, d_{k-1}) < d_1 = f(e_0, c) < b_1 = f(a_0, e_0) < d_2 = h(e_1, a_0) < b_2 = h(a_1, e_1) < \cdots d_k = h(e_{k-1}, a_{k-2}) < b_k = h(a_{k-1}, e_{k-1}) < t$, vertex d_{k+1} misses $c, a_0, \ldots, a_{k-1}, e_0, \ldots, e_{k-1}$. Since d_{k+1} misses e_{k-1}, we have $e_k \neq e_{k-1}$. If d_{k+1} sees b_k, then e_k sees the extremities of the chordless path $\mathcal{W}_k \cup \{d_{k+1}\}$ without seeing b_{k-1}, a contradiction to Lemma 3.1, so d_{k+1} misses b_k.

Suppose e_k sees d_{k-1}. Consider the general step of the algorithm when b_{k-1} is chosen. Since $L_{d_k}(e_k) < L_{d_k}(d_k) = L_{d_k}(b_{k-1})$, we have $L_{b_{k-1}}(e_k) < L_{b_{k-1}}(b_{k-1})$, by Lemma 2.1. Since $L'_{d_k}(a_{k-1}) < L'_{d_k}(e_k)$, and $L_{d_k}(b_{k-1}) = L_{d_k}(d_k)$, we have $L'_{b_{k-1}}(a_{k-1}) < L'_{b_{k-1}}(e_k)$. Set U of step 1 of the algorithm contains e_k because $L_{b_{k-1}}(e_k) < L_{b_{k-1}}(b_{k-1})$. Since $L'_{b_{k-1}}(a_{k-1}) < L'_{b_{k-1}}(e_k)$, vertex e_k is selected from U at step 2.1 before a_{k-1}. Then at step 2.2, $A \cap N(e_k)$ must be empty, for otherwise b_{k-1} is removed from A and b_{k-1} is not the selected vertex at step 3. Since vertex d_{k-1} is in $N(e_k)$, it has been removed earlier from A by a vertex u with $L'_{b_{k-1}}(e_k) \leq L'_{b_{k-1}}(u)$. Since $L'_{b_{k-1}}(u) \geq L'_{b_{k-1}}(e_k) > L'_{b_{k-1}}(a_{k-1})$, we have $u \neq a_{k-1}$. This contradicts the definition of a_{k-1}, so e_k misses d_{k-1}.

If e_k sees e_{k-1}, then d_{k-1}-$e_{k-1}d_ke_k$-d_{k+1} is a bull, so e_k misses e_{k-1}. If e_k sees one of $b_0, \dots, b_{k-2}$, $d_0, \dots, d_{k-2}$, $c, a_0, \dots, a_{k-1}, e_0, \dots, e_{k-2}$, then $\mathcal{W}_k \cup \{s\}$ contains a hole of length > 5, so e_k missees $b_0, \dots, b_{k-2}, d_0, \dots, d_{k-2}$, $c, a_0, \dots, a_{k-2}, e_0, \dots, e_{k-2}$. Thus the claim holds. □

Claim 10. $a_{k-1} < e_k$.

Proof. Suppose the claim is false and $e_k < a_{k-1}$. Since d_{k+1} sees e_k, misses a_{k-1}, and $e_k < a_{k-1} < d_{k+1}$, we have $L_{a_{k-1}}(e_k) \neq L_{a_{k-1}}(a_{k-1})$. Apply Lemma 2.2 to define $u = f(a_{k-1}, e_k)$. Vertex u sees a_{k-1}, misses e_k, and $d_{k+1} < u$. Since u sees a_{k-1}, misses e_k, $d_{k+1} = h(e_k, a_{k-1}) < u$, and $e_k = g(d_k, b_{k-1})$, vertex u sees d_k, b_{k-1}. Since u sees the extremities of the chordless path $\mathcal{W}_k \setminus \{b_k\}$, by Lemma 3.1 it must see all the vertices of $\mathcal{W}_k \setminus \{b_k\}$. But then a_{k-1}-$ue_{k-1}d_k$-e_k is a bull, a contradiction. Thus the claim holds. □

Claims 8, 9, and 10, and the definition of e_k, d_{k+1}, show that the sequences $(a_i)_{i<k}$, $(b_i)_{i\le k}$, $(d_i)_{i\le k+1}$, $(e_i)_{i<k+1}$ are well defined and satisfy the following properties:

- $c < e_0 < a_0 < \cdots < e_{k-1} < a_{k-1} < e_k < d_0 < b_0 < \cdots < d_k < b_k < d_{k+1}$.
- $L_{d_k}(b_0) = \cdots = L_{d_k}(b_{k-1}) = L_{d_k}(d_0) = \cdots = L_{d_k}(d_k)$.
- $\mathcal{W}_k$-e_k-d_{k+1} is a chordless path.

The same type of proof can be done (and we omit the details) to define vertices $a_k = g(b_k, d_k)$ and $b_{k+1} = h(a_k, e_k)$ and to show that they satisfy the following properties:

- $c < e_0 < a_0 < \cdots < e_k < a_k < d_0 < b_0 < \cdots < d_{k+1} < b_{k+1}$.
- $L_{b_k}(b_0) = \cdots = L_{b_k}(b_k) = L_{b_k}(d_0) = \cdots = L_{b_k}(d_k)$.
- b_{k+1}-a_k-$\mathcal{W}_k$-e_k-d_{k+1} is a chordless path.

This means that a-b-c-d-e admits an extension of order $k+1$. This is a contradiction to the definition of k. This completes the proof of the theorem. □

4. Algorithm Cosine*.

Algorithm Cosine* is a particular case of Algorithm Cosine due to Hertz [20], which is an $\mathcal{O}(nm)$ algorithm for optimally coloring the vertices of a Meyniel graph. The difference between Cosine and Cosine* is that the input graph of Cosine* has an ordering σ on its vertices and ties are broken according to this ordering.

Colors are viewed as integers $1, 2, \dots, \ell$. Algorithm Cosine* constructs the color classes iteratively. To construct the class of color c, the algorithm selects vertices until all the vertices of the graph have a neighbor colored c. At each step, the vertex that is selected and colored c is the vertex that has no neighbor already colored c and has the maximum number of uncolored neighbors in common with the vertices already colored c, with ties being broken by taking such a vertex that minimizes σ. More formally:

```
Algorithm Cosine*
Input: A graph $G$ on $n$ vertices and an ordering $\sigma$ on its vertices.
Output: A coloring of the vertices of $G$.
Initialization: $c = 1$;
General step: While there exist uncolored vertices do:
1. While there exist uncolored vertices that have no neighbor colored
$c$ do:
    1.1. Let $A$ be the set of uncolored vertices that have a neighbor
    colored $c$;
    1.2. Select an uncolored vertex $u$ that has no neighbor colored
    $c$ and has the maximum number of neighbors in $A$, with ties being
    broken by taking such a vertex that is minimum for $\sigma$;
    1.3. Color $u$ with $c$;
```

2. $c := c + 1$.

One may remark that the original formulation of Algorithm COSINE in [20] is different. Hertz explains his algorithm in terms of vertex contraction. We prefer to modify the formulation of the algorithm to simplify the algorithmic concepts. To prove the optimality of the algorithm, we need to introduce the notion of contraction, which is done in the next section.

Complexity analysis. To analyze the complexity of algorithm COSINE*, we will assume that the input graph is connected; thus if n is the number of vertices and m the number of edges of the graph, we have $m \geq n - 1$. If the graph is not connected, then it suffices to apply the algorithm on each of its components. Breaking the ties in COSINE* does not increase the complexity of Algorithm COSINE, that is, it can be implemented in time $\mathcal{O}(nm)$ as follows. Updating the set A at step 1.1 can be done in time $\mathcal{O}(d(u))$ whenever a new vertex u is colored at step 1.3, by adding the uncolored neighbors of u to A. For one given color c, this procedure takes time $\mathcal{O}(n + m)$, so the total cost is $\mathcal{O}(nm)$ over all colors. To compute step 1.2 efficiently, we use for each vertex a counter that represents the number of its neighbors in A. Every time a vertex is added to A we update the counter of the other vertices; this can also be done in time $\mathcal{O}(n + m)$ for a given color and so in time $\mathcal{O}(nm)$ over all colors. Then we search all the vertices in time $\mathcal{O}(n)$ to find the uncolored vertex that has the maximum counter and is minimum for σ. After each such search, one vertex is colored, so the total cost of all such searches is $\mathcal{O}(n^2)$. Therefore, the total running time of Algorithm COSINE* is $\mathcal{O}(nm)$.

5. Even pairs contraction. An *even pair* in a graph G is a pair of nonadjacent vertices such that every chordless path between them has even length. A survey on even pairs is given in [12]. Given two nonadjacent vertices x, y in G, the operation of *contracting* them means removing x and y and adding one vertex with an edge to each vertex of $N(x) \cup N(y)$. The following lemmas state essential results about even pairs.

LEMMA 5.1 (see [13, 29]). *For any graph G, the graph G' obtained from G by contracting an even pair of G satisfies $\omega(G') = \omega(G)$ and $\chi(G') = \chi(G)$.*

LEMMA 5.2 (see [12]). *If a graph G contains no odd hole, then the graph G' obtained from G by contracting an even pair contains no odd hole.*

LEMMA 5.3 (see [12]). *If a graph G contains no antihole, then the graph G' obtained from G by contracting an even pair contains no antihole different from $\overline{C}_6$.*

Following Bertschi [4], a graph G is called *even contractile* if it is either a clique or it contains an even pair whose contraction yields an even contractile graph, and G is *perfectly contractile* if every induced subgraph of G is even contractile. See [12] for a survey on perfectly contractile graphs.

We need to define a superclass of $\mathcal{B}$. Let us say that a graph G is a *quasi-$\mathcal{B}$* graph if G is a Berge graph that contains no antihole of length at least five and G has a vertex, called a *pivot*, that is an ear of every bull of G. (This definition can be compared with the definition of quasi-Meyniel graphs in [20].) We observe that every graph in class $\mathcal{B}$ is a quasi-$\mathcal{B}$ graph (and in such a graph, every vertex is a pivot), and if G is a quasi-$\mathcal{B}$ graph and z is a pivot, then $G \setminus z$ is in class $\mathcal{B}$.

We prove that, for every graph G in class $\mathcal{B}$, Algorithm LEXBFS* applied on $\overline{G}$ followed by Algorithm COSINE* applied on G produces a coloring of the vertices of G with $\omega(G)$ colors, where $\omega(G)$ is the maximum size of a clique in G. This will prove the optimality of this algorithm on the class $\mathcal{B}$. Our proof follows the same steps as Hertz's proof [20] that his algorithm COSINE is optimal on quasi-Meyniel graphs. Just like in [20], the optimality of our algorithm will follow from the fact that each color class produced by the algorithm corresponds to the contraction of even pairs.

The following lemma generalizes Lemma 3.1 to quasi-$\mathcal{B}$ graphs.

LEMMA 5.4. *In a* quasi-$\mathcal{B}$ *graph* G, *let* $P = a_0\text{-}a_1\text{-}\cdots\text{-}a_r$ *be a chordless odd path with* $r \geq 5$, *where* a_0 *is a pivot of* G, *and let* u *be a vertex that sees the two endvertices* a_0, a_r *of* P. *Then* u *sees* a_2.

Proof. Suppose the lemma is false and u misses a_2. If u sees a_1, then $a_r\text{-}ua_0a_1\text{-}a_2$ is a bull of which a_0 is not an ear, a contradiction. So u misses a_1. Denote a *segment* as any subpath of P, of length at least one, whose endvertices see u and interior vertices do not. So P is (edgewise) partitioned into its segments. Since G is odd-hole-free, every segment has length one or even length. Since P is odd, there is a least one segment of length one. Let i be the smallest integer such that u sees a_i and a_{i+1}. Since u misses a_1, a_2, we have $i \geq 3$. Then $a_{i-1}\text{-}a_ia_{i+1}u\text{-}a_0$ is a bull of which a_0 is not an ear, a contradiction. □

Now we prove the following theorem, which implies the optimality of our coloring algorithm.

THEOREM 5.5. *Let* G *be in class* $\mathcal{B}$. *Then the coloring obtained by Algorithm* LEXBFS* *applied on* $\overline{G}$ *followed by Algorithm* COSINE* *applied on* G *uses exactly* $\omega(G)$ *colors.*

Proof of Theorem 5.5. Let ℓ be the total number of colors used by the algorithm. For each color $c \in \{1, \ldots, \ell\}$ let k_c be the number of vertices colored c. Therefore every vertex of G can be renamed x_c^i, where $c \in \{1, \ldots, \ell\}$ is the color assigned to the vertex by the algorithm and $i \in \{1, \ldots, k_c\}$ is the integer such that x_c^i is the ith vertex colored c. Thus $V(G) = \{x_1^1, x_1^2, \ldots, x_1^{k_1}, x_2^1, \ldots, x_2^{k_2}, \ldots, x_\ell^1, \ldots, x_\ell^{k_\ell}\}$.

Define a sequence of graphs and vertices as follows. Put $G_1^1 = G$ and $w_1^1 = x_1^1$ (that is a pivot of G). For $i = 2, \ldots, k_1$, call G_1^i the graph obtained from G_1^{i-1} by contracting w_1^{i-1} and x_1^i into a new vertex w_1^i colored with the color one. In the graph $G_1^{k_1}$, we remark that $w_1^{k_1}$ is adjacent to all other vertices of $G_1^{k_1}$; for otherwise, there is a vertex y that is not adjacent to $w_1^{k_1}$, that means that y has no neighbor of color one, so the algorithm should have colored more vertices with color one; a contradiction. More simply, let us call w_1 the vertex $w_1^{k_1}$.

The sequence continues as follows. For each $c \in \{2, \ldots, \ell\}$, put $G_c^1 = G_{c-1}^{k_{c-1}}$ and $w_c^1 = x_c^1$. For $i = 2, \ldots, k_c$, call G_c^i the graph obtained from G_c^{i-1} by contracting vertices w_c^{i-1} and x_c^i into a new vertex w_c^i colored with the color c. In $G_c^{k_c}$, we can again remark that $w_c^{k_c}$ is adjacent to all other vertices of $G_c^{k_c}$, for the same reason as above, and we simply call w_c the vertex $w_c^{k_c}$. So the last graph in the sequence, $G_\ell^{k_\ell}$, is a clique of size l with vertices $w_1, \ldots, w_\ell$, where each w_c is obtained by the contraction of the vertices of color c.

CLAIM 1. *For every color* $c \in \{1, \ldots, \ell\}$ *and integer* $i \in \{1, \ldots, k_c - 1\}$, *if* G_c^i *is a* quasi-$\mathcal{B}$ *graph,* w_c^i *is a pivot, and not the top of a house of* G_c^i, *then there is no chordless odd path from* w_c^i *to* x_c^{i+1} *in* G_c^i.

Proof. Suppose on the contrary that there exists a chordless odd path $P = a_0\text{-}a_1\text{-}\cdots\text{-}a_{r-1}\text{-}a_r$ from $a_0 = w_c^i$ to $a_r = x_c^{i+1}$ in G_c^i. We have $r \geq 3$ since w_c^i, x_c^{i+1} are not adjacent. Note that every vertex of P has a nonneighbor in G_c^i. Put $W_1 = \emptyset$ and $W_c = \{w_1, \ldots, w_{c-1}\}$ if $c \geq 2$, and recall that any $w \in W_c$ is a vertex of G_c^i that is adjacent to all vertices of $G_c^i \setminus w$. So P contains no vertex of W_c. We know that every vertex of $G_c^i \setminus W_c$ will have a color from $\{c, c+1, \ldots, \ell\}$ when the algorithm terminates.

Let us consider the situation when Algorithm COSINE* selects x_c^{i+1}. Let A be the set defined at step 1.1 of the algorithm. Vertex a_1 is in A and a_2 is not in A. Let $T = N(x_c^{i+1}) \cap A$. Every vertex of T is adjacent to at least one vertex colored c in G and thus is adjacent to w_c^i in G_c^i.

Suppose that there exists a vertex $t \in T$ that misses a_2. If $r = 3$, then either t misses a_1 and then u, a_0, a_1, a_2, a_3 induce an odd hole, or t sees a_1 and then a_0 is the top of a house, in either case a contradiction. So $r \geq 5$. Vertex t sees both extremities of the chordless odd path P without seeing a_2, a contradiction to Lemma 5.4. So every vertex of T sees a_2. Then $T \cup \{a_1\} \subset N(a_2) \cap A$, and so a_2 has strictly more neighbors in A than x_c^{i+1}, which contradicts the fact that x_c^{i+1} is selected at step 1.2. Thus the claims holds. □

Claim 2. *For every color $c \in \{1, \ldots, \ell\}$ and integer $i \in \{0, 1, \ldots, k_c - 1\}$, the following two properties hold*:

(A_i) *If $i \geq 1$, then w_c^i and x_c^{i+1} form an even pair of G_c^i.*

(B_i) 1. *G_c^{i+1} is a quasi-$\mathcal{B}$ graph.*
2. *w_c^{i+1} is a pivot of G_c^{i+1}.*
3. *w_c^{i+1} is not the top of a house of G_c^{i+1}.*

Proof. Let $c \in \{1, \ldots, \ell\}$. We show by induction on i that (A_i) and (B_i) hold.

Property (A_0) holds by vacuity. Graph G_1^1 is in $\mathcal{B}$, so w_c^1 is a pivot of this graph, and so (1) and (2) are satisfied when $c = 1$ and $i = 0$. To prove item 3, consider the beginning of Algorithm Cosine*: The set A of step 1.1 is empty, so w_1^1 is the minimum vertex of σ. Since the ordering σ was obtained by Algorithm LexBFS* applied on $\overline{G}$, Theorem 3.2 ensures that w_1^1 is not the middle of a P_5 in $\overline{G_1^1}$, so w_1^1 is not the top of a house in G_1^1.

Suppose $c \geq 2$. In the graph G_c^1, every vertex w_h with $h \in \{1, \ldots, c-1\}$ is adjacent to all other vertices of the graph; moreover, $G_c^1 \setminus \{w_1, \ldots, w_{c-1}\}$ is in $\mathcal{B}$, since it is a subgraph of G. It follows that G_c^1 is actually in $\mathcal{B}$, and so w_c^1 is a pivot of this graph. At this step of Algorithm Cosine* the set A of step 1.1 is empty, so at step 1.2 every vertex of $G_c^1 \setminus \{w_1, \ldots, w_{c-1}\}$ has no neighbor colored c and has the maximum number of neighbors in A, so the vertex $w_c^1 = x_c^1$ that is selected is the minimum for σ in $G_c^1 \setminus \{w_1, \ldots, w_{c-1}\}$, and Theorem 3.2 ensures that this vertex is not the top of a house in $G_c^1 \setminus \{w_1, \ldots, w_{c-1}\}$. Since every vertex w_h with $h \in \{1, \ldots, c-1\}$ is adjacent to all other vertices of the graph, it follows that w_c^1 is not the top of a house in G_c^1.

Now suppose that $i \geq 1$ and that (A_{i-1}) and (B_{i-1}) hold. Claim 1 implies immediately that (A_i) holds. It remains to prove (B_i). By (A_i), (B_{i-1}), and Lemmas 5.2 and 5.3, the graph G_c^{i+1} contains no odd hole and no antihole different from $\overline{C}_6$.

Suppose that G_c^{i+1} contains a $\overline{C}_6$, with vertices $a_1, a_2, a_3, a_4, a_5, a_6$ and nonedges a_1a_2, a_2a_3, a_3a_4, a_4a_5, a_5a_6, a_6a_1. If w_c^{i+1} is not one of the a_i's, then this $\overline{C}_6$ is also contained in G_c^i, a contradiction. So, by symmetry, we may assume that $w_c^{i+1} = a_1$. By the definition of contraction, both w_c^i, x_c^{i+1} miss a_6 and a_2, and each of a_3, a_4, a_5 sees at least one of w_c^i, x_c^{i+1}. At least one of w_c^i, x_c^{i+1} sees both a_3, a_5, for otherwise either w_c^i-a_3-a_5-x_c^{i+1} or w_c^i-a_5-a_3-x_c^{i+1} is a chordless path between w_c^i and x_c^{i+1}, a contradiction to (A_i). Call u a vertex of w_c^i, x_c^{i+1} that sees both a_3, a_5, and call v the other one. None of u, v sees all of a_3, a_4, a_5, for otherwise a $\overline{C}_6$ is contained in G_c^i. So u misses a_4, and so v sees a_4 and misses at least one of a_3, a_5. By symmetry we can assume that v misses a_3. But then v-$a_4a_2a_6$-a_3 is a bull of G_c^i of which w_c^i is not an ear, a contradiction. So G_c^{i+1} contains no $\overline{C}_6$.

Suppose that G_c^{i+1} contains a bull a_1-$a_2a_3a_4$-a_5 such that w_c^{i+1} is not an ear of this bull. If w_c^{i+1} is not in the bull, then the bull is also contained in G_c^i and w_c^i is not in it, which contradicts the fact that w_c^i is a pivot of G_c^i. So, by symmetry, we may assume that $w_c^{i+1} = a_1$ or $w_c^{i+1} = a_3$. If $w_c^{i+1} = a_1$, then w_c^i, x_c^{i+1} miss all of a_3, a_4, a_5, and at least one of w_c^i, x_c^{i+1} sees a_2; but this yields a bull in G_c^i of which w_c^i is not an ear, a contradiction. If $w_c^{i+1} = a_3$, then both w_c^i, x_c^{i+1} miss both a_1, a_5, and at least one of w_c^i, x_c^{i+1} sees both a_2, a_4, for otherwise either w_c^i-a_2-a_4-x_c^{i+1} or w_c^i-a_4-a_2-x_c^{i+1} is a chordless path between w_c^i and x_c^{i+1}, a contradiction to (A_i). But this yields a bull in G_c^i of which w_c^i is not an ear, a contradiction.

It follows from the preceding two paragraphs that G_c^{i+1} is a quasi-$\mathcal{B}$ graph and that w_c^{i+1} is a pivot of G_c^{i+1}.

Now suppose that w_c^{i+1} is the top of a house in G_c^{i+1} with vertices a_1, a_2, a_3, a_4, a_5 and nonedges $a_1a_2, a_2a_3, a_3a_4, a_4a_5$. So $w_c^{i+1} = a_3$. In G_c^i, both w_c^i, x_c^{i+1} miss a_2, a_4. Vertex w_c^i misses at least one of a_1, a_5, for otherwise it is the top of a house in G_c^i, a contradiction to (B_{i-1}). By symmetry, we may assume that w_c^i misses a_5, and so x_c^{i+1} sees a_5. Then x_c^{i+1} also sees a_1, for otherwise w_c^i-a_1-a_5-x_c^{i+1} is a path that contradicts (A_i). Then w_c^i misses a_1, for otherwise w_c^i-$a_1x_c^{i+1}a_5$-a_2 is a bull in G_c^i of which w_c^i is not an ear. Note that, in G_c^i, vertices $a_1, a_2, x_c^{i+1}, a_4, a_5$ induce a house, of which x_c^{i+1} is the top, and w_c^i misses all of them. Let us consider the situation when Algorithm COSINE* selects x_c^{i+1}. Let A be the set defined at step 1.1 of the algorithm. Since w_c^i misses all of the a_i's, none of them is in A. Let $T = N(x_c^{i+1}) \cap A$, and consider any vertex t of T. By the definition of T, vertex t sees x_c^{i+1} and w_c^i in G_c^i. If t misses both a_1, a_5, then t sees a_4, for otherwise t-$x_c^{i+1}a_5a_1$-a_4 is a bull in G_c^i of which w_c^i is not an ear, and similarly t sees a_2, but then w_c^i-ta_4a_2-a_5 is a bull in G_c^i of which w_c^i is not an ear. So t sees at least one of a_1, a_5, say a_1. Then t sees a_4, for otherwise w_c^i-$tx_c^{i+1}a_1$-a_4 is a bull in G_c^i of which w_c^i is not an ear. Then t sees a_2, for otherwise w_c^i-ta_1a_4-a_2 is a bull in G_c^i of which w_c^i is not an ear. Then t sees a_5, for otherwise w_c^i-ta_4a_2-a_5 is a bull in G_c^i of which w_c^i is not an ear. So every vertex of T sees a_1, a_2, a_4, a_5. Now a_1, a_2, a_4, a_5 are all uncolored vertices that have no neighbor colored c and have at least as many neighbors in A as x_c^{i+1}, so they have the maximum number of neighbors in A, and according to the ordering σ we have $x_c^{i+1} < \min\{a_1, a_2, a_4, a_5\}$. By Theorem 3.2, x_c^{i+1} is not the top of a house, a contradiction. Thus the claim holds. □

Claim 2 implies that in the sequence $G = G_1^1, \dots, G_\ell^{k_\ell}$, each graph other than the first one is obtained from its predecessor by contracting an even pair of the predecessor. Then Lemma 5.1 applied successively along the sequence implies that $\omega(G) = \omega(G_\ell^{k_\ell})$ and $\chi(G) = \chi(G_\ell^{k_\ell})$; but $\chi(G_\ell^{k_\ell}) = \omega(G_\ell^{k_\ell}) = \ell$ since $G_\ell^{k_\ell}$ is a clique of size ℓ; so the algorithm does color the input graph optimally with $\omega(G)$ colors. This completes the proof of the theorem. □

Coloring a graph in $\mathcal{B}$ takes time $\mathcal{O}(nm)$ since algorithm LEXBFS* applied on $\overline{G}$ has complexity $\mathcal{O}(nm)$ and Algorithm COSINE* too.

6. Finding a maximum clique. We can extend the preceding algorithms by another greedy algorithm, which, in the case of a graph in class $\mathcal{B}$, will produce in linear time a clique of maximum size. Let G be any graph given with a coloring of its vertices using ℓ colors. Then we can apply the following algorithm to build a set Q:

```
ALGORITHM CLIQUE
Input: A graph $G$ and a coloring of its vertices using $\ell$ colors.
Output: A set $Q$ that consists of $\ell$ vertices of $G$.
Initialization: Set $Q := \emptyset$, $c := \ell$, and for every vertex $x$ set $q(x) := 0$;
General step: While $c \neq 0$ do:
Pick a vertex $x$ of color $c$ that maximizes $q(x)$, do $Q := Q \cup \{x\}$, for
```

every neighbor y of x do $q(y) := q(y) + 1$, and do $c := c - 1$.
Algorithm Clique can be implemented in time $\mathcal{O}(m + n)$. To do this, at the step where the vertices of color c are examined, keep one vertex of color c that maximizes the counter q, and update the counter of the neighbors of that vertex.

We claim that when the input consists of a graph G in class $\mathcal{B}$, with the coloring produced by Algorithm LexBFS* followed by Algorithm Cosine*, the output Q of Algorithm Clique is a clique of size ℓ. Actually this will be true in a more general framework.

Lemma 6.1. *Let G be a graph given with a coloring of its vertices using ℓ colors. Call its vertices $x_1^1, x_1^2, \dots, x_1^{k_1}, x_2^1, \dots, x_2^{k_2}, \dots, x_\ell^1, \dots, x_\ell^{k_\ell}$, so that vertices of subscript c have color c. Define the corresponding sequence of graphs G_c^i and vertices w_c^i $(1 \le c \le \ell,\ 1 \le i \le k_c)$ obtained by successive contractions as in the preceding section. Suppose that for each color $c = 1, \dots, \ell - 1$, we have the following:*

(i) *Every vertex of color strictly greater than c has a neighbor of color c.*

(ii) *For each $i = 1, \dots, k_c - 1$, the graph G_c^i contains no chordless path on four vertices whose endvertices are w_c^i and x_c^{i+1}.*

Let Q be a clique whose vertices have colors strictly greater than c for some $c \in \{1, \dots, \ell - 1\}$. Then there is a vertex of color c that is adjacent to all of Q.

Proof. For $i = 1, \dots, k_c$, consider the following Property P_i: "In the graph G_c^i, vertex w_c^i is adjacent to all of Q." Note that Property P_{k_c} holds by (i) and by the definition of $w_c^{k_c}$. We may assume that Property P_1 does not hold, for otherwise the lemma holds with vertex $x_c^1 = w_c^1$. So there is an integer $i \in \{2, \dots, k_c\}$ such that P_i holds and P_{i-1} does not. Then, in the graph G_c^{i-1}, vertex x_c^i must be adjacent to all of Q, for otherwise Q contains vertices a, b such that a is adjacent to w_c^{i-1} and not to x_c^i and b is adjacent to x_c^i and not to w_c^{i-1}, and then the path w_c^{i-1}-a-b-x_c^i contradicts (ii). So the lemma holds with vertex x_c^i. □

Lemma 6.2. *Let G be a graph in class $\mathcal{B}$, and let $x_1^1, x_1^2, \dots, x_1^{k_1}, x_2^1, \dots, x_2^{k_2}, \dots, x_\ell^1, \dots, x_\ell^{k_\ell}$ be a coloring produced by Algorithm* LexBFS* *applied on $\overline{G}$ followed by Algorithm* Cosine* *applied on G. Then, when Algorithm* Clique *is run on this input it produces a clique of size $\omega(G)$.*

Proof. Consider the set Q maintained during Algorithm Clique. We claim that, for each $c = \ell, \ell - 1, \dots, 1$, at the end of step c the set Q is a clique of size $\ell - c + 1$ that contains one vertex of each color $c, \dots, \ell$. This is clear when $c = \ell$. At the general step, Lemma 6.1 ensures that there exists a vertex of color $c - 1$ that is adjacent to all of Q. So Algorithm Clique will select such a vertex, add it to Q, and so the claim remains true at the end of that step. Thus the algorithm ends with a clique Q of size ℓ. Since G admits a coloring of size ℓ, we have $\ell = \chi(G) = \omega(G)$. □

7. Comments. We observe that the hypothesis of Lemma 6.2 actually yields some slightly stronger properties:

(a) For any color c, every vertex of color c lies in a clique of size c; and more generally, every clique whose smallest color is c is included in a clique that contains a vertex of each color $1, \dots, c$. This is a consequence of Lemma 6.1 that can be derived just like Lemma 6.2. A coloring that has this property is called *strongly canonical* in [22].

(b) The set of vertices of color 1 is a stable set that intersects all maximal cliques of G. This too can be derived easily from Lemma 6.1. Such a set is called a *strong stable set* in [23]. Thus every graph G in class $\mathcal{B}$ is *strongly perfect* (i.e., every induced subgraph of G has a strong stable set), which was also a corollary of Hayward's result [18]. Moreover, using for graphs in $\mathcal{B}$ the idea from Hoàng [24, Theorem 2.1], this implies that one can find a minimum weighted coloring and a maximum weighted clique for a graph in $\mathcal{B}$ in time $O(n^2m)$.

The coloring algorithm is "robust" [30] in the sense that the input graph can be any graph G, and if G is not in $\mathcal{B}$ and the output coloring is not optimal, it can detect this fault. To do this we apply Algorithm LexBFS* on $\overline{G}$ followed by Algorithm Cosine* and Algorithm Clique on G, and we need only check whether Q is a clique (which can be done in linear time). If Q is a clique, then the coloring is optimal since it uses ℓ colors and Q has size ℓ. If Q is not a clique, then we know that the input graph is not in $\mathcal{B}$.

Since every graph in $\mathcal{B}$ admits a perfect ordering, as proved in [18], one may wonder whether the ordering in which the vertices are colored by Algorithm LexBFS* applied on $\overline{G}$ followed by Algorithm Cosine* applied on G gives such a perfect order. But here is a counterexample. Let G be the graph on six vertices a, b, c, d, e, f, where a-b-c-d-e is a path on five vertices and f is adjacent to a, c, d, e. Then Algorithm LexBFS* applied on $\overline{G}$ can produce the ordering $f < b < c < e < d < a$ and Algorithm Cosine* can color the vertices in the ordering $f < b < c < e < a < d$. This is not a perfect ordering for G since the four vertices b, c, d, e form an "obstruction" [6] since $b < c$ and $e < d$.

REFERENCES

[1] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Menlo Park, CA, 1974.

[2] C. Berge, *Les problèmes de coloration en théorie des graphes*, Publ. Inst. Statist. Univ. Paris, 9 (1960), pp. 123–160.

[3] C. Berge, *Graphs*, 2nd ed., North-Holland, Amsterdam, New York, 1985.

[4] M. E. Bertschi, *Perfectly contractile graphs*, J. Combin. Theory B, 50 (1990), pp. 222–230.

[5] M. Chudnovsky, N. Robertson, P. Seymour, and R. Thomas, *The strong perfect graph theorem*, Ann. Math. (2), 164 (2006), pp. 51–229.

[6] V. Chvátal, *Perfectly ordered graphs*, in Topics on Perfect Graphs, C. Berge and V. Chvátal, eds., North-Holland Math. Stud., 88, North-Holland, Amsterdam, New York, 1984, pp. 63–65.

[7] V. Chvátal, *A Class of Perfectly Orderable Graphs*, Rpt. 89573-OR, Forschungsinstitut für Diskrete Mathematik, Bonn, 1989.

[8] V. Chvátal and N. Sbihi, *Bull-free Berge graphs are perfect*, Graphs Combin., 3 (1987), pp. 127–139.

[9] C. M. H. de Figueiredo and F. Maffray, *Optimizing bull-free perfect graphs*, SIAM J. Discrete Math., 18 (2004), pp. 226–240.

[10] C. M. H. de Figueiredo, F. Maffray, and O. Porto, *On the structure of bull-free perfect graphs*, Graphs Combin., 13 (1997), pp. 31–55.

[11] C. M. H. de Figueiredo, F. Maffray, and O. Porto, *On the structure of bull-free perfect graphs*, 2: *The weakly chordal case*, Graphs Combin., 17 (2001), pp. 435–456.

[12] H. Everett, C. M. H. de Figueiredo, C. Linhares Sales, F. Maffray, O. Porto, and B. A. Reed, *Even pairs*, in Perfect Graphs, J. L. Ramírez-Alfonsín and B. A. Reed, eds., Wiley Interscience, Chichester, UK, 2001, pp. 67–92.

[13] J. Fonlupt and J. P. Uhry, *Transformations that preserve perfectness and h-perfectness of graphs*, Ann. Discrete Math., 16 (1982), pp. 83–95.

[14] T. Gallai, *Transitiv Orientierbare Graphen*, Acta Math. Acad. Sci. Hungar., 18 (1967), pp. 25–66.

[15] M. Gröstchel, L. Lovász, and A. Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 169–197.

[16] M. Habib, F. de Montgolfier, and C. Paul, *A simple linear-time modular decomposition algorithm for graphs, using order extension*, in Proceedings of SWAT 2004, Lecture Notes in Comput. Sci. 3111, 2004, pp. 187–198.

[17] R. B. Hayward, *Weakly triangulated graphs*, J. Combin. Theory B, 39 (1985), pp. 200–208.

[18] R. B. Hayward, *Bull-free weakly chordal perfectly orderable graphs*, Graphs Combin., 17 (2000), pp. 479–500.

[19] R. B. Hayward, J. Spinrad, and R. Sritharan, *Weakly chordal graph algorithms via handles*, in Proceedings of 11th Annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, ACM Press, New York, 2000, pp. 42–49.

[20] A. Hertz, *A fast algorithm for coloring Meyniel graphs*, J. Combin. Theory B, 50 (1990), pp. 231–240.

[21] A. Hertz, C*osine*: *A new graph coloring algorithm*, Oper. Res. Lett., 10 (1991), pp. 411–415.

[22] A. Hertz and D. de Werra, *Connected sequential colorings*, Discrete Math., 74 (1989), pp. 51–59.

[23] C. T. Hoàng, *On a conjecture of Meyniel*, J. Combin. Theory B, 42 (1987), pp. 302–312.

[24] C. T. Hoàng, *Efficient algorithms for minimum weighted colouring of some classes of perfect graphs*, Discrete Appl. Math., 55 (1994), pp. 133–143.

[25] B. Lévêque, F. Maffray, B. A. Reed, and N. Trotignon, *Coloring Artemis Graphs*, Res. Rep. 135, Leibniz Laboratory, Grenoble, France, 2005.

[26] A. Lubiw, *Doubly lexical ordering of matrices*, SIAM J. Comput., 16 (1987), pp. 854–879.

[27] C. Linares Sales and B. A. Reed, eds., *Recent Advances in Algorithms and Combinatorics,* CMS Books Math./Ovurages Math. SMC 11, Springer-Verlag, New York, 2003.

[28] F. Maffray and M. Preissmann, *A translation of Tibor Gallai's paper: Transitiv orientierbare Graphen*, in Perfect Graphs, J. L. Ramírez-Alfonsín and B. A. Reed, eds., Wiley Interscience, Chichester, UK, 2001, pp. 25–66.

[29] H. Meyniel, *A new property of critical imperfect graphs and some consequences*, European J. Combin., 8 (1987), pp. 313–316.

[30] V. Raghavan and J. Spinrad, *Robust algorithms for restricted domains*, J. Algorithms, 48 (2003), pp. 160–172.

[31] B. A. Reed and N. Sbihi, *Recognizing bull-free perfect graphs*, Graphs Combin., 11 (1995), pp. 171–178.

[32] D. J. Rose, R. E. Tarjan, and G. S. Lueker, *Algorithmic aspects of vertex elimination on graphs*, SIAM J. Comput., 5 (1976), pp. 266–283.

SIAM J. DISCRETE MATH.
Vol. 21, No. 4, pp. 1019–1034

COMBINATORIAL OPTIMIZATION WITH EXPLICIT DELINEATION OF THE GROUND SET BY A COLLECTION OF SUBSETS*

MOSHE DROR† AND JAMES B. ORLIN‡

Abstract. We examine a selective list of combinatorial optimization problems in NP with respect to inapproximability (Arora and Lund (1997)) given that the ground set of elements N has additional characteristics. For each problem in this paper, the set N is expressed explicitly by subsets of N either as a partition or in the form of a cover. The problems examined are generalizations of well-known classical graph problems and include the minimal spanning tree problem, a number of elementary machine scheduling problems, the bin-packing problem, and the travelling salesman problem (TSP). We conclude that for all these generalized problems the existence of a polynomial time approximation scheme (PTAS) is impossible unless P=NP. This suggests a partial characterization for a family of inapproximable problems. For the generalized Euclidean TSP we prove inapproximability even if the subsets are of cardinality 2.

Key words. inapproximability, approximation algorithms, subset travelling salesman, generalized bin packing, generalized scheduling problems

AMS subject classifications. 68Q17, 68Q25, 68W25, 90B35, 90C27

DOI. 10.1137/050636589

1. Introduction.

When attempting to define combinatorial optimization, the combinatorial optimization community commonly states it as "the mathematical study of the arrangement, grouping, ordering, or selection of discrete objects, usually finite in number" (see Lawler (1976)). In the majority of the many books about combinatorial optimization topics, there is no attempt (or need) to define it. In the well-acknowledged book by Nemhauser and Wolsey (1988), the authors propose the following generic description of a combinatorial optimization problem.

Let $N = \{1, 2, \ldots, n\}$ be a finite set (referred to as the ground set of elements) and let $c = (c_1, c_2, \ldots, c_n)$ be an n-vector. For $F \subseteq N$, define $c(F) = \sum_{j \in F} c_j$. Given a collection of subsets $\mathcal{F}$ of N, the *combinatorial optimization* problem is

$$\text{(CP)} \qquad \max\{c(F) : F \in \mathcal{F}\}.$$

With such a generic description of combinatorial optimization problems, a characterization of a problem is largely determined by the description of the collection $\mathcal{F}$ of subsets on N. For instance, in the case of the travelling salesman problem (TSP), the collection of subsets $\mathcal{F}$ corresponds to the edges of a given graph (arcs, for the directed TSP), where a walk over each such subset forms a cyclic permutation of the

*Received by the editors July 21, 2005; accepted for publication (in revised form) June 8, 2007; published electronically January 9, 2008.
http://www.siam.org/journals/sidma/21-4/63658.html

†MIS Department, College of Business and Public Administration, University of Arizona, Tucson, AZ 85721, and Operations Research Center, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139 (mdror@eller.arizona.edu). The research of this author was partially supported through grant ORP-0303 from the Chinese Academy of Sciences' Key Lab of Complex Systems and Intelligence Science, Beijing.

‡Sloan School of Management, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139 (jorlin@mit.edu). The research of this author was partially supported through NSF grant DMI-0217123 and ONR grant N00014-98-1-0317.

nodes of the given graph. The nodes and the edges of the graph are the elements of the ground set.

This paper focuses on a number of combinatorial optimization problems characterized by a further delineation regarding the ground set of elements. As an example of what is meant here by a delineation of the ground set, we describe a version of the classical TSP known as the *generalized TSP* (GTSP).

One is given a graph $G = (V, E)$ of $n > 3$ nodes together with distances (weights) for the edges in E. The set of n nodes (the set V) is partitioned into k $(3 < k \leq n)$ nonempty subsets, and the "travelling salesman" tour for the GTSP has to visit each subset exactly once. That is, the tour visits one node in each of the subsets. For the GTSP, the optimal solution forms a distance minimizing circuit (if one exists) of k nodes with one node from each subset.

In Dror and Haouari (2000) an attempt was made to respond intuitively to a hypothetical question about the hardness relationship of finding solutions for a problem and its generalized version. In the combinatorial problems discussed in this paper, we assume either an explicit set partition of the set N or an explicit collection of subsets of N that cover N. We provide the motivation for a number of the problems. However, many problems are of interest because of their respective places in the combinatorial literature. The problems that were proven NP-hard in Dror and Haouari (2000) are (the generalized) versions of the minimal spanning tree (MST) problem, the assignment problem (AP), the Chinese postman problem (CPP), some machine scheduling problems such as the two-machine flow-shop problem, the bin-packing (BP) problem, and the TSP. In this paper we show that the above generalizations lead to inapproximability results as defined in Arora and Lund (1997). We also examine here the generalized version of a number of other problems. It is not clear how one would formally frame this set of generalized problems as a well-defined problem class, and we will not attempt to do so here. We simply prove an appropriate inapproximability result for the problems listed above. We begin by examining some classical graph problems in section 2, starting with the generalized minimum spanning tree (GMST) problem in section 2.1. In section 2.2 we examine path and circuit problems. In section 2.3 we turn to matching problems, including the generalized Chinese postman problem (GCPP). In section 2.4 we continue with acyclic graph problems. In the process of examining the different problems and the complexity of their generalizations, it seems appropriate to separate the subset partition generalizations from subset covering generalizations since the hardness of the later class is not surprising. Thus, a subset covering generalization—a subset bin-packing (SBP) problem is examined in section 3. Section 4 examines a number of machine scheduling problems. In section 5 we return to the GTSP and prove that even the Euclidean planar GTSP with subsets of cardinality 2 is inapproximable. Short conclusions follow in section 6.

A list in chronological order of new inapproximability results is as follows: (1) the Generalized Acyclic Graph Problem, (2) the SBP problem, (3) the Generalized Due Date Problem, (4) the generalized Euclidean planar TSP with subsets of cardinality 2. We also conjecture that the Generalized Weighted Sum of Completion Times Problem is inapproximable.

2. Graph problems.

2.1. The generalized minimum spanning tree (GMST) problem.

Given a connected undirected graph $G = (V, E)$ with a node set V and an edge set E, a *tree* in G is a connected subgraph $T = (V', E')$ containing no cycles. If $V' = V$, then T is a *spanning tree* for the graph G. Given a weight function $w : E \to \mathcal{Z}^+$ ($\mathcal{Z}^+$

denotes the set of positive integers), a minimal weight spanning tree is a spanning tree T^* for which $\sum_{e \in T^*} w_e \leq \sum_{e \in T} w_e$ for all $T : T$ spanning trees of G (see Magnanti and Wolsey (1995) for an in-depth overview of combinatorial optimization focused on trees).

Here we examine the GMST problem in which one requires a tree to contain at least one node out of every subset of nodes (node partition) in a graph G. This problem arises in irrigation network design in desert areas where parcels of land share a common water source (see Dror, Haouari, and Chaouachi (2000)). It is also the problem of designing a backbone (communication, high voltage, sewage, etc.) network at minimum cost where there is flexibility in locating the nodes of the network. The GMST was independently introduced by Myung, Lee, and Tcha (1995) and Dror, Haouari, and Chaouachi (2000).

DEFINITION. *Given a connected graph $G = (V, E)$, a positive integer K (≥ 3), and a partition of the node set V, where $V = \bigcup_{k=1}^{K} V_k$, $V_i \cap V_j = \emptyset$ for $i \neq j$, and $V_i \neq \emptyset$, $i = 1, \ldots, K$, together with an edge weight w_e, $e \in E$, a* GMST *is a connected subgraph of G of minimal weight with no cycles, denoted here by $GT = (V^{`}, E^{`})$, such that $V^{`} \cap V_i \neq \emptyset$, $i = 1, \ldots, K$.*

Note that in the case where $V = \bigcup_{k=1}^{K} V_k$, $V_k \subset V$ for all k (the V_k subsets cover V but are not necessarily disjoint), the construction of a minimal weight tree spanning with at least one node from each V_k can be easily transformed into an equivalent GMST problem with disjoint node subsets by making duplicates of nodes.

Complexity of the GMST. If $|V_k| = 1$, $k = 1, \ldots, K$, then the GMST problem reduces to the classical MST. A "greedy" solution procedure (for instance, Kruskal's algorithm) constructs a minimal weight spanning tree by examining the edges one at a time in nondecreasing order of weight, rejecting an edge if it forms a cycle with those already chosen. Moreover, when $K = 1$ or 2, the GMST is trivial. In general, the GMST is NP-hard in the strong sense (Dror, Haouari, and Chaouachi (2000), Myung, Lee, and Tcha (1995). The two papers present different proofs for this fact. Myung, Lee, and Tcha (1995) provide an inappoximabilty result for the GMST restricted to graphs where the nodes in each of the node subsets (node partition) are not incident to each other, asking for a spanning tree that spans exactly one node from each subset of nodes. Thus, the GMST problem can be stated as an existence decision problem. They prove that for this GMST, there is no polynomial time heuristic with a fixed performance guarantee. A more direct prove of the GMST inapproximability result is presented in Dror, Haouari, and Chaouachi (2000) by a transformation from the Steiner tree problem, which is known to be APX-complete, proving that the GMST is also APX-complete (see Ausiello, Crescenzi, and Protasi (1995) for complexity concepts).

2.2. Generalized path and circuit problems. Generalized path, circuit, or matching problems are fundamental problems in numerous applications. Here we state formally a few of the basic problems and provide their computational hardness.

Generalized Path Problem with Forbidden Pairs. We restate an "old" (1976) classical problem which can be trivially recast in our framework of subset delineation of the ground set.

INSTANCE. A directed graph $G = (V, A)$ and m pairs of arcs (m subsets of A), $|V| = n$ nodes, and two designated nodes s and t.

QUESTION. Is there a path from s to t that includes at most one arc of every pair?

This problem is equivalent to a problem of determining the existence of a path with forbidden pairs that was proven NP-complete by transformation from 3SAT in Gabow, Maheshwari, and Osterweil (1976). The optimization version of this problem (minimizing the number of paired arcs in such path) is NP-hard and inapproximable.

Cycle with at most one node per subset problem—Generalized Circuit Problem 1.

INSTANCE. A directed graph $G = (V, A)$ with arc set A and disjoint subsets of nodes $S_1, \dots, S_K$, $S_i \subset V$, $i = 1, \dots, K$.

QUESTION. Is there a directed cycle that includes at most one node of any subset?

This problem was proven NP-complete by Thompson and Orlin (1989). We restate the proof below.

THEOREM 1. *The cycle with at most one node per subset problem is NP-complete.*

Proof. The proof is based on transformation from the Hamiltonian path problem. One is given a directed graph $G = (V, A)$ with $|V| = n$ nodes. Establishing the existence of a Hamiltonian path on G is NP-complete (see Garey and Johnson (1979)).

The transformation is as follows: For each node $i \in V$ create a set S_i of n labelled copies. That is, $S_i = \{i^1, \dots, i^n\}$. For every arc $(i, j) \in A$, create a set of $n-1$ copies $\{(i^k, j^{k+1}) : k = 1, \dots, n-1\}$. Also add arcs from j^n i^1 for all i, j, $i \neq j$. Observe that any path from i^1 to j^n corresponds to a walk of $n-1$ arcs in the original graph. If we require that the path has at most one node from each subset S_i, then any feasible path corresponds to a Hamiltonian path. Clearly the other direction holds as well. Thus the equivalence. □

Again, the optimization version of this cycle with at most one node per subset problem is NP-hard and inapproximable. That is, in this case of a directed graph, there is no polynomial time approximation scheme for the problem of constructing a minimal weight cycle which contains at most one node from each subset unless P=NP.

The next problem is similar to the node subsets replaced by arc subsets. Complexity implications directly follow.

Cycle with at most one arc per subset problem—Generalized Circuit Problem 2.

INSTANCE. A directed graph $G = (V, A)$ with node set V and arc set A decomposed into disjoint subsets of arcs.

QUESTION. Is there a directed cycle that includes at most one arc from each subset of arcs?

One can transform Generalized Circuit Problem 1 into this problem by performing node splitting on each node (see Ahuja, Magnanti, and Orlin (1993)) and then decomposing the node-splitting arcs into subsets.

2.3. Generalized matching.

Matching problems play an important role in combinatorial optimization. Below we present two generalized matching problems—one hard and one easy, followed by the generalized Chinese postman problem (GCPP).

3-*Dimensional Matching Optimization.*

INSTANCE. A collection of triples $T \subseteq X \times Y \times Z$, where $X = \{x_1, \dots, x_n\}$, $Y = \{y_1, \dots, y_n\}$, $Z = \{z_1, \dots, z_n\}$, and a positive integer K.

PROBLEM. Find a maximum cardinality subset $M \subseteq T$ so that no two elements of M agree on any coordinate.

Generalized Matching Problem 1.

INSTANCE. An undirected multigraph $G = (V, E)$, where $E = \bigcup_{i=1}^{n} E_i$, $E_i \cap E_j = \emptyset$, $i \neq j$.

PROBLEM. Find the maximum cardinality subset M' so that M' is a matching in G and so that $|M' \cap E_k| \leq 1$ for all $k = 1, \dots, n$.

THEOREM 2. *The generalized matching problem is APX-complete.*

Proof. The generalized matching problem is in APX because the greedy algorithm is guaranteed to produce a matching that is within 1/3 of being maximum. We will provide a transformation from the 3-Dimensional Matching Optimization Problem, which was proved to be APX-complete by Kann (1991).

Let $T \subseteq X \times Y \times Z$ be an instance of the 3-dimensional matching problem. We construct an undirected graph $G = (V, E)$ together with a partition of E into disjoint sets $E_1, \ldots, E_n$. For each triple $(i, j, k) \in T$, create an edge $(i, j) \in E_k$.

Now suppose that $M \subseteq T$ is a feasible solution for the 3-dimensional matching problem. We construct a matching M' for the generalized matching problem as follows: For each triple $(i, j, k) \in M$, let M' contain arc $(i, j) \in E_k$. It follows that M' is a generalized matching and $|M'| = |M|$.

Conversely, suppose M' is a generalized matching. Then $|M' \cap E_k| \leq 1$ for each k. We create a 3-dimensional matching as follows: if $(i, j) \in M' \cap E_k$, then $(i, j, k) \in M$. It follows that M is a 3-dimensional matching, and $|M'| = |M|$. □

We break the routine of inapproximable problems by examining a following subset version of a maximal weight matching problem.

Node-based Generalized Matching Problem.

INSTANCE. An undirected graph $G = (V, E)$, with node partition $V = \bigcup_{i=1}^{r} V_i$, $V_i \cap V_j = \emptyset$, $i \neq j$. Let w_i be a real value weight for node $i \in V$ and an integer $K > 0$.

PROBLEM. Find a matching M and a $V' \subseteq V$ such that every node in V' is matched in M, $|V' \cap V_j| \leq 1$ for all $i = 1, \ldots, r$, and $\sum_{i \in V'} w_i \geq K$.

This problem is solvable using weighted matroid intersection. The collection V'' of nodes is independent in the first matroid if there is a matching M that matches each node of V''. A collection V'' of nodes is independent in the second matriod if it contains at most one node of V_j for each j. The answer to the instance of node-based generalized matching is yes if and only if the max weight matroid intersection has weight of at least M.

Note that we do not require that all nodes of M be included in V''.

The GCPP. The classical *Chinese postman problem* (CPP) is a problem that focuses on traversals of the graph's edges. That is, given a connected undirected graph $G = (V, E)$ with a nonnegative bounded weight for each edge, the quest is to construct a minimum weight circuit that traverses each edge in E at least once. Edmonds and Johnson (1973) have shown that the CPP can be solved efficiently (in polynomial time). The most commonly stated time complexity for the CPP is $O(|V|^3)$, even though the problem can be solved faster for sparse graphs (for more details see Dror (2000)).

Now suppose that the graph G is partitioned with respect to the edge set E into nonempty connected subgraphs $G_1, \ldots, G_K$. That is, $G_1 = (V_1, E_1), \ldots, G_K = (V_K, E_K)$, where each subgraph G_i is connected and $E_i \cap E_j = \emptyset$, $i \neq j$. The *GCPP* is defined as the problem of constructing a minimal weight circuit on G which visits at least one edge from each subgraph. To our knowledge, this problem was first introduced in Dror and Haouari (2000). The proof that the GCPP is NP-hard can be obtained very simply by transformation from a routing problem referred to in the literature as the *rural postman problem* (RPP). In fact, a special case of the GCPP with all but one of the subgraphs G_i being of one edge is already equivalent to the RPP. An additional fact (see Lenstra and Rinnooy Kan (1976)) is that the travelling salesman problem (TSP) is a special case of the RPP. Given this sequence of transformations and special cases together with the more recent results by Arora et al. (1998) and Trevisan (2000) that the TSP is inapproximable (even in Euclidean space of dimension $\log n$), the GCPP is also inapproximable.

2.4. Generalized Acyclic Graph Problem. In this subsection we examine generalizations of a number of acyclic graph problems. For more details and applications of acyclic graph problems we refer the reader to Ahuja, Magnanti, and Orlin (1993).

INSTANCE. A directed graph $G = (V, A)$, $|A| = m$, and m even. The arcs in A are partitioned into pairs (the pairs are disjoint). The graph is permitted to have multiple arcs with the same head and tail—a multigraph.

QUESTION. Is there a subset of arcs $A' \subset A$, $|A'| = m/2$, with one arc from each pair, such that A' is acyclic?

THEOREM 3. *The Generalized Acyclic Graph Problem is NP-complete.*

Proof. The proof is by transformation from 3-Satisfiability. □

3-*Satisfiability (*3*SAT).*

INSTANCE. List of literals $U = \{u_1, \bar{u}_1, u_2, \bar{u}_2, \ldots, u_n, \bar{u}_n\}$ and sequence of (conjunctive) clauses $C = (C_1, C_2, \ldots, C_m)$, where each clause C_i is a subset of U of cardinality 3.

QUESTION. Is there a truth assignment for the literals $u_1, \ldots, u_n$ that satisfies all the clauses in C, that is, a subset $U' \subseteq U$ such that $|U' \cap \{u_i, \bar{u}_i\}| = 1$, $1 \leq i \leq n$, and such that $|U' \cap C_i| \geq 1$, $1 \leq i \leq m$?

A truth assignment is an n-dimensional vector x^* of variables. If $x_i^* = 1$, then variable x_i is true. If $x_i^* = 0$, then variable x_i is false. In an instance of 3-satisfiability, we let clause $C_j = \{c_{j1}, c_{j2}, c_{j3}\}$, where c_{jr} is a literal for $r = 1$, 2, and 3. For example, if $C_j = (x_3, \bar{x}_5, \bar{x}_9)$, then $c_{j1} = x_3$, $c_{j2} = \bar{x}_5$, and $c_{j3} = \bar{x}_9$. In this case, clause C_j is *satisfied* by truth assignment x^* if $x_3^* = 1$ or $x_5^* = 0$ or $x_9^* = 0$. The collection C is satisfied if every clause in C is satisfied.

Let $C = \{C_1, \ldots, C_m\}$ be a collection of clauses for 3-SAT defined on a set $X = \{x_1, \ldots, x_n\}$ of variables. We define an instance of the Generalized Acyclic Graph Problem as follows.

The node set $V = \{s\} \cup \{x_1, \ldots, x_n\} \cup \{\bar{x}_1, \ldots, \bar{x}_n\} \cup \{w_{j1}, w_{j2}, w_{j3} : j = 1, \ldots, m\}$. That is, $|V| = 2n+1+3m$, with one node for each variable, one node for each variable complement, a node s, and $3m$ additional nodes ($w_{\cdot,\cdot}$) numbered in 1-1 correspondence with the literals in the m clauses.

For each clause (c_{j1}, c_{j2}, c_{j3}), we let $(\bar{c}_{j1}, \bar{c}_{j2}, \bar{c}_{j3})$ denote the complements of the literals in the clause.

The pairs of arcs are defined as follows:

1. For each variable x_j, there is a pair of arcs (s, x_j) and $(s, \bar{x}_j)$.
2. For each clause C_j with literals c_{j1}, c_{j2}, and c_{j3}, there are three pairs of arcs $\{(w_{j1}, w_{j2}), (\bar{c}_{j1}, s)\}$, $\{(w_{j2}, w_{j3}), (\bar{c}_{j2}, s)\}$, and $\{(w_{j3}, w_{j1}), (\bar{c}_{j3}, s)\}$.

When we refer to the node c_{jk} or $\bar{c}_{jk}$ above, we are referring to a node in $\{x_1, \ldots, x_n\} \cup \{\bar{x}_1, \ldots, \bar{x}_n\}$. For example, if $c_{j1} = \bar{x}_5$, then $(\bar{c}_{j1}, s)$ denotes the arc (x_5, s).

We will show that the instance of 3-SAT is satisfiable if and only if there is a subset A' of arcs, one from each pair, such that A' is acyclic.

We first suppose that there is a truth assignment for the instance of 3-SAT. We assume without loss of generality that the literal c_{j1} is satisfied for $j = 1, \ldots, m$. We obtain a collection A' of arcs, one arc from each pair as follows:

1. If x_j is true, then $(s, x_j) \in A'$; if x_j is false, then $(s, \bar{x}_j) \in A'$.
2. For each clause C_j, we select arcs $(\bar{c}_{j1}, s)$, (w_{j2}, w_{j3}), and (w_{j3}, w_{j1}).

We note that the arc $(\bar{c}_{j1}, s)$ does not create a directed cycle. Consider, for example, the case that $\bar{c}_{j1} = x_6$, and thus $(x_6, s) \in A'$. In this case, $c_{j1} = \bar{x}_6$, and by assumption, x_6 is false, and so $(s, \bar{x}_6) \in A'$. In particular, $(s, x_6) \notin A'$, and so (x_6, s) is not part of a directed cycle. Finally, the arcs (w_{j2}, w_{j3}) and (w_{j3}, w_{j1}) do not create a directed cycle.

Conversely, suppose that A' is a collection of arcs, one from each pair. And suppose that A' is acyclic. We create a truth assignment as follows:

1. If $(s, x_j) \in A'$, then x_j is true; if $(s, \bar{x}_j) \in A'$, then x_j is false.

For each clause C_j, we cannot select all three of (w_{j1}, w_{j2}), (w_{j2}, w_{j3}), and (w_{j3}, w_{j1}) since it would create a directed cycle. It follows that for each j, A' has an arc $(\bar{c}_{jr}, s)$ for some $r \in \{1, 2, 3\}$. It follows that $(s, \bar{c}_{jr}) \notin A'$, and so literal c_{jr} is satisfied. We conclude that x is a truth assignment. □

The problem of identifying the largest acyclic graph is already APX-hard (Papadimitriou and Yannakakis (1991)). Thus, identifying the largest acyclic graph with at most one arc from each pair is APX-hard.

Remark. In the above proof we permit multiple parallel arcs. We can get rid of the need for parallel arcs as follows:

First of all, we restrict the attention to 3SAT problems in which at least one clause contains x_j and at least one clause contains $\bar{x}_j$. Therefore, $n(j) \geq 2$, where $n(j)$ counts the number of clauses with literal j. For each variable x_j, we create nodes $\{x_{j1}, \dots, x_{j,n(j)}\}$ and $\{\bar{x}_{j1}, \dots, \bar{x}_{j,n(j)}\}$. We also create the following $2n(j) - 1$ pairs of arcs:

$$\begin{aligned}\{(s, x_{j1}), (s, \bar{x}_{j1})\} &\cup \{(s, x_{jr}), (x_{j1}, x_{jr}) : r = 2, \dots, n(j)\}\\ &\cup \{(s, \bar{x}_{jr}), (\bar{x}_{j1}, \bar{x}_{jr}) : r = 2, \dots, n(j)\}.\end{aligned}$$

Note that if arc $(s, x_{j1}) \in A'$, then there is a path from s to x_{jk}, $k = 1, \dots, n(j)$. If arc $(s, \bar{x}_{j1})$ is selected, then there is a path from s to $\bar{x}_{jk}$, $k = 1, \dots, n(j)$.

If $c_{jr} = \bar{x}_j$ and if it is the kth occurrence of the variable x_j in the clauses, the corresponding arc in the graph would be (x_{jk}, s). If $c_{jr} = x_j$ and if it is the kth occurrence of the variable x_j in the clauses, the corresponding arc in the graph would be $(\bar{x}_{jk}, s)$.

The transformation still carries through, and there are no multiple copies of any arc in the network.

3. Subset covering generalization—subset bin packing. Consider as before a set of elements $U = \{u_1, \dots, u_n\}$ together with m "covering" subsets $U_i \subset U$, $1 \leq i \leq m$. That is, $\bigcup_{i=1}^{m} U_i = U$. The size of a subset U_i is simply measured by the cardinality of this subset, implying that the different elements in U are of identical "volume" (the same size). In addition, there are an infinite number of bins of size Q—a positive integer. Following Coffman and Dror (1992), we ask what the minimal number of bins sufficient to pack all the subsets U_i, $1 \leq i \leq m$, is and call this problem the *subset bin-packing* (SBP) problem. Packing a subset U_i implies that the subset is contained (in its entirety) in at least one bin. Coffman and Dror (1992) (reproduced in Dror and Haouari (2000)) examine many well-known "good" heuristics for bin packing and show that they produce results with no bound guarantees when extended to the SBP problem. This leads to questions about hardness of the SBP problem, which we address below. But first we examine a related problem for the knapsack.

A similar problem was stated by Goldschmidt, Nehme, and Yu (1994) for the knapsack version of this problem referred to as the *set-union knapsack problem* (SKP), referred to here as a *subset knapsack problem* (SKP). Goldschmidt et al. also described a version for the bin packing referred to as the *set-union bin-packing* (SBP) *problem.* As opposed to the SBP problem as defined in the previous paragraph, they had positive integer (≥ 1) valued sizes associated with the elements in U.

A decision version of the SKP with unit size items can be stated in the form of a graph problem as follows.

INSTANCE. Given a bipartite graph $G = (V_1, V_2; E)$, denote by $v^1(i)$ all the nodes in V_1 adjacent to node $v_i \in V_2$. Let $0 < K \leq |V_1|$, $0 < B \leq |V_2|$ be two given integers.

QUESTION. Is there a subset $V_2' \subset V_2$ such that $|\bigcup_{i \in V_2'} v^1(i)| \leq K$ and $|V_2'| \geq B$?

Obviously, the optimization version of this problem is asking for maximizing the cardinality of the subset V_2' subject to $|\bigcup_{i \in V_2'} v^1(i)| \leq K$.

The unit size items SKP has been examined in the literature before. For instance, Khuller, Moss, and Naor (1999) and Moss (2001) provide detailed analysis of this and the more general nonidentical unit size SKP under the heading of *the budgeted maximum coverage problem.* They first provide a heuristic with constant approximation factor that is later improved to a factor of $(1 - 1/e)$ and prove that no approximation algorithm with performance guarantee of $1 - 1/e + \epsilon$ is likely to exist for any $\epsilon > 0$ unless $P \subseteq DTIME(n^{O(\log \log n)})$. This result has been first stated in Feige (1998). That is, these references establish that the unit size items SKP is also inapproximable. We do not reproduce their proof here. One of the immediate corollaries of this result is that the SBP problem is also inapproximable because if the number of subsets U_i (U_i corresponds to a node $i \in V_2$) that fit into one bin of size K is greater than B, then the rest of the subsets $(m - B)$ will fit into a "small" number of other bins.

Note that a stronger inapproximability result can be deduced from an appropriate transformation of a the set covering problem to a special case of the SBP problem.

4. Machine scheduling.

This section contains a number of generalized scheduling problems. We start by considering a single machine problem, which also introduces the notation.

Initially consider a set J of n jobs to be scheduled for processing on a single machine. Let $J = \bigcup_{i=1}^m J_i$, where each J_i is a nonempty subset of J. Each job $j \in J$ has a positive integer processing time requirement denoted by p_j. Assume that all the jobs are ready for processing at time zero. The machine can process at most one job at a time, and preemption of jobs is not allowed. Given a subset of jobs $\sigma \subset J$, let $C(\sigma)$ denote the minimal completion time of the jobs in σ. Given the subset representation or cover $\{J_i\}_{i=1}^m$ for the jobs in J, one would like to process some jobs and complete such a processing as soon as possible while processing at least one job from each subset. In other words, select a subset $\sigma^* \subset J$ such that σ^* shares at least one job with each of the subsets and its completion time is minimized. That is, $C(\sigma^*) = \min_{\sigma \subset J} C(\sigma)$; $\sigma \cap J_i \neq \emptyset$, $1 \leq i \leq m$. The corresponding scheduling problem which is that of minimizing makespan for the jobs in subset σ can be expressed as

$$\min_{\sigma} \{\max C_j : j \in \sigma,\ \sigma \cap J_i \neq \emptyset,\ 1 \leq i \leq m\}.$$

In terms of the 3-field notational convention in scheduling, this problem can be denoted as $1|ss|C_{max}$ ($\sum_j C_j$ if total flow time is the criterion), where ss stands for subset selection.

The interesting case in this scheduling problem is when the subsets J_i, $1 \leq i \leq m$, have pairwise nonempty intersections since when all the subsets are disjoint, the problem is solvable using the greedy algorithm. In Dror and Haouari (2000), it is proven that $1|ss|C_{max}$ ($\sum_j C_j$) are NP-hard in the strong sense by its equivalence with the *hitting set problem* (see Garey and Johnson (1979)). Below we restate the 0-1 integer programming formulation of the $1|ss|C_{max}$ problem taken from Dror and Haouari (2000) given that $|J| = n$:

$$\text{minimize} \sum_{j=1}^{n} p_j x_j \tag{4.1}$$

subject to

$$\sum_{j \in J_i} x_j \geq 1, \quad 1 \leq i \leq m, \tag{4.2}$$

$$x_j \in \{0, 1\}, \quad 1 \leq j \leq n. \tag{4.3}$$

Since the above formulation also corresponds to the SETCOVER problem, which is one of the representative inapproximable problems in Arora and Lund (1997), the inapproximability (which cannot be approximated to within a factor $o(\log n)$; see Raz and Safra (1997)) for the $1|ss|C_{max}$ follows.

This inapproximability result for the $1|ss|C_{max}$ problem implies similar results for a number of other machine scheduling problems. For instance, the two-machine flow-shop problem with the subsets J_i, $1 \leq i \leq m$, as a partition of the job set J denoted as $F2|ss^{\emptyset}|C_{max}$ (the $\emptyset$ symbol denotes the fact that the pairwise subset intersections of J_i's are empty) is also proven NP-hard in Dror and Haouari (2000).

Next, we examine two inapproximable generalizations of single machine scheduling problems not mentioned in Dror and Haouari (2000).

4.1. Generalized Due Date Problem.

INSTANCE. n pairwise disjoint subsets of jobs, labelled $J_1, \ldots, J_n$. The processing time of the jth job in J_i is p_{ij}, and its due date is d_{ij}.

PROBLEM. What is the minimum number of late jobs in a feasible generalized (selecting at least one job from each subset) schedule?

THEOREM 4. *The Generalized Due Date Problem is strongly NP-hard and inapproximable.*

Proof. We first prove that the decision version of this problem is strongly NP-complete. The proof is by transformation from 3-Partition.

3-*Partition.*

INSTANCE. Integers $a_1, \ldots, a_{3n}, b$, where $\frac{b}{4} < a_i < \frac{b}{2}$, $i = 1, \ldots, 3n$.

QUESTION. Is there a partition of the integers into n subsets each summing up to b, where $b = \sum_{i=1}^{3n} \frac{a_i}{n}$?

We create an instance of the Generalized Due Date Problem as follows. The subsets of jobs are labelled $J_1, \ldots, J_{3n}$, and each subset has exactly n jobs. Let J_{ij} be the jth job from subset J_i. For each job J_{ij}, let $p_{ij} = ja_i$, and let $d_{ij} = bj(j+1)/2$ for $j = 1, \ldots, n$. We refer to the collection $\{J_{ij} : i = 1, \ldots, 3n\}$ of jth jobs in each subset as the type j jobs.

We claim that it is possible to schedule all jobs on time if and only if there is a solution to the 3-Partition problem.

Suppose first that there is a solution for the 3-Partition problem. Let the certificate be $S_1, \ldots, S_n$, where each subset S_i has three integers summing to b, and $\bigcup_{i=1}^{n} S_i = \{a_1, \ldots, a_{3n}\}$. We now select jobs in the Generalized Due Date Problem as follows: If $a_i \in S_j$, then we select J_{ij}.

We observe that exactly one job is selected from J_i. We also note that the processing times of the selected type j jobs sums to jb, and their due dates are $j(j+1)b/2$. If we schedule the jobs in order of their type, then all jobs are scheduled on time. If the answer to the 3-Partition problem is yes, then so is the answer to the scheduling problem.

Before we prove the converse, we state and prove a lemma.

LEMMA 1. *Consider the following linear program:*

$$\text{minimize} \sum_{i=1}^{n} q_i \tag{4.4}$$

subject to

$$\sum_{i=1}^{j} q_i \quad \le j(j+1)/2, \quad j = 1, \ldots, n-1, \tag{4.5}$$

$$\sum_{j=1}^{n} q_j/j = n \tag{4.6}$$

The unique optimal solution is $q_j = j$, $j = 1, \ldots, n$, and the optimal objective value is $n(n+1)/2$.

Proof. We first note that the solution $q_j = j$ for each j is a feasible solution and that all constraints hold with equality. We also note that it is the unique solution where all inequalities are tight, and the objective value is $n(n+1)/2$.

Let q' be an optimal solution. We suppose that not all constraints are tight, and we will derive a contradiction. Let k be the minimum index of a constraint that is not tight. That is, $\sum_{i=1}^{j} q'_i = j(j+1)/2$ for $j < k$, and $\sum_{i=1}^{k} q'_i < k(k+1)/2$. Therefore $q'_k < k$. But $\sum_{j=1}^{n} q_j/j = n$, and so there must be a first index $l > k$ such that $q'_l > 0$. But then we can improve the objective function and maintain feasibility by increasing q'_k by $k\epsilon$ and decreasing q'_l by $l\epsilon$ for some very small positive ϵ. Thus q' is not optimal. $\square$

Returning to the scheduling problem, we now consider the converse. Suppose that there is a certificate for the scheduling problem. Let J^* be the selected subset of jobs. If $J_{ij} \in J^*$, then $a_i \in S_j$. We may assume that the jobs are scheduled in order of increasing due dates, and thus the jobs are scheduled in order of their type.

Let r_j be the sum of the processing times of the type j jobs, and let $q_j = r_j/b$. Then the q's satisfy constraints (4.5) and (4.6), and the time that the type n jobs complete is $b\sum_{i=1}^{n} q_i$. By Lemma 1, the minimum completion time is $bn(n+1)/2$, and for this completion time to occur, it must be true that $q_j = j$ for each j, and thus $r_j = jb$ for each j. Thus, there is a feasible schedule if and only if there is a solution to the Number Partition problem.

This proves that the problem of determining if the tardiness is zero (no tardy jobs) is NP-complete. Thus, the minimization problem is inapproximable. $\square$

As for the problem of maximizing the number of jobs that are on time, we do not know if this problem is APX-complete.

THEOREM 5. *The Generalized Due Date Problem is strongly NP-hard and inapproximable even if each subset contains two jobs.*

Proof. We extend the transformation from 3-Partition given in the proof of Theorem 4. The numbers for 3-Partition are $a_1, \ldots, a_{3n}$ and b. We assume that $a_i \ge a_j$ for $i < j$.

We create $3n^2$ pairs of jobs, one pair for each job in the transformation used in the proof of Theorem 6. The pairs are labelled P_{ij}, $i = 1, \ldots, 3n$; $j = 1, \ldots, n$. Let pair P_{ij} have two jobs, one called the small job and one called the big job. The small job of P_{ij} has a processing time of 1 and a due date of $i(n-1)$. The big job has a processing time of $n^2 j a_i$, and a due date of $M + n^2 bj(j+1)/2$, where $M = 3n(n-1)$.

We claim that there is a solution to this instance of a Generalized Due Date Problem if and only if there is a solution to the instance of the 3-Partition problem.

Suppose first that there is a solution for the 3-Partition problem. Let the subsets of numbers be $S_1, \ldots, S_n$. We now select jobs in the Generalized Due Date Problem as follows: If $a_i \in S_j$, then we select the large job from pair P_{ij} of jobs. Otherwise, we take the small job from the pair. For each $i = 1, \ldots, n$, let $P_i = \{P_{ij} : j = 1, \ldots, n\}$. We select $n-1$ small jobs from the n pairs in P_i. It is easy to verify that the last scheduled small job in P_i finishes at time $(n-1)i$, and thus the last small job completes at time $M = 3n(n-1)$. As in the proof of Theorem 1, the last scheduled large job in P_i completes at time $M + n^2 bj(j+1)/2$, and thus all large jobs are completed on time.

We now consider the converse. Without loss of generality, we may assume that there are $3n(n-1)$ small jobs selected. (If there were a feasible solution in which fewer small jobs were selected, then one could exchange a large job for a small job and still meet all of the due dates.) We also assume without loss of generality that exactly $n-1$ small jobs of P_i are selected for each i. (Otherwise, we could let k be the least index such that fewer than $n-1$ small jobs of P_k are selected, and at least two large jobs, say, from P_{kr} and P_{ks}. But then there is an index $l > k$ such that n small jobs of P_l are selected. But we could create an alternative solution that meets all of the due dates if we exchange the small job from P_{kr} for the large job and simultaneously exchange the large job from P_{lr} for the small job.

So, the feasible solution has exactly one large job from each P_i. We are now in the same situation considered in the proof of Theorem 6, except that the large jobs start at M, the due dates are all translated by M, and each processing time and due date is multiplied by a factor of n^2. This completes the proof. □

Next, we prove that another generalized scheduling problem is NP-complete.

Generalized Weighted Sum of Completion Times Problem.

INSTANCE. n pairs of jobs $J_1, \ldots, J_n$ and an integer K. The processing time of the jth job in J_i is p_{ij} and its weight w_{ij}, $i = 1, \ldots, n$; $j = 1$ or 2.

QUESTION. Is it possible to select one job from each pair so that the weighted sum of completion times is at most K?

THEOREM 6. *The Generalized Weighted Sum of Completion Times Problem is weakly NP-complete.*

Proof. The proof is by transformation from Number Partition. □

Number Partition.

INSTANCE. Integers $a_1, \ldots, a_n$.

QUESTION. Is there a 0-1 vector x such that $\sum_{i=1}^n a_i x_i = b$, where $b = \sum_{i=1}^n a_i/2$?

Given an instance of the Number Partition problem, choose n pairs of jobs as follows. The first job of J_i has processing time $p_{i1} = a_i$ and a weight $w_{i1} = 2a_i$. We refer to the first job in each pair as its *short* job. The second job J_i has a processing time $p_{i2} = 2a_i$ and a weight $w_{i2} = a_i$. We refer to this job as its *long* job. We let $K = 3b^2 + \sum_{i=1}^n (a_i)^2$.

We claim that the answer to the Number Partition problem is yes if and only if the answer to the Generalized Weighted Sum of Completion Times Problem is also yes.

First we establish an elementary fact about this instance of the Generalized Weighted Sum of Completion Times Problem. Let E be the set of indices of the pairs whose short job is selected, and let L be the indices of the pairs of jobs whose long job is selected. Let $a(E) = \sum_{i \in E} a_i$, and let $a(L) = \sum_{i \in L} a_i$.

Lemma 2. *The optimum weighted sum of completion times for this schedule is* $\sum_{i=1}^{n} (a_i)^2 + 4b^2 - a(E)a(L)$.

Proof. It is well known that the optimum solution to the Weighted Sum of Completion Times Problem is to schedule jobs in nondecreasing order of the ratio of the processing times to the weights. Thus, the order in which the jobs are scheduled are (1) the short jobs (in any order) followed by (2) the long jobs (in any order).

The weighted sum of completion times for the short jobs is $\sum_{i,j \in E: i \leq j} a_i(2a_j) = \sum_{i \in E} (a_i)^2 + a(E)^2$. The weighted sum of completion times for the long jobs is $\sum_{i,j \in L: i \leq j} 2a_i a_j + a(E)a(L) = \sum_{i \in L} (a_i)^2 + a(L)^2 + a(E)a(L)$. If we sum the weighted sum of completion times of the short and long jobs, we get $\sum_{i=1}^{n} (a_i)^2 + a(E)^2 + a(L)^2 + a(E)a(L) = \sum_{i=1}^{n} (a_i)^2 + (a(E) + a(L))^2 - a(E)a(L) = (\sum_{i=1}^{n} (a_i)^2) + 4b^2 - a(E)a(L)$.

We now consider the case that there is a subset Q such that $a(Q) = b$. In this case, let $E = Q$ and let $L = \{1, \ldots, n\} \setminus Q$. By Lemma 2, the weighted sum of processing times is $\sum_{i=1}^{n} (a_i)^2) + 4b^2 - a(E)a(L) = (\sum_{i=1}^{n} (a_i)^2) + 3b^2 = K$. Thus, whenever the answer is yes for the number partition problem, the answer is also yes for the scheduling problem.

We now consider the case that there is a subset E of pairs of jobs such that scheduling early jobs from E and the late jobs from $\{1, \ldots, n\} \setminus E$ results in a weighted sum of completion times of at most $(\sum_{i=1}^{n} (a_i)^2) + 3b^2 = K$. It follows from Lemma 2 that $a(E) = a(L) = b$, and thus there is a solution to the Number Partition problem. This completes the proof. □

Open problem. For this problem (the Generalized Weighted Sum of Completion Times Problem) we were able only to establish ordinary NP-completeness. We do not know whether this problem is strongly NP-complete, nor do we know whether the minimization problem is inapproximable.

5. The generalized TSP (GTSP). The last problem on the list of Dror and Haouari (2000) is the GTSP. For the classical TSP, given triangle inequality, the best heuristic for a very long time was that of Christofides (1976) which assured a worst case bound of no more than 3/2. All other known heuristics yielded a bound of at least 2. More recently, Arora (1996) and Mitchell (1996) developed a polynomial time approximation scheme (PTAS) for the planar Euclidean TSP. For the GTSP with triangle inequality, a Christofides-type heuristic gives at best a performance guarantee of 2 (Dror and Haouari (2000)). For the general GTSP, inapproximability is obtained simply as a consequence of the TSP being MAXSNP-hard (Papadimitriou and Yannakakis (1993)). The following is an interesting question: Is the planar Euclidean GTSP inapproximable? We address this question below.

The GTSP can be stated somewhat differently in the format resembling that of the prize collecting TSP. That is, we want to construct a closed tour (circuit) of minimal cost with K nodes (each node represents a profit of one unit) and one node from each nonempty subset of nodes V_i, $i = 1, \ldots, K$. Given a triangle inequality of the cost matrix and relaxing the profit requirement to at least K units does not change the problem. This problem can be viewed as a generalization of a similar problem that requires the construction of a prize collecting tree. This problem is referred to as the K-MST problem. An instance of this problem includes an undirected graph $G = (V, E)$, with edge costs $c : E \to Q^+$, a specified root $r \in V$, and a positive integer K. The objective is to find a minimum cost subtree T of G spanning at least K nodes and containing r. For the K-MST problem in the Euclidean plane a polynomial time approximation scheme is presented in Arora (1996) and Mitchell (1996, 1999). However, in our case K points are picked from K nonempty subsets of V, requiring one point to be selected from each subset. We note (see a comment at the end of this section) that this Euclidean plane generalized K-MST does not have a PTAS unless P=NP.

In the TSP we are given n (> 3) locations to be visited by a salesman in a cyclic fashion (a TSP tour). In the geometric TSP, the n locations lie in a Euclidean space. Even in the special case of the Euclidean plane, finding the minimal distance TSP tour is NP-hard (Garey, Graham, and Johnson (1976), Papadimitriou (1977)). If the n locations lie in a Euclidean space of dimension $\log n$, finding the minimal distance TSP tour in any l_p norm is Max SNP-hard. That is, it is NP-hard to approximate an optimal TSP tour within some constant $r > 1$ (Trevisan (2000)). However, if finding an optimal TSP tour is restricted to the Euclidean plane, then there exists a PTAS for this problem. Even if it is hard to find an optimal TSP tour, a $1+\epsilon$ approximation of an optimal tour can be constructed in polynomial time for any $\epsilon > 0$ (Arora (1996), Mitchell (1996)).

A GTSP is similar to the TSP; however, the set of locations is partitioned into $k > 2$ nonempty subsets, and the TSP tour has to visit (select) exactly one location from each subset. If each subset contains only one point, the problem reverts to the original TSP. (In the literature, versions of this problem are referred to as the Group-TSP, One-of-a-set TSP, or TSP with neighborhoods.) Computational analysis for this problem in the Euclidean plane dates back to Arkin and Hassin (1994), who describe a constant approximation ratio algorithm if each group partition (the neighborhoods) in the problem consist of discs, parallel segments of equal length, and translates of convex region. For more about the evolution of results for the different problem variants see de Berg et al. (2002), Dumitrescu and Mitchell (2001), and Safra and Schwartz (2002). Safra and Schwartz (2002) have proven that the GTSP in the plane is NP-hard to approximate to within any constant factor. However, their proof requires that the subsets' size be ≥ 4. Here we examine the GTSP in the plane with subsets of size two each.

THEOREM 7. *The planar Euclidean GTSP with subsets of cardinality* 2 *is inapproximable.*

Proof. The proof is by transformation from Vertex Cover. Berman and Karpinski (1998) have shown that Vertex Cover is hard to approximate within a factor of 79/78 in polynomial time unless P=NP, even when the degree of each vertex is bounded by four. □

Vertex Cover.

INSTANCE. An undirected graph $G = (V, E)$, where $|V| = n$, and a positive integer K. (We assume that $\sqrt{n} \leq K \leq n$; the vertex cover is inapproximable in this range.)

QUESTION. Is there a subset $S \subseteq V$ such that each edge $e \in E$ is incident to a vertex in S and $|S| \leq K$?

Transformation from Vertex Cover to the GTSP in the plane with subsets of cardinality 2 is as follows: For each edge $(i, j) \in E$, we create $3n$ pairs of points. The rth pair of points for (i, j) is $((r, i), (r, j))$, $r = 1, \ldots, 3n$. We set the first coordinate as the x coordinate and the second coordinate as the y coordinate.

Claim 1. *Let K be the value of some solution for the Vertex Cover problem, and suppose K is even. Then there is a solution to the GTSP that has a length at most $3nK + 2n$.*

Proof. Let S be the optimal vertex cover. For each $i \in S$, include in the tour the line segment consisting of all points whose y coordinate is i. Then connect up to K line segments into a tour with additional patching length of at most $2n$, using points with x coordinate 1 or $3n$. $\square$

Claim 2. *Suppose there is a solution T to the GTSP of value $\leq 3nK + 2n$ for K even. Then there is a solution to the Vertex Cover problem of value $\leq K$.*

Proof. Since all lengths are ≥ 1, there are at most $3nK + 2n$ nodes in the tour. By the pigeonhole principle, there is some value r of the x coordinate such that T has at most K nodes whose x coordinate is r. Let $V = \{i : (r, i) \in T\}$. This V is a vertex cover. $\square$

To complete the proof of the theorem, let K' be the optimal value of the Vertex Cover problem, and let $\hat{K}$ be the optimal value of the GTSP tour. Using the results of Claims 1 and 2, we conclude that

$$\left\lfloor \frac{\hat{K}}{3n} \right\rfloor - 1 \leq K' < \left\lceil \frac{\hat{K}}{3n} \right\rceil + 1.$$

By assumption, $K' > \sqrt{n}$. So, any PTAS for the GTSP in the plane will yield a PTAS for the Vertex Cover problem. $\square$

Note that the above transformation for the Euclidean GTSP can be used to prove that the generalized version of the Euclidean plane K-MST is inapproximable.

6. Conclusion. In this paper we have examined generalized versions of a number of classical combinatorial optimization problems. We have focused on generalizations for which the ground set of elements N is explicitly expressed as a partition ($N = \bigcup_{i=1}^{m} N_i$, $N_i \cap N_j = \emptyset$ for $i \neq j$) in some cases and a cover in the other cases ($N = \bigcup_{i=1}^{m} N_i$). For most of the problems examined (the GMST problem, subset scheduling, the GCPP, the SBP problem, and the GTSP) we have proven that there is no PTAS for these problems unless P=NP. This suggests that in the field of combinatorial optimization the elementary objects of the ground set over which the combinatorial search is conducted might require some sort of characterization. We have not attempted to do so in this paper and only suggest it as an interesting research question.

REFERENCES

R. K. Ahuja, T. L. Magnanti, and J. B. Orlin (1993), *Network Flows, Theory, Algorithms, and Applications*, Prentice–Hall, Upper Saddle River, NJ.

E. M. Arkin and R. Hassin (1994), *Approximation algorithms for the geometric covering salesman problem*, Discrete Appl. Math., 55, pp. 197–218.

S. Arora (1996), *Polynomial time approximation schemes for Euclidean TSP and other geometric problems*, in Proceedings of the 37th IEEE Symposium on Foundations of Computer Science, pp. 2–11.

S. Arora and C. Lund (1997), *Hardness of approximations*, in Approximation Algorithms for NP-Hard Problems, D. S. Hochbaum, ed., PWS, Boston, MA, pp. 399–446.

S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy (1992), *Proof verification and intractability of approximation problems*, in Proceedings of the 33rd IEEE Symposium on Foundations of Computer Science, pp. 13–22.

S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy (1998), *Proof verification and hardness of approximation problems*, J. ACM, 45, pp. 501–555.

G. Ausiello, P. Crescenzi, and M. Protasi (1995), *Approximate solution of NP optimization problems*, Theoret. Comput. Sci., 150, pp. 1–55.

M. de Berg, J. Gudmundsson, M. J. Katz, C. Levcopoulos, M. H. Overmars, and A. F. van der Stappen (2002), *TSP with neighborhoods of varying size*, in Algorithms—ESA 2002, Lecture Notes in Comput. Sci. 2461, R. Moring and R. Raman, eds., Springer-Verlag, Berlin, pp. 186–199.

P. Berman and M. Karpinski (1998), *On Some Tighter Inapproximability Results*, Technical Report 98-029, ECCC.

N. Christofides (1976), *Worst-Case Analysis of a New Heuristic for the Traveling Salesman Problem*, Management Sciences Research Report 388, Carnegie Mellon University, Pittsburgh, PA.

E. G. Coffman, Jr., and M. Dror (1992), *Bin Packing with Subsets of a Set*, unpublished notes.

M. Dror, ed. (2000), *Arc Routing: Theory, Solutions, and Applications*, Kluwer Academic Publishers, Norwell, MA.

M. Dror and M. Haouari (2000), *Generalized Steiner problems and other variants*, J. Comb. Optim., 4, pp. 415–436.

M. Dror, M. Haouari, and J. Chaouachi (2000), *Generalized spanning trees*, European J. Oper. Res., 120, pp. 583–592.

A. Dumitrescu and J. S. B. Mitchell (2001), *Approximation algorithms for TSP with neighborhoods in the plane*, in Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 38–46.

J. Edmonds and E. L. Johnson (1973), *Matching, Euler tours and the Chinese postman*, Math. Programming, 5, pp. 88–124.

U. Feige (1998), *A threshold of* $\ln(n)$ *for approximating set cover*, J. ACM, 45, pp. 634–652.

H. N. Gabow, S. N. Maheshwari, and L. J. Osterweil (1976), *On two problems in the generation of program test paths*, IEEE Trans. Software Engrg., 2, pp. 227–231.

M. R. Garey and D. S. Johnson (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, CA.

M. R. Garey, R. L. Graham, and D. S. Johnson (1976), *Some NP-complete geometric problems*, in Conference Record of the Eighth Annual ACM Symposium on Theory of Computing, Hershey, PA, pp. 10–22.

O. Goldschmidt, D. Nehme, and G. Yu (1994), *Note: On the set-union knapsack problem*, Naval Res. Logist., 41, pp. 833–842.

V. Kann (1991), *Maximum bounded* 3*-dimensional matching in MAX SNP-complete*, Inform. Process. Lett., 37, pp. 27–35.

S. Khuller, A. Moss, and J. Naor (1999), *The budgeted maximum coverage problem*, Inform. Process. Lett., 70, pp. 39–45.

E. Lawler (1976), *Combinatorial Optimization, Networks, and Matroids*, Holt, Rinehart and Winston, New York.

J. K. Lenstra and A. H. G. Rinnooy Kan (1976), *On general routing problems*, Networks, 6, pp. 273–280.

T. L. Magnanti and L. A. Wolsey (1995), *Optimal trees*, in Network Models, Handbooks Oper. Res. Management Sci. 7, M. O. Ball, T. L. Magnanti, C. L. Monma, and G. L. Nemhauser, eds., North–Holland, Amsterdam, pp. 503–615.

J. S. B. Mitchell (1996), *Guillotine subdivisions approximate polygonal subdivisions: A simple new method for the geometric k-MST problem*, in Proceedings of the Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, Atlanta, GA, pp. 402–408.

J. S. B. Mitchell (1999), *Guillotine subdivisions approximate polygonal subdivisions: A simple polynomial-time approximation scheme for geometric TSP, k-MST, and related problems*, SIAM J. Comput., 28, pp. 1298–1309.

A. Moss (2001), *High Profit for Low Cost Approximation Algorithms in Node-Weighted Graphs*, Ph.D. thesis, Computer Science Department, Technion, Haifa, Israel.

Y.-S. Myung, C.-H. Lee, and D.-W. Tcha (1995), *On the generalized minimum spanning tree problem*, Networks, 26, pp. 231–241.

G. L. Nemhauser and L. A. Wolsey (1988), *Integer and Combinatorial Optimization*, John Wiley and Sons, New York.

C. H. Papadimitriou (1977), *Euclidean TSP is NP-complete*, Theoret. Comput. Sci., 4, pp. 237–244.

C. H. Papadimitriou and M. Yannakakis (1991), *Optimization, approximation, and complexity classes*, J. Comput. System Sci., 43, pp. 425–440.

C. H. Papadimitriou and M. Yannakakis (1993), *The traveling salesman problem with distances one and two*, Math. Oper. Res., 18, pp. 1–11.

R. Raz and S. Safra (1997), *A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP*, in Proceedings of the 29th Annual ACM Symposium of Theory of Computing, pp. 475–484.

S. Safra and O. Schwartz (2003), *On the complexity of approximating TSP with neighborhoods* *and related problems*, in Algorithms—ESA 2003, Lecture Notes in Comput. Sci. 2832, Springer-Verlag, Berlin, pp. 446–458.

P. M. Thompson and J. B. Orlin (1989), *The Theory of Cyclic Transfers*, working paper, Operations Research Center, MIT, Cambridge, MA.

L. Trevisan (2000), *When Hamming meets Euclid: The approximability of geometric TSP and Steiner tree*, SIAM J. Comput., 30, pp. 475–485.

SIAM J. Discrete Math.
Vol. 21, No. 4, pp. 1035–1052

AVERAGE DISTANCE AND EDGE-CONNECTIVITY II[*]

PETER DANKELMANN[†], SIMON MUKWEMBI[†], AND HENDA C. SWART[†]

Abstract. The average distance $\mu(G)$ of a connected graph G of order n is the average of the distances between all pairs of vertices of G. We prove that for a 3-edge-connected graph G of order n the inequality $\mu(G) \le n/6 + 24$ on the average distance holds. Our bound is shown to be best possible even if G is 4-edge-connected, and our results answer, in part, a question of Plesník [*J. Graph Theory*, 8 (1984), pp. 1–24].

Key words. distance, average distance, edge-connectivity

AMS subject classification. 05C12

DOI. 10.1137/06065653X

1. Introduction. For a given graph G, n, $\mu(G)$, and $\lambda(G)$ denote the order, average distance, and edge-connectivity of G, respectively. In this paper the present authors continue their investigations into the average distance of graphs of given edge-connectivity begun in [2], where the bounds (i) $\mu(G) \le 2n/15 + 9$, if $\lambda = 5, 6$, (ii) $\mu(G) \le n/9 + 10$, if $\lambda = 7$, and (iii) $\mu(G) \le n/(\lambda + 1) + 5$, if $\lambda \ge 8$, were proved. Bounding the average distance of 3-edge-connected graphs turns out to be harder and requires some additional ideas. We will therefore consider this problem separately as the subject of this article. Thus here we completely solve the problem of determining a sharp upper bound on the average distance in terms of order and edge-connectivity posed in [7]. We prove that the bound $\mu(G) \le n/6 + 24$ holds for a 3-edge-connected or 4-edge-connected graph G of order n.

We will use the terminology and notation in [2]. Briefly, let $G = (V, E)$ be a connected simple graph. For a subset $S \subseteq V$, $G[S]$ denotes the subgraph induced by S in G, and $diam(S)$ is the maximum value of $d_G(x, y)$, $x, y \in S$. For $v \in V$, $ex_G(v)$ and $\sigma_G(v)$ denote the eccentricity and distance of v in G, respectively. $\sigma(G)$ denotes the distance of G. For subsets $M_1, M_2 \subseteq V$, $\sigma_{M_1}(v)$ denotes the distance (i.e., $\sum_{x\in M_1} d_G(v, x)$) of v with respect to M_1, whereas $\sigma_{M_1}(M_2)$ denotes the distance (i.e., $\sum_{x\in M_2} \sigma_{M_1}(x)$) of M_2 with respect to M_1. The distance between v and M_1 (i.e., $\min_{u\in M_1} d_G(u, v)$) is denoted by $d(v, M_1)$. $E(M_1, M_2)$ denotes the set of edges $\{ab \in E \mid a \in M_1, b \in M_2\}$. $N_G(v)$ denotes the set of all vertices adjacent to v in G, and its cardinality is the degree of v in G and is denoted by $deg_G(v)$. For a subset $S \subset V$, $N_S(v)$ denotes the set of neighbors of v in S, and its cardinality is denoted by $deg_S(v)$. For a positive integer i, $N_i(v)$ denotes the ith distance layer of v, and $k_i(v)$ denotes the cardinality of $N_i(v)$. We denote the set $\cup_{0\le j\le i} N_j(v)$ by $N_{\le i}(v)$ and the set $\cup_{j\ge i} N_j(v)$ by $N_{\ge i}(v)$. $N(S)$ denotes the neighborhood of subset $S \subseteq V$, namely, $\cup_{u\in S} N_G(u)$. $\overline{N}(S)$ denotes the closed neighborhood of S, i.e., $\overline{N}(S) = N(S) \cup S$, whereas $N_{\le i}(S)$ denotes the ith neighborhood of $S \subseteq V$, namely, $\cup_{u\in S} N_{\le i}(u)$. By $S = V_1 \uplus V_2$, we mean that $S = V_1 \cup V_2$ and $V_1 \cap V_2 = \emptyset$, where $S, V_1, V_2 \subseteq V$. The graph $G_1 + G_2 + \cdots + G_k$ denotes the sequential join of the vertex disjoint graphs $G_1, G_2, \ldots, G_k$.

[*] Received by the editors April 6, 2006; accepted for publication (in revised form) May 11, 2007; published electronically January 22, 2008. Financial support by the National Research Foundation and the University of KwaZulu-Natal is gratefully acknowledged. The results in this paper are part of the second author's Ph.D. thesis.

http://www.siam.org/journals/sidma/21-4/65653.html

[†] School of Mathematical Sciences, University of KwaZulu-Natal, Durban, 4041 South Africa (dankelma@ukzn.ac.za, mukwembi@ukzn.ac.za, swarth@ukzn.ac.za).

2. Preliminary results. Plesník [7] showed that the distance of an arbitrary vertex in a 2-edge-connected graph of order n is at most $\lfloor \frac{1}{3}(n^2-n) \rfloor$. We begin by improving this result for 3-edge-connected graphs.

LEMMA 1. *Let G be a 3-edge-connected graph of order $n \geq 4$. Then for any vertex v of G, we have $\sigma_G(v) \leq \frac{1}{4}(n^2-n)+\frac{1}{2}$. Moreover, $\sigma_G(v) \leq \frac{1}{4}(n^2-2n)+1$ if G is a block with $n \geq 6$.*

Proof. Let v be a vertex of G, and let e be the eccentricity of v. Note that

$$\sigma_G(v) = 1k_1 + 2k_2 + 3k_3 + 4k_4 + \cdots + ek_e \tag{1}$$

for $i = 0, 1, \ldots, e-1$, $E(N_i, N_{i+1})$ is a disconnecting set of G; hence $k_i k_{i+1} \geq \lambda(G) \geq 3$. Clearly, $k_i \geq 1$ for all $i = 0, 1, \ldots, e$. Thus,

$$k_i + k_{i+1} \geq 4 \text{ for all } i = 0, 1, \ldots, e-1. \tag{2}$$

We maximize (1) subject to (2) and the condition $\sum_{i=0}^{e} k_i = n$. Clearly for fixed e, (1) is maximized for

$$(k_0, k_1, \ldots, k_{e-1}) = \begin{cases} (1,3,1,3,\ldots,1,3) & \text{if } e \text{ is even,} \\ (1,3,1,3,\ldots,1,3,1) & \text{if } e \text{ is odd,} \end{cases}$$

and $k_e = n-2e$ $(n-2e+1)$ if e is even (odd). Hence

$$\begin{aligned} &\sigma_G(v) \leq \\ &\begin{cases} 1\cdot 3 + [2\cdot 1 + 3\cdot 3] + \cdots + [(e-2)\cdot 1 + (e-1)\cdot 3] \\ +e(n-2e) & \text{if } e \text{ is even,} \\ \\ 1\cdot 3 + [2\cdot 1 + 3\cdot 3] + \cdots + [(e-3)\cdot 1 + (e-2)\cdot 3] \\ +(e-1)\cdot 1 + e(n-2e+1) & \text{if } e \text{ is odd} \end{cases} \\ &\leq en - e^2 - \frac{1}{2}e + \frac{1}{2}. \end{aligned} \tag{3}$$

Summing (2) over all i yields $n \geq 2e+1$ and thus $e \leq \frac{n-1}{2}$. A simple differentiation shows that (3) is maximized, subject to the constraint $e \leq \frac{n-1}{2}$, for $e = \frac{n-1}{2}$ and we obtain

$$\sigma_G(v) \leq \left(\frac{n-1}{2}\right)n - \left(\frac{n-1}{2}\right)^2 - \frac{1}{2}\left(\frac{n-1}{2}\right) + \frac{1}{2} = \frac{1}{4}(n^2-n) + \frac{1}{2},$$

as desired.

If G is a 3-edge-connected block with $n \geq 6$, we have

$$k_0 = 1,\ k_1 \geq 3,\ k_i \geq 2 \text{ for } i = 2,3,4,\ldots,e-1 \text{ and } k_{e-1}+k_e \geq 4. \tag{4}$$

We maximize (1) subject to (4) and the condition $\sum_{i=0}^{e} k_i = n$. For fixed e, (1) is maximized for $(k_0, k_1, \ldots, k_{e-1}) = (1,3,2,2,\ldots,2)$. Thus,

$$\begin{aligned} \sigma_G(v) &\leq 1\cdot 3 + 2\cdot 2 + 3\cdot 2 + \cdots + (e-2)\cdot 2 \\ &\quad +(e-1)\cdot 2 + e(n-4-2(e-2)) \\ &= e(n-1-e)+1. \end{aligned} \tag{5}$$

By (4), $n = \sum_{i=0}^{e} k_i \geq 2e+2$. Hence $e \leq \frac{n-2}{2}$. A simple differentiation shows that (5) is maximized, subject to the constraint $e \leq \frac{n-2}{2}$, for $e = \frac{n-2}{2}$, and we obtain $\sigma_G(v) \leq \frac{1}{4}(n^2-2n)+1$, as desired. □

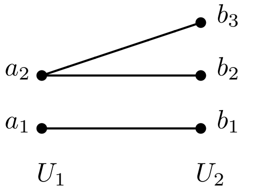


FIG. 1. *The graph induced by $E(U_1, U_2)$ of Fact* 1.

Summing $\sigma_G(v)$ for all $v \in V(G)$ and an application of Lemma 1 implies the following bound, which we will use to prove our main bound for small values of n.

COROLLARY 2. *Let G be a 3-edge-connected graph of order $n \geq 4$. Then $\sigma(G) \leq \frac{1}{4}n(n^2 - n) + \frac{1}{2}n$.*

3. Distance layers of an arbitrary vertex in 3-edge-connected blocks. In this section, we consider only 3-edge-connected blocks, unless otherwise specified. Some of these graphs have connectivity 2; they can be disconnected by removing only two vertices u, v. If such vertices u, v are close to one another, then this will allow us to bound the average distance. Here we study some of the situations which are sufficient for such 2-element vertex cut sets to occur in distance layers.

PROPOSITION 1. *Let G be a 3-edge-connected graph, v a vertex of G, and $1 \leq l \leq ex_G(v) - 1$. If $(k_l, k_{l+1}) = (2, 2)$, then $diam(N_l), diam(N_{l+1}) \leq 2$.*

Proof. Assume that $N_l = \{u_1, u_2\}$ and $N_{l+1} = \{v_1, v_2\}$. Let H be the graph induced by $E(N_l, N_{l+1})$. Then H is a bipartite graph with vertices u_1, u_2, v_1, v_2 and at least $\lambda(G) \geq 3$ edges. Hence H is isomorphic to either P_4 or $K_{2,2}$ and has diameter at most 3. Since u_1 and u_2 (v_1 and v_2) are in the same partition set, $d_H(u_1, u_2)$ ($d_H(v_1, v_2)$) is even. Hence $d_H(u_1, u_2) = 2$ ($d_H(v_1, v_2) = 2$) and thus $d_G(u_1, u_2) \leq 2$ ($d_G(v_1, v_2) \leq 2$). □

DEFINITION 1. *Let G be a 3-edge-connected block and v a vertex of G. We say that $\mathbf{T} = (N_l, N_{l+1})$, $1 \leq l \leq ex_G(v) - 2$, is a separating pair of v if $(k_l, k_{l+1}) = (2, 2)$.*

We shall use the following fact repeatedly.

FACT 1. *Let G be a 3-edge-connected block with two disjoint separating sets $U_1, U_2 \subseteq V(G)$, $U_1 = \{a_1, a_2\}$ and $U_2 = \{b_1, b_2, b_3\}$, where $E(U_1, U_2)$ is a disconnecting set. If $d(a_1, a_2) > 2$, then the graph induced by $E(U_1, U_2)$ is isomorphic to the graph $(U_1 \cup U_2, \{a_1b_1, a_2b_2, a_2b_3\})$.*

Proof. Let H be the graph induced by $E(U_1, U_2)$. Then H is a bipartite graph with vertices a_1, a_2, b_1, b_2, b_3, partition sets U_1, U_2, and at least $\lambda(G) \geq 3$ edges. From $d(a_1, a_2) > 2$, it follows that $N(a_1) \cap N(a_2) = \emptyset$; hence $deg_{U_1}(x) \leq 1$ for $x \in U_2$. Thus since $|U_2| = 3$ and $|E(H)| \geq 3$, we have $deg_{U_1}(x) = 1$ for each $x \in U_2$. Hence H is isomorphic to either $(U_1 \cup U_2, \{a_1b_1,\ a_1b_2,\ a_1b_3\})$ or $(U_1 \cup U_2, \{a_1b_1,\ a_2b_2,\ a_2b_3\})$. In the former case, $E(U_1, U_2) = \{a_1b_1,\ a_1b_2,\ a_1b_3\}$ disconnects G; hence a_1 is a cut vertex, contradicting the fact that G is a block. Therefore, H is isomorphic to $(U_1 \cup U_2, \{a_1b_1, a_2b_2, a_2b_3\})$, (see Figure 1) as desired. □

PROPOSITION 2. *Let G be a 3-edge-connected block, v a vertex of G, and $1 \leq l \leq ex_G(v) - 3$.*

(i) *If $(k_l, k_{l+1}, k_{l+2}) = (3, 2, 3)$, then $diam(N_{l+1}) \leq 2$.*

(ii) *If $(k_l, k_{l+1}, k_{l+2}) = (2, 3, 2)$, then $diam(N_l) \leq 3$ or $diam(N_{l+2}) \leq 3$.*

Proof. (i) Let $N_l = \{u_1, u_2, u_3\}$, $N_{l+1} = \{v_1, v_2\}$, and $N_{l+2} = \{w_1, w_2, w_3\}$. Suppose to the contrary that $d(v_1, v_2) \geq 3$. By Fact 1, assume without loss of generality that the graph induced by $E(N_l, N_{l+1})$ has edge set $\{u_1v_1,\ u_2v_1,\ u_3v_2\}$. Since $d(v_1, v_2) \geq 3$, $v_2v_1 \notin E(G)$. Hence from $deg(v_2) \geq 3$ and $N(v_2) \subseteq \bigcup_{i=l}^{l+2} N_i$, assume

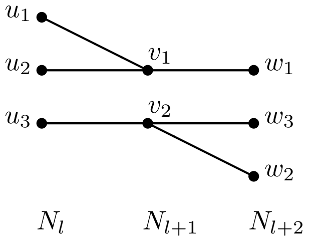


FIG. 2. *Proof of Proposition* 2(i).

without loss of generality that $v_2w_2,\ v_2w_3 \in E(G)$. Applying Fact 1, we see that the graph induced by $E(N_{l+1}, N_{l+2})$ has edge set $\{v_1w_1,\ v_2w_2,\ v_2w_3\}$. Refer to Figure 2.

Since $v_1v_2 \notin E(G)$, a path from v_1 to v_2 contains either vertices from N_l or vertices from N_{l+2}, in which case it contains the edge u_3v_2 or v_1w_1. Therefore, there are at most 2 edge-disjoint paths joining v_1 and v_2, contradicting the fact that G is 3-edge-connected.

(ii) Let $N_l = \{u_1, u_2\}$, $N_{l+1} = \{v_1, v_2, v_3\}$, and $N_{l+2} = \{w_1, w_2\}$. Suppose to the contrary that $d(u_1, u_2),\ d(w_1, w_2) \geq 4$. By Fact 1, assume without loss of generality that the graph induced by $E(N_l, N_{l+1})$ has edge set $\{u_1v_1, u_2v_2, u_2v_3\}$. It follows from the assumption on $d(u_1, u_2)$ and $d(w_1, w_2)$ that $v_1v_2, v_1v_3 \notin E(G)$ and that v_1 can be adjacent only to at most one of w_1, w_2. Hence from $N(v_1) \subseteq \bigcup_{i=l}^{l+2} N_i$, $\ deg(v_1) \leq 2$, a contradiction. ∎

DEFINITION 2. *Let G be a 3-edge-connected block and v a vertex of G. We say that $\mathbf{T} = (N_l, N_{l+1}, N_{l+2})$, $1 \leq l \leq ex_G(v) - 3$, is a separating triple of v if $(k_l, k_{l+1}, k_{l+2}) \in \{(3, 2, 3),\ (2, 3, 2)\}$.*

PROPOSITION 3. *Let G be a 3-edge-connected block and v a vertex of G. If $(k_l, k_{l+1}, k_{l+2}, k_{l+3}) = (2, 3, 3, 2)$ for some integer l, $1 \leq l \leq ex_G(v) - 4$, then $diam(N_l) \leq 4$ or $diam(N_{l+3}) \leq 4$.*

Proof. Let $N_l = \{a_1, a_2\}$, $N_{l+1} = \{b_1, b_2, b_3\}$, $N_{l+2} = \{c_1, c_2, c_3\}$, and $N_{l+3} = \{d_1, d_2\}$. Suppose to the contrary that $d(a_1, a_2), d(d_1, d_2) > 4$. By Fact 1, assume without loss of generality that the graph induced by $E(N_l, N_{l+1})$ $(E(N_{l+2}, N_{l+3}))$ has edge set $\{a_1b_1, a_2b_2, a_2b_3\}$ $(\{c_1d_1, c_2d_2, c_3d_2\})$. Hence, by our assumption on $d(a_1, a_2)$ and $d(d_1, d_2)$, $b_1b_2, b_1b_3 \notin E(G)$ and b_1 cannot be adjacent to both c_1 and $c_2(c_3)$. Therefore, since $deg(b_1) \geq 3$ and $N(b_1) \subseteq \bigcup_{i=l}^{l+2} N_i$, $b_1c_2, b_1c_3 \in E(G)$. Similarly, $c_1b_2, c_1b_3 \in E(G)$. Letting $S_1 = \{a_1, b_1, c_2, c_3, d_2\}$ and $S_2 = \{a_2, b_2, b_3, c_1, d_1\}$, our assumption on $d(a_1, a_2)$ and $d(d_1, d_2)$ implies that $E(S_1, S_2) = \emptyset$. A path from N_0 to N_e thus contains either edges from $\{a_2b_2, a_2b_3\}$ and from $\{c_1d_1\}$ or edges from $\{a_1b_1\}$ and from $\{c_3d_2, c_2d_2\}$. Thus $\{a_1b_1, c_1d_1\}$ is a disconnecting set of G, contradicting the fact that G is 3-edge-connected. ∎

DEFINITION 3. *Let G be a 3-edge-connected block and v a vertex of G. We say that $\mathbf{T} = (N_l, N_{l+1}, N_{l+2}, N_{l+3})$, $1 \leq l \leq ex_G(v) - 4$, is a separating quadruple of v if $(k_l, k_{l+1}, k_{l+2}, k_{l+3}) = (2, 3, 3, 2)$.*

LEMMA 3. *Let G be a 3-edge-connected block, v a vertex of G, and $x \geq 2$ an integer. If for some l, $1 \leq l \leq ex_G(v) - 5$, we have $(k_l, k_{l+1}, k_{l+2}, k_{l+3}, k_{l+4}) \in \{(2, 3, x, 2, 3),\ (2, 3, x, 3, 2),\ (3, 2, x, 2, 3),\ (3, 2, x, 3, 2)\}$, then there exists i, $l \leq i \leq l + 4$, for which $diam(N_i) \leq 8$ and $k_i = 2$.*

Proof. Let $S = \bigcup_{i=l}^{l+4} N_i$. Suppose to the contrary that

$$(*) \qquad \text{there is no } i \in \{l, \ldots, l + 4\} \text{ for which } diam(N_i) \leq 8 \text{ and } k_i = 2.$$

By Fact 1, let $N_{l+3} \cup N_{l+4} = \{x_{l+3}, y_{l+3}, x_{l+4}, y_{l+4}, a\}$, where $x_{l+3}, y_{l+3} \in N_{l+3}$, $x_{l+4}, y_{l+4} \in N_{l+4}$, $x_{l+3}x_{l+4}, y_{l+3}y_{l+4} \in E(G)$, and a is adjacent to either $\{x_{l+3}, x_{l+4}\}$ or $\{y_{l+3}, y_{l+4}\}$ but not both. Assume, without loss of generality, that a is adjacent to $\{y_{l+3}, y_{l+4}\}$. For $i = 1, 2, \dots, l+2$, let $vx_1x_2 \dots x_{l+2}x_{l+3}$ $(vy_1y_2 \dots y_{l+2}y_{l+3})$ be a $v - x_{l+3}$ $(v - y_{l+3})$ shortest path. It follows that for $i = l, l+1, l+2$ we have that $x_i \neq y_i$; otherwise $diam(N_i) \leq 8$ for $i = l+3$ or $i = l+4$ with $k_i = 2$, a contradiction to (*). Therefore, with the above notation, write $N_l \cup N_{l+1} = \{x_l, y_l, x_{l+1}, y_{l+1}, b\}$, and from Fact 1, b is adjacent to $\{x_l, x_{l+1}\}$ or $\{y_l, y_{l+1}\}$ but not both. We distinguish two cases.

Case A: $b \in N_l$. It follows that $k_{l+1} = 2$ so that every vertex in $\bigcup_{i=l+1}^{l+4} N_i$ is in $N_{\leq 3}(x_{l+1})$ or $N_{\leq 3}(y_{l+1})$. Moreover, by Fact 1, b is adjacent to N_{l+1}. By (*), $d(x_{l+1}, y_{l+1}) \geq 9$; hence $S = (S \cap N_{\leq 3}(x_{l+1})) \biguplus (S \cap N_{\leq 3}(y_{l+1}))$ is a disjoint union, and the sets $(S \cap N_{\leq 3}(x_{l+1}))$ and $(S \cap N_{\leq 3}(y_{l+1}))$ are nonadjacent. A path from N_0 to N_e thus contains either edges from $E(N_l, N_{l+1}) \cap E(G[S \cap N_{\leq 3}(x_{l+1})])$ and from $E(N_{l+3}, N_{l+4}) \cap E(G[S \cap N_{\leq 3}(x_{l+1})])$ or edges from $E(N_l, N_{l+1}) \cap E(G[S \cap N_{\leq 3}(y_{l+1})])$ and from $E(N_{l+3}, N_{l+4}) \cap E(G[S \cap N_{\leq 3}(y_{l+1})])$. Hence, the union of any two of the four sets $E(N_l, N_{l+1}) \cap E(G[S \cap N_{\leq 3}(x_{l+1})])$, $E(N_{l+3}, N_{l+4}) \cap E(G[S \cap N_{\leq 3}(x_{l+1})])$, $E(N_l, N_{l+1}) \cap E(G[S \cap N_{\leq 3}(y_{l+1})])$, $E(N_{l+3}, N_{l+4}) \cap E(G[S \cap N_{\leq 3}(y_{l+1})])$ is a disconnecting set of G. Note that $x_{l+3}x_{l+4}$ is the only edge in $E(N_{l+3}, N_{l+4}) \cap E(G[S \cap N_{\leq 3}(x_{l+1})])$. Recall that $by_{l+1} \in E(G)$ or $bx_{l+1} \in E(G)$. The former implies that x_lx_{l+1} is the only edge in $E(N_l, N_{l+1}) \cap E(G[S \cap N_{\leq 3}(x_{l+1})])$, so that the removal of x_lx_{l+1} and $x_{l+3}x_{l+4}$ separates v and x_{l+4} and thus disconnects G, a contradiction. The latter implies that y_ly_{l+1} is the only edge in $E(N_l, N_{l+1}) \cap E(G[S \cap N_{\leq 3}(y_{l+1})])$, so that the removal of y_ly_{l+1} and $x_{l+3}x_{l+4}$ separates v and x_{l+4} and thus disconnects G, again a contradiction.

Case B: $b \in N_{l+1}$. We first show that

$$(**) \qquad S = (S \cap N_{\leq 4}(x_l)) \biguplus (S \cap N_{\leq 4}(y_l)).$$

Note first that, since $b \in N_{l+1}$, we have $N_l = \{x_l, y_l\}$, and every vertex in S is within distance 4 of N_l; hence $S = (S \cap N_{\leq 4}(x_l)) \cup (S \cap N_{\leq 4}(y_l))$ in (**) follows. From (*), $d(x_l, y_l) \geq 9$; hence the union in (**) is disjoint. Next we show that

$$(***) \qquad E\left(S \cap N_{\leq 4}(x_l), S \cap N_{\leq 4}(y_l)\right) = \emptyset.$$

Now suppose that $E\left(S \cap N_{\leq 4}(x_l), S \cap N_{\leq 4}(y_l)\right)$ is nonempty and contains an edge pq. From $d(x_l, y_l) \geq 9$ and $pq \in E(S \cap N_{\leq 4}(x_l), S \cap N_{\leq 4}(y_l))$ we can assume without loss of generality that $p \in N_{l+4} \cap N_{\leq 4}(x_l)$ and $q \in N_{l+4} \cap N_{\leq 4}(y_l)$. If, on one hand, $a \in N_{l+3}$, then $N_{l+4} \cap N_{\leq 4}(x_l) = \{x_{l+4}\}$ and $N_{l+4} \cap N_{\leq 4}(y_l) = \{y_{l+4}\}$, and by (*), $d(x_{l+4}, y_{l+4}) \geq 9$; so $x_{l+4}y_{l+4} \notin E(G)$. If, on the other hand, $a \in N_{l+4}$, then $N_{l+4} \cap N_{\leq 4}(x_l) = \{x_{l+4}\}$ and $N_{l+4} \cap N_{\leq 4}(y_l) = \{y_{l+4}, a\}$. By Fact 1, $ay_{l+3} \in E(G)$. By (*), $d(x_{l+3}, y_{l+3}) \geq 9$; hence $\{y_{l+4}, a\}$ and $\{x_{l+4}\}$ are nonadjacent. Thus (***) is established.

A path from N_0 to N_e thus contains either edges from $E(N_l, N_{l+1}) \cap E(G[S \cap N_{\leq 4}(x_l)])$ and from $E(N_{l+3}, N_{l+4}) \cap E(G[S \cap N_{\leq 4}(x_l)])$ or edges from $E(N_l, N_{l+1}) \cap E(G[S \cap N_{\leq 4}(y_l)])$ and from $E(N_{l+3}, N_{l+4}) \cap E(G[S \cap N_{\leq 4}(y_l)])$. Hence, the union of any two of the four sets $E(N_l, N_{l+1}) \cap E(G[S \cap N_{\leq 4}(x_l)])$, $E(N_{l+3}, N_{l+4}) \cap E(G[S \cap N_{\leq 4}(x_l)])$, $E(N_l, N_{l+1}) \cap E(G[S \cap N_{\leq 4}(y_l)])$, $E(N_{l+3}, N_{l+4}) \cap E(G[S \cap N_{\leq 4}(y_l)])$ is a disconnecting set of G. Now first observe from Fact 1 that $x_{l+3}x_{l+4}$ is the only edge in $E(N_{l+3}, N_{l+4}) \cap E\left(G\left[S \cap N_{\leq 4}(x_l)\right]\right)$. Also $x_lb \in E(G)$ or $y_lb \in E(G)$. In the

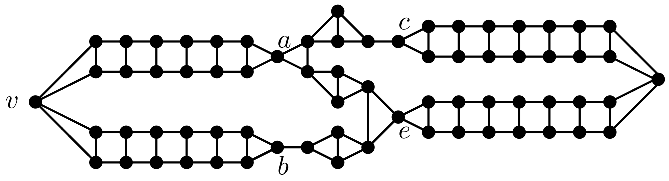


FIG. 3. *A separating quintuple.*

former case, one gets that $y_l y_{l+1}$ is the only edge in $E(N_l, N_{l+1}) \cap E\left(G\left[S \cap N_{\leq 4}(y_l)\right]\right)$; hence $\{y_l y_{l+1}, x_{l+3} x_{l+4}\}$ separates v and x_{l+4} and is thus a disconnecting set of G, a contradiction. In the latter case, one gets that $x_l x_{l+1}$ is the only edge in $E(N_l, N_{l+1}) \cap E\left(G\left[S \cap N_{\leq 4}(x_l)\right]\right)$; hence $\{x_l x_{l+1}, x_{l+3} x_{l+4}\}$ separates v and x_{l+4} and is thus a disconnecting set of G, again a contradiction. This completes the proof of the lemma. □

DEFINITION 4. *Let G be a 3-edge-connected block and v a vertex of G. We say that $\mathbf{T} = (N_l, N_{l+1}, N_{l+2}, N_{l+3}, N_{l+4})$, $1 \leq l \leq ex_G(v) - 5$, is a separating quintuple of v if $(k_l, k_{l+1}, k_{l+2}, k_{l+3}, k_{l+4}) \in \{(2,3,x,2,3),\ (2,3,x,3,2),\ (3,2,x,2,3),\ (3,2,x,3,2)\}$ for some integer $x \geq 2$.*

Example 1. If the hypothesis of Lemma 3 is satisfied, then Lemma 3 guarantees the existence of a 2-element distance layer N_i of diameter at most 8. To see that the value 8 is close to best possible, consider the 3-edge-connected graph of Figure 3. Then $\mathbf{T} = (N_7, N_8, N_9, N_{10}, N_{11})$ is a separating quintuple of v, and $d(a,b) = 7 = d(c,e)$.

3.1. Forbidden separators. Let G be a 3-edge-connected block and v be a vertex of G with $ex_G(v) = e \geq 63$. Assume that $\mathbf{T}$ is a separating pair, triple, quadruple, or quintuple of v. By Propositions 1, 2, and 3 and Lemma 3, $\mathbf{T}$ contains a distance layer $N_\alpha = \{u_1, u_2\}$, say, of v with $diam(N_\alpha) \leq 8$. Assume that one wishes to estimate, using N_α, the distance between two vertices x and y, where x is a vertex close to u_1 and y is a vertex close to u_2. Roughly, if α is large enough, the bound

$$d(x,y) \leq d(x,u_1) + d(u_1,u_2) + d(u_2,y) \leq d(x,u_1) + 8 + d(u_2,y)$$

is better than

$$d(x,y) \leq d(x,u_1) + d(u_1,u_2) + d(u_2,y) \leq d(x,u_1) + 2\alpha + d(u_2,y).$$

Thus, roughly speaking, for α large, a small diameter of N_α reduces distances, and, for our purposes, if $\mathbf{T}$ is such that $31 \leq \alpha \leq e - 31$, we say that $\mathbf{T}$ is a *forbidden separator* of v. More formally, we state the following.

DEFINITION 5. *Let G be a 3-edge-connected block and v be a vertex of G with $ex_G(v) = e$.*

(a) *A separating pair (N_l, N_{l+1}) of v is called a forbidden separating pair of v if $31 \leq l \leq e - 32$.*

(b) *A separating triple (N_l, N_{l+1}, N_{l+2}) of v is called a forbidden separating triple of v if $31 \leq l \leq e - 33$.*

(c) *A separating quadruple $(N_l, N_{l+1}, N_{l+2}, N_{l+3})$ of v is called a forbidden separating quadruple of v if $31 \leq l \leq e - 34$.*

(d) *A separating quintuple $(N_l, N_{l+1}, N_{l+2}, N_{l+3}, N_{l+4})$ of v is called a forbidden separating quintuple of v if $31 \leq l \leq e - 35$.*

We say that v has a forbidden separator if v has a forbidden separating pair, a forbidden separating triple, a forbidden separating quadruple, or a forbidden separating quintuple.

Claim 1. *Let G be a 3-edge-connected block and v a vertex of G with $ex_G(v) = e \geq 63$. If v has no forbidden separator, then $k_i + k_{i+1} \geq 5$ for all $i = 31, \ldots, e-32$.*

Proof. Note that, since G is 2-connected, we have $k_i \geq 2$ for all $i = 1, \ldots, e-1$. If for some $i \in \{31, \ldots, e-32\}$, we have $k_i + k_{i+1} = 4$, then clearly, $k_i = 2 = k_{i+1}$ and v has a forbidden separating pair (N_i, N_{i+1}), a contradiction. □

Definition 6. *Let G be a 3-edge-connected block and v a vertex of G with $ex_G(v) = e \geq 63$. Assume that v has no forbidden separator. If for $l \in \{31, \ldots, e-32\}$ we have $k_l + k_{l+1} = 5$, then we say that (N_l, N_{l+1}) is a 5_v-class.*

We remark that if (N_l, N_{l+1}) is a 5_v-class, then either $k_l = 2$ and $k_{l+1} = 3$ or $k_l = 3$ and $k_{l+1} = 2$, since otherwise, if $k_l = 1$ or $k_{l+1} = 1$, G has a cut vertex.

Claim 2. *Let G be a 3-edge-connected block, let v be a vertex of G with no forbidden separator, and let (N_l, N_{l+1}) and (N_{l+e_1}, N_{l+e_1+1}) be two 5_v-classes where e_1 is a positive integer. Then $e_1 \geq 4$.*

Proof. Assume that (N_l, N_{l+1}) is a 5_v-class. We first show that (N_{l+1}, N_{l+2}) cannot be a 5_v-class. If, on one hand, $k_l = 2$ and $k_{l+1} = 3$, then $k_{l+2} > 2$; otherwise v has a forbidden separating triple (N_l, N_{l+1}, N_{l+2}), a contradiction. Thus (N_{l+1}, N_{l+2}) cannot be a 5_v-class. If, on the other hand, $k_l = 3, k_{l+1} = 2$, then $k_{l+2} \geq 4$; otherwise v has a forbidden separating pair (N_{l+1}, N_{l+2}) or a forbidden separating triple (N_l, N_{l+1}, N_{l+2}), which is a contradiction. Therefore, (N_{l+1}, N_{l+2}) is not a 5_v-class.

Next, we show that (N_{l+2}, N_{l+3}) cannot be a 5_v-class. We have seen that $k_{l+2} \geq 3$ and equality holds only if $k_l = 2$ and $k_{l+1} = 3$. Therefore, (N_{l+2}, N_{l+3}) can be a 5_v-class only if $k_{l+2} = 3$ and $k_{l+3} = 2$. However, this implies that v has a forbidden separating quadruple $(N_l, N_{l+1}, N_{l+2}, N_{l+3})$, contradicting the fact that v has no forbidden separator. Thus (N_{l+2}, N_{l+3}) cannot be a 5_v-class.

Last, e_1 cannot be 3, since otherwise v has a forbidden separating quintuple $(N_l, N_{l+1}, N_{l+2}, N_{l+3}, N_{l+4})$. □

Claim 3. *Let G be a 3-edge-connected block and v be a vertex of G with $ex_G(v) = e \geq 65$. Assume that v has no forbidden separator. If (N_l, N_{l+1}), $32 \leq l \leq e-33$, is a 5_v-class, then*

$$k_{l-1} + k_l + k_{l+1} + k_{l+2} \geq 12.$$

Proof. If $k_l = 2$ and $k_{l+1} = 3$, then, as in the proof of Claim 2, $k_{l-1} \geq 4$ and $k_{l+2} \geq 3$. Therefore,

$$k_{l-1} + k_l + k_{l+1} + k_{l+2} \geq 4 + 2 + 3 + 3 = 12,$$

as desired. The case $k_l = 3$ and $k_{l+1} = 2$ follows analogously. □

Claim 4. *Let G be a 3-edge-connected block and v be a vertex of G with $ex_G(v) = e \geq 63$. Assume that v has no forbidden separator. If (N_l, N_{l+1}) is not a 5_v-class, where $l \in \{31, \ldots, e-32\}$, then $k_l + k_{l+1} \geq 6$.*

Proof. By Claim 1, $k_l + k_{l+1} \geq 5$. If $k_l + k_{l+1} = 5$, then (N_l, N_{l+1}) is a 5_v-class, a contradiction. Therefore, $k_l + k_{l+1} \geq 6$, as desired. □

The following lemma guarantees that if the distance layers between two consecutive 5_v-classes have on average less than three vertices, then one of these distance layers has exactly two vertices and small diameter.

Lemma 4. *Let G be a 3-edge-connected block, and let v be a vertex of G with no forbidden separator. Let (N_l, N_{l+1}) be a 5_v-class and e_1 be the smallest integer greater than 4 such that (N_{l+e_1}, N_{l+e_1+1}) is a 5_v-class. If*

$$\sum_{i=l+3}^{l+e_1-2} k_i < 3(e_1 - 4),$$

then

(i) e_1 *is odd,* $(k_{l+3}, k_{l+4}, k_{l+5}, k_{l+6}, \dots, k_{l+e_1-2}) = (2, 4, 2, 4, \dots, 2)$*, and*

(ii) *there exists* i*,* $l \leq i \leq l + e_1 + 1$*, such that* $diam(N_i) \leq 8$ *and* $k_i = 2$*.*

Proof. (i) By the definition of e_1, (N_i, N_{i+1}) is not a 5_v-class for $i = l + 3, \dots, l + e_1 - 3$. Therefore, by Claim 4, $k_i + k_{i+1} \geq 6$. If e_1 is even, then

$$\sum_{i=l+3}^{l+e_1-2} k_i = (k_{l+3} + k_{l+4}) + \cdots + (k_{l+e_1-3} + k_{l+e_1-2})$$
$$\geq 6 + \cdots + 6 = 3(e_1 - 4),$$

a contradiction to our hypothesis. Therefore, e_1 is odd. Now, $\sum_{i=l+3}^{l+e_1-2} k_i = (k_{l+3} + k_{l+4}) + \cdots + (k_{l+e_1-4} + k_{l+e_1-3}) + k_{l+e_1-2} \geq (\frac{e_1-5}{2})6 + 2$ and $\sum_{i=l+3}^{l+e_1-2} k_i = k_{l+3} + (k_{l+4} + k_{l+5}) + \cdots + (k_{l+e_1-3} + k_{l+e_1-2}) \geq 2 + (\frac{e_1-5}{2})6$. If the above inequalities are strict, then $\sum_{i=l+3}^{l+e_1-2} k_i > (\frac{e_1-5}{2})6 + 2 = 3e_1 - 13$, a contradiction. Therefore, both inequalities hold with equality. Then $k_{l+3} = k_{l+e_1-2} = 2$ and $k_i + k_{i+1} = 6$ for all $i \in \{l + 3, l + 4, \dots, l + e_1 - 3\}$. Hence $(k_{l+3}, k_{l+4}, k_{l+5}, k_{l+6}, \dots, k_{l+e_1-2}) = (2, 4, 2, 4, \dots, 2)$, as desired.

(ii) *Observation* 1. From (i) it is immediate that

(a) for every $j = l, l + 1, \dots, l + e_1 - 2$, at least one of $k_{j+1}, k_{j+2}, k_{j+3}$ is equal to 2;

(b) for all $x \in N_i$, $l + 2 \leq i \leq l + e_1 + 1$, there exists $\theta \in \{i - 2, i - 1, i\}$, with $k_\theta = 2$. Hence $d_G(x, N_\theta) \leq 2$.

Suppose to the contrary that (ii) does not hold, i.e.,

$(*)$ there is no $i \in \{l, l+1, \dots, l+e_1+1\}$ with $diam(N_i) \leq 8$ and $k_i = 2$.

Let $N_{l+e_1} \cup N_{l+e_1+1} = \{x_{l+e_1}, y_{l+e_1}, x_{l+e_1+1}, y_{l+e_1+1}, a\}$, where $x_{l+e_1}, y_{l+e_1} \in N_{l+e_1}$, $x_{l+e_1+1}, y_{l+e_1+1} \in N_{l+e_1+1}$. Because of (*), by Fact 1, we can assume without loss of generality that $x_{l+e_1}x_{l+e_1+1}$, $y_{l+e_1}y_{l+e_1+1} \in E(G)$ and a is adjacent to $\{y_{l+e_1}, y_{l+e_1+1}\}$. For $i = 1, 2, \dots, l + e_1 - 1$, let $vx_1x_2 \dots x_{l+e_1-1}x_{l+e_1}$ $(vy_1y_2 \dots y_{l+e_1-1}y_{l+e_1})$ be a $v{-}x_{l+e_1}$ $(v{-}y_{l+e_1})$ shortest path, and denote by P_1 (P_2) its $x_l{-}x_{l+e_1+1}$ $(y_l{-}y_{l+e_1+1})$ section. We first show that

$(**)$ $x_i \neq y_i$ for all $i = l, l + 1, \dots, l + e_1 - 1$.

Suppose that (**) is false, and let i, $i \leq l + e_1$, be the largest value for which $x_i = y_i$. First note that $x_{l+e_1-1} \neq y_{l+e_1-1}$, since otherwise $diam(N_r) \leq 4 < 8$ for some $r \in \{l+e_1, l+e_1+1\}$, with $k_r = 2$, a contradiction to $(*)$. Thus $l \leq i \leq l+e_1-2$. By Observation 1(a), let $N_\theta = \{x_\theta, y_\theta\}$ be such that $k_\theta = 2$ and $\theta \in \{i+1, i+2, i+3\}$. Then $d(x_\theta, y_\theta) \leq d(x_\theta, x_i) + d(y_i, y_\theta) = 3 + 3 < 8$, a contradiction to (*). This establishes (**).

With the above notation, noting Fact 1 and letting $N_l \cup N_{l+1} = \{x_l, y_l, x_{l+1}, y_{l+1}, b\}$, we have that b is adjacent to $\{x_l, x_{l+1}\}$ or $\{y_l, y_{l+1}\}$ but not both. Let $S = \cup_{i=l}^{l+e_1+1} N_i$ and $S \cap N_{\leq 2}(V(P_i)) =: S_i$ for $i = 1, 2$. By Observation 1(b) and Fact 1, $S = S_1 \cup S_2$. We show that, in fact,

$(***)$ S_1 and S_2 are disjoint and not joined by an edge.

Suppose that the two sets intersect and that $u \in S_1 \cap S_2$, $u \in N_i$. We first show that $u \notin N_l \cup N_{l+1}$. Assume without loss of generality that b is adjacent to $\{y_l, y_{l+1}\}$. If, on

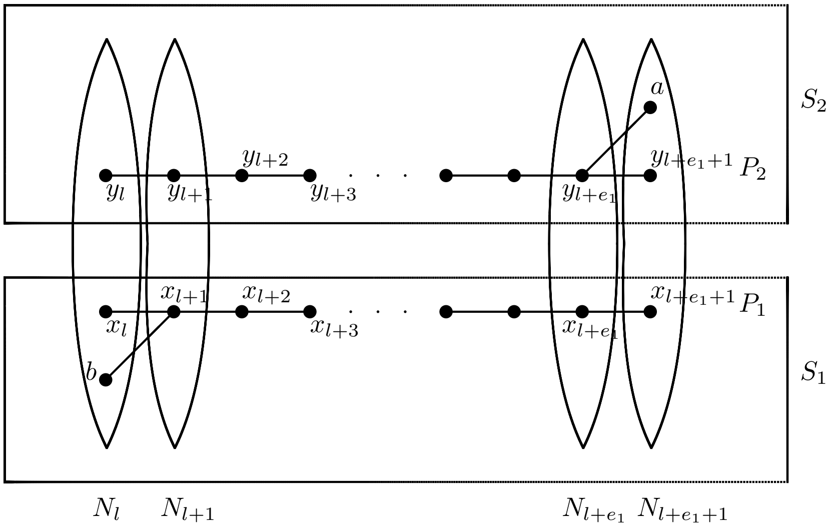


Fig. 4.

one hand, $u \in \{x_l, x_{l+1}\}$ is within distance 2 of $V(P_2)$, then the vertices in $N_l \cup N_{l+1}$ are connected by a path of order at most 8. If, on the other hand, $u \in \{b, y_l, y_{l+1}\}$ is within distance 2 of $V(P_1)$, then again the vertices of $N_l \cup N_{l+1}$ are connected by a path of order at most 8. In either case, we have that $diam(N_r) \leq 8$ for some $r \in \{l, l+1\}$, with $k_r = 2$, a contradiction to $(*)$. Hence $u \notin N_l \cup N_{l+1}$ and thus $l+2 \leq i \leq l+e_1+1$. By Observation 1(b), there exists θ, $i-2 \leq \theta \leq i$, with $d(u, N_\theta) \leq 2$ and $k_\theta = 2$. Without loss of generality we can assume that $d(x_\theta, u) \leq d(y_\theta, u)$ and thus $d(x_\theta, u) \leq 2$. From $u \in S_2$, let y_r be such that $d(u, y_r) \leq 2$. Hence $i-2 \leq r \leq i+2$. Thus $d(x_\theta, y_\theta) \leq d(x_\theta, u) + d(u, y_r) + d(y_r, y_\theta) \leq 2+2+4 = 8$, a contradiction to (*). Therefore, S_1 and S_2 are disjoint.

Now suppose to the contrary that pq joins S_1 and S_2, $p \in N_i$ and $q \in N_j$, and hence $|i-j| \leq 1$. We first show that $i, j \geq l+2$. Assume, without loss of generality, that b is adjacent to $\{y_l, y_{l+1}\}$. Since S_1 and S_2 are disjoint, no vertex from $\overline{N}(\{x_l, x_{l+1}\})$ is adjacent to a vertex from $\overline{N}(\{b, y_l, y_{l+1}\})$, and hence $i, j \geq l+2$. By Observation 1(b) let $N_\theta = \{x_\theta, y_\theta\}$, $i-2 \leq \theta \leq i$, be such that $d(p, N_\theta) \leq 2$. Therefore, since S_1 and S_2 are disjoint, $d(p, x_\theta) \leq 2$. Again by Observation 1(b) let $N_r = \{x_r, y_r\}$, $j-2 \leq r \leq j$, be such that $d(q, N_r) \leq 2$. Therefore, since S_1 and S_2 are disjoint, $d(q, y_r) \leq 2$. The fact that $|i-j| \leq 1$, $i-2 \leq \theta \leq i$, in conjunction with $j-2 \leq r \leq j$ gives $d(y_r, y_\theta) \leq 3$. It follows that $d(x_\theta, y_\theta) \leq d(x_\theta, p) + d(p, q) + d(q, y_r) + d(y_r, y_\theta) \leq 2+1+2+3 = 8$, a contradiction to (*). Therefore, S_1 and S_2 are not joined by an edge and (***) is proved. (Refer to Figure 4 showing the paths P_1 and P_2 and all of the edges in $E(N_l, N_{l+1})$ and $E(N_{l+e_1}, N_{l+e_1+1})$ for the case where b is in N_l and adjacent to $\{x_l, x_{l+1}\}$ and a is in N_{l+e_1+1} and adjacent to $\{y_{l+e_1}, y_{l+e_1+1}\}$.)

By (***), a path from N_0 to N_e thus contains either edges from $E(N_l, N_{l+1}) \cap E(G[S_1])$ and from $E(N_{l+e_1}, N_{l+e_1+1}) \cap E(G[S_1])$ or edges from $E(N_l, N_{l+1}) \cap E(G[S_2])$ and from $E(N_{l+e_1}, N_{l+e_1+1}) \cap E(G[S_2])$. Hence, the union of any two of the four sets $E(N_l, N_{l+1}) \cap E(G[S_1])$, $E(N_{l+e_1}, N_{l+e_1+1}) \cap E(G[S_1])$, $E(N_l, N_{l+1}) \cap E(G[S_2])$, $E(N_{l+e_1}, N_{l+e_1+1}) \cap E(G[S_2])$ is a disconnecting set of G.

Recall by Fact 1 that $x_{l+e_1}x_{l+e_1+1}$ is the only edge in $E(N_{l+e_1}, N_{l+e_1+1}) \cap E(G[S_1])$. Also b is adjacent to $\{x_l, x_{l+1}\}$ or b is adjacent to $\{y_l, y_{l+1}\}$ but not both. In the former case, $y_l y_{l+1}$ is the only edge in $E(N_l, N_{l+1}) \cap E(G[S_2])$; hence $\{y_l y_{l+1}, x_{l+e_1} x_{l+e_1+1}\}$ is a disconnecting set of G, a contradiction. In the latter case, $x_l x_{l+1}$ is the only edge in $E(N_l, N_{l+1}) \cap E(G[S_1])$; hence $\{x_l x_{l+1}, x_{l+e_1} x_{l+e_1+1}\}$ is a disconnecting set of G, again a contradiction. This completes the proof of the lemma. ◻

3.2. An upper bound on distance in 3-edge-connected blocks. In this subsection, we begin by establishing an upper bound on the distance of a vertex v of a 3-edge-connected block G, given that v has no forbidden separator and that for any two consecutive 5_v-classes (N_l, N_{l+1}) and (N_{l+e_1}, N_{l+e_1+1}) we have $\sum_{i=l+3}^{l+e_1-2} k_i \geq 3(e_1 - 4)$. First we need the following definition about sequences which carry some information about the vertex v.

DEFINITION 7. *Let n and e, $e \geq 67$, be positive integers. We say that a sequence of numbers $(a_i) = (a_0, a_1, \dots, a_e)$ is (n,e)-realizable if (a_i) satisfies the following conditions:*

(A) $\sum_{i=0}^{e} a_i = n$.

(B) $a_0 = 1,\ a_1 \geq 3,\ a_i \geq 2$ *for* $i \in \{2, 3, \dots, e-1\}$ *and* $a_{e-1} + a_e \geq 4$.

(C) $a_i + a_{i+1} \geq 5$ *for* $i \in \{31, \dots, e-32\}$.

(D) *If* $i, j \in \{31, \dots, e-32\}$, $i < j$, *and* $a_i + a_{i+1} = 5 = a_j + a_{j+1}$, *then* $j - i \geq 4$.

(E) *If* $i_1 < i_2 < \cdots < i_t$, $i_1 \geq 31$, $i_t \leq e-32$, *are consecutive integers for which* $a_{i_1} + a_{i_1+1} = a_{i_2} + a_{i_2+1} = \cdots = a_{i_t} + a_{i_t+1} = 5$, *then the following hold:*

(i) $\sum_{r=i_j+3}^{i_{j+1}-2} a_r \geq 3(i_{j+1} - i_j - 4)$ *for* $j = 1, 2, \dots, t-1$.

(ii) $\sum_{r=i_j-1}^{i_j+2} a_r \geq 12$ *for* $j = 2, \dots, t-1$.

(iii) *If* $i_1 = 31$ $(i_t = e-32)$, *then* $\sum_{r=i_1}^{i_1+2} a_r$ $(\sum_{r=i_t-1}^{i_t+1} a_r) \geq 8$ *and if* $i_1 \geq 32$ $(i_t \leq e-33)$, *then* $\sum_{r=i_1-1}^{i_1+2} a_r (\sum_{r=i_t-1}^{i_t+2} a_r) \geq 12$.

LEMMA 5. *Let n and e, $e \geq 67$, be integers and (a_i) be an (n,e)-realizable sequence. Then there exists a sequence $(b_i) = (b_0, b_1, \dots, b_e)$ which satisfies the following conditions:*

(i) $b_0 = 1,\ b_1 \geq 3,\ b_i \geq 2$ *for* $i \in \{2, 3, \dots 30\} \cup \{e-30, e-29, \dots, e-1\}$ *and* $b_{e-1} + b_e \geq 4$;

(ii) $b_i \geq 3$ *for all values of* $i \in \{31, \dots, e-31\}$ *with the exception of at most two values of* i *for which* b_i *may assume the value* 2;

(iii) $\sum_{i=0}^{e} b_i = n$;

(iv) $\sum_{i=1}^{e} i a_i \leq \sum_{i=1}^{e} i b_i + \frac{e-61}{2}$.

Proof. First we prove the statement for the case that $a_{i-1} + a_i \geq 6$ for all $i \in \{32, \dots, e-31\}$. We consider two cases separately. If, on one hand, e is odd, then let

$$b_i = \begin{cases} a_i & \text{if } i \in \{0, 1, \dots, 30\} \cup \{e-30, \dots, e\}, \\ 3 & \text{if } i \in \{31, \dots, e-31\},\ i \text{ odd}, \\ a_{i-1} + a_i - 3 & \text{if } i \in \{31, \dots, e-31\},\ i \text{ even}. \end{cases}$$

Clearly, since (a_i) is (n,e)-realizable, (i) holds for (b_i), and, moreover since $a_{i-1} + a_i \geq 6$ for all $\{32, \dots, e-31\}$, we have $b_i \geq 3$ for all $i \in \{31, \dots, e-31\}$ and thus (ii) holds for (b_i). For all $q = 15, 16, \dots, (e-33)/2$, we have $b_{2q+1} + b_{2q+2} = a_{2q+1} + a_{2q+2}$, and hence from $\sum_{i=0}^{e} a_i = n$ it follows that $\sum_{i=0}^{e} b_i = n$; that is, (iii) holds for (b_i). Moreover,

$$\begin{aligned} (2q+1)a_{2q+1} + (2q+2)a_{2q+2} &= (2q+1)b_{2q+1} + (2q+2)b_{2q+2} + 3 - a_{2q+1} \\ &\leq (2q+1)b_{2q+1} + (2q+2)b_{2q+2} + 1. \end{aligned}$$

Thus,

$$\begin{aligned}\sum_{i=31}^{e-31} ia_i &= \sum_{q=15}^{\frac{e-33}{2}} [(2q+1)a_{2q+1} + (2q+2)a_{2q+2}] \\ &\leq \sum_{q=15}^{\frac{e-33}{2}} [(2q+1)b_{2q+1} + (2q+2)b_{2q+2} + 1] = \sum_{i=31}^{e-31} ib_i + \frac{e-61}{2}.\end{aligned}$$

Hence $\sum_{i=1}^{e} ia_i \leq \sum_{i=1}^{e} ib_i + \frac{e-61}{2}$, and (b_i) is the required sequence. If, on the other hand, e is even, then let

$$b_i = \begin{cases} a_i & \text{if } i \in \{0,1,\dots,30\} \cup \{e-31,\dots,e\}, \\ 3 & \text{if } i \in \{31,\dots,e-32\},\ i \text{ odd}, \\ a_{i-1}+a_i-3 & \text{if } i \in \{31,\dots,e-32\},\ i \text{ even}. \end{cases}$$

Clearly, as above, (b_i) is the required sequence with a slightly stronger bound $\sum_{i=1}^{e} ia_i \leq \sum_{i=1}^{e} ib_i + \frac{e-62}{2}$.

Hence we can assume that there is at least one member i, $i \in \{31,\dots,e-32\}$, for which $a_i + a_{i+1} = 5$. Let $i_1 < i_2 < \cdots < i_t$ be the collection of these values of i. Define the sets $A := \{0,1,\dots,30\}$, $J_1 := \{31,32,\dots,i_1+2\}$, for $j = 1,2,\dots,t-1$, $R_j := \{i_j+3, i_j+4, \dots, i_{j+1}-2\}$, for $j = 2,3,\dots,t-1$, $S_j := \{i_j-1, i_j, i_j+1, i_j+2\}$, $J_2 := \{i_t-1, i_t, \dots, e-31\}$, and $B := \{e-30,\dots,e\}$. Then $\{0,1,\dots,e\} = A \cup J_1 \cup (\bigcup_{j=1}^{t-1} R_j) \cup (\bigcup_{j=2}^{t-1} S_j) \cup J_2 \cup B$. We define the sequence (b_i) on each of these sets. We will define b_i so that the equation $\sum_{h\in S} a_h = \sum_{h\in S} b_h$ for all $S \in \{A, J_1, R_j, S_j, J_2, B\}$ can easily be verified, mostly by manipulations of pairs of the sequences as done above. For $i \in A \cup B$, let $b_i = a_i$. Hence (b_i) satisfies (i).

Let $h \in R_j$. We consider three cases. For $i_{j+1} - i_j$ even, let

$$b_h = \begin{cases} 3 & \text{if } h = i_j + r,\ r \text{ odd}, \\ a_{h-1} + a_h - 3 & \text{if } h = i_j + r + 1, r \text{ odd}. \end{cases}$$

Since $a_{h-1} + a_h \geq 6$, we have $b_h \geq 3$ for all $h \in R_j$. For $q = 1,2,\dots,(i_{j+1} - i_j - 4)/2$, we have

$$\begin{aligned}(i_j+2q+1)a_{i_j+2q+1} &+ (i_j+2q+2)a_{i_j+2q+2} \\ &= (i_j+2q+1)b_{i_j+2q+1} + (i_j+2q+2)b_{i_j+2q+2} \\ &\quad +3 - a_{i_j+2q+1} \\ &\leq (i_j+2q+1)b_{i_j+2q+1} + (i_j+2q+2)b_{i_j+2q+2} + 1.\end{aligned}$$

Hence for all $j = 1,2,\dots,t-1$, we obtain $\sum_{h\in R_j} ha_h \leq \sum_{h\in R_j} hb_h + (i_{j+1} - i_j - 4)/2$. For $i_{j+1} - i_j$ odd and $a_{i_{j+1}-2} \geq 3$, let

$$b_h = \begin{cases} 3 & \text{if } h = i_j + r, r \text{ odd}, h \neq i_{j+1} - 2, \\ a_{h-1} + a_h - 3 & \text{if } h = i_j + r + 1, r \text{ odd}, \\ a_h & \text{if } h = i_{j+1} - 2. \end{cases}$$

Clearly, as above, $b_h \geq 3$, and, for all $j = 1,2,\dots,t-1$, we have $\sum_{h\in R_j} ha_h \leq \sum_{h\in R_j} hb_h + (i_{j+1} - i_j - 5)/2$. For $i_{j+1} - i_j$ odd and $a_{i_{j+1}-2} = 2$, we proceed as follows. From $\sum_{r=i_j+3}^{i_{j+1}-2} a_r \geq 3(i_{j+1}-i_j-4)$, there exists $\theta = i_j + r$, r odd, such that $a_\theta + a_{\theta+1} \geq 7$. We define b_h as follows: Let

$$b_h = \begin{cases} 3 & \text{if } h = i_j + r, r \text{ odd}, \\ a_{h-1} + a_h - 3 & \text{if } h = i_j + r + 1, r \text{ odd}, h \neq \theta + 1, \\ a_\theta + a_{\theta+1} - 4 & \text{if } h = \theta + 1. \end{cases}$$

Clearly, $b_h \geq 3$ for all $h \in R_j$. Denote the set $\{\theta, \theta+1, i_{j+1}-2\}$ by A_j. Since $\theta - i_{j+1} \leq -4$, we have $\sum_{r \in A_j} ra_r = \sum_{r \in A_j} rb_r + \theta - i_{j+1} + 6 - a_\theta \leq \sum_{r \in A_j} rb_r$. A simple calculation as above shows that $\sum_{r \in R_j \setminus A_j} ra_r \leq \sum_{r \in R_j \setminus A_j} rb_r + (i_{j+1} - i_j - 7)/2$. Hence for all $j = 1, 2, \ldots, t-1$, we have $\sum_{h \in R_j} ha_h \leq \sum_{h \in R_j} hb_h + (i_{j+1} - i_j - 7)/2$. We conclude that in all cases, for $h \in R_j$, $j = 1, 2 \ldots, t-1$, we have defined $b_h \geq 3$ so that

$$\sum_{h \in R_j} ha_h \leq \sum_{h \in R_j} hb_h + \frac{i_{j+1} - i_j - 4}{2}. \tag{6}$$

Recall that $S_j = \{i_j - 1, i_j, i_j + 1, i_j + 2\}$ for $j = 2, 3, \ldots, t-1$. For $h \in S_j$, let $b_h = 3$ if $h \neq i_j + 2$ and $b_h = \sum_{r \in S_j} a_r - 9$ if $h = i_j + 2$. Since $\sum_{r \in S_j} a_r \geq 12$, we have $b_h \geq 3$ for all $h \in S_j$. From $a_{i_j-1} + a_{i_j} \geq 6$, $a_{i_j} + a_{i_j+1} = 5$, $a_{i_j-1} \geq 3$,

$$\sum_{h \in S_j} ha_h = \sum_{h \in S_j} hb_h + 13 - (a_{i_j-1} + a_{i_j}) - 2a_{i_j-1} \leq \sum_{h \in S_j} hb_h + 1. \tag{7}$$

Recall that $J_1 = \{31, \ldots, i_1 + 2\}$ and $J_2 = \{i_t - 1, \ldots, e - 31\}$. For $h \in J_1 \cup J_2$, we define b_h as follows. If $i_1 = 31$ ($i_t = e - 32$), then $b_h = a_h$ for $h \in J_1$ ($h \in J_2$). Thus $b_h \geq 3$ for all $h \in J_1$ ($h \in J_2$) with the exception of at most one value of h. If, however, $i_1 > 31$, let

$$b_h = \begin{cases} 3 & \text{if } h = i_1 - 1, i_1, i_1 + 1, \\ a_{i_1-1} + a_{i_1} + a_{i_1+1} + a_{i_1+2} - 9 & \text{if } h = i_1 + 2, \end{cases}$$

and, for $h \in \{31, \ldots, i_1 - 2\}$ and i_1 even, let

$$b_h = \begin{cases} 3 & \text{if } h \text{ is odd}, \\ a_{h-1} + a_h - 3 & \text{if } h \text{ is even}, \end{cases}$$

whereas for $h \in \{31, \ldots, i_1 - 2\}$ and i_1 odd, set

$$b_h = \begin{cases} 3 & \text{if } h \text{ is odd } h \neq i_1 - 2, \\ a_{h-1} + a_h - 3 & \text{if } h \text{ is even}, \\ a_h & \text{if } h = i_1 - 2. \end{cases}$$

Clearly, for $i_1 > 31$ we have $b_h \geq 3$ for all $h \in J_1$ with the exception of at most one value of h. Moreover, similar calculations as above give $\sum_{h \in J_1} ha_h \leq \sum_{h \in J_1} hb_h + (i_1 - 30)/2$. If $i_t \leq e - 33$, let

$$b_h = \begin{cases} 3 & \text{if } h = i_t - 1, i_t, i_t + 1, \\ a_{i_t-1} + a_{i_t} + a_{i_t+1} + a_{i_t+2} - 9 & \text{if } h = i_t + 2, \end{cases}$$

and, for $h \in \{i_t + 3, \ldots, e - 31\}$ and $e - i_t$ odd, let

$$b_h = \begin{cases} 3 & \text{if } h = i_t + r \text{ , } r \text{ is odd}, \\ a_{h-1} + a_h - 3 & \text{if } h = i_t + r + 1, \ r \text{ is odd}, \end{cases}$$

whereas, for $h \in \{i_t+3,\dots,e-31\}$ and $e-i_t$ even, set

$$b_h = \begin{cases} 3 & \text{if } h = i_t + r,\ r \text{ is odd } h \neq e-31, \\ a_{h-1}+a_h-3 & \text{if } h = i_t+r+1,\ r \text{ odd}, \\ a_h & \text{if } h = e-31. \end{cases}$$

Clearly, for $i_t \le e-33$ we have $b_h \ge 3$ for all $h \in J_2$ with the exception of at most one value of h. Moreover, similar calculations as above give $\sum_{h\in J_2} ha_h \le \sum_{h \in J_2} hb_h + (e-i_t-31)/2$. Thus,

$$\sum_{h\in J_1\cup J_2} ha_h \le \sum_{h\in J_1 \cup J_2} hb_h + \frac{i_1-30}{2} + \frac{e-i_t-31}{2}. \tag{8}$$

Recall that $\{0,1,\dots,e\} = A \cup J_1 \cup (\bigcup_{j=1}^{t-1} R_j) \cup (\bigcup_{j=2}^{t-1} S_j) \cup J_2 \cup B$. Hence the fact that $\sum_{h\in S} a_h = \sum_{h \in S} b_h$ if S equals A, J_1, J_2, B and if S equals R_j, S_j for all j yields $n = \sum_{i=0}^{e} a_i = \sum_{i=0}^{e} b_i$. The fact that $b_i = a_i$ for all $i \in A\cup B$, in conjunction with (6), (7), and (8), yields

$$\begin{aligned}\sum_{r=1}^{e} ra_r &\le \sum_{r=1}^{e} rb_r + \sum_{j=1}^{t-1}\left(\frac{i_{j+1}-i_j-4}{2}\right) + \sum_{j=2}^{t-1} 1 + \frac{i_1-30}{2} + \frac{e-i_t-31}{2} \\ &= \sum_{r=1}^{e} rb_r + \frac{e-61}{2} - t < \sum_{r=1}^{e} rb_r + \frac{e-61}{2},\end{aligned}$$

as desired. This completes the proof of the lemma. □

Lemma 6. *Let G be a 3-edge-connected block, and let v be a vertex of G with $ex_G(v) = e \ge 67$. Assume that v has no forbidden separator. If for any two consecutive 5_v-classes (N_l, N_{l+1}) and (N_{l+e_1}, N_{l+e_1+1}),*

$$\sum_{i=l+3}^{l+e_1-2} k_i \ge 3(e_1-4),$$

then

$$\sigma_G(v) \le \frac{1}{6}(n^2+64n+469).$$

Proof. Since G is a 3-edge-connected block, we have $k_0 = 1, k_1 \ge 3,\ k_i \ge 2$ for $i \in \{2,3,\dots,30\} \cup \{e-30,\dots,e-1\}$ and $k_{e-1}+k_e \ge 4$. Hence our hypothesis, in conjunction with Claims 1, 2, 3, and 4, yields that the sequence $(k_0,k_1,\dots,k_e)$ is (n,e)-realizable.

By Lemma 5, let $(b_i) = (b_0,b_1,\dots,b_e)$ be a sequence with

(i) $b_0 = 1,\ b_1 \ge 3,\ b_i \ge 2$ for $i \in \{2,3,\dots,e-1\}$ and $b_{e-1}+b_e \ge 4$,

(ii) $b_i \ge 3$ for all values of $i \in \{31,\dots,e-31\}$ with the exception of at most two values of i,

(iii) $\sum_{i=0}^{e} b_i = n$, and

(iv) $\sum_{i=1}^{e} ik_i \le \sum_{i=1}^{e} ib_i + \frac{e-61}{2}$.

We first find an upper bound on $\sum_{i=1}^{e} ib_i$ by maximizing $\sum_{i=1}^{e} ib_i$ subject to the constraints (i), (ii), and (iii) above. Clearly, subject to these conditions, $\sum_{i=1}^{e} ib_i$ is maximum for $b_0 = 1, b_1 = 3,\ b_i = 2$ for $i \in \{2,3,\dots,32\} \cup \{e-30,\dots,e-1\}$, $b_i = 3$

for $i = 33, 34, \dots, e-31$, and $b_e = n - 3e + 63$. Thus

$$\begin{aligned}\sum_{i=1}^{e} ib_i &\le 3 + 2(2+3+\cdots+32) + 3(33+34+\cdots+e-31)\\ &\quad + 2(e-30+\cdots+e-1) + e(n-3e+63)\\ &= en - \frac{3}{2}e^2 + \frac{63}{2}e - 62;\end{aligned}$$

hence from (iv), we have

$$\sigma_G(v) = \sum_{i=1}^{e} ik_i \le en - \frac{3}{2}e^2 + 32e - \frac{185}{2}. \tag{9}$$

From the hypothesis of the lemma, and a little calculation, we have $n = \sum_{i=0}^{e} k_i \ge 3e - 61$, that is, $e \le \frac{n+61}{3}$. Subject to this condition and by differentiation, (9) is maximized for $e = \frac{n+32}{3}$ to yield $\sigma_G(v) \le \frac{1}{6}(n^2 + 64n + 469)$, as desired. □

Lemma 7. *Let G be a 3-edge-connected block of order n. Then*

$$\sigma(G) \le n(n-1)\left(\frac{n}{6} + 24\right).$$

Proof. Let $v \in V(G)$ be a vertex of largest distance, and let e be its eccentricity. If $e \le 66$, we distinguish two cases. For $n \le 133$, Corollary 2 gives the result, whereas, if $n \ge 134$, the following holds:

$$\sigma_G(v) \le 3 + 2(2+\cdots+65) + 66(n-1-3-2\cdot 64) = 66n - 4421;$$

hence

$$\sigma(G) \le n\sigma_G(v) \le n(66n - 4421) \le n(n-1)\left(\frac{n}{6} + 24\right)$$

for all n. Thus assume that $e \ge 67$, and hence $n \ge 136$. If v has no forbidden separator and for any two consecutive 5_v-classes (N_l, N_{l+1}) and (N_{l+e_1}, N_{l+e_1+1}), $\sum_{i=l+3}^{l+e_1-2} k_i \ge 3(e_1 - 4)$, then by Lemma 6, $\sigma_G(v) \le \frac{1}{6}(n^2 + 64n + 469)$. Summing $\sigma_G(v)$ for all v and since $n > 8$, we obtain $\sigma(G) \le \frac{1}{6}n(n^2 + 64n + 469) \le n(n-1)(\frac{n}{6} + 24)$, as desired.

If, however, v has a forbidden separator or there exists two consecutive 5_v-classes (N_l, N_{l+1}) and (N_{l+e_1}, N_{l+e_1+1}) such that $\sum_{i=l+3}^{l+e_1-2} k_i < 3(e_1 - 4)$, then we proceed with our proof by induction on the order n of G. For $n < 288$, the result follows by Corollary 2. We assume that $n \ge 288$ and that, for any 3-edge-connected block with less than n vertices, the result holds. Since v has a forbidden separator or there are two consecutive 5_v-classes (N_l, N_{l+1}) and (N_{l+e_1}, N_{l+e_1+1}) such that $\sum_{i=l+3}^{l+e_1-2} k_i < 3(e_1 - 4)$, then by Propositions 1–3, Lemmas 3 and 4, let $N_\alpha = \{u_1, u_2\}$, $31 \le \alpha \le e - 31$, be a distance layer with $d(u_1, u_2) \le 8$.

We form a new graph H from G as follows: Consider the sequential join $K = \sum_{i=1}^{7} H_i$, where

$$H_i = \begin{cases} K_3 & \text{if } i \text{ is odd,}\\ K_1 & \text{if } i \text{ is even,}\end{cases}$$

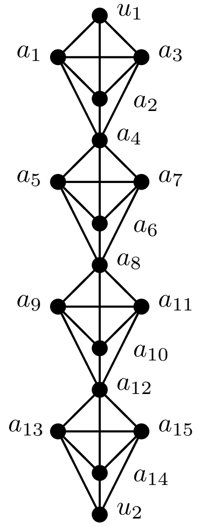


FIG. 5. *Attaching graph K to $N_\alpha = \{u_1, u_2\}$.*

and $V(K) \cap V(G) = \emptyset$. We attach the graph K to G to obtain the graph H by joining u_1 (respectively, u_2) to every vertex of H_1 (respectively, H_7).

Denote the vertex set of K by $V(K) = A = \{a_1, a_2, \ldots, a_{15}\}$. (See Figure 5).

Clearly, in $H[N_\alpha \cup A]$, $d(u_1, u_2) = 8$, unless $u_1u_2 \in E(G)$; hence, since $d_G(u_1, u_2) \leq 8$, we have $d_G(u_1, u_2) = d_H(u_1, u_2)$. This, in conjunction with the fact that new vertices and new edges have been attached exclusively to u_1 and u_2, yields

$$d_G(x, y) = d_H(x, y) \text{ for all } x, y \in V(G). \tag{10}$$

Let $G_1 = H[N_{\leq\alpha} \cup A]$ and $G_2 = H[N_{\geq\alpha} \cup A]$. Thus G_1 and G_2 have in common $N_\alpha \cup A$. Hence, letting $|V(G_i)| = n_i$ for $i = 1, 2$, we obtain

$$n_1 + n_2 - 32 = n. \tag{11}$$

Note that $|N_{\leq\alpha}| \geq \sum_{i=0}^{31} k_i \geq 64$ and $|N_{\geq\alpha}| \geq \sum_{e-31}^{e} k_i \geq 64$. Hence, $n_1 \geq |N_{\leq\alpha}| + |A| \geq 64 + 15 = 79$ and $n_2 \geq |N_{\geq\alpha}| + |A| \geq 64 + 15 = 79$. Thus from (11), we have $n_1 \leq n - 47$. We are going to bound $\sigma(G)$ using upper bounds on $\sigma(G_1)$ and $\sigma(G_2)$. Hence, the following two claims, which are easy to verify, are important.

CLAIM 5. *G_1 and G_2 are 3-edge-connected blocks.*

CLAIM 6. *Let $X = \sum_{i=1}^{15} \sigma_{G_1}(a_i)$ and $Y = \sum_{i=1}^{15} \sigma_{G_2}(a_i)$. Then*

$$\begin{aligned} \sigma(G) \quad = \quad & \sigma(G_1) + \sigma(G_2) - 2(X+Y) + 2\sigma_A(A) \\ & + 2 \sum_{x \in N_{<\alpha}} \sum_{y \in N_{>\alpha}} d_H(x, y) - \sum_{(x,y) \in N_\alpha \times N_\alpha} d_H(x, y). \end{aligned}$$

Clearly,

$$\sigma_A(A) \leq \sigma(K) = 568 \text{ and } \sum_{(x,y) \in N_\alpha \times N_\alpha} d_H(x, y) = 2d_G(u_1, u_2) \geq 2. \tag{12}$$

Now in H,

$$\begin{aligned}2\left(\sum_{x\in N_{<\alpha},y\in N_{>\alpha}} d_H(x,y)\right) &\le 2\sum_{x\in N_{<\alpha}}\sum_{y\in N_{>\alpha}}[d_H(x,u_1)+d_H(u_1,y)]\\ &= 2\sum_{x\in N_{<\alpha}}[|N_{>\alpha}|d_H(x,u_1)+\sigma_{G_2}(u_1)-\sigma_{A\cup\{u_2\}}(u_1)]\\ &= 2[|N_{>\alpha}|(\sigma_{G_1}(u_1)-\sigma_{A\cup\{u_2\}}(u_1))+|N_{<\alpha}|(\sigma_{G_2}(u_1)-\sigma_{A\cup\{u_2\}}(u_1))].\end{aligned}$$

Since $d_G(u_1,u_2)\ge 1$, we have $\sigma_{A\cup\{u_2\}}(u_1)\ge 40$. Thus,

$$2\sum_{x\in N_{<\alpha},y\in N_{>\alpha}} d_H(x,y)\le 2[(n_2-17)(\sigma_{G_1}(u_1)-40)+(n_1-17)(\sigma_{G_2}(u_1)-40)]. \tag{13}$$

We find lower bounds on X and Y. Note that

$$\begin{aligned}\sigma_{G_1}(u_1)=\sum_{x\in V(G_1)} d(u_1,x) &\le \sum_{x\in V(G_1)}[d(u_1,a_1)+d(a_1,x)]\\ &= \sum_{x\in V(G_1)}[1+d(a_1,x)] = n_1+\sigma_{G_1}(a_1).\end{aligned}$$

Therefore, $\sigma_{G_1}(a_1)\ge\sigma_{G_1}(u_1)-n_1$. Similarly, $\sigma_{G_1}(a_2),\sigma_{G_1}(a_3)\ge\sigma_{G_1}(u_1)-n_1$; $\sigma_{G_1}(a_4)\ge\sigma_{G_1}(u_1)-2n_1$; $\sigma_{G_1}(a_5),\sigma_{G_1}(a_6),\sigma_{G_1}(a_7)\ge\sigma_{G_1}(u_1)-3n_1$; $\sigma_{G_1}(a_8)\ge\sigma_{G_1}(u_1)-4n_1$; $\sigma_{G_1}(a_9),\sigma_{G_1}(a_{10}),\sigma_{G_1}(a_{11})\ge\sigma_{G_1}(u_1)-5n_1$; $\sigma_{G_1}(a_{12})\ge\sigma_{G_1}(u_1)-6n_1$; $\sigma_{G_1}(a_{13}),\sigma_{G_1}(a_{14}),\sigma_{G_1}(a_{15})\ge\sigma_{G_1}(u_1)-7n_1$. Summing yields

$$X\ge 15\sigma_{G_1}(u_1)-60n_1. \tag{14}$$

Analogously,

$$Y\ge 15\sigma_{G_2}(u_1)-60n_2. \tag{15}$$

By Claim 5, we use induction to bound $\sigma(G_1)$ and $\sigma(G_2)$. Hence $\sigma(G_i)\le n_i(n_i-1)\left(\frac{1}{6}n_i+24\right)$ for $i=1,2$. Combining this with Claim 6 and (12), (13), (14), and (15) yields

$$\begin{aligned}\sigma(G)\le n_1(n_1-1)\left(\frac{1}{6}n_1+24\right)+n_2(n_2-1)\left(\frac{1}{6}n_2+24\right)\\ +2(n_2-32)\sigma_{G_1}(u_1)+2(n_1-32)\sigma_{G_2}(u_1)+40n+5134.\end{aligned}$$

This, in conjunction with Lemma 1 and (11), yields

$$\sigma(G)\le n(n-1)\left(\frac{1}{6}n+24\right)+\left\{\frac{3142}{3}(n_1+n_2)-\frac{53}{3}n_1n_2+40n-14706\right\}.$$

Denote the term in curly brackets of the right-hand side of the inequality by $f(n_1,n_2)$. We show that $f(n_1,n_2)\le 0$. Recall that $n_1,n_2\ge 79$, and thus $(n_1-79)(n_2-79)\ge 0$. It follows that $n_1n_2\ge 79(n_1+n_2)-6241$. This, in conjunction with (11), yields $f(n_1,n_2)\le 84405-\frac{925}{3}n\le 0$ for all $n\ge 274$, as desired. This completes the proof of Lemma 7. □

4. An upper bound on average distance in 3-edge-connected graphs. We now show that the upper bound on the distance, proved in the previous section for 3-edge-connected blocks, holds for all 3-edge-connected graphs.

THEOREM 8. *Let G be a 3-edge-connected graph of order n. Then*

$$\mu(G) \leq \frac{1}{6}n + 24.$$

Apart from the additive constant, this inequality is best possible.

Proof. We prove this theorem by induction on the number of blocks in G. If G is a block, then the theorem follows by Lemma 7. So assume that G has at least two blocks. We prove the equivalent statement

$$\sigma(G) \leq n(n-1)\left(\frac{1}{6}n + 24\right).$$

Let G_1 be an end block, G_2 be the union of the other blocks, and u be the unique common vertex of G_1 and G_2. Hence, letting $n_i = |V(G_i)|$ for $i = 1, 2$, we obtain $n_1 + n_2 - 1 = n$. Note that

$$\begin{aligned}\sigma(G) &= \sum_{(x,y)\in V(G)\times V(G)} d(x,y)\\ &= \sigma(G_1) + 2\sum_{x\in V(G_1)-\{u\}}\ \sum_{y\in V(G_2)-\{u\}} d(x,y) + \sigma(G_2)\\ &= \sigma(G_1) + 2[(n_2-1)\sigma_{G_1}(u) + (n_1-1)\sigma_{G_2}(u)] + \sigma(G_2). \qquad (16)\end{aligned}$$

It is easy to verify that G_1 and G_2 are 3-edge-connected. Hence, by the induction hypothesis, we get $\sigma(G_i) \leq n_i(n_i-1)(\frac{1}{6}n_i + 24)$ for $i = 1, 2$. By Lemma 1, we have $\sigma_{G_i}(u) \leq \frac{1}{4}(n_i^2 - n_i) + \frac{1}{2}$ for $i = 1, 2$. This, in conjunction with $n_1 + n_2 - 1 = n$ and (16), yields

$$\sigma(G) \leq n(n-1)\left(\frac{1}{6}n + 24\right) + \left\{\frac{146}{3}(n_1+n_2) - \frac{143}{3}n_1n_2 - \frac{149}{3}\right\}.$$

Denote the term in curly brackets of the right-hand side of the inequality by $f(n_1, n_2)$. We show that $f(n_1, n_2) \leq 0$. Since $n_1, n_2 \geq 4$, we have $(n_1 - 4)(n_2 - 4) \geq 0$, that is, $n_1n_2 \geq 4(n_1 + n_2) - 16$. This, in conjunction with $n_1 + n_2 = n + 1$, yields $f(n_1, n_2) \leq 571 - 142n \leq 0$ for all $n \geq 5$. Observe that if $n = 4$, then $G = K_4$, and our result clearly holds.

It remains to show that, apart from an additive constant, the bound is best possible. Let n be an even integer, and let $G_i = K_2$, where $i \in \mathbf{N}$. Let $G_{n,3} = G_1 + G_2 + \cdots + G_k$, where $k = \frac{n}{2}$. Clearly, $G_{n,3}$ is 3-edge-connected and $\mu(G_{n,3}) = \frac{1}{6}n + \frac{1}{6} + \frac{1}{2(n-1)}$, as desired. □

THEOREM 9. *Let G be a 4-edge-connected graph of order n. Then*

$$\mu(G) \leq \frac{1}{6}n + 24.$$

Apart from the additive constant, this inequality is best possible.

Proof. Since a 4-edge-connected graph is also 3-edge-connected, the previous theorem applies. To see that, apart from the additive constant, the bound is best possible, consider the graph $G_{n,4} = G_1 + G_2 + G_3 + \cdots + G_k$, where $G_i = K_1$, for

$i = 1, k$, $G_i = K_2$ for $i = 3, 4, \ldots, k-2$, $G_i = K_4$, for $i = 2, k-1$, and $k = \frac{n-2}{2}$, n an even integer. Clearly, $G_{n,4}$ is 4-edge-connected and

$$\mu(G) = \frac{1}{6}n + \frac{1}{6} + \frac{100 - 21n}{2n(n-1)},$$

as desired. □

REFERENCES

[1] P. Dankelmann and R. C. Entringer, *Average distance, minimum degree and spanning trees*, J. Graph Theory, 33 (2000), pp. 1–13.

[2] P. Dankelmann, S. Mukwembi, and H. C. Swart, *Average distance and edge-connectivity* I, SIAM J. Discrete Math., to appear.

[3] J. K. Doyle and J. E. Graver, *Mean distance in a graph*, Discrete Math., 7 (1977), pp. 147–154.

[4] R. C. Entringer, D. E. Jackson, and D. A. Snyder, *Distance in graphs*, Czechoslovak Math. J., 26 (1976), pp. 283–296.

[5] L. Lovász, *Combinatorial Problems and Exercises*, Akadémiai Kiadó, Budapest, 1979.

[6] S. Mukwembi, *Bounds on Distances in Graphs*, Ph.D. thesis, University of KwaZulu-Natal, 2006.

[7] J. Plesník, *On the sum of all distances in a graph or digraph*, J. Graph Theory, 8 (1984), pp. 1–24.

SIAM J. DISCRETE MATH.
Vol. 21, No. 4, pp. 1053–1060

EXPLICIT CONSTRUCTION OF SMALL FOLKMAN GRAPHS*

LINYUAN LU†

Abstract. A Folkman graph is a K_4-free graph G such that if the edges of G are 2-colored, then there exists a monochromatic triangle. Erdős offered a prize for proving the existence of a Folkman graph with at most 1 million vertices. In this paper, we construct several "small" Folkman graphs within this limit. In particular, there exists a Folkman graph on 9697 vertices.

Key words. Folkman graph, spectrum, K_4-free, monochromatic triangle, circulant graph

AMS subject classifications. 05C55, 05C35, 05D10

DOI. 10.1137/070686743

1. Introduction.

For two graphs G and H, the Rado arrow notation $G \to (H)_p$ is the statement that if the edges of G are p-colored, then there exists a monochromatic subgraph of G isomorphic to H. In 1967 Erdős and Hajnal [2] (also see [3]) conjectured that for each p there exists a graph G, containing no K_4, which has the property that $G \to (K_3)_p$. This conjecture was proved by Folkman [4] for $p = 2$. A Folkman graph is a K_4-free graph G with $G \to (K_3)_2$. Nešetřil and Rödl [9] proved the conjecture for general p. In particular, for any $k_1 < k_2$ and any $p \geq 2$, one could ask what is the smallest integer n such that there is a K_{k_2}-free graph G on n vertices satisfying

$$G \to (K_{k_1})_p.$$

Let $f(p, k_1, k_2)$ denote this smallest integer n. Graham [6] proved that $f(2, 3, 6) = 8$ by showing

$$K_8 \setminus C_5 \to (K_3)_2.$$

Irving [7] proved that $f(2, 3, 5) \leq 18$, and it was further improved by Khadzhiivanov and Nenov [8] to $f(2, 3, 5) \leq 16$. Finally, Piwakowski, Radziszowski, and Urbański [13] and Nenov [12] proved $f(2, 3, 5) = 15$. However, both upper bounds of Folkman and of Nešetřil and Rödl for $f(2, 3, 4)$ are extremely large. Frankl and Rödl [5] first gave a reasonable bound

$$f(2, 3, 4) \leq 7 \times 10^{11}.$$

Erdős set a prize of \$100 for the challenge $f(2, 3, 4) \leq 10^{10}$. This reward was claimed by Spencer [10, 11], who proved that

$$f(2, 3, 4) < 3 \times 10^9.$$

Erdős then offered another \$100 prize (see [1, page 46]) for the new challenge

$$f(2, 3, 4) < 10^6.$$

*Received by the editors March 29, 2007; accepted for publication (in revised form) August 20, 2007; published electronically January 22, 2008.
http://www.siam.org/journals/sidma/21-4/68674.html

†Department of Mathematics, University of South Carolina, Columbia, SC 29208 (lu@math.sc.edu). This author was supported in part by NSF grant DMS 0701111.

Here we claim the reward.

THEOREM 1.

$$f(2,3,4) \leq 9697.$$

In fact, we construct several "small" Folkman graphs. This paper is organized as follows. In section 2, we use spectral analysis to establish a sufficient condition for $G \to (K_3)_2$. This allows us to test a set of graphs efficiently. In section 3, we examine a special class of graphs and find four "small" Folkman graphs.

2. Spectral analysis.

2.1. Localization. Our starting point is the following lemma from Spencer [10]. We will use the following notation.

For any graph H and a vertex-set partition $V(H) = X \cup Y$, let $e(X,Y)$ be the number of edges in H with one end in X and the other end in Y. Let $b(H)$ be the maximum of $e(X,Y)$ among all partition $V(H) = X \cup Y$.

Consider a random partition $V(H) = X \cup Y$ by putting each vertex independently into X or Y with equal probability. The expected number of $e(X,Y)$ is exactly $\frac{1}{2}|E(H)|$. Thus we have

$$b(H) \geq \frac{1}{2}|E(H)|.$$

DEFINITION 1. *For $0 < \delta < \frac{1}{2}$, a graph H is said to be* δ-fair *if $b(H) < (\frac{1}{2} + \delta)|E(H)|$.*

Supposing $G \not\to (K_3)_2$, we see that the edges of G can be colored in red and blue with no monochromatic triangle. For each triangle, there are two possible colorings (two red edges and a blue edge or vice versa). Each triangle has two vertices incident with a red edge and a blue edge. Thus

$$|\{xyz\colon xy \text{ is a red edge}, xz \text{ is a blue edge, and } yz \text{ is an edge}\}| = 2|\{\text{all triangles}\}|.$$

For any vertex $v \in V(G)$, let $\Gamma(v)$ be the set of neighbors of v in G. Let G_v be the induced subgraph on $\Gamma(v)$. The left-hand side of the above equation is at most $\sum_v b(G_v)$ while the right-hand side is exactly $\frac{2}{3}\sum_v |E(G_v)|$. This observation leads to the following lemma.

LEMMA 1 (see Spencer [10]). *If $\sum_v b(G_v) < \frac{2}{3}\sum_v |E(G_v)|$, then $G \to (K_3)_2$.*

COROLLARY 1. *Suppose for each vertex v the local graph G_v is $\frac{1}{6}$-fair. Then*

$$G \to (K_3)_2.$$

If in addition G_v is triangle-free for each v, then G is a Folkman graph.

2.2. δ-fair graphs. Suppose H is a graph on vertices $v_1, v_2, \ldots, v_n$. Let $A = (a_{ij})$ be the adjacency matrix of H so that

$$a_{ij} = \begin{cases} 1 & v_iv_j \text{ is an edge of } H; \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathbf{1}$ denote the n-dimensional column vector with all entries 1. Let $\mathbf{d} = (d_1, d_2, \ldots, d_n)'$ be the column vector of degrees. Here d_i is the degree of vertex v_i. By definition, we have

$$\mathbf{d} = A \cdot \mathbf{1}. \tag{1}$$

For any set $S \subset V(H)$, the volume of S is defined as

$$\mathrm{Vol}(S) = \sum_{v \in S} d_v.$$

We write $\mathrm{Vol}(H) = \mathrm{Vol}(V(H)) = \sum_v d_v = 2|E(H)|$. Let $\bar{d} = \frac{\mathrm{Vol}(H)}{n}$ be the average degree of H.

LEMMA 2. *If the smallest eigenvalue of* $M = A - \frac{1}{\mathrm{Vol}(H)} \mathbf{d} \cdot \mathbf{d}'$ *is greater than* $-2\delta\bar{d}$, *then* H *is* δ*-fair.*

Proof. For any partition of the vertex set $V(H) = X \cup Y$, let $\mathbf{1}_X$ be the n-dimensional column vector whose entries are 1 if the index is in X and 0 otherwise. The vector $\mathbf{1}_Y$ is defined similarly. By definition, we have

$$\mathbf{1}_X + \mathbf{1}_Y = \mathbf{1}. \tag{2}$$

From (1), we have

$$\begin{aligned} M \cdot \mathbf{1} &= \left(A - \frac{1}{\mathrm{Vol}(H)} \mathbf{d} \cdot \mathbf{d}' \right) \cdot \mathbf{1} \\ &= A \cdot \mathbf{1} - \frac{1}{\mathrm{Vol}(H)} \mathbf{d} \cdot \mathbf{d}' \cdot \mathbf{1} \\ &= \mathbf{d} - \frac{1}{\mathrm{Vol}(H)} \mathbf{d} \mathrm{Vol}(H) \\ &= 0. \end{aligned}$$

Thus, 0 is always an eigenvalue of M and $\mathbf{1}$ is the corresponding eigenvector.

Let $\alpha(t) = (1-t)\mathbf{1}_X - t\mathbf{1}_Y$. For any t, we claim

$$\alpha(t)' \cdot M \cdot \alpha(t) = -e(X,Y) + \frac{1}{\mathrm{Vol}(H)} \mathrm{Vol}(X)\mathrm{Vol}(Y).$$

From (2), we can rewrite

$$\alpha(t) = \mathbf{1}_X - t\mathbf{1} = -\mathbf{1}_Y + (1-t)\mathbf{1}.$$

We have

$$\begin{aligned} \alpha(t)' \cdot M \cdot \alpha(t) &= (\mathbf{1}_X - t\mathbf{1})' \cdot M \cdot (-\mathbf{1}_Y + (1-t)\mathbf{1}) \\ &= -\mathbf{1}_X' \cdot M \cdot \mathbf{1}_Y \\ &= -\mathbf{1}_X' \cdot A \cdot \mathbf{1}_Y + \frac{1}{\mathrm{Vol}(H)} \mathbf{1}_X' \cdot \mathbf{d} \cdot \mathbf{d}' \cdot \mathbf{1}_Y \\ &= -e(X,Y) + \frac{\mathrm{Vol}(X)\mathrm{Vol}(Y)}{\mathrm{Vol}(H)}. \end{aligned}$$

Here we use the fact that $M \cdot \mathbf{1} = 0$.

Let ρ be the largest eigenvalue of $-M$. By assumption, $\rho < 2\delta\bar{d}$. We have

$$\begin{aligned} e(X,Y) - \frac{1}{\mathrm{Vol}(H)} \mathrm{Vol}(X)\mathrm{Vol}(Y) &= \alpha(t)' \cdot (-M) \cdot \alpha(t) \\ &\leq \rho \|\alpha(t)\|^2. \end{aligned}$$

Choose $t = \frac{|X|}{n}$ so that $\|\alpha(t)\|^2$ reaches its minimum $\frac{|X||Y|}{n}$. We have

$$e(X,Y) - \frac{\mathrm{Vol}(X)\mathrm{Vol}(Y)}{\mathrm{Vol}(H)} \leq \rho \frac{|X||Y|}{n}.$$

Apply the Cauchy–Schwarz inequalities to $\mathrm{Vol}(X)\mathrm{Vol}(Y)$ and to $|X||Y|$. We have

$$\begin{aligned} e(X,Y) &\leq \frac{\mathrm{Vol}(X)\mathrm{Vol}(Y)}{\mathrm{Vol}(H)} + \rho \frac{|X||Y|}{n}. \\ &\leq \frac{(\mathrm{Vol}(X)+\mathrm{Vol}(Y))^2}{4\mathrm{Vol}(H)} + \rho \frac{(|X|+|Y|)^2}{4n} \\ &= \frac{\mathrm{Vol}(H)}{4} + \rho \frac{n}{4} \\ &< \frac{\mathrm{Vol}(H)}{4} + 2\delta \bar{d} \frac{n}{4} \\ &= (1+2\delta)\frac{\mathrm{Vol}(H)}{4} \\ &= (\frac{1}{2}+\delta)|E(H)|. \end{aligned}$$

Since this holds for any partition $X \cup Y$, we have

$$b(H) \leq \left(\frac{1}{2}+\delta\right)|E(H)|.$$

H is δ-fair as claimed. □

COROLLARY 2. *Suppose H is a d-regular graph and that the smallest eigenvalue of its adjacency matrix A is greater than $-2\delta d$. Then H is δ-fair.*

Proof. Since H is d-regular, we have $\mathbf{d} = d\mathbf{1}$ and $\mathrm{Vol}(H) = nd$. Thus,

$$M = A - \frac{d}{n}\mathbf{1} \cdot \mathbf{1}'.$$

Note that $\mathbf{1}$ is the eigenvector of A with respect to the eigenvalue d. Suppose α is another eigenvector of A with respect to an eigenvalue λ ($\lambda \neq d$). The eigenvector α is orthogonal to $\mathbf{1}$. We have $M\alpha = A\alpha = \lambda\alpha$. Suppose A has eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = d$. Then M has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{n-1}$, and 0. In particular, the smallest eigenvalue of M equals the smallest eigenvalue of A. The conclusion follows from Lemma 2. □

Remark. The largest Laplacian eigenvalue of graph H can also be used to derive the δ-fairness of H. However, in practice, it is not as effective as the matrix M.

2.3. The spectrum of circulant graphs. Let $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$ be the cyclic group of order n. A circulant graph H generated by a subset $S \subset \mathbb{Z}_n$ is a graph with the vertex set $V(H) = \mathbb{Z}_n$ and the edge set $E(H) = \{xy \mid x - y \in S\}$. Here $S \subset \mathbb{Z}_n$ is a subset satisfying that

- if $s \in S$, then $-s \in S$;
- $0 \notin S$.

The following lemma determines the spectrum of circulant graphs.

LEMMA 3. *The eigenvalues of the adjacency matrix for the circulant graph generated by $S \subset \mathbb{Z}_n$ are*

$$\sum_{s \in S} \cos \frac{2\pi i s}{n}$$

for $i = 0, \dots, n-1$.

Proof. Let $J = (J_{ij})$ be the adjacency matrix of the directed cycle on n vertices. Namely, $J_{ij} = 1$ if $j - i \equiv 1 \mod n$, and 0 otherwise. The adjacency matrix of the circulant graph generated by $(\mathbb{Z}_n, S)$ can be expressed as

$$A = \sum_{s\in S} J^s.$$

We identify elements $\mathbb{Z}_n$ with $0, 1, 2, \dots, n-1$ and define a polynomial $f(x) = \sum_{s\in S} x^s$. Note that $A = f(J)$. The eigenvalues of A are completely determined by the eigenvalues of J and the polynomial $f(x)$.

Let $\rho = e^{\frac{2\pi i}{n}}$ denote the primitive nth unit root. We observe that J has eigenvalues

$$1, \rho, \rho^2, \dots, \rho^{n-1}.$$

Thus, the eigenvalues of A are

$$f(1), f(\rho), \dots, f(\rho^{n-1}).$$

Since A is symmetric, the above eigenvalues are all real. For $i = 0, 1, 2, \dots, n-1$, we have

$$f(\rho^i) = \Re(f(\rho^i)) = \sum_{s\in S} \cos\frac{2\pi i s}{n}. \qquad \square$$

3. Graph $\boldsymbol{L(m,s)}$. The previous section allows us to test a special class of graphs efficiently.

Suppose m is an odd positive integer and $s < m$ is another positive integer relatively prime to m. Let $\phi(m)$ be the totient function of m, which is the number of positive integers not exceeding m and relatively prime to m. By Euler's theorem, we have $s^{\phi(m)} \equiv 1 \mod m$. Let n be the smallest positive integer satisfying $s^n \equiv 1 \mod m$. In particular, n is a factor of $\phi(m)$. Define a subset $S = S(s) \subset \mathbb{Z}_m$ as

$$S = \{s^i \mod m \mid i = 0, 1, 2, \dots, n-1\}.$$

We observe that

- if $-1 \in S$, then for any $t \in S$, $-t \in S$;
- with inherited multiplication from $\mathbb{Z}_m$, S forms an abelian group isomorphic to $\mathbb{Z}_n$.

DEFINITION 2. *We define* **graph $\mathbf{L(m,s)}$** *to be the circulant graph on m vertices generated by $S = S(s)$ provided $-1 \in S$.*

The graph $G = L(m,s)$ is a vertex-transitive graph on m vertices. All local graphs G_v are isomorphic to each other. The following lemma shows that G_v is also a circulant graph under isomorphism.

LEMMA 4. *The unique local graph of $L(m,s)$ is isomorphic to a circulant graph of order n.*

Proof. The local graph H of $L(m,s)$ can be described as follows.

1. $V(H) = S$.
2. $E(H) = \{xy \mid x \in S,\ \ y \in S, \text{ and } x - y \in S\}$.

We define a bijection $f : \mathbb{Z}_n \to S$ which maps i to $s^i \mod m$. This is a well-defined map since $s^n \equiv 1 \mod m$. The map f is a group isomorphism from $\mathbb{Z}_n$ to S:

$$f(i + j) = f(i)f(j).$$

We define $T \subset \mathbb{Z}_n$ as

$$T = \{i \mid f(i) - 1 \in S\}.$$

Let H' be the circulant graph generated by $(\mathbb{Z}_n, T)$. If suffices to show f is a graph homomorphism mapping H' to H.

On the one hand, for any edge $jk \in E(H')$, we have $j - k \in T$. Thus,

$$f(j-k) - 1 \in S.$$

Since $f(j) - f(k) = f(k)(f(j-k)-1)$ and S is a group, we conclude that $f(j) - f(k) \in S$. Equivalently, $f(j)f(k)$ is an edge of H.

On the other hand, for any edge $f(j)f(k) \in E(H)$, we have $f(j) - f(k) \in S$. Note that $f(-k)$ is the inverse of $f(k)$ in S. We conclude that

$$f(j-k) - 1 = f(-k)(f(j) - f(k)) \in S.$$

Thus, $j - k \in T$ and jk is an edge of H'. □

3.1. Results from computation. For a fixed pair (m, s), let H be the local graph of $L(m, s)$ and A the adjacency matrix of H. Let $\sigma = \sigma(m, s)$ be the ratio of the smallest eigenvalue and the largest eigenvalue of A. If $\sigma > -\frac{1}{3}$, then H is $\frac{1}{6}$-fair from Corollary 2. Thus, from Corollary 1, $L(m, s) \to (K_3)_2$. Table 1 (except for the last row) shows graphs $L(m, s)$ satisfying that

1. $L(m, s)$ is K_4-free;
2. $\sigma = \sigma(m, s)$ is maximized in the sense that $\sigma(m, s) > \sigma(m', s')$, for all pairs (m', s') in the table and $m' < m$.

We note that $\sigma > -\frac{1}{3}$ in the last four rows of Table 1. Thus, $L(9697, 4)$, $L(30193, 53)$, $L(33121, 2)$, and $L(57401, 7)$ are Folkman graphs.

Table 1
A set of candidates for Folkman graphs.

$L(m,s)$	σ
$L(17,2)$	$-0.8047\cdots$
$L(61,8)$	$-0.7826\cdots$
$L(79,12)$	$-0.7625\cdots$
$L(127,5)$	$-0.6363\cdots$
$L(421,7)$	$-0.6253\cdots$
$L(457,6)$	-0.6
$L(631,24)$	$-0.5749\cdots$
$L(761,3)$	$-0.5613\cdots$
$L(785,53)$	$-0.5404\cdots$
$L(941,12)$	$-0.5376\cdots$
$L(1777,53)$	$-0.5216\cdots$
$L(1801,125)$	$-0.4912\cdots$
$L(2641,2)$	$-0.4275\cdots$
$L(9697,4)$	$-0.3307\cdots$
$L(30193,53)$	$-0.3094\cdots$
$L(33121,2)$	$-0.2665\cdots$
$L(57401,7)$	$-0.3289\cdots$

Proof of Theorem 1. It suffices to show that $G = L(9697, 4)$ is a Folkman graph. The local graph of G is a circulant graph H generated by $T \subset \mathbb{Z}_n$. Here $n = 1212$

and

$$\begin{aligned} T = \{&3, 9, 46, 57, 62, 70, 81, 91, 98, 115, 141, 166, 202, 204, 233, 271,\\ &286, 301, 325, 342, 372, 376, 383, 396, 397, 403, 411, 428, 430, 436,\\ &448, 450, 456, 471, 472, 479, 489, 516, 522, 532, 556, 564, 566, 588,\\ &593, 595, 617, 619, 624, 646, 648, 656, 680, 690, 696, 723, 733, 740,\\ &741, 756, 762, 764, 776, 782, 784, 801, 809, 815, 816, 829, 836, 840,\\ &870, 887, 911, 926, 941, 979, 1008, 1010, 1046, 1071, 1097, 1114,\\ &1121, 1131, 1142, 1150, 1155, 1166, 1203, 1209\}.\end{aligned}$$

An easy calculation (by Maple) shows that H has the following properties:

1. H is a 92-regular and triangle-free graph.
2. The smallest eigenvalue of the adjacency matrix of H is

$$\sum_{t\in T} \cos \frac{2\pi \cdot 502t}{1212} \approx -30.43170597\ldots.$$

Since $30.43170597\ldots < \frac{92}{3}$, H is $\frac{1}{6}$-fair. Thus, $L(9697, 4)$ is a Folkman graph on 9697 vertices. □

Remark 1. We say G is a *strong* Folkman graph if G is K_4-free and $G \to (K_4-e)_2$. Here $K_4 - e$ is the graph obtained by removing one edge from K_4. We can show that both $L(30193, 53)$ and $L(33121, 2)$ are strong Folkman graphs.

Remark 2. Graphs with relatively large σ (as shown in Table 1) are good candidates for Folkman graphs. Recently Exoo showed that $L(17, 2)$, $L(61, 8)$, $L(79, 12)$, $L(421, 7)$, and $L(631, 24)$ are not Folkman graphs. Little is known for other graphs. For example, is $L(2641, 2)$ a Folkman graph?

Remark 3. Exoo (see [14]) conjectured that $L(127, 5)$ is a Folkman graph. The set $S \subset \mathbb{Z}_{127}$ generated by 5 is precisely all nonzero cubes in $\mathbb{Z}_{127}$. Exoo did extensive computation on this graph. If his conjecture is true, then it implies $f(2, 3, 4) \leq 127$.

Remark 4. Recently, Dudek and Rödl independently proved $f(2, 3, 4) < 130000$.

REFERENCES

[1] F. Chung and R. Graham, *Erdős on Graphs. His Legacy of Unsolved Problems*, A. K. Peters, Ltd., Wellesly, MA, 1998.

[2] P. Erdős and A. Hajnal, *Research problem* 2.5, J. Combinatorial Theory, 2 (1967), p. 105.

[3] P. Erdős, *Problems and results on finite and infinite graphs*, in Graph Theory, Proc. 2nd Czechoslovak Sympos. (Prague, 1974), Academia, Prague, 1975, pp. 183–192 (loose errata).

[4] J. Folkman, *Graphs with monochromatic complete subgraphs in every edge coloring*, SIAM J. Appl. Math., 18 (1970), pp. 19–24.

[5] P. Frankl and Rödl, *Large triangle-free subgraphs in graphs without* K_4, Graphs Combin., 2 (1986), pp. 135–144.

[6] R. L. Graham, *On edgewise* 2*-colored graphs with monochromatic triangles and containing no complete hexagon*, J. Combinatorial Theory, 4 (1968), p. 300.

[7] R. W. Irving, *On a bound of Graham and Spencer for a graph coloring constant*, J. Combin. Theory Ser. B, (1973), pp. 200–203.

[8] N. G. Khadzhiivanov and N. D. Nenov, *An example of a* 16*-vertex Ramsey* (3, 3)*-graph with clique number* 4, Serdica, 9 (1983), pp. 74–78 (in Russian).

[9] J. Nešetřil and V. Rödl, *The Ramsey property for graphs with forbidden complete subgraphs*, J. Combin. Theory Ser. B, 20 (1976), pp. 243–249.

[10] J. H. Spencer, *Three hundred million points suffice*, J. Combin. Theory Ser. A, 49 (1988), pp. 210–217.

[11] J. H. Spencer, *Erratum to three hundred million points suffice*, J. Combin. Theory Ser. A, 50 (1989), p. 323.

[12] N. Nenov, *An example of a* 15*-vertex* $(3,3)$*-Ramsey graph with clique number* 4, C. R. Acad. Bulgare Sci., 34 (1981), pp. 1487–1489 (in Russian).

[13] K. Piwakowski, S. P. Radziszowski, and S. Urbański, *Computation of the Folkman number* $F_e(3,3;5)$, J. Graph Theory, 32 (1999), pp. 41–49.

[14] S. P. Radziszowski and X. Xu, *On the most wanted Folkman graph*, Geombinatorics, 16 (2007), pp. 367–381.

SIAM J. Discrete Math.
Vol. 21, No. 4, pp. 1061–1070

NONSEPARATING INDUCED CYCLES CONSISTING OF CONTRACTIBLE EDGES IN k-CONNECTED GRAPHS*

YOSHIMI EGAWA[†], KATSUMI INOUE[†], AND KEN-ICHI KAWARABAYASHI[‡]

Abstract. Egawa and Saito proved that every k-connected graph with girth at least 4 has an induced cycle C such that $G - V(C)$ is $(k-3)$-connected, and every edge of C is contractible. This means that we can find not only a nonseparating cycle C but also one that consists of contractible edges. Motivated by this result, we prove that if G is a k-connected graph which does not contain K_4^-, then G has an induced cycle C such that $G - V(C)$ is $(k-2)$-connected and either every edge of C is k-contractible or C is a triangle. As a corollary of this result, we get the following result: Every k-connected graph with girth at least 4 has an induced cycle C such that $G - V(C)$ is $(k-2)$-connected, and every edge of C is contractible. This theorem is a generalization of some known theorems. In particular, this generalizes the above-mentioned result proved by Egawa and Saito and the result of Egawa which says that a k-connected graph with girth at least 4 has an induced cycle C such that $G - V(C)$ is $(k-2)$-connected.

Key words. nonseparating cycle, contractible edge, k-connected graphs

AMS subject classification. 05C40

DOI. 10.1137/060665956

1. Introduction.

Let $k \geq 2$ be an integer. An edge e of a k-connected graph is said to be *k-contractible* if the graph obtained from G by contracting e (and replacing each of the resulting pairs of double edges by a single edge) is still k-connected.

The study of contractible edges and their applications to noseparating cycles has received much attention by many researchers; cf. [1, 2, 3, 5, 9, 10].

There are some theorems concerning nonseparating induced cycles in k-connected graphs. In [9], Thomassen proved the following "fundamental" theorem.

Theorem 1. *Let G be a k-connected graph. Then G has an induced cycle C such that $G - V(C)$ is $(k-3)$-connected.*

Later, Egawa [2, 3] considered the cases of girth 4 and girth 5 in Theorem 1 and proved the following theorems.

Theorem 2. *Let G be a k-connected graph with girth at least 4. Then G has an induced cycle C such that $G - V(C)$ is $(k-2)$-connected.*

Theorem 3. *Let G be a k-connected graph with girth at least 5. Then G has an induced cycle C such that $G - V(C)$ is $(k-1)$-connected.*

Let K_4^- be the graph obtained from the complete graph of order 4 by deleting an edge. Recently, Kawarabayashi [6] proved the following theorem, which is stronger than Theorem 2.

Theorem 4. *Let G be a k-connected graph which does not contain a K_4^-. Then G has an induced cycle C such that $G - V(C)$ is $(k-2)$-connected.*

*Received by the editors July 25, 2006; accepted for publication (in revised form) July 25, 2007; published electronically January 25, 2008.
http://www.siam.org/journals/sidma/21-4/66595.html

[†] Department of Applied Mathematics, Science University of Tokyo, 1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan.

[‡] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan (k_keniti@nii.ac.jp). The research of this author was partly supported by the Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research, and by the Sumitomo Foundation, the Inamori Foundation, and the Kayamori Foundation.

The key to the proofs of the above results is to use a k-contractible edge. In fact, what we need to do is apply induction on the number of vertices. For instance, let us consider a proof of Theorem 1. If there is a triangle T in G, then we just delete this triangle T. This decreases the connectivity to at most 3, so we are done. So we may assume that G has no triangles. Then Thomassen [9] proved that there is a k-contractible edge e in G. We contract this edge e and apply induction (on the number of vertices). This is possible since the resulting graph is still k-connected. This is, roughly, how the proof in [9] goes. As we see here, the existence of k-contractible edges plays an important role in this proof. This is also the case in [2, 3, 6].

But if we look at k-connected graphs with girth at least 4, then G must contain a k-contractible edge by the result of Thomassen [9]. In fact, there are many k-contractible edges in these graphs; see [4]. So one natural question is whether we can find a k-contractible edge in some certain configuration. In particular, since we know that such graphs have nonseparating cycles by Theorems 2, 3, and 4, thus one natural question is whether we can find a nonseparating cycle containing a k-contractible edge.

Motivated by this question, several results have appeared in the literature. In [1], Dean proved the following theorem.

THEOREM 5. *Let G be a k-connected graph with girth at least* 4*. Then G has an induced cycle C such that each edge of C is a k-contractible edge and $G - V(C)$ is connected.*

As a generalization of Theorem 5, Egawa and Saito [5] proved the following theorem.

THEOREM 6. *Let G be a k-connected graph with girth at least* 4*. Then G has an induced cycle C such that each edge of C is a k-contractible edge and $G - V(C)$ is $(k-3)$-connected.*

In this paper, we prove the following theorem, which is a common refinement of Theorems 4 and 6.

THEOREM 7. *Let G be a k-connected graph which does not contain a K_4^-. Then G has an induced cycle C such that $G - V(C)$ is $(k-2)$-connected and either C is a triangle or each edge of C is a k-contractible edge.*

It is easy to see that the following corollary immediately follows from Theorem 7.

COROLLARY 8. *Let G be a k-connected graph with girth at least* 4*. Then G has an induced cycle C such that each edge of C is a k-contractible edge and $G - V(C)$ is $(k-2)$-connected.*

In this paper, all graphs considered are finite, undirected, and without loops or multiple edges. For a graph G, $V(G)$, $E(G)$, and $\delta(G)$ denote the set of vertices, the set of edges, and the minimum degree of G, respectively.

For a vertex $x \in V(G)$, let $N(x) = N_G(x)$ denote the neighborhood of x in G, and let $d_G(x) = |N_G(x)|$. For a subset S of $V(G)$, $N(S) = N_G(S)$ denotes the union of $N(x)$ as x ranges over S.

For a subset S of $V(G)$, the subgraph induced by S is denoted by $\langle S \rangle$. For a subgraph H of G and a vertex $v \in V(G)$, we let $N_H(v) = N_G(v) \cap V(H)$.

A cutset consisting of k vertices is called a *k-cutset.*

In our proof of Theorem 7, when there is no confusion we sometimes write $H \cap A$ for a subgraph H and a vertex set A.

2. Proof of Theorem 7.

Throughout this section, we assume that G is k-connected and that G does not contain K_4^-. Then it is easy to see that, for any cycle C of order 3 or 4, $|N_C(v)| \leq 2$ for $v \in V(G) - V(C)$.

We first note that, when $k = 3$, Thomassen and Toft [10] proved that, for any 3-connected graph of order at least 5, G has an induced cycle C such that each edge of C is 3-contractible and $G - V(C)$ is connected. Hence, we may assume $k \geq 4$.

For a technical reason, we prove the following statement, which immediately implies Theorem 7.

(I) Let G be a k-connected graph that does not contain K_4^-. Then G has an induced cycle C such that $G - V(C)$ is $(k-2)$-connected and either each edge of C is k-contractible, or C is a triangle. Moreover, for any vertex $v \in V(G) - V(C)$, $|N_C(v)| \leq 2$.

If there exists a triangle T such that $G - V(T)$ is $(k-2)$-connected, then (I) holds. Thus we may assume that there are no such triangles. This implies that every triangle is contained in a k-cutset, which in particular implies that no k-contractible edge is contained in a triangle. Suppose that there is no induced cycle C, as described in (I). First, we prove (I) for the following case.

Case 1. Every k-contractible edge is contained in a cycle of length 4 (this includes the case where G has no k-contractible edge).

Recall that every triangle is contained in a k-cutset, and no k-contractible edge is contained in a triangle. Since we assume that (I) fails, any C_4 consisting of k-contractible edges only is contained in a $(k+1)$-cutset. Note that, for any C_4 and for any vertex $v \in V(G) - V(C_4)$, $|N_G(v) \cap V(C_4)| \leq 2$, and for any triangle T and for any vertex $v \in V(G) - V(T)$, $|N_G(v) \cap V(T)| \leq 1$.

Let A_1, A_2, and A_3 denote, respectively, the set of k-cutsets containing a non-k-contractible edge which is not contained in any triangle, k-cutsets containing a triangle, and $(k+1)$-cutsets containing a C_4, each edge of which is k-contractible. We claim that the assumption of Case 1 and the observation in the preceding paragraph imply that $A_1 \cup A_2 \cup A_3 \neq \emptyset$. If there is a triangle, then clearly the assertion holds. Otherwise, every edge is not contained in any triangles. If there is an edge that is not k-contractible, then clearly the assertion holds. On the other hand, if every edge is k-contractible, then the assumption of Case 1 implies that there must be a C_4 in which each edge consists of k-contractible edges. Since we assume (I) fails, this C_4 is contained in a $(k+1)$-cutset of G. Hence we may assume that $A_1 \cup A_2 \cup A_3 \neq \emptyset$.

Let $A \in A_1 \cup A_2 \cup A_3$ and let H be a component in $G - A$. Let $W = G - A - H$. First, we prove the following lemmas.

LEMMA 1. *If $A \in A_1 \cup A_2$, then $|H| \geq k-1$. (Thus, $|W| \geq k-1$.) If $A \in A_3$, then $|H| \geq k-2$. (Thus, $|W| \geq k-2$.)*

Proof. It is easy to see that there exists an edge zw in H. Since G does not contain a K_4^-, we have $|N_G(z) \cap N_G(w)| \leq 1$. Hence, $|N_G(z) \cup N_G(w)| \geq 2k-1$, which implies $|H| \geq 2k-1-|A|$. Hence when $A \in A_1 \cup A_2$, $|H| \geq k-1$; when $A \in A_3$, $|H| \geq k-2$. □

Hence, if there is no confusion, we may write $H \cap A$ instead of $V(H) \cap A$. We may also write $W \cap A, W \cap A'$, etc., if there is no confusion. These may be applied to proofs of Lemmas 3, 8, 14, etc.

LEMMA 2. *Let $A \in A_1 \cup A_2$, $A' \in A_1 \cup A_2 \cup A_3$, and let H be a component in $G - A$. Then $H \not\subseteq A'$.*

Proof. Suppose $H \subseteq A'$. Let $W = G - A - H$. Let H' be a component in $G - A'$, and let W' denote $G - A' - H'$. By Lemma 1, $|H|, |W| \geq k-1$ and $|H'|, |W'| \geq k-2$. Let H_1, H_2, and H_3 denote $H \cap H'$, $H \cap A'$, and $H \cap W'$, respectively. Actually, by our assumption, $H \subseteq A'$, $H_1 = H_3 = \emptyset$. Also, let W_1, W_2, and W_3 denote $W \cap H'$, $W \cap A'$, and $W \cap W'$, respectively. Let Q_1, Q_2, and Q_3 denote $A \cap H'$, $A \cap A'$, and $A \cap W'$, respectively. Suppose $A' \in A_3$. Since $|H| \geq k-1$, $|H_2| \geq k-1$. Hence $|W_2 \cup Q_2| \leq 2$. Since $|W| \geq k-1$ and $|W_2| \leq 2$, $W_1 \neq \emptyset$ or $W_3 \neq \emptyset$. By choosing a different component of $G - A'$, if necessary, we may assume that W_1 and W_3 are symmetric. So, without loss of generality, we may assume $W_3 \neq \emptyset$. Then $W_2 \cup Q_2 \cup Q_3$ is a cutset, and hence $|W_2|+|Q_2|+|Q_3| \geq k$. Since $|W_2|+|Q_2| \leq 2$, this implies $|Q_3| \geq k-2$. Now if $W_1 \neq \emptyset$, we similarly obtain $|Q_1| \geq k-2$; if $W_1 = \emptyset$, we have $|Q_1| = |H'| \geq k-2$. Thus $|Q_1| \geq k-2$. Consequently, $2(k-2) \leq |Q_1| + |Q_3| = |A| - |Q_2| = k - |Q_2|$. Since $k \geq 4$, this forces $k = 4$, $|Q_2| = 0$, and $|Q_1| = |Q_3| = 2$, which implies $|W_2| = 2$, $|H_2| = 3$, and $A' = W_2 \cup H_2$. Since there is no edge between W_2 and H_2, this contradicts the fact that A' contains a cycle of length 4. This completes the proof for the case where $A' \in A_3$, and the case where $A' \in A_1 \cup A_2$ can be settled in a similar way. □

LEMMA 3. *Let $A \in A_1 \cup A_2 \cup A_3$, $A' \in A_1 \cup A_2$, and let H be a component in $G - A$. Then $H \not\subseteq A'$.*

Proof. Suppose $H \subseteq A'$. Let $W = G - A - H$. Let H' be a component in $G - A'$, and also let W' denote $G - A' - H'$. By Lemma 2, $|H|, |W| \geq k-2$ and $|H'|, |W'| \geq k-1$. Let H_1, H_2, and H_3 denote $H \cap H'$, $H \cap A'$, and $H \cap W'$, respectively. Actually, by our assumption, $H \subseteq A'$, $H_1 = H_3 = \emptyset$. Also, let W_1, W_2, and W_3 denote $W \cap H'$, $W \cap A'$, and $W \cap W'$, respectively. Let Q_1, Q_2, and Q_3 denote $A \cap H'$, $A \cap A'$, and $A \cap W'$, respectively. In view of Lemma 2, we may assume $A \in A_3$. Since $|A| = k+1$ and $|A'| = k$, $2k+1 = |A| + |A'| = \sum_{i=1}^{3} |Q_i| + |W_2| + |Q_2| + |H_2|$. Since $|H| \geq k-2$, $|H_2| \geq k-2$. Hence $|W_2 \cup Q_2| \leq 2$.

Applying Lemma 2 with the roles of A and A' interchanged, we see from $H_1 = H_3 = \emptyset$ that $W_1 \neq \emptyset$ and $W_3 \neq \emptyset$, and hence both $W_2 \cup Q_2 \cup Q_3$ and $W_2 \cup Q_1 \cup Q_2$ are cutsets. Since A contains a k-contractible edge, this implies $|W_2 \cup Q_2 \cup Q_3| + |W_2 \cup Q_2 \cup Q_1| \geq 2k+1$. Also, since $|H_2| \geq k-2$, $2k+1 = |A| + |A'| = \sum_{i=1}^{3} |Q_i| + |W_2| + |Q_2| + |H_2| = |W_2 \cup Q_2 \cup Q_3| + |W_2 \cup Q_2 \cup Q_1| + |H_2| - |W_2| \geq 2k+1+k-2-|W_2| \geq 3k-3$. This holds only if $k = 4$ and $|W_2| = 2$. This implies $|H_2| = 2$ and $|Q_2| = 0$, which in turn implies that Q_1 or Q_3 contains a cycle of length 4. Hence, either $|Q_1| \geq 4$ or $|Q_3| \geq 4$, and thus $|Q_1| \leq 1$ or $|Q_3| \leq 1$. But then, either $|W_2 \cup Q_2 \cup Q_3| \leq 3$ or $|W_2 \cup Q_2 \cup Q_1| \leq 3$, which is contrary to the fact that G has connectivity $k \geq 4$. □

We henceforth assume that we have chosen $A \in (A_1 \cup A_2 \cup A_3)$ and a component H in $G - A$ so that $|H|$ is least possible.

Recall that $|H| \geq k-2$, and hence $E(H) \neq \emptyset$. Note also that if an edge e with $V(e) \cap H \neq \emptyset$ is not k-contractible, then there exists a k-cutset A' containing e (so $A' \cap H \neq \emptyset$) such that $A' \in A_2$ or $A' \in A_1$, depending on whether e is contained in a triangle or not.

In the rest of the proof, we let $W = G - A - H$, and we let C' denote the set of these cycles T' of length 4 such that each edge of T' is k-contractible, and such that $|T' \cap H| \geq 2$ (so $T' \subseteq A \cup H$). Further, we use the following notation for $A' \in A_1 \cup A_2 \cup A_3$: Let H' be a component in $G - A'$, and let W' denote $G - A' - H'$. Let H_1, H_2, and H_3 denote $H \cap H'$, $H \cap A'$, and $H \cap W'$, respectively. Also, let W_1, W_2, and W_3 denote $W \cap H'$, $W \cap A'$, and $W \cap W'$, respectively. Let Q_1, Q_2, and Q_3 denote $A \cap H'$, $A \cap A'$, and $A \cap W'$, respectively.

We need the following lemma, which is due to Mader [7, 8].

LEMMA 4. *Let G be a k-connected graph. Let B be a k-cutset containing either an edge not contained in any triangle or a triangle. Let S be a component in $G - B$. We choose B and S so that $|S|$ is least possible. Assume there exists an edge uv with $v \in S$ and $u \in B \cup S$ such that either vu is not contained in any triangle and vu is* *not k-contractible, or uv is contained in a triangle which is contained in a k-cutset. Then $|S| \leq \frac{1}{2}k$.*

We prove the following lemma.

LEMMA 5. *$A \in A_3$.*

Proof. Assume $A \in A_1 \cup A_2$. Since $k-1 > \frac{1}{2}k$ for $k \geq 4$, it follows from Lemmas 1 and 4, and the assumption of Case 1, that the following hold:

(1) No vertex in H is contained in any triangle.

(2) For any edge e such that $V(e) \cap H \neq \emptyset$, e is k-contractible and contained in some C_4.

We first prove that $C' \neq \emptyset$. Assume that $C' = \emptyset$. We prove the following claim.

CLAIM 1. *Let vv_1 be an edge in H. (Note that vv_1 is a k-contractible edge by* (2)*.) Then there exists an edge $v_2v_3 \in E(A)$ such that $v_2 \in N_G(v_1)$ and $v_3 \in N_G(v)$.*

Proof. By (2), vv_1 must be contained in a C_4. Let $vv_1v_2v_3v$ be a C_4. By (2), vv_2 and v_1v_3 are k-contractible edges. If $v_2v_3 \notin E(\langle A \rangle)$, then again by (2) v_2v_3 must be a k-contractible edge, and hence $vv_1v_2v_3v \in C'$. Hence v_2v_3 must be in A. □

Claim 1 also implies that for any $v \in V(H)$, $N_G(v) \cap A \neq \emptyset$. Next, we prove the following claim.

CLAIM 2. *Let xyz be a P_3 in H. Then $N_G(x) \cap N_G(z) = \{y\}$.*

Proof. Note that both xy and yz are k-contractible edges. Assume that the claim is false, and let $w \in N_G(x) \cap N_G(z) - \{y\}$. Then both xw and zw are k-contractible edges. Hence $xyzwx \in C'$, a contradiction. □

Take a vertex $x \in H$, and write $N_G(x) \cap H = \{x_1, \ldots, x_m\}$, where $m = |N_G(x) \cap H|$. Then by Claim 1, $N_G(x_i) \cap A \neq \emptyset$ for all i. Since x is not contained in any triangle, $N_G(x) \cap N_G(x_i) = \emptyset$ for any i. Also, by Claim 2, $N_G(x_i) \cap N_G(x_j) = \{x\}$ for all i, j with $i \neq j$. Since $|A| = k$ and $m + |N_G(x) \cap A| = |N_G(x)| \geq k$, this implies that $m = k - |N_G(x) \cap A|$ and $|N_G(x_i) \cap A| = 1$ for all i, since $N_G(x_i) \cap A \neq \emptyset$ and $N_G(x_i) \cap N_G(x_j) = \{x\}$ for all i, j with $i \neq j$. Since x was taken arbitrarily, we can apply the above argument to each x_i to get $|N_G(x) \cap A| = 1$, and hence $m = k-1$. Write $N_G(x) \cap A = \{c\}$ and $N_G(x_i) \cap A = \{b_i\}$. Then it is easy to see $A = \{c, b_1, \ldots, b_{k-1}\}$. By Claim 1, $cb_i \in E(\langle A \rangle)$ for all i. Again, since x was taken arbitrarily, we can apply the above argument to each x_i to see $b_ib_j \in E(\langle A \rangle)$ for all i, j with $i \neq j$. Therefore, we can conclude that $\langle A \rangle$ is a complete graph, and since $k \geq 4$, $\langle A \rangle$ contains a K_4^-, a contradiction. This proves $C' \neq \emptyset$.

Let $A' \in A_3$ be a $(k+1)$-cutset containing a member T' of C'. By Lemma 2, $H_1 \neq \emptyset$ or $H_3 \neq \emptyset$. Without loss of generality, we may assume $H_1 \neq \emptyset$. Since $|A| = k$ and $|A'| = k+1$, $|A| + |A'| = \sum_{i=1}^{3} |Q_i| + |W_2| + |Q_2| + |H_2| = 2k+1$.

We claim $W_3 = \emptyset$. Assume it does not. Then, by the connectivity of G, $|W_2 \cup Q_2 \cup Q_3| \geq k$. Since $A' \cap H \neq \emptyset$ and $T' \subseteq Q_2 \cup H_2$, by the minimality of H, $|Q_1 \cup Q_2 \cup H_2| \geq k+2$. But then $2k+1 = \sum_{i=1}^{3} |Q_i| + |W_2| + |Q_2| + |H_2| = |W_2 \cup Q_2 \cup Q_3| + |Q_1 \cup Q_2 \cup H_2| \geq k + k + 2 = 2k+2$, a contradiction. Thus, $W_3 = \emptyset$.

On the other hand, from $|Q_1 \cup Q_2 \cup H_2| \geq k+2$ and $|Q_1 \cup Q_2 \cup Q_3| = k$, we obtain $|Q_3| < |H_2|$. Consequently, $|W'| = |Q_3 \cup H_3| < |H_2 \cup H_3| < |H|$, which contradicts the minimality of $|H|$. This proves Lemma 5. □

By arguing as in the last part of the proof of Lemma 5, but now Lemma 3 in place of Lemma 2, we can obtain the following lemma.

LEMMA 6. *Let T' be a triangle or an edge such that $V(T') \cap H \neq \emptyset$. Then there is no $A' \in A_1 \cup A_2$ such that $V(T') \subseteq A'$.*

Sketch of proof. Suppose that there exists such an A'. By Lemma 3, we may assume $H_1 \neq \emptyset$. Suppose that $W_3 \neq \emptyset$. Let T be a cycle of length 4 in A, each edge of which is k-contractible. If $T \subseteq Q_1 \cup Q_2$, then $|Q_1 \cup Q_2 \cup H_2| \geq k+2$ and $|W_2 \cup Q_2 \cup Q_3| \geq k$; if $T \not\subseteq Q_1 \cup Q_2$, then $E(Q_2 \cup Q_3) \cap E(T) \neq \emptyset$, and hence $|Q_1 \cup Q_2 \cup H_2| \geq k+1$ and $|W_2 \cup Q_2 \cup Q_3| \geq k+1$. In either case, $2k+1 = |W_2 \cup Q_2 \cup Q_3| + |Q_1 \cup Q_2 \cup H_2| \geq 2k+2$, a contradiction. Thus $W_3 = \emptyset$. But then $|W'| = |Q_3 \cup H_3| \leq |H_2 \cup H_3| < |H|$, a contradiction. □

By Lemma 6 and the observation made at the beginning of Case 1, we know the following:

(1) No vertex in H is contained in any triangle.

(2) For any edge $e = ab$ such that $\{a, b\} \cap V(H) \neq \emptyset$, e is k-contractible and contained in some C_4.

Take an edge $e = xy$ in H. By (1), $N_G(x) \cap N_G(y) = \emptyset$. Hence, we know that $|H| \geq 2k - k - 1 \geq k - 1$.

Lemma 7. *Let $A' \in A_3$. Then $H \not\subseteq A'$.*

Proof. Suppose $H \subseteq A'$. Since $|A| = |A'| = k+1$, $|A| + |A'| = \sum_{i=1}^{3} |Q_i| + |W_2| + |Q_2| + |H_2| = 2k+2$. Also, by the assumption, $H_1 = H_3 = \emptyset$. Since $|H| \geq k-1$, $|H_2| \geq k-1$. Hence $|W_2 \cup Q_2| \leq 2$. Since $|W| \geq k-1$ and $|W_2| \leq 2$, $W_1 \neq \emptyset$ or $W_3 \neq \emptyset$. Assume $W_1 = \emptyset$. Then $|Q_1| \geq k-1$ and $W_3 \neq \emptyset$. Hence $|Q_2 \cup Q_3| \leq 2$. Recall that $|W_2 \cup Q_2| \leq 2$. Since $W_3 \neq \emptyset$, $W_2 \cup Q_2 \cup Q_3$ is a cutset and its cardinality is at most $|Q_2 \cup Q_3| + |W_2 \cup Q_2| \leq 4$. This holds only if $k = 4$, $|H_2| = |Q_1| = 3$, $Q_2 = \emptyset$, and $|Q_3| = |W_2| = 2$. But since A contains a cycle of length 4 and $Q_2 = \emptyset$, $|Q_1| \geq 4$ or $|Q_3| \geq 4$, a contradiction.

Next, assume $W_1 \neq \emptyset$ and $W_3 \neq \emptyset$. Then, $W_2 \cup Q_1 \cup Q_2$ and $W_2 \cup Q_2 \cup Q_3$ are cutsets, and since A contains a k-contractible edge, $|W_2 \cup Q_1 \cup Q_2| + |W_2 \cup Q_2 \cup Q_3| \geq 2k+1$. Then, since $|H_2| \geq k-1$, $2k+2 = |A| + |A'| = \sum_{i=1}^{3} |Q_i| + |W_2| + |Q_2| + |H_2| \geq |W_2 \cup Q_1 \cup Q_2| + |W_2 \cup Q_2 \cup Q_3| + |H_2| - |W_2| \geq 2k+1+k-1-2 \geq 3k-2$. This forces $k = 4$, $|H_2| = 3$, $|W_2| = 2$, and $Q_2 = \emptyset$, which contradicts the fact that A' contains a cycle of length 4.

Finally, assume $W_3 = \emptyset$. This case can be settled by an argument similar to the proof of the case $W_1 = \emptyset$. □

Recall that C' denotes the set of cycles T' of length 4 such that each edge of T' is k-contractible and $|T' \cap V(H)| \geq 2$. Set $H'' = V(C') \cap V(H)$ and $H_0 = H - H''$, where $V(C')$ denotes the union of the vertex sets of the members of C'.

Next, we consider the structure of H_0 and we prove the following lemmas.

Lemma 8. *H_0 does not contain a P_3.*

Proof. Assume not. First, we claim the following.

Claim 3. *$|N_G(x) \cap A| \geq 2$ for any $x \in V(H_0)$.*

Proof. By (2) and an argument similar to the proof of Claim 1, we can conclude $|N_G(x) \cap A| \geq 1$ for any $x \in V(H_0)$. Suppose that, for some $x \in V(H_0)$, $|N_G(x) \cap A| = 1$. Let $N_G(x) \cap V(H) = \{x_1, \ldots, x_m\}$, where $m = |N_G(x) \cap H|$. Then by an argument similar to the proof of Claim 1, we can conclude $N_G(x_i) \cap A \neq \emptyset$ for all i. Write $N_G(x) \cap A = \{c\}$ and take $b_i \in N_G(x_i) \cap A$. By an argument similar to the proofs of Claims 1 and 2 (applied to each x_i), we can conclude $cb_i \in E(A)$ for all i, and $b_i \neq b_j$ for all i, j with $i \neq j$. Since $|N_G(x) \cap A| = 1$, $|\{c, b_1, \ldots, b_m\}| = |N_G(x)| \geq k$. Let T be a C_4 in A, each edge of which is k-contractible. Since $|A| = k+1$, this implies that $|V(T) \cap \{c, b_1, \ldots, b_m\}| \geq 3$. If $c \notin V(T)$, then $\langle V(T) \cup \{c\} \rangle$ contains a K_4^-, a contradiction. Hence $c \in V(T)$. Assume $b_n \in V(T)$. If $cb_n \notin E(C)$, then $\langle V(T) \rangle$ contains a K_4^-, a contradiction. Recall that C is the cycle contained in A. Hence $cb_n \in E(C)$. But then since cb_n is a k-contractible edge, xx_nb_ncx is a C_4, each edge of which is k-contractible, which is contrary to the fact that $x \in V(H_0)$. □

It suffices to prove that for any $x \in V(H_0)$, $|N_G(x) \cap V(H_0)| \leq 1$. Since x is not in any C_4, by the definition of H_0 we have $N_G(y) \cap N_G(z) \cap A = \emptyset$ for any $y, z \in N_G(x) \cap V(H)$ with $y \neq z$.

Since x is not contained in any triangle, $N_G(x) \cap N_G(y) = \emptyset$ for any $y \in N_G(x) \cap V(H)$. By using the same argument as in the proof of Claim 1, $N_G(z) \cap A \neq \emptyset$ for any $z \in N_G(x) \cap H''$. Also, by Claim 3, $|N_G(y) \cap A| \geq 2$ for any $y \in N_G(x) \cap V(H_0)$. Hence $|A| \geq |N_G(x) \cap A| + |N_G(x) \cap H''| + 2|N_G(x) \cap V(H_0)|$. Since $|A| = k+1$, this implies $|N_G(x) \cap V(H_0)| \leq 1$. □

Lemma 8 also implies that $H'' \neq \emptyset$; otherwise, H does not contain a $P_3 \in H$, which is impossible because $k \geq 4$. Note that $|H| \geq k-1$ by the remark just before Lemma 7. Hence there exists $A' \in A_3$ such that $A' \cap V(H) = \emptyset$ and A' contains a member of C'.

Lemma 9. *If H has an edge not contained in any member of C', then $|H| \geq 2k-2$.*

Proof. Let $e = xy$ be an edge of H not contained in any member of C'. It is easy to see that there exists a vertex $z \in H$ such that $A' \cap V(H) = \emptyset$ and $z \in N_G(x)$ or $z \in N_G(y)$. We may assume $z \in N_G(y)$. By using the same argument as in the proof of Claim 2, $N_G(x) \cap N_G(z) = \{y\}$. Also, since y is not contained in any triangle, $N_G(x) \cap N_G(y) = N_G(y) \cap N_G(z) = \emptyset$. Hence $|H| \geq 3k - 1 - (k+1) = 2k-2$. □

In what follows, we let $A' \in A_3$ denote a $(k+1)$-cutset containing a cycle T' of length 4 such that each edge of T' is k-contractible, and such that $T' \subseteq A \cup H$, $T' \cap H \neq \emptyset$, and $T' \cap W = \emptyset$.

Lemma 10. *Suppose $Q_1 \cup Q_2$ contains a k-contractible edge. Then $Q_2 \cup Q_3$ does not contain a k-contractible edge.*

Proof. Assume that both $Q_1 \cup Q_2$ and $Q_2 \cup Q_3$ contain k-contractible edges. By Lemma 7, $H_1 \neq \emptyset$ or $H_3 \neq \emptyset$. Without loss of generality, we may assume $H_1 \neq \emptyset$. Since $|A| = |A'| = k+1$, $|A| + |A'| = \sum_{i=1}^{3} |Q_i| + |W_2| + |Q_2| + |H_2| = 2k+2$.

We claim $W_3 = \emptyset$. Assume it does not. Then, by the connectivity of G, $|W_2 \cup Q_2 \cup Q_3| \geq k+1$ since $Q_2 \cup Q_3$ contains a k-contractible edge. By the minimality of H, $|Q_1 \cup Q_2 \cup H_2| \geq k+2$. But $2k+2 = \sum_{i=1}^{3} |Q_i| + |W_2| + |Q_2| + |H_2| = |W_2 \cup Q_2 \cup Q_3| + |Q_1 \cup Q_2 \cup H_2| \geq k+1+k+2 = 2k+3$, a contradiction. Thus, $W_3 = \emptyset$. From $|Q_1 \cup Q_2 \cup H_2| \geq k+2$ and $|Q_1 \cup Q_2 \cup Q_3| = k+1$, we also obtain $|Q_3| < |H_2|$. Consequently, $|W'| = |Q_3 \cup H_3| < |H_2 \cup H_3|$, which contradicts the minimality of $|H|$. □

Let T be a cycle of length 4 in A, each edge of which is k-contractible. By Lemma 10, T is contained in either $Q_1 \cup Q_2$ or $Q_2 \cup Q_3$, and $E(T) \cap E(Q_2) = \emptyset$. Without loss of generality, we may assume that T is contained in $Q_1 \cup Q_2$.

Lemma 11. $H_3 = \emptyset$.

Proof. Assume $H_3 \neq \emptyset$. Then by the minimality of H, $|H_2 \cup Q_2 \cup Q_3| \geq k+2$. If $W_1 \neq \emptyset$, then since $Q_1 \cup Q_2 \cup W_2$ is a cutset containing T, $|Q_1 \cup Q_2 \cup W_2| \geq k+1$. But $2k+2 = \sum_{i=1}^{3} |Q_i| + |W_2| + |Q_2| + |H_2| = |H_2 \cup Q_2 \cup Q_3| + |Q_1 \cup Q_2 \cup W_2| \geq k+1+k+2 = 2k+3$, a contradiction. Thus, $W_1 = \emptyset$. Since $|H_2 \cup Q_2 \cup Q_3| \geq k+2$ and $|Q_1 \cup Q_2 \cup Q_3| = k+1$, we have $|H_2| \geq |Q_1| + 1$. But then $|Q_1 \cup H_1| < |H_1 \cup H_2|$, which contradicts the minimality of $|H|$. □

Lemma 12. $|Q_1 \cup Q_2 \cup H_2| = k+2$ *and* $|W_2 \cup Q_2 \cup Q_3| = k$.

Proof. By Lemmas 7 and 11, we have $H_1 \neq \emptyset$. Hence by the minimality of $|H|$, we have $|Q_1 \cup Q_2 \cup H_2| \geq k+2$, and hence $|H_2| > |Q_3|$. Now if $W_3 = \emptyset$, then $|W'| = |Q_3 \cup H_3| < |H_2 \cup H_3| < |H|$, which contradicts the minimality of $|H|$. Thus $W_3 \neq \emptyset$. Since G is k-connected, this implies $|W_2 \cup Q_2 \cup Q_3| \geq k$. But

$2k+2 = \sum_{i=1}^{3} |Q_i| + |W_2| + |Q_2| + |H_2| = |Q_1 \cup Q_2 \cup H_2| + |W_2 \cup Q_2 \cup Q_3|$, and hence equality holds. □

Since $|Q_1 \cup Q_2 \cup Q_3| = k+1$, by Lemma 12, we have $|H_2| = |Q_3| + 1$. Further if $T' \in C'$, then $|H_2| \geq 2$, and hence we have $|Q_3| \geq 1$.

LEMMA 13. *For any $v \in H$, $d_T(v) \leq 1$.*

Proof. Suppose not and let $vv_1, vv_3 \in E(G)$, where $v_1v_2v_3v_4v_1$ is T. Then, since vv_1, vv_3 are k-contractible edges, $vv_1v_2v_3v$ is a C_4, each edge of which is k-contractible, and hence there exists a $(k+1)$-cutset A' containing $vv_1v_2v_3v$. But then $A \cap A'$ contains k-contractible edges, v_1v_2 and v_2v_3, a contradiction to Lemma 10. □

LEMMA 14. $|N(U) \cap H| \geq |U| + 1$ *for all nonempty subsets U of $A - T$.*

Proof. Suppose there exists a nonempty subset U of $A-T$ with $|N(U) \cap H| \leq |U|$. Then, $|U| \leq |A - T| \leq k-2 < k-1 \leq |H|$. Hence, $H - N(U) \neq \emptyset$. Since each edge of T is k-contractible, this implies that $(A-U) \cup (N(U) \cap H)$ is a $(k+1)$-cutset containing T and separating $H - N(U)$ from $W \cup U$. But, since $|H - N(U)| < |H|$, this contradicts the minimality of $|H|$. □

By Lemma 14, since $N(Q_3) \cap H \subseteq H_2$ and $|H_2| = |Q_3| + 1$, we have $N(Q_3) \cap H = H_2$. Also, recall that if $T' \in C'$, then $|H_2| \geq 2$, and hence we have $|Q_3| \geq 1$.

By Lemmas 11–14, we can obtain the following fact.

Let $\{T_1, \dots, T_j\} \subseteq C'$. Then each T_i is contained in some $(k+1)$-cutset $A^i \in A_3$. Let H^i be a component in $G - A^i$, and let W^i denote $G - A^i - H^i$. Let H_1^i, H_2^i, and H_3^i denote $H \cap H^i$, $H \cap A^i$, and $H \cap W^i$, respectively. Also, let W_1^i, W_2^i, and W_3^i denote $W \cap H^i$, $W \cap A^i$, and $W \cap W^i$, respectively. Let Q_1^i, Q_2^i, and Q_3^i denote $A \cap H^i$, $A \cap A^i$, and $A \cap W^i$, respectively. We may assume $T \subseteq Q_1^i \cup Q_2^i$. Then $H_3^i = \emptyset$, $|H_2^i| = |Q_3^i| + 1$, and $N(Q_3^i) \cap V(H) = H_2^i$.

LEMMA 15.

(1) $|\bigcup_{i=1}^{j} N(Q_3^i) \cap V(H)| \leq 2|\bigcup_{i=1}^{j} Q_3^i|$.

(2) *If $(N(Q_3^h) \cap V(H)) \cap (\bigcup_{i=1}^{h-1} N(Q_3^i) \cap V(H)) \neq \emptyset$ for each h with $1 \leq h \leq j$, then $|\bigcup_{i=1}^{j} N(Q_3^i) \cap V(H)| \leq |\bigcup_{i=1}^{j} Q_3^i| + 1$.*

Proof. We prove the lemma by induction on j. Suppose $j = 1$. Then since $N(Q_3^1) \cap V(H) = H_2$ and $|H_2| = |Q_3^1| + 1$, the result follows. Assume $j \geq 2$. Let $R = Q_3^j \cap \bigcup_{i=1}^{j-1} Q_3^i$. We first prove (2). If $R \neq \emptyset$, then by Lemma 14, $|(\bigcup_{i=1}^{j-1} N(Q_3^i)) \cap N(Q_3^j) \cap V(H)| \geq |N(R) \cap V(H)| \geq |R| + 1$. Thus, whether or not $R = \emptyset$, we have $|(\bigcup_{i=1}^{j-1} N(Q_3^i)) \cap N(Q_3^j) \cap V(H)| \geq |R| + 1$ because $(\bigcup_{i=1}^{j-1} N(Q_3^i)) \cap N(Q_3^j) \cap V(H) \neq \emptyset$ by the assumption in (2). Hence by the induction hypothesis, $|\bigcup_{i=1}^{j} N(Q_3^i) \cap V(H)| \leq |\bigcup_{i=1}^{j-1} Q_3^i| + 1 + |Q_3^j| + 1 - |R| - 1 = |\bigcup_{i=1}^{j} Q_3^i| + 1$, as desired.

We now prove (1). If $R \neq \emptyset$, then $|(\bigcup_{i=1}^{j-1} N(Q_3^i)) \cap N(Q_3^j) \cap V(H)| \geq |R| + 1$, and hence $|\bigcup_{i=1}^{j} N(Q_3^i) \cap V(H)| \leq 2|\bigcup_{i=1}^{j-1} Q_3^i| + |Q_3^j| + 1 - |R| - 1 \leq 2|\bigcup_{i=1}^{j-1} Q_3^i| + 2(|Q_3^j - |R|) = 2|\bigcup_{i=1}^{j} Q_3^i|$; if $R = \emptyset$, then $|\bigcup_{i=1}^{j} N(Q_3^i) \cap V(H)| \leq |\bigcup_{i=1}^{j-1} N(Q_3^i) \cap V(H)| + |N(Q_3^j) \cap V(H)| \leq 2|\bigcup_{i=1}^{j-1} Q_3^i| + 2|Q_3^j| = 2|\bigcup_{i=1}^{j} Q_3^i|$. □

LEMMA 16. *If $|H_0| \geq 3$, then $|H''| \geq 2k-5$.*

Proof. Suppose there exists an edge $x_1x_2 \in E(\langle H_0 \rangle)$. Take a vertex $x_3 \in V(H_0) - \{x_1, x_2\}$. By Lemmas 8 and 13, for each i with $1 \leq i \leq 3$, $|N_G(x_i) \cap (V(H'') \cup A - T)| \geq k-2$. It is easy to see $N_G(x_1) \cap N_G(x_2) = \emptyset$. Also, $|N_G(x_i) \cap N_G(x_3)| \leq 1$ for $i = 1, 2$. Otherwise, there exists a C_4 containing x_i, x_3, each edge of which is a k-contractible edge for $i = 1$ or $i = 2$.

Hence $|H''| \geq \sum_{i=1}^{3} |N_G(x_i) \cap (H'' \cup A - T)| - \sum_{i=1}^{2} |N_G(x_i) \cap N_G(x_3)| - |A - T| \geq 3k - 6 - 2 - (k-3) = 2k - 5$.

Next, assume that there is no edge in H_0. Take three vertices $x_1, x_2, x_3 \in V(H_0)$. By the same argument as in the preceding paragraph, we know $|N_G(x_i) \cap N_G(x_j)| \leq 1$ for $i, j = 1, 2, 3$ with $i \neq j$. Hence, $|H''| \geq \sum_{i=1}^{3} d_G(x_i) - 3 - (k+1) \geq 2k - 4$. □

Suppose H has an edge not contained in any member of C'. Then $|H| \geq 2k-2$ by Lemma 9. Let $\{T_1, \ldots, T_j\} = C'$. Then since $T_i \cap H \subseteq H_2^i = N(Q_3^i) \cap V(H)$ for each i, $H'' \subseteq \bigcup_{i=1}^{j} N(Q_3^i) \cap V(H)$, and hence by Lemma 15(1), $|H''| \leq 2|V(A) - T| = 2k-6$. Hence, we have $|H_0| \geq 4$ and thus $|H''| \geq 2k-5$. This is a contradiction to the assertion that $|H''| \leq 2k-6$.

Finally, suppose that every edge of H is contained in a member of C'. In this case, we define $\{T_1, \ldots, T_j\} = C'$ by the following procedure. Let e_1 be any edge of H, let T_1 be a member of C' containing e_1, and let $A^1 \in A_3$ be a $(k+1)$-cutset containing T_1. Assume that we have defined T_i and A^i for $1 \leq i \leq l-1$. If $H = \bigcup_{i=1}^{l-1} H_2^i$, then we let $j = l-1$ and terminate our procedure; if $H \neq V(\bigcup_{i=1}^{l-1} H_2^i)$, then we let e_l be an edge of H joining $\bigcup_{i=1}^{l-1} V(H_2^i)$ to $V(H) - \bigcup_{i=1}^{l-1} V(H_2^i)$ (such an edge exists because H is connected), let T_l be a member of C' containing e_l, and let $A^l \in A_3$ be a $(k+1)$-cutset containing T_l. Then since $H_2^i = N(Q_3^i) \cap V(H)$ for each $1 \leq i \leq j$, we have $(N(Q_3^k) \cap V(H)) \cap (\bigcup_{i=1}^{k-1} N(Q_3^i) \cap V(H)) \neq \emptyset$ for each $2 \leq k \leq j$, and $\bigcup_{i=1}^{j} N(Q_3^i) \cap V(H) = V(H)$. Hence, by Lemma 15(2), we have $|H| \leq |\bigcup_{i=1}^{j} Q_3^i| + 1 \leq |A - T| + 1 \leq k-2$. But, since $|H| \geq k-1$ by the remark just after Lemma 7, this is impossible. □

Case 2. There exists a k-contractible edge which is not contained in a cycle of length 4.

We need the following proposition.

PROPOSITION 1. *Let G be a k-connected graph and xy be a k-contractible edge. Also, let G' be the graph obtained from G by contracting xy and z be the vertex of G' which comes from the contraction of xy.*

(1) *If $ab \in E(G')$ is a k-contractible edge in G' and $z \neq a, b$, and if ab is not contained in any triangle in G', then ab is also k-contractible in G.*

(2) *If $za \in E(G')$ is a k-contractible edge in G' and if xy is not contained in any triangle in G, then either xa or ya is a k-contractible edge.*

Proof. (1) Assume ab is a k-contractible edge in G' but not in G. This is possible only when $\{a, b\} \subset A$ for some k-cutset A such that $A \cap \{x, y\} \neq \emptyset$, and $\{x, y\} - A$ is a connected component in $G - A$. Then $\langle z, a, b \rangle$ is a triangle which contains ab, which is contrary to the fact that ab is not contained in any triangle. Similarly, we can prove (2). Note that since $\langle x, y, a \rangle$ cannot be a triangle, one of xa and ya cannot be an edge in G. □

Now, we can finish the proof for Case 2. By the assumption of Case 2, there exists a k-contractible edge $e = xy$ which is not contained in a cycle of length 4. Then the graph G' obtained from G by the contraction of e does not contain a K_4^-. Recall that no k-contractible edge is contained in any triangle in G; in particular, e is not contained in a triangle. Let v be the vertex of G', which comes from the contraction of xy.

By the induction hypothesis, G' contains an induced cycle C such that $G' - V(C)$ is $(k-2)$-connected and each edge of C is k-contractible or C is a triangle, and for any vertex $x \in V(G') - V(C)$, $|N_C(x)| \leq 2$ (if G' has a k-contractible edge contained in a triangle, then we take such a triangle as C). First, we consider the case $v \notin V(C)$. We can regard C as an induced cycle in G, and we can easily see that $|N_C(x)| \leq 2$ and $|N_C(y)| \leq 2$, and hence $G - V(C)$ is $(k-2)$-connected. If C is a triangle, then we are done. If C is not a triangle, then by the choice of C, no edge of C is contained in any triangle in G', and hence by Proposition 1, we know that each edge of C is also k-contractible in G. Hence C is a desired induced cycle.

Finally, suppose $v \in V(C)$. Now, we let C_1 be the unique cycle of G which contains the path $C-v$ and either one or both of x, y. Since C is an induced cycle and e is not contained in any triangle, we see that $|N_{C_1}(z)| \leq 2$ for all $z \in V(G)-V(C_1)$, and $G-V(C_1)$ is $(k-2)$-connected. If C is a triangle, then since e is not contained in a cycle of length 4, C_1 is also a triangle; if C is not a triangle, then by Proposition 1, each edge of C' is k-contractible in G.

This completes the proof of Theorem 7. □

3. Concluding remarks. In [6], Kawarabayashi proved the following theorem.

THEOREM 9. *Let $k \geq 3$ be an integer, and let G be a k-connected graph which does not contain K_4^-. Then there exists a k-contractible edge which is not contained in a triangle, or there exists a k-contractible triangle, where a triangle of a k-connected graph is said to be a k-contractible triangle if the graph obtained from G by contracting the triangle (and replacing each of the resulting pairs of double edges by a single edge) is still k-connected.*

Considering Theorem 9, the following stronger statement will most likely hold.

CONJECTURE 1. *Let G be a k-connected graph which does not contain a K_4^-. Then G has an induced cycle such that $G-V(C)$ is $(k-1)$-connected or $G-V(C)$ is $(k-2)$-connected and each edge of C is k-contractible.*

Acknowledgments. We thank both referees for many helpful suggestions which improved the presentation of this paper.

REFERENCES

[1] N. DEAN, *Distribution of contractible edges in k-connected graphs*, J. Combin. Theory Ser. B, 48 (1990), pp. 1–5.

[2] Y. EGAWA, *Cycles in k-connected graphs whose deletion results in a $(k-2)$-connected graph*, J. Combin. Theory Ser. B, 42 (1987), pp. 371–377.

[3] Y. EGAWA, *Contractible cycles in graphs with girth at least* 5, J. Combin. Theory Ser. B, 74 (1998), pp. 213–264.

[4] Y. EGAWA, H. ENOMOTO, AND A. SAITO, *Contractible edges in triangle-free graphs*, Combinatorica, 6 (1986), pp. 269–274.

[5] Y. EGAWA AND A. SAITO, *Contractible edges in nonseparating cycles*, Combinatorica, 11 (1991), pp. 389–392.

[6] K. KAWARABAYASHI, *Contractible edges and triangles in k-connected graphs*, J. Combin. Theory Ser. B, 85 (2002), pp. 207–221.

[7] W. MADER, *Disjunkte Fragmente in kritisch n-fach zusammenhängenden Graphen*, European J. Combin., 6 (1985), pp. 353–359.

[8] W. MADER, *Generalizations of critical connectivity of graphs*, Discrete Math., 72 (1988), pp. 267–283.

[9] C. THOMASSEN, *Nonseparating cycles in k-connected graphs*, J. Graph Theory, 5 (1981), pp. 351–354.

[10] C. THOMASSEN AND B. TOFT, *Nonseparating induced cycles in graphs*, J. Combin. Theory Ser. B, 31 (1981), pp. 199–224.

SIAM J. DISCRETE MATH.
Vol. 21, No. 4, pp. 1071–1082

MATCHED-FACTOR d-DOMATIC COLORING OF GRAPHS*

K. S. SUDEEP[†] AND SUNDAR VISHWANATHAN[†]

Abstract. Consider a graph G and a collection of connected spanning subgraphs $G_1, G_2, \ldots, G_k$, not necessarily edge-disjoint. A subset U_i of the vertex set is said to *d-dominate* G_i if in G_i, all the vertices are at distance at most d from some vertex in U_i. Alon et al. [*Discrete Math.*, 262 (2003), pp. 17–25] introduced and studied a function $\mu(k)$, which is defined as the minimum radius of domination d such that the vertex set of every graph with a collection of k spanning subgraphs can be partitioned into $U_1, U_2, \ldots, U_k$ such that U_i d-dominates G_i. They proved that $\mu(k) < \frac{3}{2}k$, and the proof yields a polynomial time algorithm for the same. We prove that the problem is $\mathcal{NP}$-*complete*, and we also answer a question from their paper by improving their bound to $(\frac{3}{2} - \epsilon)k$. We also present an algorithm which finds such a coloring in polynomial time.

Key words. graph theory, algorithms, domination, matched-factor domatic number

AMS subject classification. 05C69

DOI. 10.1137/060662307

1. Introduction.

The concept of factor domination was introduced by Brigham and Dutton [3]. They defined *k-factoring* as a decomposition of a graph into k edge-disjoint spanning subgraphs. Later, Alon et al. [1] considered problems related to domination where all the subgraphs were required to be connected.

For the rest of this paper a *factor* means a connected spanning subgraph. Consider a connected graph $G(V, E)$. A *k-factorization* of G is a set of k connected spanning subgraphs $S_1, S_2, \ldots, S_k$ of G whose union is G. We do not require these subgraphs to be edge-disjoint. We denote by $d_G(u, v)$ the distance between vertices u and v in graph G. The neighborhood of a vertex v in graph G is the set of vertices adjacent to v along with the vertex v itself and can be represented as $\{x \in V : d_G(v, x) \leq 1\}$. Generalizing this, the *d-neighborhood* of v in G is $\{x \in V : d_G(v, x) \leq d\}$. These are the set of vertices at distance at most d from v in G.

A *d-dominating set* of vertices in graph G is a set $S \subseteq V$ such that every vertex in V is in the d-neighborhood of some element of S. A *d-domatic coloring* of G is a partition of the vertex set V into color classes such that each color class constitutes a d-dominating set of G. Note that it need not be a proper vertex coloring in the sense that adjacent vertices can have the same color. The maximum number of colors in any d-domatic coloring of a fixed graph G is called the *d-domatic number* of G. A 1-dominating set is commonly known as a *dominating set*, and the 1-domatic number is called the *domatic number*. Approximation algorithms for the domatic number were studied in [7].

We consider a related concept introduced by Alon et al. [1]. Let G be a graph and let $S_1, S_2, \ldots, S_k$ be k connected spanning subgraphs of G whose union is G. A vertex coloring of G, where the vertices of color i constitute a d-dominating set for the subgraph S_i, is called a *matched-factor d-domatic coloring of G with respect to* $S_1, S_2, \ldots, S_k$. A coloring is called an *all-factor d-domatic coloring of G with respect to* $S_1, S_2, \ldots, S_k$ if the vertices of each color constitute a d-dominating set in each S_j

*Received by the editors June 7, 2006; accepted for publication (in revised form) July 28, 2007; published electronically January 25, 2008.

http://www.siam.org/journals/sidma/21-4/66230.html

[†]Department of Computer Science and Engineering, Indian Institute of Technology, Bombay, India (sudeep@cse.iitb.ac.in, sundar@cse.iitb.ac.in). This research was supported by an Infosys Fellowship.

for $1 \leq j \leq k$. Note that both of them are always valid d-domatic colorings of the graph G.

Given an integer k, we are interested in the minimum $d(k)$ such that every k-factorization of every graph on at least k vertices admits a matched-factor $d(k)$-domatic coloring. Note that any partition of the vertex set into k sets is a matched-factor R-domatic coloring for some R. Following Alon et al. [1] we denote this minimum $d(k)$ by $\mu(k)$. It was proved in [1] that $k \leq \mu(k) \leq \lceil \frac{3}{2}(k-1) \rceil$.

The problem we address is formally stated below.

Input: A graph $G(V, E)$ and a collection of connected spanning subgraphs $S_1, S_2, \dots, S_k$.

Output: A coloring of the vertices of G using k colors such that the vertices of color i constitute an R-dominating set for the subgraph S_i and R is minimum.

1.1. Summary of results. We first show that the corresponding decision problem is $\mathcal{NP}$*-complete* for $k = 2$, even if we take the simple case of $d = 1$, and restrict the connected spanning subgraphs to be trees.

It is proved in [1] that for every $k \geq 2$, $\mu(k) \leq \lceil \frac{3}{2}(k-1) \rceil$. Their proof yields a polynomial time algorithm for the task. We present an improved upper bound of $(\frac{3}{2} - \epsilon)k$ for a fixed positive constant ϵ. One feature of our proof is that it uses the Lovász local lemma [4]. We then use techniques from Molloy and Reed [9] to get a randomized algorithm that has an expected polynomial running time to find a $(\frac{3}{2} - \epsilon)k$-domatic coloring. It is derandomized later to get a deterministic algorithm that runs in polynomial time.

2. $\mathcal{NP}$-completeness.

Given two trees on the same vertex set, we show that the problem of finding whether the vertex set can be partitioned into two parts such that vertices of one part are a dominating set in one of the trees and those of the other part are a dominating set in the other tree is $\mathcal{NP}$*-complete*. For that we reduce 3-*SAT* to the problem defined above.

2.1. Reduction of 3-*SAT* to the problem.

The problem 3-*SAT* [8] is as follows:

- *Instance*: A Boolean expression C in conjunctive normal form (CNF) in n variables and m clauses such that each clause has exactly three literals. $C = C_1 \wedge C_2 \wedge \cdots \wedge C_m$, where $C_i = w_{i1} \vee w_{i2} \vee w_{i3}$; $w_{ij} \in \{u_1, u'_1, u_2, u'_2, \dots, u_n, u'_n\}$, the set of literals. u'_i denotes the negation of u_i.
- *Question*: Is there a truth assignment to all the variables in the set $U = \{u_1, u_2, \dots, u_n\}$ such that C evaluates to *true*?

We construct an instance of our problem from 3-*SAT* as follows. As the truth setting component, for every variable u_i $(1 \leq i \leq n)$ we have $2m$ vertices $u_{i1}, u_{i2}, \dots, u_{im}$, $u'_{i1}, u'_{i2}, \dots, u'_{im}$ in V, one copy each, of the variable and its negation, for every clause. In addition to that we have four more vertices x_i, x'_i, y_i, and y'_i for each u_i. In the satisfiability component, for every clause C_j $(1 \leq j \leq m)$ we have six vertices $C_j, C'_j, C''_j, v_{j1}, v_{j2}$, and v_{j3}.

We construct the tree T_1 as follows:

(i) For every variable u_i $(1 \leq i \leq n)$, there is an edge between vertices x_i and x'_i. Vertices $u_{i1}, u_{i2}, \dots, u_{im}$, along with y_i, are adjacent to x_i. Similarly, vertices $u'_{i1}, u'_{i2}, \dots, u'_{im}$, y'_i are adjacent to x'_i.

(ii) For every clause C_j, there is a star that connects vertices $C'_j, C''_j, v_{j1}, v_{j2}$, and v_{j3} to C_j.

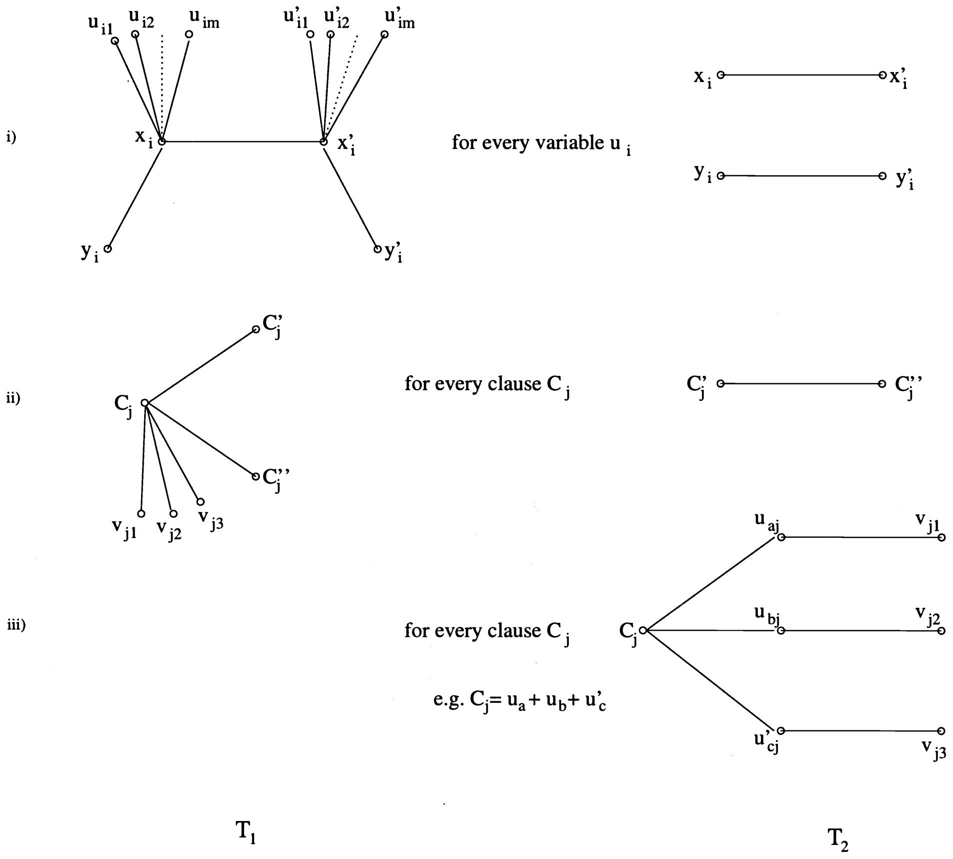


FIG. 1. *Construction of* T_1 *and* T_2 *from* 3-*SAT*.

In T_2, we have (i) edges $x_i x'_i$ and $y_i y'_i$ for every variable u_i. For every clause C_j we have (ii) an edge $C'_j C''_j$, along with (iii) a tree in which vertices that correspond to the jth copy of the three literals w_{j1}, w_{j2}, and w_{j3} of the clause are adjacent to the root vertex C_j. There are three more vertices v_{j1}, v_{j2}, and v_{j3} in the tree, and v_{jx} is adjacent to w_{jx} ($x = 1, 2, 3$). For example, for a clause $C_j = u_a + u_b + u'_c$, we have u_{aj}, u_{bj}, and u'_{cj} adjacent to C_j, and edges that connect v_{j1} to u_{aj}, v_{j2} to u_{bj}, and v_{j3} to u'_{cj}.

Four vertices r_1, r_2, a, and b are added on both sides to connect the subtrees. A path $r_1 r_2 ab$, with r_2 connected to one vertex each in every subtree, is added to T_2. While in T_1, r_1 is adjacent to one vertex each in every subtree and a, b, and r_1 are adjacent to r_2. The construction of the components of trees T_1 and T_2 that correspond to the variables and clauses in 3-*SAT* is illustrated in Figure 1. How the subtrees in the two trees are connected using vertices r_1, r_2, a, and b is shown in Figure 2.

Now we ask the following question: Does there exist a partition of the vertex set V into two color classes V_1 and V_2 such that V_1 is a dominating set in T_1 and V_2 is a dominating set in T_2?

We claim that there exists such a partition if and only if there is a satisfying assignment for the corresponding 3-*SAT* instance.

A broad outline is as follows. The details follow immediately afterwards. The variable u_i being true will correspond to the vertex x_i being picked in V_1. We will

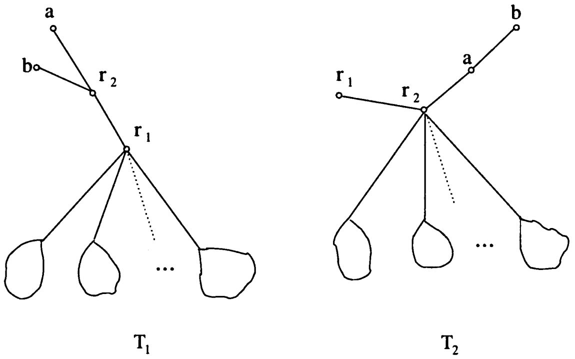


FIG. 2. *Vertices $r_1, r_2, a,$ and b that connect the subtrees in T_1 and T_2.*

ensure that both x_i and x'_i do not get picked in V_1, ensuring the truth setting. Also when this happens, we will force all of the u'_{ij}'s to be in V_1. For the satisfaction of the clauses, we will ensure that for each j, the vertex C_j corresponding to clause j will be picked in V_1. This ensures that at least one u_{fj} is in V_2. This will force the corresponding x_t to be in V_1. The other vertices and edges are present to ensure such an outcome.

Proof. The problem is in $\mathcal{NP}$ because if we are given a partition $\{V_1, V_2\}$ of V claiming that V_1 and V_2 are dominating sets in T_1 and T_2, respectively, then we can verify that claim in polynomial time.

We first make the observation that the four extra vertices that connect the subtrees are added in such a way that it does not affect the partitions. For V_2 to be a dominating set in T_2, one or both of $\{a, b\}$ is in V_2, and hence $r_2 \in V_1$ and $r_1 \in V_2$.

Now we assume we have a partition $\{V_1, V_2\}$ such that V_1 is a dominating set in T_1 and V_2 dominates T_2. We will show that from V_1 and V_2 we can get a satisfying assignment for C.

For any variable $u_i \in U$, the construction (i) of T_2 ensures that x_i and x'_i cannot both be in V_1. This is because if they are, then in T_2, neither x_i nor x'_i will be in the neighborhood of any vertex in V_2. The same holds for y_i and y'_i. Further, (i) of T_1 insists that both y_i and y'_i cannot be in V_2 either, since at most one of x_i and x'_i can be in V_1. Thus we have either $x_i, y'_i \in V_2$ and $y_i, x'_i, u_{i1}, u_{i2}, \ldots, u_{im} \in V_1$, or $y_i, x'_i \in V_2$ and $x_i, y'_i, u'_{i1}, u'_{i2}, \ldots, u'_{im} \in V_1$.

By (ii) of T_2, both C'_j and C''_j cannot be in V_1. This means that $C_j \in V_1$, by virtue of construction (ii) of T_1. Now, (iii) of T_2 leaves us with the only choice of having at least one of the three neighbors of C_j in T_2 be in V_2 since V_2 is a dominating set in T_2. If $u_{aj} \in V_2$, it cannot be the case that $y_a, x'_a, u_{a1}, u_{a2}, \ldots, u_{am} \in V_1$. This forces x_a to be in V_1. Similarly, if $u'_{cj} \in V_2$, it cannot be that $x_c, y'_c, u'_{c1}, u'_{c2}, \ldots, u'_{cm} \in V_1$, and we have $x_c \in V_2$. Hence assigning u_i to *true* if $x_i \in V_1$ and *false* otherwise, we get an assignment that makes at least one of the three literals in every clause C_j true, giving us a satisfiable assignment for C.

Conversely, if we have a truth assignment of variables in C, by assigning $x_i, y'_i, u'_{i1}, u'_{i2}, \ldots, u'_{im}$ to be in V_1 and $x'_i, y_i, u_{i1}, u_{i2}, \ldots, u_{im}$ to be in V_2 if u_i is *true* and swapping the partition if u_i is *false*, we get a partition that we need.

$C_j \in V_1$ for all j, and one or both of $\{C'_j, C''_j\}$ is in V_2. A vertex v_{jk} is in V_2 if the literal corresponding to its neighbor in T_2 is *false*; else it could be in V_1 or in V_2. □

3. An upper bound. We first sketch an upper bound on $\mu(k)$ from Alon et al. [1]. Their proof yields a polynomial time algorithm. Our proof builds on this proof.

THEOREM 1 (see [1]). *For every $k \geq 2$, $\mu(k) \leq \lceil \frac{3}{2}(k-1) \rceil$.*

THEOREM 2. *For a large enough k, $\mu(k) \leq (\frac{3}{2} - \epsilon)k$, where $0 < \epsilon < 1$ is a constant.*

One tool we use here is the Lovász local lemma [4]. We use a symmetric version of the lemma from Alon and Spencer [2].

LEMMA 3 (Lovász local lemma, symmetric version). *Let $A_1, A_2, \ldots, A_n$ be events in a probability space such that $Pr\,[A_i] \leq p$ for all i. Let H be a graph with vertices $A_1, A_2, \ldots, A_n$ such that for each $i = 1, 2, \ldots, n$ the event A_i is mutually independent of any combination of events that are not neighbors of A_i in H. Suppose the maximum degree of H is d, and $e \cdot p \cdot (d+1) \leq 1$, where e is the base of the natural logarithm. Then there is a positive probability that none of the events A_i occurs.*

We also use the following, simplest version of the Chernoff bounds [2].

LEMMA 4. *Let X_i, for $1 \leq i \leq n$, be mutually independent random variables with $P(X_i = 1) = P(X_i = 0) = 1/2$. Then $P(X_1 + \cdots + X_n > a) < \exp \frac{-a^2}{2n}$.*

We use the following two lemmas in the proof of Theorem 2, along with Lemma 3. Alon et al. prove the following lemma in [1]. We reproduce the simple proof for completeness.

LEMMA 5 (see [1]). *Any tree T on vertex set $V, |V| = n \geq l \geq 1$ can be decomposed into vertex-disjoint trees $T_1, T_2, \ldots, T_p$ such that $\cup V(T_i) = V$, $|T_i| \geq l$, and the diameter of T_i is at most $2l - 2$ for $i = 1, 2, \ldots, p$.*

Proof. If T has diameter at most $2l-2$, we are done. So we assume that $diameter(T) \geq 2l - 1$. It follows that T contains an edge e such that both subtrees T_1 and T_2 of $T - e$ have at least l vertices each and both T_1 and T_2 are smaller in size compared to T. Applying induction, T_1 can be further decomposed into $T_{11}, T_{12}, \ldots, T_{1a}$, and T_2 can be further decomposed into $T_{21}, T_{22}, \ldots, T_{2b}$, each of which has a diameter of at most $2l - 2$ and size at least l. $\square$

Alon et al. use this lemma to break up a big tree and deal with each small tree individually. In a subtree that has at least k vertices and whose diameter is at most $2k - 2$, they show that there are at least k vertices at a distance of at most $\frac{k-1}{2}$ from the center of the tree, and every vertex in this group is at a distance of at most $\lceil \frac{3}{2}(k-1) \rceil$ from any vertex in the tree. In other words, every vertex in this group is a $\lceil \frac{3}{2}(k-1) \rceil$-dominating set in this subtree. An auxiliary bipartite graph is then constructed. To distinguish the vertices of the original graph from those of this bipartite graph, we will call the vertices of the bipartite graph *points*.

Each of these groups of vertices from each subtree constitutes points in one partition, say B. On the other hand, we have a set of points W, each of whose point corresponds to a vertex of G. The edges of the graph are defined by inclusion. That is, there is an edge between a point in B and a point in W if the corresponding vertex is contained in the corresponding group.

A group of vertices corresponding to a point in B has exactly k vertices in it, which means that the degree of each point in B is exactly k. Each vertex of the graph is in exactly k subtrees, and hence in at most k of these groups. So the degree of a vertex in W is at most k. By Hall's theorem [6], this implies that it is possible to find a matching that saturates all points in B. In other words, we can choose one vertex from each group such that the vertices chosen are distinct. In each subtree of S_i, we assign color i to the vertex that represents the group of dominating vertices in the subtree. Thus we arrive at a coloring in which all vertices of S_i are at a distance of at most $\lceil\frac{3}{2}(k-1)\rceil$ from one of these vertices colored i. This is the essence of the Alon et al. algorithm.

We motivate our algorithm below. If all subtrees are stars, then we can choose k vertices in each star such that any vertex forms a 2-dominating set, and hence the previous algorithm yields a much better result than the one proved. The bad case is if all subtrees are paths of length $2k-2$. The algorithm picks k vertices around the center of the path. Supposing all trees were such paths, it is better to take two sets of size $k/2$, each at a distance of $k/2$ from the two endpoints. For the dominating set we need to pick one vertex from each set. In the bipartite graph, the degree of a point corresponding to these groups now falls to $k/2$. To make the matching argument go through we have to make sure that a vertex is picked at most $k/2$ times. It is not obvious how to do this. The other difficulty is that the subtrees we get could be heterogeneous. They could look like paths, stars, or anything in between. We deal with both these problems using a structural lemma which holds for all trees and randomness. We give details below.

For a tree $T(V,E)$ and $x, y \in V$, let $d_T(x,y)$ denote the distance between vertices x and y in T. For a subset $S \subseteq V$ and a vertex $v \in V$, we define $D_T(v,S) = \max_{s\in S}\{d_T(v,s)\}$.

Lemma 6 (the tree lemma). *Let $T(V,E)$ be a tree such that $|V| \geq l = (1+\delta)k$ and diameter$(T) \leq 2l - 2$. Let $D = (\frac{3}{2}-\epsilon)k$, where δ and ϵ are positive constants such that $28\epsilon + 39\delta \leq 3$. Then one of the following must hold:*

(i) *There is a subset $S \subseteq V$ such that $|S| = l$ and for all $v \in V$, $D_T(v,S) \leq D$.*

(ii) *There exist disjoint subsets $P, Q, R \subset V(T)$ such that $|P| = |Q| = |R| = \frac{l}{2}$ and for all $v \in V$, $D_T(v,P) \leq D$ and either $D_T(v,Q) \leq D$ or $D_T(v,R) \leq D$.*

Before we begin the proof, we note that case (i) of the lemma applies to trees which are star-like. The other case deals with trees which look like paths. The reason for having three sets (and not two, as one would expect) in case (ii) is technical, to facilitate the phenomenon of picking a vertex at most $k/2$ times. We will not achieve $k/2$ exactly but rather a bit more.

Proof.

Case 1. *diameter*$(T) \leq 2k(1-\sigma)$, where $\sigma = \epsilon + \frac{\delta}{2}$.

This case, as also Case 2.1 below, deals with trees that behave like stars. The rest is for trees that behave like paths. Let r be a center vertex of the tree T. Let u and v be two vertices of maximum distance in T such that $d_T(r,v) \leq d_T(r,u)$. For any vertex $x \in V$, $d_T(r,x) \leq d_T(r,u) \leq k(1-\sigma)$. Let $S^* = \{x : x \in V \text{ and } d_T(r,x) \leq \lceil\frac{l}{2}\rceil\}$. We claim that S^* has at least l elements in it. This holds when $d_T(r,u) \leq \frac{l}{2}$, in which case S^* is the entire vertex set V, since no vertex in the tree is farther from r than u. Otherwise, since $d_T(r,v) \geq d_T(r,u) - 1$, S^* contains at least $\frac{l}{2}$ elements each (other than r) on the unique paths from r to u and v, thus ensuring at least l elements in S^*. The distance from any vertex $y \in V$ to any vertex $s \in S^*$ is at most $d_T(y,r) + d_T(r,s)$. This sum is at most $k(1-\sigma) + \frac{k}{2}(1+\delta) = (\frac{3}{2} - (\sigma - \frac{\delta}{2}))k = (\frac{3}{2}-\epsilon)k$. Any size-$l$ subset of S^* qualifies as subset S in the lemma.

Case 2. *diameter*$(T) > 2k(1-\sigma)$.

Case 2.1. There are at least l vertices at a distance of at most $\frac{k}{2}(1-2\epsilon-2\delta)$ from the center r.

Pick any l of those vertices as a set S. For $y \in V$ and $s \in S, d_T(y,s) \leq d_T(y,r) + d_T(r,s) \leq k(1+\delta) + \frac{k}{2}(1-2\epsilon-2\delta) = (\frac{3}{2}-\epsilon)k$.

Case 2.2. There are fewer than l vertices at a distance at most $\frac{k}{2}(1-2\epsilon-2\delta)$

from the center r.

We pick subsets P, Q, and R as follows. We let the middle $\frac{l}{2}$ vertices around the center vertex r on the (u, v)-path be the subset P. The $\frac{l}{2}$ vertices, each preceding and succeeding P on the path (u, v), are picked as Q and R, respectively. We argue below that this is always possible by showing that there are at least $\frac{3}{2}l$ vertices on the path. The path length is at least $2k(1 - \sigma)$. We will show that the choice of σ is such that $2k(1 - \sigma) \geq \frac{3}{2}l$. Note that these sets are disjoint by construction. By recalling the value of σ and easy manipulations, $2k(1 - \sigma) = \frac{3k}{2}\frac{4(1-\epsilon-\delta/2)}{3}$. As long as $5\delta + 4\epsilon \leq 1$, we have $\frac{4(1-\epsilon-\delta/2)}{3} \geq 1 + \delta$, which finishes this claim.

It remains to be shown that these sets satisfy the distance constraints. The distance from any vertex $y \in V$ to any vertex $p \in P$ is at most $d_T(y, r) + d_T(r, p) \leq k(1 + \delta) + \frac{k}{4}(1 + \delta) \leq \frac{5k}{4}(1 + \delta) = (\frac{3}{2} - \frac{1}{4} + \frac{5}{4}\delta)k \leq (\frac{3}{2} - \epsilon)k$, as we have chosen δ such that $4\epsilon + 5\delta \leq 1$. It is enough to prove that a vertex $y \in V$ is close to the vertices in one of the two sets Q and R. Let the path from y to r touch the (u, v)-path at x. Without loss of generality, we may assume that x lies between r and u on the path (u, v). Let x be at a distance of d from r. Also, let w be the point at a distance of $\frac{k}{2}(1 - 2\epsilon - 2\delta)$ from r on the path from y to r.

Case 2.2.1. w lies between x and r on the path from y to r.

We have $d_T(x, r) > d_T(w, r)$ or $d > \frac{k}{2}(1 - 2\epsilon - 2\delta)$.

Case 2.2.2. w lies between y and x on the path from y to r.

We know that there are fewer than l vertices within a distance of $\frac{k}{2}(1 - 2\epsilon - 2\delta)$ from r (Case 2.2). Adding up the number of vertices that lie within that distance from r on paths (r, v), (r, u), and (w, x), we get $\frac{3k}{2}(1 - 2\epsilon - 2\delta) - d < k(1 + \delta)$ or $d > \frac{k}{2}(1 - 6\epsilon - 8\delta)$. Note that this inequality holds even in Case 2.2.1.

For any element z in Q, if z lies between x and r in the path from u to r, $d_T(y, z) = d_T(y, r) - d_T(r, z)$. But T is such that $d_T(y, r) < l$ and Q is chosen so that for $z \in Q$, $d_T(r, z) \geq \frac{l}{4}$. That is, $d_T(y, z) < l - \frac{l}{4} = \frac{3}{4}k(1 + \delta) < (\frac{3}{2} - \epsilon)k$. In a less trivial case of x falling between r and z on the path, $d_T(y, z) \leq d_T(y, x) + d_T(x, z) = (d_T(y, r) - d) + (d_T(r, z) - d) \leq (l - d) + (\frac{3}{4}l - d)$. Using the fact that $d > \frac{k}{2}(1 - 6\epsilon - 8\delta)$, we get $d_T(y, z) \leq \frac{7}{4}k(1 + \delta) - k(1 - 6\epsilon - 8\delta) = (\frac{3}{4} + \frac{39}{4}\delta + 6\epsilon)k \leq (\frac{3}{2} - \epsilon)k$ when $39\delta + 28\epsilon \leq 3$. Note that ϵ could be made as big as $\frac{3}{28}$. □

With these lemmas in place, we are now prepared for a proof of the theorem.

Proof of Theorem 2. Let $S_1, S_2, \ldots, S_k$ be a k-factorization of a graph G on $n \geq k$ vertices. Let T_i be a spanning subtree of S_i for $i = 1, 2, \ldots, k$. By Lemma 5, each T_i can be decomposed into vertex-disjoint trees $T_{i1}, T_{i2}, \ldots, T_{ip_i}$ such that $\cup V(T_{ij}) = V(T_i)$, $|T_{ij}| \geq l$, and $diameter(T_{ij}) \leq 2l - 2$ for $j = 1, 2, \ldots, p_i$.

Using Lemma 6, if in each T_{ij} we have a subset S of size l satisfying Lemma 6(i), then we pick any $\frac{l}{2}$ elements of S at random to get a subset B_{ij}. Otherwise we have sets P, Q, and R, all of size $\frac{l}{2}$. Flip a coin and pick either the subset P to be B_{ij} or the two subsets Q and R to be B_{ij1} and B_{ij2}, depending on the coin flip. When we pick a single set B_{ij}, by the tree lemma we have for all $x \in V(T_i)$ and $y \in B_{ij}$, $d_T(x, y) \leq (\frac{3}{2} - \epsilon)k$. When we pick B_{ij1} and B_{ij2}, we ensure that for all $x \in V(T_i)$, $d_T(x, y) \leq (\frac{3}{2} - \epsilon)k$ either for all $y \in B_{ij1}$ or for all $y \in B_{ij2}$. Note that in a tree, a vertex gets picked with probability at most $\frac{1}{2}$. A vertex appears in k subtrees, one each in the decompositions of every T_i, and the expected number of times a vertex appears in the picked sets is thus bounded above by $\frac{k}{2}$.

We define a *bad event* to be the event of a vertex v getting picked in at least $\frac{k}{2}(1 + \delta)$ sets. By Chernoff's bounds, the probability of such an event is at most

$e^{-\delta^2 \frac{k}{6}}$. Two events of two vertices v_1 and v_2 getting picked are dependent only if the two are in the same subtree. In addition, if (i) of Lemma 6 holds, the two events are independent unless both vertices belong to the subset S. In the case of path-like trees, i.e., when (ii) of Lemma 6 holds, the events are independent unless both vertices are in $P \cup Q \cup R$. A vertex v is in exactly k subtrees, so the event of it being picked has edges in the dependency graph with at most $k \cdot 3 \cdot \frac{k}{2}(1+\delta)$ events. These events correspond to other vertices in S or in $P \cup Q \cup R$—whichever may be the case—getting picked, in each subtree in which v lies. We may recall here that $|S| = k(1+\delta)$ and $|P \cup Q \cup R| = 3 \cdot \frac{k}{2}(1+\delta)$.

With $p \leq e^{-\delta^2 \frac{k}{6}}$ and $d \leq k \cdot 3 \cdot \frac{k}{2}(1+\delta)$, it follows from Lemma 3 that there is always a way of picking the vertices such that no vertex is picked $\frac{l}{2}$ times or more if k is big enough. We consider such a selection for the rest of our proof, since it is enough for us to prove the existence of a particular coloring in order to prove an upper bound on $\mu(k)$. We will require the problem to satisfy stricter conditions, compared to that of the local lemma, in order to get an algorithm which actually finds such a coloring. We discuss that in the next section.

Now, construct a bipartite graph H with bipartition B and W as follows. For every $1 \leq i \leq k$ and $1 \leq j \leq p_i$, the class B contains points corresponding to subset B_{ij} or two points corresponding to two subsets B_{ij1} and B_{ij2}, depending on what we have picked for the tree T_{ij}. W contains points corresponding to all vertices of G. The edges of H are defined by containment as follows. Every point $b \in B$ is connected to a point $v \in W$ if and only if $v \in b$. We know that $deg_H(b) = \frac{l}{2}$ for each element $b \in B$, and $deg_H(v) < \frac{l}{2}$ for each $v \in W$. This implies that for every $B' \subseteq B$ there are at least $|B'|$ vertices in the neighborhood of B' in W, and by Hall's theorem [6] there exists a matching M that saturates B; i.e., we have a vertex to represent each of the dominating groups of vertices in each subtree T_{ij}.

We color vertices matched to B_{ij}, B_{ij1}, or B_{ij2} in M with color i for $1 \leq i \leq k$. All remaining vertices of G are colored by any of the k colors. This way, the vertices that represent the dominating groups in every subtree of S_i are colored i. In trees T_{ij}, where we have selected one set B_{ij} such that every vertex in the tree is at a distance of at most $(\frac{3}{2} - \epsilon)k$ from any vertex in this set, we have one vertex from B_{ij} in the color class i. In trees where we have picked two sets B_{ij1} and B_{ij2} such that every vertex is close to the vertices from either of these sets, the color class i has one vertex each from both of these sets. Since T_{ij} is a subgraph of S_i, the distance between any two vertices in the tree is no less than the distance between the two vertices in S_i. This makes the coloring a matched-factor $(\frac{3}{2} - \epsilon)k$-domatic coloring of G with respect to $S_1, S_2, \ldots, S_k$. □

4. A polynomial time algorithm.

In many probabilistic proofs, there is a reasonably high probability that a randomly chosen object satisfies the desired properties. Thus these proofs often imply a randomized algorithm that finds an object that we desire in expected linear or polynomial time. However, though some tools, such as the Lovász local lemma [4], help in proving that such an object exists, the probability with which the object occurs is very small. It becomes difficult to turn the proofs of existence into efficient algorithms. Beck was the first to come up with polynomial time algorithms for certain applications of the local lemma. A parallel and simpler version appears in [2]. Molloy and Reed [9] present a general framework for the application of these techniques and we follow their framework.

Using these standard techniques, we get a randomized algorithm that finds a $(\frac{3}{2} - \epsilon)k$-domatic coloring in expected polynomial time. We later remove the randomness using the method of conditional probabilities introduced by Erdős and Selfridge [5, 2] and now widely used.

We state the general technique from [9] first and then apply it to the problem at hand. Let $\mathcal{F} = \{f_1, \ldots, f_m\}$ be a set of independent random trials. Let $\mathcal{A} = \{A_1, \ldots, A_n\}$ be a set of events such that each A_i is determined by the outcome of the trials in $F_i \subseteq \mathcal{F}$. We say that A_i *intersects* A_j if $F_i \cap F_j \neq \phi$.

For any $f_{j_1}, \ldots, f_{j_k} \in F_i$ and any $w_{j_1}, \ldots, w_{j_k}$ in the domains of $f_{j_1}, \ldots, f_{j_k}$, respectively, we define $\mathbf{Pr}^*(A_i | f_{j_1} \rightarrow w_{j_1}, \ldots, f_{j_k} \rightarrow w_{j_k})$ to be the probability of A_i conditional on the event that the outcomes of $f_{j_1}, \ldots, f_{j_k}$ are $w_{j_1}, \ldots, w_{j_k}$, respectively. If $f_{j_1}, \ldots, f_{j_k}$ are already carried out and $w_{j_1}, \ldots, w_{j_k}$ are their outcomes, we sometimes just say $\mathbf{Pr}^*(A_i)$ when there is no ambiguity. When $k = 0$, $\mathbf{Pr}^*(A_i) = \mathbf{Pr}(A_i)$.

THEOREM 7 (see [9]). *If we have that*

1. *for each* $1 \leq i \leq n, Pr(A_i) \leq p$*;*
2. *each* F_i *intersects at most* d *other* F_j*'s;*
3. $pd^9 < \frac{1}{8}$*;*
4. *for each* $1 \leq i \leq n$, $|F_i| \leq \omega$*;*
5. *for each* $1 \leq i \leq m$*, the size of the domain of* f_i *is at most* γ*, and we can carry out the random trial in time* t_1*;*
6. *for each* $1 \leq i \leq n, f_{j_1}, \ldots, f_{j_k} \in F_i$*, and* $w_{j_1}, \ldots, w_{j_k}$ *in the domains of* $f_{j_1}, \ldots, f_{j_k}$*, respectively, we can compute* $\mathbf{Pr}^*(A_i)$ *in time* t_2*,*

then we have a randomized $O(m \times d \times (t1 + t2) + m \times \gamma^{\omega d \log \log m})$*-time algorithm which will determine outcomes of* $f_1, \ldots, f_m$ *such that none of the events in* $\mathcal{A}$ *hold.*

In this instance, depending on whether case (i) or (ii) of the tree lemma applies, $\mathcal{F} = \{f_1, \ldots, f_m\}$ is the set of independent random trials in which we pick a set of vertices B_{ij} of size $\frac{l}{2}$, or sets B_{ij1} and B_{ij2} of size $\frac{l}{2}$, where each vertex is in T_{ij}. We may denote by f_{ij} the trial that corresponds to the subtree T_{ij}. Since we are working with k connected spanning subgraphs S_1 through S_k and since each subtree T_{ij} is of size at least l, $|\mathcal{F}| = m$ is bounded above by $\frac{kn}{l} < n$.

For every vertex v, A_v is the bad event that the vertex is picked at least $\frac{l}{2}$ times. This is determined by F_v, the set of trials of picking a set B_{ij} or sets B_{ij1} and B_{ij2} from subtrees T_{ij} in which v appears. Thus, $f_{ij} \in F_v$ if and only if $v \in V(T_{ij})$.

We know that $\mathbf{Pr}(A_v) \leq p = e^{-\delta^2 \frac{k}{6}}$ and d of condition 2 is bounded above by $k \cdot 3 \cdot \frac{k}{2}(1 + \delta)$. These values of p and d imply condition 3 for large enough k. In condition 4, for each $1 \leq i \leq n$, $|F_v| = k = \omega$. For star-like trees the set S of the tree lemma can be partitioned into two equal parts S_a and S_b, and the choice narrows down to picking one of these two sets as B_{ij}. For the remaining trees the choice is between picking the subset P as B_{ij} and picking two subsets Q and R as B_{ij1} and B_{ij2}. Thus, the domain size of a trial, denoted by γ in condition 5, is 2. The random trial can be carried out in constant time, which means t_1 is $O(1)$. Given any set of possible outcomes of the trials, whether or not A_v holds now depends on the number of trials left. This we can compute by summing certain binomial coefficients, and hence t_2 in condition 6, the time required to test each possible combinations of the remaining trials, is bounded above by $O(k)$.

While this is enough to give a randomized algorithm, to get a deterministic algorithm we need to consider the details of the specific problem at hand. Hence we need to look at the details of the proof of the above theorem as applied to our problem and show that the method of conditional probabilities works. We do that in the next two sections.

4.1. Randomized algorithm and analysis. Before we give an outline of the algorithm, we recall that the input is a graph $G(V, E)$ and a collection of connected spanning subgraphs $S_1, S_2, \ldots, S_k$, and the output that we derive is a coloring of the vertices of G using k colors such that the vertices of color i constitute a $(\frac{3}{2} - \epsilon)k$-dominating set for the subgraph S_i for a positive constant ϵ.

The algorithm is as follows. The spanning subgraphs S_i are decomposed first into subtrees T_{ij} ($1 \leq i \leq k$ and $1 \leq j \leq p_i$) with the help of Lemma 5.

In the first sweep, the subtrees T_{ij} are considered in sequential order. If condition (i) of Lemma 6 holds in T_{ij}, we pick any $\frac{l}{2}$ elements of the set S at random as B_{ij}. Otherwise, we flip a coin and pick either the subset P to be B_{ij} or subsets Q and R to be B_{ij1} and B_{ij2}. We compute $\mathbf{Pr}^*(A_v)$ for each vertex v in the subtree T_{ij}. This is the probability that vertex v gets picked at least $\frac{l}{2}$ times, given the subsets that have already been picked, including the one(s) that got picked in this step. If $\mathbf{Pr}^*(A_v) > p^{2/3}$, then we say that A_v is *dangerous*, and we undo the trial and freeze the subtree and all other subtrees containing v so that we do not consider them again in this sweep.

At the end of the first sweep, $\mathbf{Pr}^*(A_v) \leq p^{2/3}$ for all v, implying that by the local lemma, there is a solution in which none of the A_v's occurs, extending the partial solution given by the trials already carried out. It turns out that not many of these events become dangerous. Specifically, for each vertex v, the probability p_d that A_v becomes dangerous in the first sweep is at most $p^{1/3}$. To see this, we recall that by the definition of a dangerous event A_v, $\mathbf{Pr}^*(A_v) > p^{2/3}$ and $\mathbf{Pr}(A_v) \geq p_d \cdot \mathbf{Pr}^*(A_v)$. This would exceed p if $p_d > p^{1/3}$.

In addition, dangerous vertices are distributed in such a way that we can carry out the frozen trials independently on small components. This makes it nearly feasible to find a good set of outcomes for the remaining trials using an exhaustive search. Before doing an exhaustive search we may repeat the first sweep at most a constant number of times on any big components remaining. To show this in quantitative terms, we use the notations from [9], which we describe below.

We denote by $\mathcal{H}$ the hypergraph with $V(\mathcal{H}) = \mathcal{F}$, and $E(\mathcal{H}) = \{F_1, \ldots, F_n\}$, and we let $\mathcal{L}$ be the line graph of $\mathcal{H}$. Note that the vertices of $\mathcal{L}$ correspond to the vertices of G and there is an edge between two vertices if the corresponding vertices in G fall in the same subtree T_{ij}. We denote by $\mathcal{L}^{(a,b)}$ the graph with a vertex set that is the same as that of $\mathcal{L}$, and where two vertices are adjacent if they are at a distance of a or b in $\mathcal{L}$. $T \subseteq E(\mathcal{H})$ is called a (1,2)-*tree* if the subgraph induced by T in $\mathcal{L}^{(1,2)}$ is connected. We call $T \subseteq E(\mathcal{H})$ a (2,3)-*tree* if the subgraph induced by T in $\mathcal{L}^{(2,3)}$ is connected *and* no two vertices of T are adjacent in $\mathcal{L}$ (the corresponding edges do not intersect in $\mathcal{H}$). We call a tree *dangerous* if all of its vertices correspond to dangerous events.

The main observation is that no A_i intersects two events A_j and A_k which belong to two different maximal dangerous $(1, 2)$-trees. This is because it would mean that (A_i, A_j) and (A_i, A_k) are edges in $\mathcal{L}$, and in turn there is an edge that connects A_j and A_k in $\mathcal{L}^{(1,2)}$. This makes it possible to deal with the frozen trials independently in each maximal dangerous $(1, 2)$-tree.

CLAIM 1. *With probability at least $\frac{1}{2}$, there are no dangerous $(1, 2)$-trees of size greater than $d \log_2 m$.*

Proof. We reproduce the proof of the claim from [9].

Note that every dangerous $(1, 2)$-tree of size dK contains a dangerous $(2, 3)$-tree of size K. For each f_{ij} (that corresponds to subtree T_{ij}), the number of $(2, 3)$-trees of size K in $\mathcal{H}$ in which f_{ij} lies is at most $(ed^3)^K$. The hyperedges F_v of $\mathcal{H}$, which correspond to events A_v, lying in any such tree are disjoint (by definition of a $(2,3)$-tree). Since an A_v becomes dangerous in the first sweep with probability at most $p^{1/3}$, the probability that all events A_i corresponding to these hyperedges become dangerous is at most $(p^{1/3})^K$. There are m trials in $\mathcal{F}$, and the expected total number of dangerous $(2,3)$-trees of size K is thus at most $m(ed^3p^{1/3})^K$. This is less than $\frac{1}{2}$ for $K = \log m$. That is, the expected number of $(1,2)$-trees of size $d\log m$ is less than $\frac{1}{2}$. It follows from the definition of an expected value that the probability that there is at least one dangerous $(1,2)$-tree of size $d\log m$ is less than $\frac{1}{2}$. □

After the first pass, if there is any dangerous $(1,2)$-tree of size $d\log_2 m$, then we repeat the first pass. Each repetition takes $O(m \times d \times 2^k)$ time, and the expected number of repetitions is constant. An exhaustive search for the satisfactory outcomes to the frozen trials corresponding to each dangerous $(1,2)$-tree now takes time $O(\gamma^{\omega d \log m})$. Since γ is constant, this is polynomial in the problem size as long as k is $O(\log n)$. The running time can be improved by having a second sweep in the same manner as the first sweep, where we carry out frozen events of the first sweep in sequence and declare an event dangerous when conditional probability exceeds $p^{1/3}$. Within an expected linear number of repetitions of the second pass, there will be no dangerous $(1,2)$-trees of size greater than $d\log\log m$, and the exhaustive search in each subtree takes time $O(\gamma^{\omega d \log\log m}) = O(2^{\frac{3}{2}k^3 \log\log n})$. Thus the total expected running time is $O(n \times k^2 \times 2^k + n \times \sqrt{8}^{k^3 \log\log n})$.

4.2. Derandomization.

Although the algorithm above is a general framework for a variety of problems, it turns out that, in the particular case of the problem at hand, it can be derandomized using the method of conditional probabilities due to Erdős and Selfridge [5] to get a deterministic algorithm that runs in polynomial time.

THEOREM 8. *There is a deterministic polynomial time algorithm, which when provided with a graph G and a k-factorization $S_1, S_2, \ldots, S_k$ of G, finds a matched-factor $(\frac{3}{2} - \epsilon)k$-domatic coloring when $k \geq k_0$, where k_0 is a constant.*

Note that we need k to be $O(\log n)$.

Proof. We eliminate the randomness in the algorithm by replacing each random trial—in this case picking either a set B_{ij} of size $\frac{l}{2}$ or two sets B_{ij1} and B_{ij2} of size $\frac{l}{2}$ each—with making a choice that minimizes the expected size of the biggest dangerous $(2,3)$-tree that remains after the first sweep. This is to make sure the dangerous $(1,2)$-trees that remain are small enough, as every dangerous $(1,2)$-tree of size dK contains a dangerous $(2,3)$-tree of size K. Here also we mark an event A_v dangerous and freeze all the trials that belong to A_v at the point when $\mathbf{Pr}^*(A_v)$ crosses $p^{2/3}$, as we did in the randomized algorithm. As before, at the end of the first sweep we have that $\mathbf{Pr}^*(A_v) \leq p^{2/3}$ for all v.

As we have noted earlier, the domain size of a trial $\gamma = 2$. For each choice we are to make, between S_a and S_b or between P and (Q,R) depending on which case of the tree lemma applies to the subtree, we do not pick one of the two at random. Instead we estimate the expected number of dangerous $(2,3)$-trees of size $K = \log m$ or more for both choices. How we do it is explained below.

We cannot make a general estimate on the probability that any $(2,3)$-tree of size K becomes dangerous, as in the case of the randomized version, because at this point, the choice has already been made on a certain number of subtrees. In other words, a certain number of trials has already been carried out. So we have to look at every candidate that could possibly become a dangerous $(2,3)$-tree at the end of the first sweep and estimate the probability that it actually becomes dangerous. We have already seen that for each $f_{ij} \in \mathcal{F}$ the number of $(2,3)$-trees of size K in $\mathcal{H}$ in which f_{ij} lies is at most $(ed^3)^K$, and any two vertices in such a tree T, which are hyperedges F_v (that correspond to events A_v) of $\mathcal{H}$, are disjoint. Since the events in T are independent of each other, the probability that all events A_v in T become dangerous is the same as the product of probabilities that A_v becomes dangerous if the remaining trials are carried out at random. In each T, this product can be computed in time $O(\log m \times t_2)$, as $|T| = K = \log m$. For each such T, we determine the probability of it becoming dangerous. The expected number of dangerous $(2,3)$-trees of size K is the sum of these probabilities over all such T. This can be calculated in time $O(\log m \times t_2 \times m(ed^3)^{\log m})$ because the number of such sets T is bounded above by $m(ed^3)^K$.

Of the two choices that we have before us, we make a choice for which this expected number is minimum. We know that originally this value is less than $\frac{1}{2}$, and since the expected value is an average of all possibilities at a particular point of time, we can be sure that at each point we can make a choice that does not increase the expected value. Thus we end up with a sequence of choices that ensures there is no dangerous $(2,3)$-tree of size $\log m$, and hence no $(1,2)$-tree of size $d \log m$.

As in the randomized case, we do an exhaustive search for the satisfactory outcomes to the frozen trials corresponding to each dangerous $(1,2)$-tree, which is possible in time polynomial in n, the number of vertices in G.

The time required to carry out the random trial f_i in Theorem 7, t_1, is replaced by $O(\log m \times 2^k \times (ed^3)^{\log m})$. This is the time required to calculate the conditional probabilities of each $(2,3)$-tree becoming dangerous in a trial f_{ij} if the remaining trials are carried out at random. Thus the running time of the deterministic algorithm that uses a single sweep is $O(n \times k^2 \times \log n \times 2^k \times (\frac{27}{8}ek^6)^{\log n} + n \times \sqrt{8}^{k^3 \log n})$. A second sweep reduces the running time, but we do not attempt to analyze it here. $\square$

5. Open problems. The combinatorial open problem is to determine $\mu(k)$ exactly. The approximation ratio of the problem is wide open. The best lower bound we have is two, given by the $\mathcal{NP}$-completeness proof.

REFERENCES

[1] N. Alon, G. Fertin, A. L. Liestman, T. C. Shermer, and L. Stacho, *Factor d-domatic colorings of graphs*, Discrete Math., 262 (2003), pp. 17–25.

[2] N. Alon and J. H. Spencer, *The Probabilistic Method*, 2nd ed., John Wiley & Sons, Inc., New York, 2000.

[3] R. C. Brigham and R. D. Dutton, *Factor domination in graphs*, Discrete Math., 86 (1990), pp. 127–136.

[4] P. Erdős and L. Lovász, *Problems and results on* 3*-chromatic hypergraphs and some related questions*, in Infinite and Finite Sets, Vol. II, Colloq. Math. Soc. Janos. Bolyai 10, North–Holland, Amsterdam, 1975, pp. 609–627.

[5] P. Erdős and J. Selfridge, *On a combinatorial game*, J. Combin. Theory Ser. A, 14 (1973), pp. 298–301.

[6] P. Hall, *On representation of subsets*, J. London Math. Soc., 10 (1935), pp. 26–30.

[7] U. Feige, M. M. Halldórsson, G. Kortsarz, and A. Srinivasan, *Approximating the domatic number*, SIAM J. Comput., 32 (2002), pp. 172–195.

[8] M. R. Garey and D. S. Johnson, *Computers and Intractability*, W. H. Freeman, San Francisco, 1979.

[9] M. Molloy and B. Reed, *Further algorithmic aspects of the local lemma*, in Proceedings of the 30th Annual ACM Symposium on the Theory of Computing, ACM, New York, 1998, pp. 524–529.

SIAM J. Discrete Math.
Vol. 21, No. 4, pp. 1083–1092

CYCLE SYSTEMS IN THE COMPLETE BIPARTITE GRAPH PLUS A ONE-FACTOR[*]

LIQUN PU[†], HAO SHEN[‡], JUN MA[‡], AND SAN LING[§]

Abstract. Let $K_{n,n}$ denote the complete bipartite graph with n vertices in each partite set and $K_{n,n}+I$ denote $K_{n,n}$ with a one-factor added. It is proved in this paper that there exists an m-cycle system of $K_{n,n}+I$ if and only if $n \equiv 1 \pmod 2$, $m \equiv 0 \pmod 2$, $4 \leq m \leq 2n$, and $n(n+1) \equiv 0 \pmod{m}$.

Key words. complete bipartite graph, one-factor, cycle system

AMS subject classification. 05C38

DOI. 10.1137/06065461X

1. Introduction.

Let G be a graph with vertex set $V(G)$ and edge set $E(G)$. An *m-cycle system* of G is a collection T of m-cycles such that each edge of G is contained in a unique m-cycle of T.

It is easy to get the necessary conditions for the existence of an m-cycle system of G:

$$\begin{cases} 3 \leq m \leq |V(G)|; \\ |E(G)| \equiv 0 \pmod{m}; \\ d(u) \equiv 0 \pmod{2} \text{ for each } u \in V(G), \end{cases}$$

where $d(u)$ denotes the number of edges incident with u in G.

Let K_n denote the complete graph of order n, and let $K_{x,y}$ denote a complete bipartite graph with partite sets of sizes x and y. For $G = K_n$ or $K_{n,n}$, let $G+I$ denote G with a one-factor added and $G-I$ denote G with a one-factor removed. The existence of m-cycle systems has been studied extensively, and the following results are known.

THEOREM 1.1 (see [1, 6]). *Let m and n be positive integers. Then there exists an m-cycle system of K_n if and only if $n \equiv 1 \pmod 2$, $3 \leq m \leq n$, and $n(n-1) \equiv 0 \pmod{2m}$.*

THEOREM 1.2 (see [7]). *Let $m \equiv 0 \pmod 2$ and $m \geq 4$. Then there exists an m-cycle system of $K_{x,y}$ if and only if $x, y \geq \frac{1}{2}m$, $x \equiv y \equiv 0 \pmod 2$, and $xy \equiv 0 \pmod{m}$.*

THEOREM 1.3 (see [5]). *Let n be an even integer and m be an integer in the range $3 \leq m \leq n$. Then there exists an m-cycle system of $K_n + I$ if and only if m divides $\frac{n^2}{2}$.*

[*] Received by the editors March 19, 2006; accepted for publication (in revised form) July 29, 2007; published electronically January 25, 2008. The research of Liqun Pu and San Ling was supported by NTU Research grant M58110003, Singapore.

http://www.siam.org/journals/sidma/21-4/65461.html

[†] Department of Mathematics, Zhengzhou University, Zhengzhou 450001, China, and School of Physical and Mathematical Sciences, Nanyang Technology University, Singapore 637616 (liqunpu@yahoo.com.cn).

[‡] Department of Mathematics, Shanghai Jiao Tong University, Shanghai 200240, China (haoshen@sjtu.edu.cn, mj904@sjtu.edu.cn).

[§] School of Physical and Mathematical Sciences, Nanyang Technology University, Singapore 637616 (lingsan@ntu.edu.sg).

THEOREM 1.4 (see [1, 6]). *Let m and n be positive integers. Then there exists an m-cycle system of $K_n - I$ if and only if $n \equiv 0 \pmod 2$, $3 \le m \le n$, and $n(n-2) \equiv 0 \pmod{2m}$.*

THEOREM 1.5 (see [2, 4]). *Let m and n be positive integers. Then there exists an m-cycle system of $K_{n,n} - I$ if and only if $n \equiv 1 \pmod 2$, $m \equiv 0 \pmod 2$, $4 \le m \le 2n$, and $n(n-1) \equiv 0 \pmod m$.*

In this paper, we study the existence and construction of m-cycle systems for the bipartite graph $K_{n,n} + I$. Since in $K_{n,n} + I$ there are $2n$ vertices, $n^2 + n$ edges, $\mathrm{d}(u) = n + 1$ for each vertex u, and m must be even, we have the following necessary conditions for the existence of an m-cycle system of $K_{n,n} + I$.

LEMMA 1.6. *If there exists an m-cycle system of $K_{n,n} + I$, then*

$$\begin{cases} n \equiv 1 \pmod 2, \\ m \equiv 0 \pmod 2 \text{ and } 4 \le m \le 2n, \\ n(n+1) \equiv 0 \pmod m. \end{cases}$$

The purpose of this paper is to prove that these conditions are also sufficient for the existence of an m-cycle system of $K_{n,n} + I$. This is an extension of the result in [4].

2. Construction techniques.

A *cycle on m vertices* is denoted by C_m. A C_n *in a graph with n vertices* is called a Hamilton cycle. If there exists an m-cycle system of G, then G is C_m-decomposable and is denoted by $C_m | G$.

In this section, we will provide some construction techniques for m-cycle systems of $K_{n,n} + I$. For our first construction, we need the following result.

LEMMA 2.1 (see [3]). *Let n be an integer, $n \ge 3$. Then there exists an n-cycle system of K_n if and only if $n \equiv 1 \pmod 2$.*

When m is even, we can construct m-cycle systems of $K_{\frac{1}{2}m,\frac{1}{2}m} + I$ by applying $\frac{1}{2}m$-cycle systems of $K_{\frac{1}{2}m}$.

THEOREM 2.2. *Let m be a positive integer such that $m \equiv 2 \pmod 4$ and $m \ge 6$. Then $C_m | K_{\frac{1}{2}m,\frac{1}{2}m} + I$.*

Proof. Let $V(K_{\frac{1}{2}m,\frac{1}{2}m}) = \{u_0, u_1, \dots, u_{\frac{1}{2}m-1}\} \cup \{v_0, v_1, \dots, v_{\frac{1}{2}m-1}\}$. Since $m \equiv 2 \pmod 4$ and $m \ge 6$, we have $\frac{1}{2}m \equiv 1 \pmod 2$ and $\frac{1}{2}m \ge 3$. Hence by Lemma 2.1, $K_{\frac{1}{2}m}$ has a $\frac{1}{2}m$-cycle system, denoted by T. Let $V(K_{\frac{1}{2}m}) = \{w_0, w_1, \dots, w_{\frac{1}{2}m-1}\}$.

For

$$C' = \left(w_{j_0}, w_{j_1}, w_{j_2}, w_{j_3}, \dots, w_{j_{\frac{1}{2}m-1}}\right) \in T,$$

let

$$C'^{1*} = \left(u_{j_0}, v_{j_0}, u_{j_1}, v_{j_1}, u_{j_2}, v_{j_2}, u_{j_3}, v_{j_3}, \dots, u_{j_{\frac{1}{2}m-1}}, v_{j_{\frac{1}{2}m-1}}\right)$$

and

$$C'^{2*} = \left(v_{j_0}, u_{j_0}, v_{j_1}, u_{j_1}, v_{j_2}, u_{j_2}, v_{j_3}, u_{j_3}, \dots, v_{j_{\frac{1}{2}m-1}}, u_{j_{\frac{1}{2}m-1}}\right).$$

For each

$$C = \left(w_{i_0}, w_{i_1}, w_{i_2}, w_{i_3}, \dots, w_{i_{\frac{1}{2}m-1}}\right) \in T \setminus \{C'\},$$

let

$$C^* = \left(u_{i_0}, v_{i_1}, u_{i_2}, v_{i_3}, \dots, u_{i_{\frac{1}{2}m-1}}, v_{i_0}, u_{i_1}, v_{i_2}, u_{i_3}, \dots, v_{i_{\frac{1}{2}m-1}}\right).$$

Let $T^* = \{C^* | C \in T \setminus \{C'\}\} \cup \{C'^{1*}, C'^{2*}\}$ and $I = \{u_i v_i | 0 \leq i \leq \frac{1}{2}m - 1\}$. Then T^* is an m-cycle system of $K_{\frac{1}{2}m,\frac{1}{2}m} + I$. □

Now for a positive integer n, let $D \subseteq Z_n$ and let $X(n; D)$ be a graph with vertex set $V(X(n;D)) = \{i_j | i \in Z_n, j \in Z_2\}$ and edge set $E(X(n;D)) = \{\{i_0, (i+d)_1\} | d \in D, i \in Z_n\}$. Clearly, $K_{n,n} = X(n; Z_n)$. The elements of D are called $(0,1)$-mixed differences. We say that $\{i_0, (i+d)_1\}$ is an edge of difference d.

Suppose that $C = ((i_1)_0, (i_2)_1, \ldots, (i_{m-1})_0, (i_m)_1)$ is a C_m in $X(n; D)$. For $x \in Z_n$, let $C + x = ((i_1+x)_0, (i_2+x)_1, \ldots, (i_{m-1}+x)_0, (i_m+x)_1)$. Obviously, $C + x$ is still a C_m. Let $(C) = \{C + x | x \in Z_n\}$. Here, (C) is called the orbit generated by C, and C is called a base cycle of (C).

In our proof, we denote the union of multisets by $\uplus$, for example, $\{1,1,2\} \uplus \{2,3\} = \{1,1,2,2,3\}$.

We use the difference method to give constructions of m-cycle systems of $X(n; D)$ which we need in this paper.

LEMMA 2.3. *For an even integer m, $m \geq 4$, $C_m | K_{m-1,m-1} + I$, where I is a one-factor of $K_{m-1,m-1}$.*

Proof. We view $K_{m-1,m-1}$ as $X(m-1; Z_{m-1})$ and $I = \{\{i_0, i_1\} | i \in Z_{m-1}\}$. Let $d_r \in Z_{m-1} \uplus \{0\}$ and

$$d_{r+1} = \begin{cases} r & \text{if} \quad 0 \leq r \leq \frac{1}{2}m - 1, \\ 0 & \text{if} \quad r = \frac{1}{2}m, \\ r - 1 & \text{if} \quad \frac{1}{2}m + 1 \leq r \leq m - 1. \end{cases}$$

Let $e_r = \sum_{i=1}^{r} (-1)^{i+1} d_i$ for $1 \leq r \leq m$. Then

$$e_r = e_{r-1} + (-1)^{r+1} d_r.$$

When $m \equiv 0 \pmod 4$,

$$e_i = \begin{cases} -\frac{i}{2} & \text{if} \quad i \equiv 0 \ (\text{mod } 2), 1 \leq i \leq \frac{1}{2}m; \\ \frac{i-1}{2} & \text{if} \quad i \equiv 1 \ (\text{mod } 2), 1 \leq i \leq \frac{1}{2}m; \\ -\frac{m+1-i}{2} & \text{if} \quad i \equiv 1 \ (\text{mod } 2), \frac{1}{2}m + 1 \leq i \leq m; \\ -(m - 1 - \frac{m-i}{2}) & \text{if} \quad i \equiv 0 \ (\text{mod } 2), \frac{1}{2}m + 1 \leq i \leq m. \end{cases}$$

When $m \equiv 2 \pmod 4$,

$$e_i = \begin{cases} -\frac{i}{2} & \text{if} \quad i \equiv 0 \ (\text{mod } 2), 1 \leq i \leq \frac{1}{2}m; \\ \frac{i-1}{2} & \text{if} \quad i \equiv 1 \ (\text{mod } 2), 1 \leq i \leq \frac{1}{2}m; \\ \frac{m-i}{2} & \text{if} \quad i \equiv 0 \ (\text{mod } 2), \frac{1}{2}m + 1 \leq i \leq m; \\ m - 1 - \frac{m+1-i}{2} & \text{if} \quad i \equiv 1 \ (\text{mod } 2), \frac{1}{2}m + 1 \leq i \leq m. \end{cases}$$

That is,

$$(1) \qquad e_i = e_{m+1-i} (\text{mod } (m-1)) \quad \text{for} \quad 1 \leq i \leq \frac{1}{2}m.$$

When $m \equiv 0 \pmod 4$, let C be the following closed trail:

$$((e_1)_0, (e_2)_1, (e_3)_0, \ldots, (e_{\frac{1}{2}m-2})_1, (e_{\frac{1}{2}m-1})_0, (e_{\frac{1}{2}m})_1, (e_{\frac{1}{2}m+1})_0, \ldots, (e_{m-1})_0, (e_m)_1).$$

By (1), C can also be written as

$$((e_1)_0,(e_2)_1,\ldots,(e_{\frac{1}{2}m-1})_0,(e_{\frac{1}{2}m})_1,(e_{\frac{1}{2}m})_0,(e_{\frac{1}{2}m-1})_1,\ldots,(e_2)_0,(e_1)_1).$$

The differences used in C are d_1, $d_2,\ldots,d_m$.

Since

$$0=e_1<e_3<\cdots<e_{\frac{1}{2}m-1}=\frac{1}{4}m-1$$

and

$$m-2=m-1+e_2>m-1+e_4>\cdots>m-1+e_{\frac{1}{2}m}=\frac{3}{4}m-1>0,$$

it follows that the vertices of C are distinct so that C is an m-cycle.

When $m\equiv 2 \pmod 4$, let C be the following closed trail:

$$((e_1)_0,(e_2)_1,(e_3)_0,\ldots,(e_{\frac{1}{2}m-1})_1,(e_{\frac{1}{2}m})_0,(e_{\frac{1}{2}m+1})_1,(e_{\frac{1}{2}m+2})_0,\ldots,(e_{m-1})_0,(e_m)_1).$$

By (1), C can also be written as

$$((e_1)_0,(e_2)_1,(e_3)_0,\ldots,(e_{\frac{1}{2}m-1})_1,(e_{\frac{1}{2}m})_0,(e_{\frac{1}{2}m})_1,(e_{\frac{1}{2}m-1})_0,(e_{\frac{1}{2}m-2})_1,\ldots,(e_2)_0,(e_1)_1).$$

The differences used in C are d_1, $d_2,\ldots,d_m$.

As before, it is easy to check that the vertices of C are distinct so that C is an m-cycle. Let $T=(C)$. Then T is an m-cycle system of $K_{m-1,m-1}+I$ and $C_m|K_{m-1,m-1}+I$. □

3. Cycle decomposition of $K_{n,n}+I$ with $\frac{1}{2}m<n<\frac{3}{2}m$. The main purpose of this section is to prove Theorem 3.4, which considers cycle decomposition of $K_{n,n}+I$ with $\frac{1}{2}m<n<\frac{3}{2}m$. Lemmas 3.1, 3.2, and 3.3 will be needed in the proof of Theorem 3.4. The following notation will appear in the three lemmas.

For any integer x, let

$$\varepsilon(x) = \begin{cases} 0 & \text{if} \quad x\equiv 0 \pmod 2,\\ 1 & \text{if} \quad x\equiv 1 \pmod 2.\end{cases}$$

LEMMA 3.1. *Let m and n be positive integers with $m\equiv 0 (\mathrm{mod}\ 2)$, $n\equiv 1(\mathrm{mod}\ 2)$, and $\frac{1}{2}m<n<\frac{3}{2}m$. Let $g=\gcd(m,n)>1$ and $n=s\frac{m}{g}-1$. Let $D=\{2,3,\ldots,\frac{m}{g},\frac{n}{g}+\frac{m}{2g}+1\}$. Then $C_m|X(n;D)$.*

Proof. Let $V(X(n;D))=\{i_j|i\in Z_n, j\in Z_2\}$. Let

$$d_i=\begin{cases} 0 & \text{if} \quad i\ =\ 0,\\ i+1 & \text{if} \quad 1\le i\le \frac{m}{g}-1,\\ \frac{n}{g}+\frac{m}{2g}+1 & \text{if} \quad i=\frac{m}{g}.\end{cases}$$

For $1\le i\le\frac{m}{g}$, let

$$\begin{cases} e_0=0,\\ e_i=e_{i-1}+(-1)^i d_i.\end{cases}$$

Then

$$e_i=\begin{cases} \frac{i}{2} & \text{if} \quad i\ \equiv\ 0 \pmod 2,\ 0\le i\le\frac{m}{g}-2,\\ -\frac{i+3}{2} & \text{if} \quad i\ \equiv\ 1 \pmod 2,\ 1\le i\le\frac{m}{g}-1,\\ \frac{n}{g} & \text{if} \quad i=\frac{m}{g}.\end{cases}$$

Let P be the trail of length $\frac{m}{g}$ given by

$$P = (e_0)_0, (e_1)_1, (e_2)_0, (e_3)_1, \ldots, (e_{\frac{m}{g}-2})_0, (e_{\frac{m}{g}-1})_1, (e_{\frac{m}{g}})_0.$$

The differences used in P are d_1, d_2, $d_3, \ldots, d_{\frac{m}{g}}$.

Since

$$0 = e_0 < e_2 < e_4 < \cdots < e_{\frac{m}{g}} = \frac{n}{g}$$

and

$$s\frac{m}{g} - 3 = n + e_1 > n + e_3 > n + e_5 > \cdots > n + e_{\frac{m}{g}-1} = s\frac{m}{g} - \frac{m}{2g} - 2,$$

the vertices of P are distinct so that P is a path. Moreover, the first and last vertices of P are the only ones which are congruent modulo $\frac{n}{g}$. It follows that

$$C = P \cup \left(P + \frac{n}{g}\right) \cup \left(P + \frac{2n}{g}\right) \cup \cdots \cup \left(P + \frac{(g-1)n}{g}\right)$$

is a C_m.

In C, each difference in D occurs exactly g times, and $\{i_0, (i+d)_1\}$ incident with edges of difference d are all congruent modulo $\frac{n}{g}$. Let $T = (C)$. It follows that T is an m-cycle system of $X(n; D)$ and $C_m | X(n; D)$. $\square$

LEMMA 3.2. *Let m and n be positive integers with $m \equiv 0 \pmod 4$, $n \equiv 1 \pmod 2$, and $\frac{1}{2}m < n < \frac{3}{2}m$. Let $g = \gcd(m, n) > 1$ and $n = s\frac{m}{g} - 1$. Let $D_l = \{(l-1)\frac{m}{g} + 1, (l-1)\frac{m}{g} + 2, \ldots, l\frac{m}{g} - 1, l\frac{m}{g}, (l-1)\frac{m}{g} + \frac{n}{g} + \frac{m}{2g} + \varepsilon(l)\} \setminus \{(l-2)\frac{m}{g} + \frac{m}{2g} + \frac{n}{g} + \varepsilon(l+1)\}$ for $2 \le l \le s$. Then $C_m | X(n; D_l)$.*

Proof. Let $V(X(n; D_l)) = \{i_j | i \in Z_n, j \in Z_2\}$. For $l \equiv 0 \pmod 2$, let

$$d_i = \begin{cases} 0 & \text{if} \quad i = 0, \\ (l-1)\frac{m}{g} + i & \text{if} \quad 1 \le i < \frac{n}{g} - \frac{m}{2g} + 1, \\ (l-1)\frac{m}{g} + i + 1 & \text{if} \quad \frac{n}{g} - \frac{m}{2g} + 1 \le i \le \frac{m}{g} - 1, \\ (l-1)\frac{m}{g} + \frac{n}{g} + \frac{m}{2g} & \text{if} \quad i = \frac{m}{g}. \end{cases}$$

For $1 \le i \le \frac{m}{g}$, let

$$\begin{cases} e_0 = 0, \\ e_i = e_{i-1} + (-1)^i d_i. \end{cases}$$

Then

$$e_i = \begin{cases} \frac{i}{2} & \text{if} \quad i \equiv 0 \pmod 2, 0 \le i \le \frac{n}{g} - \frac{m}{2g}, \\ -(l-1)\frac{m}{g} - \frac{i+1}{2} & \text{if} \quad i \equiv 1 \pmod 2, 0 \le i \le \frac{n}{g} - \frac{m}{2g}, \\ \frac{i}{2} + 1 & \text{if} \quad i \equiv 0 \pmod 2, \frac{n}{g} - \frac{m}{2g} + 1 \le i \le \frac{m}{g} - 1, \\ -(l-1)\frac{m}{g} - \frac{i+1}{2} & \text{if} \quad i \equiv 1 \pmod 2, \frac{n}{g} - \frac{m}{2g} + 1 \le i \le \frac{m}{g} - 1, \\ \frac{n}{g} & \text{if} \quad i = \frac{m}{g}. \end{cases}$$

Let P be the trail of length $\frac{m}{g}$ given by

$$P = (e_0)_0, (e_1)_1, (e_2)_0, (e_3)_1, \ldots, (e_{\frac{m}{g}-2})_0, (e_{\frac{m}{g}-1})_1, (e_{\frac{m}{g}})_0.$$

The differences used in P are $d_1, d_2, d_3, \ldots, d_{\frac{m}{g}}$.

Since

$$0 = e_0 < e_2 < e_4 < \cdots < e_{\frac{m}{g}} = \frac{n}{g}$$

and

$$(s-l+1)\frac{m}{g} - 2 = n + e_1 > n + e_3 > n + e_5 > \cdots > n + e_{\frac{m}{g}-1} = (s-l+1)\frac{m}{g} - \frac{m}{2g} - 1,$$

the vertices of P are distinct so that P is a path. Moreover, the first and last vertices are the only ones which are congruent modulo $\frac{n}{g}$. It follows that

$$C = P \cup \left(P + \frac{n}{g}\right) \cup \left(P + \frac{2n}{g}\right) \cup \cdots \cup \left(P + \frac{(g-1)n}{g}\right)$$

is a C_m.

In C, each difference in D occurs exactly g times, and $\{i_0, (i+d)_1\}$ incident with edges of difference d are congruent modulo $\frac{n}{g}$. Let $T = (C)$. It follows that T is an m-cycle system of $X(n; D_l)$ and $C_m | X(n; D_l)$ for l even.

For $l \equiv 1 \pmod 2$, let

$$d_i = \begin{cases} 0 & \text{if} \quad i = 0, \\ (l-1)\frac{m}{g} + i & \text{if} \quad 1 \leq i < \frac{n}{g} - \frac{m}{2g}, \\ (l-1)\frac{m}{g} + i + 1 & \text{if} \quad \frac{n}{g} - \frac{m}{2g} \leq i \leq \frac{m}{g} - 1, \\ (l-1)\frac{m}{g} + \frac{n}{g} + \frac{m}{2g} + 1 & \text{if} \quad i = \frac{m}{g}. \end{cases}$$

Then

$$e_i = \begin{cases} \frac{i}{2} & \text{if} \quad i \equiv 0 \pmod 2, 0 \leq i \leq \frac{n}{g} - \frac{m}{2g} - 1, \\ -(l-1)\frac{m}{g} - \frac{i+1}{2} & \text{if} \quad i \equiv 1 \pmod 2, 0 \leq i \leq \frac{n}{g} - \frac{m}{2g} - 1, \\ \frac{i}{2} & \text{if} \quad i \equiv 0 \pmod 2, \frac{n}{g} - \frac{m}{2g} \leq i \leq \frac{m}{g} - 1, \\ -(l-1)\frac{m}{g} - \frac{i+3}{2} & \text{if} \quad i \equiv 1 \pmod 2, \frac{n}{g} - \frac{m}{2g} \leq i \leq \frac{m}{g} - 1, \\ \frac{n}{g} & \text{if} \quad i = \frac{m}{g}. \end{cases}$$

Let P be the trail of length $\frac{m}{g}$ given by

$$P = (e_0)_0, (e_1)_1, (e_2)_0, (e_3)_1, \ldots, (e_{\frac{m}{g}-2})_0, (e_{\frac{m}{g}-1})_1, (e_{\frac{m}{g}})_0.$$

The differences used in P are $d_1, d_2, d_3, \ldots, d_{\frac{m}{g}}$.

Since

$$0 = e_0 < e_2 < e_4 < \cdots < e_{\frac{m}{g}} = \frac{n}{g}$$

and

$$(s-l+1)\frac{m}{g} - 2 = n + e_1 > n + e_3 > n + e_5 > \cdots > n + e_{\frac{m}{g}-1} = (s-l+1)\frac{m}{g} - \frac{m}{2g} - 2,$$

the vertices of P are distinct so that P is a path. Moreover, the first and last vertices are the only ones which are congruent modulo $\frac{n}{g}$. It follows that

$$C = P \cup \left(P + \frac{n}{g}\right) \cup \left(P + \frac{2n}{g}\right) \cup \cdots \cup \left(P + \frac{(g-1)n}{g}\right)$$

is a C_m.

In C, each difference in D occurs exactly g times, and $\{i_0, (i+d)_1\}$ incident with edges of difference d are congruent modulo $\frac{n}{g}$. Let $T = (C)$. It follows that T is an m-cycle system of $X(n; D_l)$ and $C_m | X(n; D_l)$ for l odd. □

LEMMA 3.3. *Let m and n be positive integers with $m \equiv 2 \pmod 4$, $n \equiv 1 \pmod 2$, and $\frac{1}{2}m < n < \frac{3}{2}m$. Let $g = \gcd(m, n) > 1$ and $n = s\frac{m}{g} - 1$. Let $D_l = \{(l-1)\frac{m}{g}+1, (l-1)\frac{m}{g}+2, \ldots, l\frac{m}{g}-1, l\frac{m}{g}, (l-1)\frac{m}{g}+\frac{n}{g}+\frac{m}{2g}+1\} \setminus \{(l-2)\frac{m}{g}+\frac{m}{2g}+\frac{n}{g}+1\}$ for $2 \le l \le s$. Then $C_m | X(n; D_l)$.*

Proof. Let $V(X(n; D)) = \{i_j | i \in Z_n, j \in Z_2\}$ and let

$$d_i = \begin{cases} 0 & \text{if} \quad i = 0, \\ (l-1)\frac{m}{g} + i & \text{if} \quad 1 \le i < \frac{n}{g} - \frac{m}{2g} + 1, \\ (l-1)\frac{m}{g} + i + 1 & \text{if} \quad \frac{n}{g} - \frac{m}{2g} + 1 \le i \le \frac{m}{g} - 1, \\ (l-1)\frac{m}{g} + \frac{n}{g} + \frac{m}{2g} + 1 & \text{if} \quad i = \frac{m}{g}. \end{cases}$$

For $1 \le i \le \frac{m}{g}$, let

$$\begin{cases} e_0 = 0, \\ e_i = e_{i-1} + (-1)^i d_i. \end{cases}$$

Then

$$e_i = \begin{cases} \frac{i}{2} & \text{if} \quad i \equiv 0 \pmod 2, 0 \le i \le \frac{n}{g} - \frac{m}{2g}, \\ -(l-1)\frac{m}{g} - \frac{i+1}{2} & \text{if} \quad i \equiv 1 \pmod 2, 0 \le i \le \frac{n}{g} - \frac{m}{2g}, \\ \frac{i}{2} & \text{if} \quad i \equiv 0 \pmod 2, \frac{n}{g} - \frac{m}{2g} + 1 \le i \le \frac{m}{g} - 1, \\ -(l-1)\frac{m}{g} - \frac{i+3}{2} & \text{if} \quad i \equiv 1 \pmod 2, \frac{n}{g} - \frac{m}{2g} + 1 \le i \le \frac{m}{g} - 1, \\ \frac{n}{g} & \text{if} \quad i = \frac{m}{g}. \end{cases}$$

Let P be the trail of length $\frac{m}{g}$ given by

$$P = (e_0)_0, (e_1)_1, (e_2)_0, (e_3)_1, \ldots, (e_{\frac{m}{g}-2})_0, (e_{\frac{m}{g}-1})_1, (e_{\frac{m}{g}})_0.$$

The differences used in P are $d_1, d_2, d_3, \ldots, d_{\frac{m}{g}}$.

Since

$$0 = e_0 < e_2 < e_4 < \cdots < e_{\frac{m}{g}} = \frac{n}{g}$$

and

$$(s-l+1)\frac{m}{g} - 2 = n + e_1 > n + e_3 > n + e_5 > \cdots > n + e_{\frac{m}{g}-1} = (s-l+1)\frac{m}{g} - \frac{m}{2g} - 2,$$

the vertices of P are distinct so that P is a path. Moreover, the first and last vertices are the only ones which are congruent modulo $\frac{n}{g}$. It follows that

$$C = P \cup \left(P + \frac{n}{g}\right) \cup \left(P + \frac{2n}{g}\right) \cup \cdots \cup \left(P + \frac{(g-1)n}{g}\right)$$

is a C_m. In C, each difference in D occurs exactly g times, and $\{i_0, (i+d)_1\}$ incident with edges of difference d are congruent modulo $\frac{n}{g}$. Let $T = (C)$. It follows that T is an m-cycle system of $X(n; D_l)$ and $C_m | X(n; D_l)$. □

With the above preparations, we now prove the following theorem.

THEOREM 3.4. *Let m be an even integer and n be an odd integer with $\frac{1}{2}m < n < \frac{3}{2}m$. Then there exists an m-cycle system of $K_{n,n}+I$ if and only if m divides n^2+n.*

Proof. The necessity is similar to that in Lemma 1.6; here we consider only the sufficiency. Let $g = \gcd\,(m,n)$. If $g=1$, then since $n(n+1) \equiv 0 \pmod{m}$ and $n+1 < 2m$, we have $n = m-1$. By Lemma 2.3, there exists an m-cycle system of $K_{m-1,m-1}+I$.

If $n \neq m-1$, then $g > 1$. Since $n(n+1) \equiv 0 \pmod{m}$, we have $n+1 = s\frac{m}{g}$. When $m \equiv 0 \pmod 4$, let

$$I = \left\{ \left\{ i_0, \left(i + (s-1)\frac{m}{g} + \frac{n}{g} + \frac{m}{2g} + \varepsilon(s) \right)_1 \right\} \middle| i \in Z_n \right\}.$$

We can put an additional difference $(s-1)\frac{m}{g} + \frac{n}{g} + \frac{m}{2g} + \varepsilon(s)$ on Z_n. Then

$$Z_n \uplus \left\{ (s-1)\frac{m}{g} + \frac{n}{g} + \frac{m}{2g} + \varepsilon(s) \right\} = \bigcup_{l=1}^{s-1} D_l \uplus D_s$$

where

$$D_1 \ = \left\{ 2, \ldots, \frac{m}{g}, \frac{n}{g} + \frac{m}{2g} + 1 \right\}$$

and

$$D_l \ = \left\{ (l-1)\frac{m}{g} + 1, (l-1)\frac{m}{g} + 2, \ldots, l\frac{m}{g} - 1, l\frac{m}{g}, (l-1)\frac{m}{g} + \frac{n}{g} + \frac{m}{2g} + \varepsilon(l) \right\}$$

$$\Big\backslash \left\{ (l-2)\frac{m}{g} + \frac{n}{g} + \frac{m}{2g} + \varepsilon(l+1) \right\}$$

for $2 \leq l \leq s$.

By Lemmas 3.1 and 3.2, there exists an m-cycle system of $K_{n,n}+I$. This completes this case.

When $m \equiv 2 \pmod 4$, let

$$I = \left\{ \left\{ i_0, \left((s-1)\frac{m}{g} + \frac{n}{g} + \frac{m}{2g} + 1 \right)_1 \right\} \middle| i \in Z_n \right\}.$$

We can put an additional difference $(s-1)\frac{m}{g} + \frac{n}{g} + \frac{m}{2g} + 1$ on Z_n. Then

$$Z_n \uplus \left\{ (s-1)\frac{m}{g} + \frac{n}{g} + \frac{m}{2g} + 1 \right\} = \bigcup_{l=1}^{s-1} D_l \uplus D_s,$$

where

$$D_1 \ = \left\{ 2, \ldots, \frac{m}{g}, \frac{n}{g} + \frac{m}{2g} + 1 \right\}$$

and

$$D_l \ = \left\{ (l-1)\frac{m}{g} + 1, (l-1)\frac{m}{g} + 2, \ldots, l\frac{m}{g} - 1, l\frac{m}{g}, (l-1)\frac{m}{g} + \frac{n}{g} + \frac{m}{2g} + 1 \right\}$$

$$\Big\backslash \left\{(l-2)\frac{m}{g}+\frac{n}{g}+\frac{m}{2g}+1\right\}$$

for $2 \le l \le s$.

By Lemmas 3.1 and 3.3, there exists an m-cycle system of $K_{n,n}+I$. This completes the proof. □

4. Main result.

Now we are in position to prove the main theorem of this paper.

THEOREM 4.1. *Let m be an even integer and n be an odd integer with $4 \le m \le 2n$. Then $K_{n,n}+I$ can be decomposed into cycles of length m if and only if m divides n^2+n.*

Proof. The necessity can be found in Lemma 1.6; we need only prove the sufficiency. For $\frac{1}{2}m \le n < \frac{3}{2}m$, we have proved the result in Theorems 2.2 and 3.4.

Now we need to consider the case $n \ge \frac{3}{2}m$. Let $n = qm+r$, where q is a positive integer and $\frac{1}{2}m \le r < \frac{3}{2}m$. Let $V(K_{n,n}) = \{v_1, v_2, \ldots, v_{qm+r}\} \cup \{u_1, u_2, \ldots, u_{qm+r}\}$, $V_i = \{v_{(i-1)m+j} | 1 \le j \le m\}$, and $U_i = \{u_{(i-1)m+j} | 1 \le j \le m\}$ for $1 \le i \le q$. Let $V_{q+1} = \{v_{qm+j} | 1 \le j \le r\}$, $U_{q+1} = \{u_{qm+j} | 1 \le j \le r\}$, and $I = \{u_i v_i | 1 \le i \le n\}$.

Let $H_{i,i}$ be a subgraph of $K_{m-1,m-1}+I$ induced by $(V_i \setminus \{v_{(i-1)m+1}\}) \cup (U_i \setminus \{u_{(i-1)m+1}\})$ for $1 \le i \le q$. Then $H_{i,i} \;=\; K_{m-1,m-1}+I_{i,i}$, where

$$I_{i,i} \;= \{u_{(i-1)m+r}v_{(i-1)m+r} | 2 \le r \le m\}.$$

By Lemma 2.3, $C_m | H_{i,i}$ for $1 \le i \le q$. Let $T_{i,i}$ be the m-cycle system of $H_{i,i}$.

Let $H_{i,j}$ be a subgraph of $K_{m,m}$ induced by $V_i \cup U_j$, where $1 \le i,j \le q$ and $i \ne j$. Then $H_{i,j} \;=\; K_{m,m}$. By Theorem 1.2, $C_m | H_{i,j}$. Let $T_{i,j}$ be the m-cycle system of $H_{i,j}$.

Let $H^i_{r+1,m}$ be a subgraph of $K_{r+1,m}$ induced by $(V_{q+1} \cup \{v_{(i-1)m+1}\}) \cup U_i$ for $1 \le i \le q$. Then $H^i_{r+1,m} \;=\; K_{r+1,m}$. By Theorem 1.2, $C_m | H^i_{r+1,m}$ for $1 \le i \le q$. Let $T^i_{r+1,m}$ be the m-cycle system of $H^i_{r+1,m}$.

Let $H^i_{m,r+1}$ be a subgraph of $K_{m,r+1}$ induced by $V_i \cup (U_{q+1} \cup \{u_{(i-1)m+1}\})$, where $1 \le i \le q$. Then $H^i_{m,r+1} \;=\; K_{m,r+1}$. By Theorem 1.2, $C_m | H^i_{m,r+1}$ for $1 \le i \le q$. Let $T^i_{m,r+1}$ be the m-cycle system of $H^i_{m,r+1}$.

Let $H_{r,r}$ be a subgraph of $K_{r,r}+I$ induced by $V_{q+1} \cup U_{q+1}$. Then $H_{r,r} = K_{r,r}+I_{r,r}$, where $I_{r,r} \;= \{u_{qm+j}v_{qm+j} | 1 \le j \le r\}$. By Theorem 1.2, $C_m | H_{r,r}$. Let $T_{r,r}$ be the m-cycle system of $H_{r,r}$.

Let

$$T = \bigcup_{1 \le i,j \le q} T_{i,j} \bigcup_{1 \le i \le q} \left(T^i_{m,r+1} \bigcup T^i_{r+1,m}\right) \bigcup T_{r,r}.$$

Then T is an m-cycle system of $K_{n,n}+I$. This concludes the proof. □

Acknowledgment. The authors would like to thank the referees for their helpful comments to improve the paper.

REFERENCES

[1] B. Alspach and H. Gavlas, *Cycle decompositions of K_n and $K_n - I$*, J. Combin. Theory Ser. B, 81 (2001), pp. 77–99.

[2] D. Archdeacon, M. Debowsky, J. Dinitz, and H. Gavlas, *Cycle systems in the complete bipartite graph minus a one factor*, Discrete Math., 284 (2004), pp. 37–43.

[3] B. Bollobas, *Modern Graph Theory*, Springer-Verlag, New York, 1998.

[4] J. Ma, L. Pu, and H. Shen, *Cycle decompositions of $K_{n,n} - I$*, SIAM J. Discrete Math., 20 (2006), pp. 603–609.

[5] M. Šajna, *Decomposition of the complete graph plus a 1-factor into cycles of equal length*, J. Combin. Des., 11 (2003), pp. 170–207.

[6] M. Šajna, *Cycle decompositions* III: *Complete graphs and fixed length cycles*, J. Combin. Des., 10 (2002), pp. 27–78.

[7] D. Sotteau, *Decompositions of $K_{m,n}(K^*_{m,n})$ into cycles (circuits) of length* $2k$, J. Combin. Theory Ser. B, 29 (1981), pp. 75–81.