SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 1–10

GIRTH, PEBBLING, AND GRID THRESHOLDS*

ANDRZEJ CZYGRINOW† AND GLENN HURLBERT†

Abstract. The pebbling number of a graph is the smallest number t such that from any initial configuration of t pebbles one can move a pebble to any prescribed vertex by a sequence of pebbling steps. It is known that graphs whose connectivity is high compared to their diameter have a pebbling number as small as possible. We will use the above result to prove two related theorems. First, answering a question of the second author, we show that there exist graphs of arbitrarily high constant girth and least possible pebbling number. In the second application, we prove that the product of two graphs of high minimum degree has a pebbling number equal to the number of vertices of the product. This shows that Graham's product conjecture is true in the case of high minimum degree graphs. In addition, we consider a probabilistic variant of the pebbling problem and establish a pebbling threshold result for products of paths. The last result shows that the sequence of paths satisfies the probabilistic analogue of Graham's product conjecture.

Key words. girth, pebbling, grids, threshold, connectivity

AMS subject classifications. 05D05, 05C35, 05A20

DOI. 10.1137/S0895480102416374

1. Introduction.

1.1. Pebbling. A *pebbling configuration* $\mathbf{C}$ on a graph G is a distribution of pebbles on the vertices of G. Given a particular configuration, one is allowed to move the pebbles about the graph according to this simple rule: if two or more vertices sit at vertex v, then one of them can be moved to a neighbor provided another is removed from v. Given a specific *root* vertex r, we say that $\mathbf{C}$ is *r-solvable* if one can move a pebble to r after a finite number of pebbling steps, and that $\mathbf{C}$ is *solvable* if it is r-solvable for every r. The pebbling number is the least number $\pi = \pi(G)$ such that every configuration of π pebbles on G is solvable.

The two most obvious pebbling facts are for complete graphs and paths. The pigeonhole principle implies that $\pi(K_n) = n$, and $\pi(P_n) = 2^{n-1}$ follows by induction or a simple weight function method. In fact, $\pi(G) \geq \min\{n(G), 2^{\mathrm{diam}(G)}\}$ for every G. Results for trees (a formula based on the maximum path partition of a tree in [13]; see also [3]), d-dimensional cubes Q^d (see [3]), and many other graphs with interesting properties are known (see surveys [11, 12]).

A probabilistic version of pebbling was introduced in [6]. Let $\mathcal{G} = (G_i)_{i=1}^{\infty}$ be a sequence of graphs with strictly increasing numbers of vertices $N = n(G_i)$. For a function $t = t(N)$ let $\mathbf{C}_t$ denote a configuration on G_i that is chosen uniformly at random from all configurations of t pebbles. The sequence $\mathcal{G}$ has a *pebbling threshold* $\tau = \tau(\mathcal{G})$ if, for every $\omega \gg 1$, (1) $\Pr[\mathbf{C}_t \text{ is solvable}] \to 0$ for $t = \tau/\omega$ and (2) $\Pr[\mathbf{C}_t \text{ is solvable}] \to 1$ for $t = \omega\tau$.

It was proved in [4] that the sequence of cliques has threshold $\tau(\mathcal{K}) = \Theta(N^{1/2})$. Bekmetjev et al. [1] showed recently that every graph sequence has a pebbling thresh-

*Received by the editors October 22, 2002; accepted for publication (in revised form) June 3, 2005; published electronically February 15, 2006.
http://www.siam.org/journals/sidma/20-1/41637.html

†Department of Mathematics and Statistics, Arizona State University, Tempe, AZ 85287-1804 (andrzej@math.la.asu.edu, hurlbert@asu.edu). The work of the second author was partially supported by National Security Agency grant MDA9040210095.

old. Bounds on the sequence of paths have undergone several improvements, the results of which are summarized as follows.

RESULT 1. *The pebbling threshold for the sequence of paths* $\mathcal{P} = (P_n)_{n=1}^{\infty}$ *satisfies*

$$\tau(\mathcal{P}) \in \Omega\left(N2^{c\sqrt{\lg N}}\right) \cap O\left(N2^{c'\sqrt{\lg N}}\right)$$

for every $c < 1/\sqrt{2}$ *and* $c' > 1$.

The lower bound is found in [1] and the upper bound is found in [9].

It is important to draw a distinction between this random pebbling model and the one in which each of t pebbles independently chooses uniformly at random a vertex on which to be placed. In the world of random graphs, the analogues of these two models are asymptotically equivalent. However, in the pebbling world they are vastly different. For example, in the independent model the pebbling threshold for paths is at most $N \lg N$ since, with more than that many pebbles, almost always every vertex already has a pebble on it.

1.2. Results. Pachter, Snevily, and Voxman [14] proved that every graph of diameter two on N vertices has a pebbling number either N or $N+1$. Graphs G with $\pi(G) = n(G)$ are called *Class* 0, and in [5] a characterization of diameter two Class 0 graphs was found and used to prove that diameter two graphs with connectivity at least 3 are Class 0. The authors also conjectured that every graph of fixed diameter and high enough connectivity was Class 0. This conjecture was proved by Czygrinow et al. [7] in the following result.

RESULT 2. *Let* d *be a positive integer and set* $k = 2^{2d+3}$. *If* G *is a graph of diameter at most* d *and connectivity at least* k, *then* G *is of Class* 0.

In this note, we present two applications of this result. Our first application concerns the following girth problem posed in [11].

QUESTION 3. *Does there exist a constant* C *such that if* G *is a connected graph on* n *vertices with* $girth(G) > C$, *then* $\pi(G) > n$?

We answer the above question in the negative. Let $g_0(n)$ denote the maximum number g such that there exists a graph G on at most n vertices with finite $girth(G) \geq g$ and $\pi(G) = n(G)$. That is, $g_0(n)$ is the highest girth, as a function of n, among all Class 0 graphs. It is easy to see that

$$g_0(n) \leq 1 + 2\lg n$$

(because the cycle on k vertices has a pebbling number at least $2^{\lfloor k/2 \rfloor}$—see [14]) and we prove the following lower bound.

THEOREM 4. *For all* $n \geq 3$ *we have*

$$g_0(n) \geq \lfloor \sqrt{(\lg n)/2 + 1/4} - 1/2 \rfloor \ .$$

We prove this theorem in section 2.1 using Result 2.

Our second application concerns the following conjecture of Graham (see [3]).

CONJECTURE 5. *Every pair of graphs* G *and* H *satisfy* $\pi(G \Box H) \leq \pi(G)\pi(H)$.

Here, the Cartesian product has vertices $V(G \Box H) = V(G) \times V(H)$ and edges $E(G \Box H) = \{u \times E(H)\}_{u \in V(G)} \cup \{E(G) \times v\}_{v \in V(H)}$. A number of theorems have been published in support of this conjecture, including the recent work of Herscovici [10] which verifies the case for all pairs of cycles. We show the following.

THEOREM 6. *Let G and H be connected graphs on n vertices with minimum degrees $\delta(G)$, $\delta(H)$ and let $\delta = \min\{\delta(G), \delta(H)\}$. If $\delta \geq 2^{12n/\delta+15}$, then $G\Box H$ is of Class 0.*

In particular, there is a constant c such that if $\delta > c\frac{n}{\lg n}$, then $G\Box H$ is of Class 0. We prove this in section 2.2, again using Result 2. As a corollary we obtain that Graham's conjecture is satisfied for graphs with minimum degree $\delta > c\frac{n}{\lg n}$.

COROLLARY 7. *Let G and H be as in Theorem 6, with $\delta \geq 2^{12n/\delta+15}$. Then $\pi(G\Box H) \leq \pi(G)\pi(H)$.*

Proof. We have $\pi(G\Box H) = n(G\Box H) = n(G)n(H) \leq \pi(G)\pi(H)$. □

Finally, in this paper we also consider the following probabilistic analogue of Graham's Conjecture 5, which we consider a correction of one from [11].

PROBLEM 8. *Let $\mathcal{G} = (G_n)_{n=1}^{\infty}$ and $\mathcal{H} = (H_n)_{n=1}^{\infty}$ be two graph sequences. Define the product sequence $\mathcal{G}\Box\mathcal{H} = (G_n\Box H_n)_{n=1}^{\infty}$. Find $\tau(\mathcal{G}\Box\mathcal{H})$.*

Let $N_1 = N(G_n)$, $N_2 = N(H_n)$ denote the number of vertices of graphs G_n and H_n from Problem 8. It would be interesting to determine for which sequences $\mathcal{G} = (G_n)_{n=1}^{\infty}$ and $\mathcal{H} = (H_n)_{n=1}^{\infty}$ we have

$$f(N_1N_2) \in O\Big(g(N_1)h(N_2)\Big), \tag{1.1}$$

where $f \in \tau(\mathcal{G}\Box\mathcal{H})$, $g \in \tau(\mathcal{G})$, and $h \in \tau(\mathcal{H})$. We call pairs of sequences which satisfy (1.1) *well behaved*. One might conjecture that all pairs of sequences are well behaved, but we believe counterexamples might exist.

We define the two-dimensional grid $P_n^2 = P_n\Box P_n$, and in general the d-dimensional grid $P_n^d = P_n\Box P_n^{d-1}$. It is easy to show that $P_n^d = P_n^{\alpha}\Box P_n^{\beta}$ for all α and β for which $\alpha + \beta = d$. If we denote $\mathcal{P}^d = (P_n^d)_{n=1}^{\infty}$, then we have $\mathcal{P}^d = \mathcal{P}^{\alpha}\Box\mathcal{P}^{\beta}$. Thus, for example, in light of Result 1, the truth of (1.1) would imply that

$$\tau(\mathcal{P}^2) \in O\left(\left(\sqrt{N}2^{c'\sqrt{\lg\sqrt{N}}}\right)^2\right) = O\left(N2^{c'\sqrt{2\lg N}}\right).$$

Here we prove the following stronger theorem.

THEOREM 9. *Let $\mathcal{P}^d = (P_n^d)_{n=1}^{\infty}$ be the sequence of d-dimensional grids, where $P_n^d = (P_n)^d$ is the Cartesian product of d paths on n vertices each, and let $N = n^d$ be the number of vertices of $\mathcal{P}_n^d$. Then*

$$\tau(\mathcal{P}^d) \subseteq \Omega\Big(N2^{c_d(\lg N)^{1/(d+1)}}\Big) \cap O\Big(N2^{c'_d(\lg N)^{1/(d+1)}}\Big)$$

for all $c_d < 2^{-d/(d+1)}$ and $c'_d > d+1$.

This verifies (1.1) in the case of grids.

COROLLARY 10. *Let α, β be any pair of positive integers; then for $\mathcal{G} = \mathcal{P}^{\alpha}$ and $\mathcal{H} = \mathcal{P}^{\beta}$, (1.1) holds.*

Proof. Indeed, if $g \in \tau(\mathcal{G})$ and $h \in \tau(\mathcal{H})$, then Theorem 9 says that

$$\begin{aligned} g(N^{\overline{\alpha}})h(N^{\overline{\beta}}) &\in \Omega\Big(N^{\overline{\alpha}}2^{c_\alpha(\lg N^{\overline{\alpha}})^{1/(\alpha+1)}}N^{\overline{\beta}}2^{c_\beta(\lg N^{\overline{\beta}})^{1/(\beta+1)}}\Big) \\ &\subseteq \Omega\Big(N2^{c(\lg N)^{1/(\gamma+1)}}\Big) \\ &\subseteq \Omega\Big(N2^{c(\lg N)^{1/(d/2+1)}}\Big), \end{aligned}$$

for some c, where $\gamma = \min\{\alpha, \beta\}$, $d = \alpha + \beta$, $\overline{\alpha} = \alpha/d$, and $\overline{\beta} = \beta/d$. On the other hand, Theorem 9 also says that

$$\tau(\mathcal{P}^{\alpha+\beta}) = \tau(\mathcal{P}^d) \in O\left(N2^{c'_d(\lg N)^{1/(d+1)}}\right),$$

which is asymptotically smaller. □

We prove Theorem 9 in section 2.3.

2. Proofs.

2.1. Proof of Theorem 4. We will make use of Mader's theorem (see [8]) below.

RESULT 11. *Every graph having average degree at least $\bar{d}$ has a subgraph of connectivity at least $\lfloor \bar{d}/4 \rfloor$.*

We will also make use of the following result from [2, Chapter III, Theorem 1.1].

RESULT 12. *For any $g \geq 3$ and $\delta \geq 3$ there exists some graph H with girth at least g, minimal degree at least δ, and no more than $(2\delta)^g$ vertices.*

Proof of Theorem 4. Set $\delta = 2^{2g+1}$ and $n = 2^{2g(g+1)}$; then $g = \lfloor \sqrt{(\lg n)/2 + 1/4} - 1/2 \rfloor$. Let H be a graph guaranteed to exist by Result 12. By Result 11, H has some subgraph, F say, which is 2^{2g-1}-connected; clearly, F also has girth at least g. Now let $\hat{F}$ be an edge-maximal graph on the same vertices as F such that F is a subgraph of $\hat{F}$ and $\hat{F}$ has girth at least g. $\hat{F}$ can have diameter no more than $g-2$, for if there existed vertices x and y in $\hat{F}$ such that the shortest path between x and y had length $g-1$ or more, adding the edge xy to $\hat{F}$ would give a graph of girth g or more, contradicting maximality. Therefore $\hat{F}$ has diameter at most $g-2$ and is 2^{2g-1}-connected, so by Result 2 it is of Class 0, and it has no more than $(2\delta)^g = 2^{2g(g+1)}$ vertices. □

2.2. Proof of Theorem 6. Theorem 6 follows from the following two lemmas and Result 2.

LEMMA 13. *Let G be a connected graph on n vertices with minimum degree δ. Then the diameter of G is at most $3\frac{n}{\delta} + 3$.*

Proof. Fix two vertices x, y in G and consider the shortest path $x = x_1, \ldots, x_k = y$ between x and y. Let $i = \lfloor \frac{k-1}{3} \rfloor$. Then $x_1, x_4, x_7, \ldots, x_{3i+1}$ must have disjoint neighborhoods, and thus $i(\delta+1) \leq n$, which yields $\frac{k-3}{3} \leq \lfloor \frac{k-1}{3} \rfloor = i \leq \frac{n}{\delta+1}$ such that $k < \frac{3n}{\delta+1} + 3 \leq \frac{3n}{\delta} + 3$. □

The next lemma was proved by Czygrinow and Kierstead. We reproduce the proof here.

LEMMA 14. *For connected graphs G and H, the product $G\Box H$ has connectivity $\kappa(G\Box H) \geq \min\{\delta(G), \delta(H)\}$.*

Proof. Set $\delta = \min\{\delta(G), \delta(H)\}$. Let $v_1 = (g, h_1), v_2 = (g, h_2), \ldots, v_\delta = (g, h_\delta)$, $w_1 = (g_1, h), w_2 = (g_2, h), \ldots, w_\delta = (g_\delta, h)$ be distinct vertices (other than perhaps $v_1 = w_1$) in $G\Box H$ that satisfy

$$dist_G(g_i, g) \leq dist_G(g_{i+1}, g) \tag{2.1}$$

and

$$dist_H(h_i, h) \leq dist_H(h_{i+1}, h) \tag{2.2}$$

for $i = 1, \ldots, \delta - 1$. We shall construct vertex-disjoint paths $P_1, \ldots, P_\delta$ such that P_i connects v_i with w_i. Construct P_1 as follows: let $g_1\bar{g}(1)\ldots\bar{g}(k)g$ be any shortest

path in G connecting g_1 with g and let $h\bar{h}(1)\dots\bar{h}(l)h_1$ be any shortest path in H connecting h with h_1. Then P_1 is the following path:

$$w_1 = (g_1, h)(g_1, \bar{h}(1))\dots(g_1, h_1)(\bar{g}(1), h_1)\dots(g, h_1) = v_1.$$

Delete v_1 and w_1 and construct $P_2, \dots, P_\delta$ inductively. We claim that $P_2, \dots, P_\delta$ are vertex-disjoint with P_1. Indeed, suppose that $V(P_j) \cap V(P_1) \neq \emptyset$ for some $j = 2, \dots, \delta$. There are two similar cases to consider. First, suppose that $(g_j, f) \in V(P_j) \cap V(P_1)$. Since $g_j \neq g_1$, $f = h_1$ and $g_j = \bar{g}(i)$ for some $i = 1, \dots, k$. Then, however,

$$dist_G(g_j, g) < dist_G(g_1, g),$$

contradicting (2.1). Similarly, if $(f, h_j) \in V(P_j) \cap V(P_1)$, then $f = g_1$ and $h_j = \bar{h}(i)$ for some $i = 1, \dots, l$, which implies that

$$dist_H(h_j, h) < dist_H(h_1, h),$$

contradicting (2.2).

By induction, paths $P_1, \dots, P_\delta$ are vertex-disjoint. Now, for any two distinct vertices $v = (g, \tilde{h}), w = (\tilde{g}, h) \in V(G\Box H)$, let $v_1 = (g, h_1), v_2 = (g, h_2), \dots, v_\delta = (g, h_\delta)$ be neighbors of v in the H-dimension, and let $w_1 = (g_1, h), w_2 = (g_2, h), \dots, w_\delta = (g_\delta, h)$ be neighbors of w in the G-dimension ordered according to (2.1) and (2.2). By the previous argument we can find vertex-disjoint paths $P_1, \dots, P_\delta$ connecting the v_is with the w_js. These paths now can be used to connect v with w by δ internally vertex-disjoint paths. Indeed, if any of the paths contains v or w, then it yields a shorter path between v and w which is disjoint with other paths. Therefore the connectivity of $G\Box H$ is at least δ. □

Proof of Theorem 6. By Lemma 13, the diameter d of $G\Box H$ is at most $6\frac{n}{\delta} + 6$, and by Lemma 14, the connectivity k of $G\Box H$ is at least δ. Since $\delta \geq 2^{12n/\delta+15}$ the assumptions of Result 2 are satisfied and so $G\Box H$ is of Class 0. □

2.3. Proof of Theorem 9. Throughout, we let $N = n^d$. Also, we define $\left\langle {a \atop b} \right\rangle = \binom{a+b-1}{b}$. Note that $\left\langle {a \atop b} \right\rangle$ is the number of ways to place b unlabeled balls into a labeled urns. For our purposes, it equals the number of configurations of b pebbles on a graph of a vertices. We will also use the fact that $\left\langle {a \atop b} \right\rangle$ counts the number of points in $\mathbb{Z}^a$ whose coordinates are nonnegative and sum to b.

We begin by proving that a configuration with relatively few pebbles almost always has no vertices having a huge number of pebbles. For natural numbers a and b, define $a^{\underline{b}} = a!/(a-b)!$. For a configuration $\mathbf{C}$ of pebbles on a graph let $\mathbf{C}(v)$ denote the number of pebbles on vertex v.

LEMMA 15. *Let $s \gg 1$ and $t = sN$. Let $\mathbf{C}$ be a random configuration of t pebbles on the vertices of P_n^d, and let $p = (1+\epsilon)s \ln N$ for some $\epsilon > 0$. Then*

$$\Pr[\mathbf{C}(v) < p \text{ for all } v] \to 1 \text{ as } n \to \infty .$$

Proof. Let q be the probability that the vertex v satisfies $\mathbf{C}(v) \geq p$. Then q is at most

$$\frac{\left\langle {N \atop t-p} \right\rangle}{\left\langle {N \atop t} \right\rangle} = \frac{t^{\underline{p}}}{(N+t-1)^{\underline{p}}}$$

$$< \left(\frac{t}{N+t-1}\right)^p$$
$$= \left(1 - \frac{1-1/N}{s+1-1/N}\right)^p$$
$$\leq e^{-p(1-1/N)/(s+1-1/N)} .$$

Hence, the probability that some vertex v satisfies $\mathbf{C}(v) \geq p$ is at most

$$Ne^{-p(1-1/N)/(s+1-1/N)} \;=\; e^{\ln N(1-\epsilon s+[(1+\epsilon)s-1]/N)/(s+1-1/N)} \to 0$$

as $n\to\infty$. Therefore the probability that every vertex v satisfies $\mathbf{C}(v) < p$ tends to 1 as $n\to\infty$. □

Next we show that a configuration with relatively few pebbles almost always has some large hole with no pebbles in it. For any set S of vertices, denote by $\mathbf{C}(S)$ the number of pebbles on its vertices.

LEMMA 16. *Let $N = n^d$, $0 < c < 2^{-d/(d+1)}$, $u = c(\lg N)^{1/(d+1)}$, $s = 2^u$, and $t = \lfloor sN \rfloor$. Write $c = ((1-\epsilon)/(2+\delta)^d)^{1/(d+1)}$ for some $\epsilon, \delta > 0$, and set $m = \lfloor (2+\delta)u \rfloor$, $M = m^d$, and $k = \lfloor n/m \rfloor^d$. Let $B_1, \ldots, B_k$ be a collection of k pairwise disjoint blocks of vertices of P_n^d, each having every side of length m. Let $\mathbf{C}$ be a random configuration of t pebbles on the vertices of P_n^d. Then*

$$\Pr[\mathbf{C}(B_h) = 0 \text{ for some } h] \to 1 \text{ as } n\to\infty .$$

Proof. The second moment method applies. Let X_h be the indicator variable for the event that the block B_h contains no pebbles, and let $X = \sum_{h=1}^k X_h$. Then Chebyshev's inequality yields

$$\Pr[X = 0] \;\leq\; \frac{\mathrm{var}[X]}{\mathbf{E}[X]^2} ,$$

and

$$\mathrm{var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2$$
$$= \sum_{h,j} \mathbf{E}[X_h X_j] - \sum_{h,j} \mathbf{E}[X_h]\mathbf{E}[X_j]$$
$$\leq \sum_h \mathbf{E}[X_h^2] ,$$

since $\mathbf{E}[X_h X_j] \leq \mathbf{E}[X_h]\mathbf{E}[X_j]$ for $h \neq j$. Hence,

$$\mathrm{var}[X] \;\leq\; \sum_h \mathbf{E}[X_h^2] \;=\; \sum_h \mathbf{E}[X_h] \;=\; \mathbf{E}[X] .$$

Moreover, we have

$$\mathbf{E}[X] = k\left\langle {N-M \atop t} \right\rangle \Big/ \left\langle {N \atop t} \right\rangle$$

$$
\begin{aligned}
&= \left\lfloor \frac{n}{m} \right\rfloor^d \frac{(N-1)^{\underline{M}}}{(N+t-1)^{\underline{M}}} \\
&\geq \left(\frac{n}{m} - 1 \right)^d \left(\frac{N-M}{N+t-M} \right)^M \\
&\gtrsim \left(\frac{N}{M} \right) \left(\frac{N-M}{(s+1)N-M} \right)^M \\
&> \left(\frac{N}{M(s+1)^M} \right) \left(1 - \frac{M}{N} \right)^M \\
&\sim \frac{N}{M(s+1)^M} \\
&\sim \frac{N}{m^d s^{m^d(1+o(1))}} \\
&= \frac{N}{m^d 2^{u^{d+1}(2+\delta)^d(1+o(1))}} \\
&= \frac{N}{m^d N^{(1-\epsilon)(1+o(1))}} \\
&\to \infty \ .
\end{aligned}
$$

Hence $\Pr[X=0] \leq \mathrm{var}[X]/\mathbf{E}[X]^2 \leq 1/\mathbf{E}[X] \to 0$ as $n \to \infty$. □

The following lemma records the structure of the d-dimensional grid in order to keep track of the results of pebbling steps.

LEMMA 17. *For any intervals $I_1, \ldots, I_d$ in $\mathbb{Z}$ such that each I_j contains r integers, let $\mathbf{B} = I_1 \times \cdots \times I_d \subseteq \mathbb{Z}^d$, and for $i > 0$, let $\mathcal{S}_i$ be the set of points in $\mathbb{Z}^d$ having distance i from $\mathbf{B}$, where distance between a pair of points in $\mathbb{Z}^d$ is defined by the sum of the absolute values of the differences of their coordinates. Then*

$$R_i := |\mathcal{S}_i| \leq \sum_{1 \leq j \leq d} \binom{d}{j} 2^j r^{d-j} \binom{i-1}{j-1}.$$

Proof. We partition $\mathbb{Z}^d$ according to the number j of coordinates in which a given point differs from its nearest neighbor in $\mathbf{B}$. Given a fixed j, there are $\binom{d}{j}$ ways to pick which j coordinates to change, each of the changed coordinates can be to either side of $\mathbf{B}$, giving 2^j possibilities, and there are r ways to pick each unchanged coordinate, giving r^{d-j} possibilities. Given this information, we can specify an element of $\mathcal{S}_i$ by specifying a j-tuple of positive integers with sum i, which can be done in $\left\langle {j \atop i-j} \right\rangle = \binom{i-1}{j-1}$ ways. □

Finally, our proof of Theorem 9 in the case of the lower bound will use this technical lemma to bound the number of pebbles that can reach the empty hole.

LEMMA 18. $\sum_{i=1}^{nd} \binom{i-1}{j-1} 2^{-i} < 1$.

Proof. It is straightforward to use generating functions or induction to prove $\sum_{i \geq 1} \binom{i-1}{j-1} 2^{-i} = 1$. □

Turning to the case of the upper bound, we show that almost every configuration with relatively many pebbles fills every reasonably large block with plenty of pebbles.

LEMMA 19. *Let $N = n^d$, $c' = d + 1 + \epsilon$ for some $\epsilon > 0$, $u' = c'(\lg N)^{1/(d+1)}$, $s' = 2^{u'}$, $t' = \lceil s'N \rceil$, $m' = \lceil (\frac{d+1}{c'})^{1/d} (\lg N)^{1/(d+1)} \rceil$, $M' = (m')^d$, and $k' = \lceil n/m' \rceil^d$. Let $B'_1, \ldots, B'_{k'}$ be a collection of k' blocks, each having every side of length m', that cover the vertices of P_n^d. Let $\mathbf{C}$ be a random configuration of t' pebbles on the vertices of P_n^d. Then*

$$\Pr[\mathbf{C}(B'_f) \geq M' 2^{dm'} \text{ for all } f] \to 1 \text{ as } n \to \infty .$$

Proof. Define Z_f to be the event that block B'_f contains fewer than $M^* = M'2^{dm'}$ pebbles and approximate the probability $\Pr[\cup_{f=1}^{k'} Z_f]$ by

$$\Pr[\cup_{f=1}^{k'} Z_f] \;\leq\; k' \sum_{f=0}^{M^*-1} \left\langle {M' \atop f} \right\rangle \left\langle {N - M' \atop t' - f} \right\rangle \Big/ \left\langle {N \atop t'} \right\rangle .$$

Now use the estimate

$$\left\langle {N - M' \atop t' - f} \right\rangle \;\leq\; \left(\frac{N}{N + t'} \right)^{M'} \left\langle {N \atop t'} \right\rangle$$

to obtain

$$\Pr[\cup Z_f] \;\leq\; k' \left(\frac{N}{N + t'} \right)^{M'} \sum_{f=0}^{M^*-1} \left\langle {M' \atop f} \right\rangle .$$

Then use the upper bound

$$\sum_{f=0}^{M^*-1} \left\langle {M' \atop f} \right\rangle \;=\; \sum_{f=0}^{M^*-1} \left\langle {f+1 \atop M'-1} \right\rangle \;=\; \sum_{j=1}^{M^*} \left\langle {j \atop M'-1} \right\rangle \;=\; \left\langle {M^* \atop M'} \right\rangle \;\leq\; M^{*M'}$$

to obtain

$$\begin{aligned}
\Pr[\cup Z_f] &\leq k' \left(\frac{N}{N + t'} \right)^{M'} M^{*M'} \\
&\lesssim \frac{N}{M'} \left(\frac{M' 2^{dm'}}{s'} \right)^{M'} \\
&= \frac{1}{M'} 2^{\lg N - M'(u' - \lg M' - dm')} \\
&= \frac{1}{M'} 2^{\lg N - (1+d)\lg N + o(\lg N) + d(\frac{1+d}{c'})^{\frac{d+1}{d}} \lg N} \\
&= \frac{1}{M' N^{d - d(\frac{1+d}{c'})^{\frac{d+1}{d}} - o(1)}} \\
&\to 0 .
\end{aligned}$$

Thus, almost surely, every f satisfies $\mathbf{C}(B'_f) \geq M'2^{dm'}$. ∎

Proof of Theorem 9. We begin with the lower bound. Given $N = n^d$ and $0 < c < 2^{-d/(d+1)}$, we write $c = ((1-\epsilon)/(2+\delta)^d)^{1/(d+1)}$ for some $\epsilon, \delta > 0$, and set $u = c(\lg N)^{1/(d+1)}$, $s = 2^u$, $t = \lfloor sN \rfloor$, $m = \lfloor (2+\delta)u \rfloor$, $M = m^d$, and $k = \lfloor n/m \rfloor^d$. Let $B_1, \ldots, B_k$ be a collection of k pairwise disjoint blocks of vertices of P_n^d, each having every side of length m. Let $\mathbf{C}$ be a random configuration of t pebbles on the vertices of P_n^d. By Lemma 16 we know that, almost surely, some block B_h has no pebbles on its vertices. By Lemma 15 we know that, almost surely, no vertex has more that p pebbles on it, where $p = (1+\epsilon)s \ln N$ for some $\epsilon > 0$.

Let $\overline{B}_h$ be the boundary of B_h. Any vertex v with $\mathbf{C}(v)$ pebbles on it can contribute at most $\mathbf{C}(v)/2^i$ pebbles to $\overline{B}_h$, where i is the distance from v to $\overline{B}_h$. Also, the number of vertices of $P_n^d - B_h$ at distance i from $\overline{B}_h$ is at most R_i. Thus, according to Lemmas 17 and 18, the number of pebbles that can be amassed on $\overline{B}_h$ via pebbling steps almost surely is less than or equal to

$$\begin{aligned}
\sum_{i=1}^{nd} pR_i/2^i &\le \sum_{i=1}^{nd} p \sum_{j=1}^{d} \binom{d}{j} 2^j m^{d-j} \binom{i-1}{j-1} 2^{-i} \\
&\le p \sum_{j=1}^{d} \binom{d}{j} 2^j m^{d-j} \sum_{i=1}^{nd} \binom{i-1}{j-1} 2^{-i} \\
&< p \sum_{j=1}^{d} \binom{d}{j} 2^j m^{d-j} \\
&< p(m+2)^d \\
&\ll 2^{m/2} \ .
\end{aligned}$$

The last line holds because the dominant term in $p(m+2)^d$ is 2^u, and we have $m = \lfloor (2+\delta)u \rfloor$. Therefore, almost surely, too few vertices are amassed on $\overline{B}_h$ to be able to move a single pebble to the center of B_h. This shows that $\tau(\mathcal{P}^d) \in \Omega(sN)$, as required.

Next we prove the upper bound. Given $N = n^d$ and $c' = d + 1 + \epsilon$ for some $\epsilon > 0$, set $u' = c'(\lg N)^{1/(d+1)}$, $s' = 2^{u'}$, $t' = \lceil s'N \rceil$, $m' = \lceil (\frac{d+1}{c'})^{1/d} (\lg N)^{1/(d+1)} \rceil$, $M' = (m')^d$, and $k' = \lceil n/m' \rceil^d$. Let $B'_1, \ldots, B'_{k'}$ be a collection of k' blocks, each having every side of length m', that cover the vertices of P_n^d. Let $\mathbf{C}$ be a random configuration of t' pebbles on the vertices of P_n^d. Then Lemma 19 states that, almost surely, every block B'_f has at least $M' 2^{dm'}$ pebbles. Since (see [6]) every graph G is solvable by $n(G)2^{\mathrm{diam}(G)}$ pebbles, any given vertex v in P_n^d almost surely is solvable by the pebbles in the block B'_f which contains v. This shows that $\tau(\mathcal{P}^d) \in O(s'N)$, as required. □

3. Remarks. Let $l = l(n)$ and $d = d(n)$, and denote by $\mathcal{P}_l^d$ the sequence of graphs $(P_{l(n)}^{d(n)})_{n=1}^{\infty}$, where $P_l^d = (P_l)^d$. For $l(n) = 2$, $\mathcal{P}_l^n = \mathcal{Q}$, which can be shown to have a threshold asymptotically less than N.

We conjecture that the same result holds for all fixed l.

CONJECTURE 20. *Let $\mathcal{P}_l$ denote the graph sequence $(P_l^n)_{n=1}^{\infty}$. Then for fixed l we have $\tau(\mathcal{P}_l) \in o(N)$.*

In contrast, we have proved that $\tau(\mathcal{P}^d) \in \omega(N)$ for fixed d. Thus we believe there

should be some relationship between two functions $l = l(n)$ and $d = d(n)$, both of which tend to infinity, for which the sequence $\mathcal{P}_l^d$ has threshold on the order of N.

PROBLEM 21. *Denote by $\mathcal{P}^d$ the graph sequence $(P_n^{d(n)})_{n=1}^{\infty}$. Find a function $d = d(n) \to \infty$ for which $\tau(\mathcal{P}^d) = \Theta(N)$. In particular, how does d compare to n?*

Acknowledgment. The authors thank one of the referees for extensive assistance in simplifying the paper.

REFERENCES

[1] A. Bekmetjev, G. Brightwell, A. Czygrinow, and G. Hurlbert, *Thresholds for families of multisets, with an application to graph pebbling*, Discrete Math., 269 (2003), pp. 21–34.
[2] B. Bollobas, *Extremal Graph Theory*, Academic Press, London, New York, 1978.
[3] F. R. K. Chung, *Pebbling in hypercubes*, SIAM J. Discrete Math., 2 (1989), pp. 467–472.
[4] T. Clarke, *Pebbling on Graphs*, Master's thesis, Arizona State University, Tempe, AZ, 1996.
[5] T. Clarke, R. Hochberg, and G. H. Hurlbert, *Pebbling in diameter two graphs and products of paths*, J. Graph Theory, 25 (1997), pp. 119–128.
[6] A. Czygrinow, N. Eaton, G. Hurlbert, and P. M. Kayll, *On pebbling threshold functions for graph sequences*, Discrete Math., 247 (2002), pp. 93–105.
[7] A. Czygrinow, G. Hurlbert, H. Kierstead, and W. T. Trotter, *A note on graph pebbling*, Graphs Combin., 18 (2002), pp. 219–225.
[8] R. Diestel, *Graph Theory*, Springer-Verlag, New York, 1997.
[9] A. Godbole, M. Jablonski, J. Salzman, and A. Wierman, *An improved upper bound for the pebbling threshold of the n-path*, Discrete Math., 275 (2004), pp. 367–373.
[10] D. Herscovici, *Graham's pebbling conjecture on products of cycles*, J. Graph Theory, 42 (2003), pp. 141–154.
[11] G. Hurlbert, *A survey of graph pebbling*, Congr. Numer., 139 (1999), pp. 41–64.
[12] G. Hurlbert, *Recent Progress in Graph Pebbling*, Graph Theory Notes of New York, to appear.
[13] D. Moews, *Pebbling graphs*, J. Combin. Theory Ser. B, 55 (1992), pp. 244–252.
[14] L. Pachter, H. S. Snevily, and B. Voxman, *On pebbling graphs*, Congr. Numer., 107 (1995), pp. 65–80.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 11–25

COMPUTING OPTIMAL MORSE MATCHINGS*

MICHAEL JOSWIG† AND MARC E. PFETSCH‡

Abstract. Morse matchings capture the essential structural information of discrete Morse functions. We show that computing optimal Morse matchings is $\mathcal{NP}$-hard and give an integer programming formulation for the problem. Then we present polyhedral results for the corresponding polytope and report on computational results.

Key words. discrete Morse function, Morse matching

AMS subject classifications. Primary, 90C27; Secondary, 06A07, 52B99, 57Q05, 57R70

DOI. 10.1137/S0895480104445885

1. Introduction. Discrete Morse theory was developed by Forman [8, 10] as a combinatorial analog to the classical smooth Morse theory. Applications to questions in combinatorial topology and related fields are numerous: e.g., Babson et al. [3], Forman [9], Shareshian [30], Batzies and Welker [4], and Jonsson [19].

It turns out that the topologically relevant information of a discrete Morse function f on a simplicial complex can be encoded as a (partial) matching in its Hasse diagram (considered as a graph), the *Morse matching* of f. A matching in the Hasse diagram is Morse if it satisfies a certain, entirely combinatorial acyclicity condition. Unmatched k-dimensional faces are called *critical*; they correspond to the critical points of index k of a smooth Morse function. The total number of noncritical faces equals twice the number of edges in the Morse matching. The purpose of this paper is to study algorithms which compute maximum Morse matchings of a given finite simplicial complex. This is equivalent to finding a Morse matching with as few critical faces as possible.

A Morse matching M can be interpreted as a discrete flow on a simplicial complex Δ. The flow indicates how Δ can be deformed into a more compact description as a CW complex with one cell for each critical face of M. Naturally one is interested in a most compact description, which leads to the combinatorial optimization problem described above. This way optimal (or even sufficiently good) Morse matchings of Δ can help to recognize the topological type of a space given as a finite simplicial complex. The latter problem is known to be undecidable even for highly structured classes of topological spaces, such as smooth 4-manifolds. We have to admit, however, that so far no new topological results have been obtained by our approach.

Optimization of discrete Morse matchings has been studied by Lewiner, Lopes, and Tavares [23, 24]. Hersh [17] investigated heuristic approaches to the maximum Morse matching problem with applications to combinatorics. Morse matchings can also be interpreted as pivoting strategies for homology computations; see [20]. Furthermore, the set of all Morse matchings of a given simplicial complex itself has the structure of a simplicial complex; see [6].

*Received by the editors August 23, 2004; accepted for publication (in revised form) July 29, 2005; published electronically February 15, 2006. The authors' research was partially supported by the DFG Research Center MATHEON in Berlin.

http://www.siam.org/journals/sidma/20-1/44588.html

†Fachbereich Mathematik, AG 7, TU Darmstadt, 64289 Darmstadt, Germany (joswig@mathematik.tu-darmstadt.de).

‡Zuse Institute Berlin, Takustr. 7, 14195 Berlin, Germany (pfetsch@zib.de).

The paper is structured as follows. First we show that computing optimal Morse matchings is $\mathcal{NP}$-hard. This issue has been addressed previously by Lewiner, Lopes, and Tavares [24], but their argument omits details which to us seem quite important to address carefully. Then we give an integer programming (IP) formulation for the problem. The formulation consists of two parts: one for the matching conditions and one for the acyclicity constraints. This turns out to be related to the acyclic subgraph problem studied by Grötschel, Jünger, and Reinelt [14]. We derive polyhedral results for the corresponding polytope. In particular, we give two different polynomial time algorithms for the separation of the acyclicity constraints. The paper closes with computational results.

Like most of discrete Morse theory, also most of our results extend to arbitrary finite regular CW-complexes. We stick to the simplicial setting, however, to simplify the presentation.

2. Discrete Morse functions and Morse matchings. We will first introduce discrete Morse functions as developed by Forman. The essential structure of discrete Morse functions is captured by so-called Morse matchings; see Forman [8] and Chari [5]. It turns out that this latter formulation directly leads to a combinatorial optimization problem in which one wants to maximize the size of a Morse matching.

We first need some notation. Let Δ be a *(finite abstract) simplicial complex*, i.e., a set of subsets of a finite set V with the following property: if $F \in \Delta$ and $G \subseteq F$, then $G \in \Delta$; in other words, Δ is an independence system with ground set V. In the following we will ignore $\varnothing$ as a member of Δ. The elements in V are called *vertices* and the sets in Δ are called *faces*. The *dimension* of a face F is $\dim F := |F| - 1$. Let $d = \max\{\dim F \,:\, F \in \mathcal{F}\}$ be the dimension of Δ. We often write i-faces for i-dimensional faces. Let $\mathcal{F}$ be the set of faces of Δ and let $f_i = f_i(\Delta)$ be the number of faces of dimension $i \geq 0$. The maximal faces with respect to inclusion are called *facets* and 1-faces are called edges. The complex Δ is *pure*, if all facets have the same dimension. For $F, G \in \Delta$, we write $F \prec G$ if $F \subset G$ and $\dim F = \dim G - 1$, i.e., "$\prec$" denotes the covering relation in the Boolean lattice. The *graph* of Δ is the (abstract) graph on V in which two vertices are connected by an edge if there exists a 1-face containing both vertices. Throughout this paper we assume that Δ is *connected*, i.e., its graph is connected. This is no loss of generality since the connected components can be treated separately.

The *size* of Δ is defined as the coding length of its face lattice, i.e., if Δ has n faces, then $\operatorname{size}\Delta = \mathcal{O}(n \cdot d \cdot \log n)$. Statements about the complexity of algorithms in the subsequent sections are always with respect to this notion of size.

A function $f : \Delta \to \mathbb{R}$ is a *discrete Morse function* if for every $G \in \Delta$ the sets

$$\{F : F \prec G,\ f(G) \leq f(F)\} \quad \text{and} \quad \{H : G \prec H,\ f(H) \leq f(G)\} \tag{2.1}$$

both have cardinality at most 1. The first set includes the faces covered by face G which are not assigned a lower value than G, while the second set includes the faces covering G which are not assigned a higher value. The face G is *critical* if both sets have cardinality 0. A simple example of a discrete Morse function can be obtained by setting $f(F) = \dim F$ for every $F \in \Delta$. With respect to this function every face is critical.

Discrete Morse functions are interesting because they can be used to deform a simplicial complex into a (smaller) CW-complex that has a cell for each critical face; see section 3.

Consider the *Hasse diagram* $H = (\mathcal{F}, A)$ of Δ, that is, a directed graph on the faces of Δ with an arc $(G, F) \in A$ if $F \prec G$; note that the arcs lead from higher to lower dimensional faces. Let $M \subset A$ be a matching in H, i.e., each face is incident to at most one arc in M. Let $H(M)$ be the directed graph obtained from H by reversing the direction of the arcs in M. Then M is a *Morse matching* of Δ if $H(M)$ does not contain directed cycles, i.e., is acyclic (in the directed sense). Morse matchings are also often called *acyclic matchings*. Given $M \subset A$, one can decide in linear time (in the size of Δ) whether it is a Morse matching: the matching conditions are trivial and acyclicity of $H(M)$ can be checked by depth first search in linear time (see, e.g., Korte and Vygen [22]).

There is the following relation between Morse functions and Morse matchings; see Forman [8] and Chari [5]. Let f be a discrete Morse function and let M be the set of arcs $(G, F) \in A$ such that $f(G) \leq f(F)$, i.e., f is not decreasing on these arcs. A simple proof shows that at most one of the sets in (2.1) can have cardinality one. This shows that M is a matching. Since the order given by f can be refined to a linear ordering of the faces of Δ, the directed graph $H(M)$ is in fact acyclic and therefore a Morse matching. To construct a discrete Morse function from a Morse matching, compute a linear ordering extending $H(M)$ (which is acyclic) and then number the faces consecutively in the reverse order.

Although we lose the concrete numbers attached to the faces when going from a discrete Morse function f to the corresponding Morse matching M, we do not lose the information about critical faces: Critical faces of f are exactly the unmatched faces of M. Hence, by maximizing $|M|$ we minimize the number of critical faces of f. In fact, the number of critical faces is $|\mathcal{F}| - 2\,|M|$. For $0 \leq j \leq d$, let $c_j = c_j(M)$ be the number of critical faces of dimension j and let $c(M)$ be the total number of critical faces.

It seems helpful to briefly describe the case of Morse matchings for a one-dimensional simplicial complex Δ. Then Δ represents the incidences of a graph G. A Morse matching M of Δ matches edges with nodes of G. Let $\tilde{G}$ be the following oriented subgraph of G: take all edges which are matched in M and orient them towards its matched node. Since M is a matching, this construction is well defined and the in-degree of each node is at most one. The acyclicity property shows that $\tilde{G}$ contains no directed cycles and hence is a branching, i.e., the underlying graph is a forest and each (weakly) connected component has a unique root. Therefore, the Morse matchings on a graph G are in one-to-one correspondence with orientations of subgraphs of G which are branchings.

Building on this idea, Lewiner, Lopes, and Tavares [23] computed maximum Morse matchings, i.e., Morse matchings with maximal cardinality, for combinatorial 2-manifolds. In [24] they developed a heuristic for computing Morse matchings for arbitrary simplicial complexes. In the general case, however, this problem is $\mathcal{NP}$-hard, as shown in section 4.

3. Properties of Morse matchings. In this section we briefly review some important properties of Morse matchings which we need in what follows.

Let F be a facet of Δ and let G be a facet of F, which is not contained in any other facet of Δ. The operation of transforming Δ to $\Delta \setminus \{F, G\}$ is called a *simplicial* or *elementary collapse*. We will simply use collapse in the following.

PROPOSITION 3.1 (see Forman [8]). *Let Δ be a simplicial complex and Σ a subcomplex of Δ. Then there exists a sequence of collapses from Δ to Σ if and only if there exists a discrete Morse function such that $\Delta \setminus \Sigma$ contains no critical face.*

Forman [8] also proved the following result, which describes one of the most interesting features of Morse matchings:

THEOREM 3.2. *Let Δ be a simplicial complex and M be a Morse matching on Δ. Then Δ is homotopy equivalent to a CW-complex containing a cell of dimension i for each critical face of dimension i.*

We refer to Munkres [27] for more information on CW-complexes. By Theorem 3.2 we can hope for a compact representation of the topology of Δ (up to homotopy) by computing a Morse matching with few critical faces. This is the main motivation for the combinatorial optimization problem studied in this paper.

Let K be a field and let $\beta_j = \beta_j(K)$ be the Betti number for dimension j over K for Δ; see again Munkres [27] for details. Forman [8] proved the following bounds on the number of critical faces c_j of a Morse matching M:

THEOREM 3.3 (*weak Morse inequalities*). *Let K be a field, Δ be a simplicial complex, and M a Morse matching for Δ. We have*

$$c_j \geq \beta_j \qquad \forall\, j = 0, \ldots, d \tag{3.1}$$

and

$$c_0 - c_1 + c_2 - \cdots + (-1)^d c_d = \beta_0 - \beta_1 + \beta_2 - \cdots + (-1)^d \beta_d. \tag{3.2}$$

The Betti numbers over $\mathbb{Q}$ and finite fields can easily be obtained in polynomial time (in the size of Δ), by computing the ranks of the boundary matrices for each dimension. Although harder to compute (see Iliopoulos [18]), the homology over $\mathbb{Z}$ can be used to choose among the finite fields or $\mathbb{Q}$, in order to obtain the strongest form of the Morse inequalities (3.1).

4. Hardness of optimal Morse matchings. In this section we prove $\mathcal{NP}$-hardness of the problem to compute a maximum Morse matching, i.e., to find a Morse matching M with maximal cardinality. As we saw previously, this is equivalent to minimizing the number of critical faces.

We want to reduce the following *collapsibility problem*, introduced by Eğecioğlu and Gonzalez [7], to the problem of finding an optimal Morse matching: Given a connected pure 2-dimensional simplicial complex Δ that is embeddable in $\mathbb{R}^3$ and an integer k, decide whether there exists a subset $\mathcal{K}$ of the facets of Δ with $|\mathcal{K}| \leq k$ such that there exists a sequence of collapses which transforms $\Delta \setminus \mathcal{K}$ to a 1-dimensional complex. Eğecioğlu and Gonzalez proved that this collapsibility problem is strongly $\mathcal{NP}$-complete. Using Proposition 3.1, this result reads as follows in terms of discrete Morse theory.

THEOREM 4.1. *Given a connected pure 2-dimensional simplicial complex Δ that is embeddable in $\mathbb{R}^3$ and a nonnegative integer k, it is $\mathcal{NP}$-complete in the strong sense to decide whether there exists a Morse matching with at most k critical 2-faces.*

When k is fixed, we can try all possible sets $\mathcal{K}$ of size at most k and then decide whether the resulting complex is collapsible to a 1-dimensional complex in polynomial time. Therefore we let k be part of the input.

We need the following construction. Consider a Morse matching M for a simplicial complex Δ, with $\dim \Delta \geq 1$. Let $\Gamma(M)$ be the graph obtained from the graph of Δ by removing all edges (1-faces) matched with 2-faces. Note that $\Gamma(M)$ contains all vertices of Δ.

LEMMA 4.2. *The graph $\Gamma(M)$ is connected.*

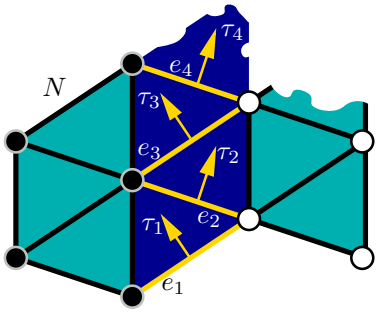


FIG. 1. *Illustration of the proof of Lemma* 4.2.

Proof. Without loss of generality we assume that $\dim \Delta \geq 2$. Otherwise, $\Gamma(M)$ coincides with the graph of Δ, which is connected (recall that Δ is connected).

Suppose that $\Gamma(M)$ is disconnected. Let N be its set of nodes in a connected component of $\Gamma(M)$, and let C be the set of *cut edges*, that is, edges of Δ with one vertex in N and one vertex in its complement. Since Δ is connected, C is not empty. By definition of $\Gamma(M)$, each edge in C is matched to a unique 2-face.

Consider the directed subgraph D of the Hasse diagram consisting of the edges in C and their matching 2-faces. The standard direction of arcs in the Hasse diagram (from the higher to the lower dimensional faces) is reversed for each matching pair of M, i.e., D is a subgraph of $H(M)$.

We construct a directed path in D as follows; see Figure 1. Start with any node of D corresponding to a cut edge e_1. Go to the node of D determined by the unique 2-face τ_1 to which e_1 is matched to. Then τ_1 contains at least one other cut edge e_2, otherwise e_1 cannot be a cut edge. Now iteratively go to e_2, then to its unique matching 2-face τ_2, choose another cut edge e_3, and so on. We observe that we obtain a directed path $e_1, \tau_1, e_2, \tau_2, \ldots$ in D, i.e., the arcs are directed in the correct direction.

Since we have a finite graph at some point the path must arrive at a node of D which we have visited already. Hence, D (and therefore also $H(M)$) contains a directed cycle, which is a contradiction since M is a Morse matching. ∎

Now pick an arbitrary node r and any spanning tree of $\Gamma(M)$ (which can be computed in polynomial time; see Korte and Vygen [22]) and direct all edges away from r. This yields a maximum Morse matching on $\Gamma(M)$; see the end of section 2. It is easy to see that replacing the part of M on $\Gamma(M)$ with this matching yields a Morse matching. This Morse matching has only one critical vertex (the root r). Note that every Morse matching contains at least one critical vertex; this can be seen from the Morse inequalities (3.1) in Theorem 3.3. Furthermore, the total number of critical faces can only decrease, since we computed an optimal Morse matching on $\Gamma(M)$. The number of critical i-faces for $i \geq 2$ stays the same. We have thus proved the following corollary, which is also implicit in Forman [8].

COROLLARY 4.3. *Let M be a Morse matching on Δ. Then we can compute a Morse matching M' in polynomial time which has exactly one critical vertex and the same number of critical faces of dimension 2 or higher as M, such that $c(M') \leq c(M)$.*

We can now prove the hardness result.

THEOREM 4.4. *Given a simplicial complex Δ and a nonnegative integer c, it is strongly $\mathcal{NP}$-complete to decide whether there exists a Morse matching with at most c*

critical faces, even if Δ is connected, pure, 2-dimensional, and can be embedded in $\mathbb{R}^3$.

Proof. Clearly this problem is in $\mathcal{NP}$. So let (Δ, k) be an input for the collapsibility problem. We claim that there exists a Morse matching with at most k critical 2-faces if and only if there exists a Morse matching with at most $g(k) := 2(k+1)-\chi(\Delta)$ critical faces altogether. Here, $\chi(\Delta) = \beta_0 - \beta_1 + \cdots + (-1)^d\beta_d$ is the Euler characteristic, which can be computed in polynomial time; see section 3. Hence g is a polynomial-time computable function. Using Theorem 4.1 then finishes the proof.

So assume that M is a Morse matching on Δ with at most k critical 2-faces. We use Corollary 4.3 to compute a Morse matching M', in polynomial time, such that $c_0(M') = 1$, $c_2(M') = c_2(M)$, and $c(M') \le c(M)$. By (3.2) of Theorem 3.3, we have $c_1(M') = c_2(M') + 1 - \chi(\Delta)$. Since $c(M') = c_0(M') + c_1(M') + c_2(M')$ it follows that

$$c_2(M) = c_2(M') = \tfrac{1}{2}(c(M') + \chi(\Delta)) - 1. \tag{4.1}$$

Solving for $c(M')$, it follows that M' has at most $2(k+1) - \chi(\Delta)$ critical faces altogether.

Conversely, assume that there exists a Morse matching M with at most $g(k)$ critical faces. Computing M' as above, we obtain by (4.1), that

$$c_2(M) = c_2(M') \le \tfrac{1}{2}(g(k) + \chi(\Delta)) - 1 = k,$$

which completes the proof. ∎

Lewiner, Lopes, and Tavares [24] showed that it is $\mathcal{NP}$-hard to compute an optimal Morse matching, but their proof omits an argument similar to Lemma 4.2 above. We therefore provided a proof for it.

Since there exists a Morse matching with at most c critical faces if and only if there exists a Morse matching of size at least $\frac{1}{2}(|\mathcal{F}| - c)$, we proved the following corollary.

COROLLARY 4.5. *Let Δ be as in Theorem 4.4 and m be a nonnegative integer. Then it is $\mathcal{NP}$-complete in the strong sense to decide whether there exists a Morse matching of size at least m.*

We do not know about the complexity status for this problem with m fixed.

Eğecioğlu and Gonzalez [7] additionally proved that the collapsibility problem is as hard to approximate as the set covering problem. In particular, the collapsibility problem cannot be approximated better than within a logarithmic factor in polynomial time, unless $\mathcal{P} = \mathcal{NP}$. Using this, Lewiner, Lopes, and Tavares [24] claimed that the problem to compute a Morse matching minimizing the number of critical faces is hard to approximate. However, the function g used in the proof above is not "approximation preserving" and we do not see how the nonapproximability result carries over.

Similarly, the problem to approximate the size of a Morse matching seems to be open.

5. An IP-formulation. In this section we introduce an integer programming formulation for the problem to compute a Morse matching of maximal size. From now on we assume that $\dim \Delta \ge 1$, since the other cases are uninteresting in our context.

We use the following notation. We depict vectors in bold font. Let $\boldsymbol{e}_i$ be the ith unit vector and let $\mathbb{1}$ be the vector of all ones. For any vector $\boldsymbol{x} \in \mathbb{R}^n$ and $I \subseteq \{1, \ldots, n\}$ we define

$$\boldsymbol{x}(I) := \sum_{i \in I} x_i.$$

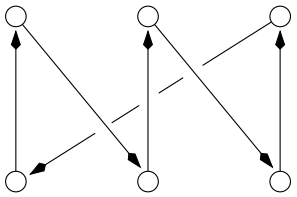


FIG. 2. *Example for a directed cycle of size* 6*; at least three arcs with reversed orientation (pointing "up") are necessary to close a* 6*-cycle in the Hasse diagram of a simplicial complex.*

Furthermore, for $S \subseteq \{1, \dots, n\}$, $\boldsymbol{I}(S) \in \mathbb{R}^n$ denotes the incidence vector of S.

For a node v in a directed graph, let $\delta(v)$ be the arcs incident to v, i.e., the arcs having v as one of their endnodes. For a subset $A' \subseteq A$, we denote by $N(A')$ the nodes incident to at least one arc in A'. Throughout this article, all directed or undirected cycles are assumed to be *simple*, i.e., without node repetitions.

For ease of notation, we consider the Hasse diagram H as directed or undirected depending on the context; we will explicitly say *directed* when we refer to the directed version.

We split H into d levels $H_0 = (\mathcal{F}^0, A_0), \dots, H_{d-1} = (\mathcal{F}^{d-1}, A_{d-1})$, where H_i denotes the level of the Hasse diagram between faces of dimension i and $i+1$. Then A is the disjoint union of $A_0, \dots, A_{d-1}$ and $\mathcal{F}^{i-1} \cap \mathcal{F}^i$ consists of the faces of dimension i. Recall that the arcs in the Hasse diagram are directed from the higher to the lower dimensional faces.

Let $M \subset A$ be a Morse matching of Δ. By definition, its incidence vector $\boldsymbol{x} = \boldsymbol{I}(M) \in \{0,1\}^A$ satisfies the *matching inequalities*

$$\boldsymbol{x}(\delta(F)) \leq 1 \quad \forall\, F \in \mathcal{F}. \tag{5.1}$$

Now assume that for some $M \subseteq A$ there exists a directed cycle D in $H(M)$. Then in D "up" and "down" arcs alternate; for an example, see Figure 2. In particular, the size of D is always even. Hence, $\frac{1}{2}|D|$ arcs are contained in M, i.e., are reversed in $H(M)$. We will use the following well-known observation.

OBSERVATION 1. *Let $M \subset A$ be a matching. If D is a directed cycle in $H(M)$, the edges in D can only belong to one level H_i ($i \in \{0, \dots, d-1\}$), i.e., we have $\{\dim F : F \in N(D)\} = \{i, i+1\}$.*

Putting these arguments together we obtain: If M is acyclic, $\boldsymbol{x} = \boldsymbol{I}(M)$ satisfies the following *cycle inequalities*:

$$\boldsymbol{x}(C) \leq \tfrac{1}{2}|C| - 1 \quad \forall\, C \in \mathcal{C}_i,\ i = 1, \dots, d-1, \tag{5.2}$$

where $\mathcal{C}_i$ are the cycles in H_i.

Conversely, it is easy to see that every $\boldsymbol{x} \in \{0,1\}^A$ which fulfills inequalities (5.1) and (5.2) is the incidence vector of a Morse matching. Hence, we arrive at the following IP formulation for the problem to find a maximum Morse matching:

$$\begin{aligned} (\text{MAXMM}) \quad \max \quad & \mathbb{1}^{\mathrm{T}} \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{x}(\delta(F)) \leq 1 && \forall\, F \in \mathcal{F} \\ & \boldsymbol{x}(C) \leq \tfrac{1}{2}|C| - 1 && \forall\, C \in \mathcal{C}_i,\ i = 1, \dots, d-1 \\ & \boldsymbol{x} \in \{0,1\}^A. \end{aligned}$$

This formulation can easily be extended to arbitrary weights on the arcs, i.e., replacing $\mathbb{1}$ in the objective function by an arbitrary nonnegative vector $\boldsymbol{w}$.

A different view on this optimization problem is to find directed spanning trees in the hypergraph defined by H_i and to patch them together (see Warme, Winter, and Zachariasen [31] for spanning trees in hypergraphs).

We define the corresponding polytope as

$$P_M = \operatorname{conv}\big\{\boldsymbol{x} \in \{0,1\}^A \,:\, \boldsymbol{x} \text{ satisfies (5.1) and (5.2)}\big\}.$$

Let M be a Morse matching and $\boldsymbol{x} = \boldsymbol{I}(M)$ be its incidence vector. Then $F \in \mathcal{F}$ is a critical face with respect to M if and only if it is unmatched by M, i.e., $\boldsymbol{x}(\delta(F)) = 0$. Hence, the total number of critical faces is

$$c(M) = \sum_{F\in\mathcal{F}} \left(1 - \sum_{a\in\delta(F)} x_a\right) = |\mathcal{F}| - 2\sum_{a\in A} x_a = |\mathcal{F}| - 2\,\mathbb{1}^{\mathrm{T}}\boldsymbol{x}, \tag{5.3}$$

since every arc is incident to exactly two nodes. Using this formula one can easily switch between the number of critical faces and the number of arcs in a Morse matching.

The LP relaxation of MaxMM can be strengthened by using the weak Morse inequalities (3.1) of Theorem 3.3. Applying (5.3), this yields the following *Betti inequality* for dimension i:

$$\sum_{F:\dim F=i} \left(1 - \sum_{a\in\delta(F)} x_a\right) \geq \beta_i \qquad \Leftrightarrow \qquad \sum_{F:\dim F=i} \sum_{a\in\delta(F)} x_a \leq f_i - \beta_i. \tag{5.4}$$

Observe that we can choose the field in Theorem 3.3 to employ the Morse inequalities in their strongest form.

Example 1. This can be illustrated by the real projective plane $\mathbb{RP}_2$. The Betti numbers with respect to $\mathbb{Q}$ and $\mathbb{Z}_2$ are $\boldsymbol{\beta}(\mathbb{Q}) = (1,0,0)$ and $\boldsymbol{\beta}(\mathbb{Z}_2) = (1,1,1)$, respectively. The resulting lower bounds are $(1,1,1)$, i.e., we have at least three critical faces in any Morse matching (this is, in fact, optimal).

Remark 1. The cycle inequalities (5.2) are similar to the cycle inequalities for the acyclic subgraph problem (ASP); see Jünger [21], and Grötschel, Jünger, and Reinelt [14]. The separation problem for (5.2), however, is more complicated than the corresponding problem for ASP; see section 5.2.

Furthermore, there is a similarity to the relation between the ASP and the linear ordering problem (see Reinelt [28], and Grötschel, Jünger, and Reinelt [13]): an alternative formulation for our problem can be obtained by modeling discrete Morse functions as linear orders on the faces, subject to matching requirements. Since this formulation is based on the relation between faces, it leads to quadratically many variables in the number of faces; therefore we have opted for the above formulation, at the cost of having to solve the separation problem for the cycle inequalities; see section 5.2.

5.1. Facial structure of P_M. It is easy to see that P_M is a full dimensional polytope and $x_a \geq 0$ defines a facet for every $a \in A$. Furthermore, P_M is monotone, since every subset of a Morse matching is a Morse matching. It is well known that this implies that every facet defining inequality $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x} \leq \beta$ not equivalent to the nonnegativity inequalities fulfills $\boldsymbol{\alpha} \geq 0$, $\beta > 0$; see Hammer, Johnson, and Peled [16].

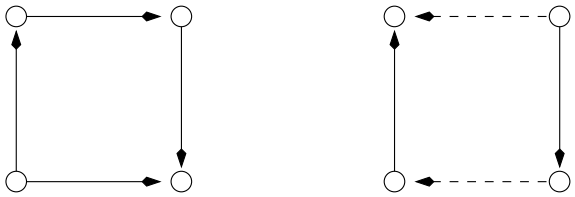


FIG. 3. *Example of a nonmonotone behavior of acyclic matchings. The directed graph on the right, obtained from the left graph by reversing the dashed arcs, is acyclic. However, if the top arc is set to its original orientation, the graph is not acyclic anymore. This shows that subsets of acyclic matchings are not necessarily acyclic.*

Interestingly, if we consider acyclic matchings as defined above for arbitrary acyclic directed graphs, the collection of such acyclic matchings is not necessarily monotone anymore; see the example in Figure 3. Therefore, the structure of the generalized problem is likely to be more complicated.

We have the following two results.

PROPOSITION 5.1. *The matching inequalities* $\boldsymbol{x}(\delta(F)) \leq 1$ *define facets of* P_M *for* $F \in \mathcal{F}$*, except if* $|\delta(F)| = 1$*, in which case* F *is a vertex.*

Proof. Let F be a face with $|\delta(F)| > 1$ (note that $|\delta(F)| = 0$ does not occur since $\dim \Delta \geq 1$ and Δ is connected). We can assume that $A = \{a_1, \ldots, a_k, a_{k+1}, \ldots, a_m\}$, where $\delta(F) = \{a_1, \ldots, a_k\}$. For $i = k+1, \ldots, m$, observe that a_i cannot be adjacent to every arc in $\delta(F)$: since $|\delta(F)| > 1$, a_i would either be incident to at least two nodes of the same dimension or to two nodes whose dimensions are two apart, which is impossible. Therefore, choose $p(i) \in \{1, \ldots, k\}$ such that a_i and $a_{p(i)}$ are not adjacent. It follows that $\boldsymbol{e}_i + \boldsymbol{e}_{p(i)} \in P_M$. Then

$$\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k, \boldsymbol{e}_{k+1} + \boldsymbol{e}_{p(k+1)}, \ldots, \boldsymbol{e}_m + \boldsymbol{e}_{p(m)}$$

are affinely independent and fulfill $\boldsymbol{x}(\delta(F)) = 1$. ∎

It follows that the inequalities $x_a \leq 1$, $a \in A$, never define facets, since each arc has a nonvertex endpoint.

THEOREM 5.2. *The cycle inequalities* (5.2) *define facets of* P_M*.*

Proof. We extend the corresponding proof by Jünger [21] for the ASP.

Let C be a cycle in H. Without loss of generality we can assume that $A = \{a_1, \ldots, a_k, a_{k+1}, \ldots, a_m\}$, where $C = (a_1, \ldots, a_k)$ and k is even. We will construct affinely independent feasible vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k, \boldsymbol{v}_{k+1}, \ldots, \boldsymbol{v}_m$ satisfying the cycle inequality corresponding to C with equality.

Let $C_1 = \{a_1, a_3, \ldots, a_{k-1}\}$ and $C_2 = \{a_2, a_4, \ldots, a_k\}$. Hence C_1 and C_2 are the "up" and "down" arcs in C.

Define

$$\boldsymbol{v}_i = \begin{cases} \boldsymbol{I}(C_1 \setminus \{a_i\}) & \text{if } a_i \in C_1 \\ \boldsymbol{I}(C_2 \setminus \{a_i\}) & \text{if } a_i \in C_2 \end{cases} \qquad \text{for } i = 1, \ldots, k.$$

Hence, for $i = 1, \ldots, k$ we have $\boldsymbol{v}_i(C) = \frac{k}{2} - 1$.

For $i = k+1, \ldots, m$, consider $a_i = \{u, v\} \notin C$. We have four cases.

▷ $u, v \in N(C)$: Let $\tilde{C} := C \setminus \big(\delta(u) \cup \delta(v)\big)$. We have that $|\tilde{C}| = k - 4$ (since there exist no odd cycles) and $\tilde{C}$ splits into two odd nonempty parts, $\tilde{C}_1$ and $\tilde{C}_2$, which are both paths. Let $k_1 := |\tilde{C}_1|$ and $k_2 := |\tilde{C}_2|$; k_1 and k_2 are odd, since u and v

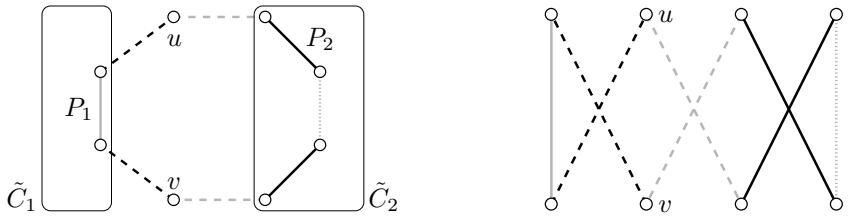


FIG. 4. *Illustration of the first case in the proof of Theorem* 5.2. *The sets P_1 and P_2 are shown by continuous lines. The edges in C_1 are drawn gray and hence $P_1 \subset C_1$; edges in C_2 are drawn black. The dashed edges incident to u and v are not considered. The right-hand side shows the graph embedded in the Hasse diagram.*

are on opposite sides of the bipartition. We choose a subset $P_1 \subset \tilde{C}_1$ by taking every second arc in order to get $|P_1| = \frac{k_1+1}{2}$; similarly we choose $P_2 \subset \tilde{C}_2$ with $|P_2| = \frac{k_2+1}{2}$. By construction either $P_i \subset C_1$ or $P_i \subset C_2$ and either $P_i \cap C_2 = \varnothing$ or $P_i \cap C_1 = \varnothing$ for $i = 1, 2$. An easy calculation shows that $|P_1 \cup P_2| = \frac{k}{2} - 1$; see Figure 4 for an illustration of this case. Then define $\boldsymbol{v}_i := \boldsymbol{I}(P_1 \cup P_2 \cup \{a_i\})$.

▷ $u \notin C,\ v \in C$: Here we define $\boldsymbol{v}_i := \boldsymbol{I}(C_1 \setminus \delta(v) \cup \{a_i\})$.

▷ $u \in C,\ v \notin C$: Define $\boldsymbol{v}_i := \boldsymbol{I}(C_1 \setminus \delta(u) \cup \{a_i\})$.

▷ $u, v \notin C$: Choose any $a \in C_1$ and define $\boldsymbol{v}_i := \boldsymbol{I}(C_1 \setminus \{a\} \cup \{a_i\})$.

It is easy to check in each case that $\boldsymbol{v}_i \in P_M$ and that $\boldsymbol{v}_i(C) = \frac{k}{2} - 1$.

It can be shown that the m vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ are affinely independent, which concludes the proof. □

The separation problem for the cycle inequalities is discussed in the next section.

5.2. Separating the cycle inequalities. Of course, there are exponentially many cycle inequalities (5.2). Hence we have to deal with the separation problem for these inequalities.

We can assume that we are given $\boldsymbol{x}^* \in [0,1]^A$, which satisfies all matching inequalities (5.1). We consider the separation problem for each graph H_i in turn, $i = 0, \ldots, d-1$. The problem is to find an undirected cycle C in H_i such that

$$\boldsymbol{x}^*(C) > \tfrac{1}{2}|C| - 1$$

or conclude that no such cycle exists. In the next sections we describe two methods to solve this problem in polynomial time.

5.2.1. Undirected shortest path with conservative weights. A well-known trick to solve the above separation problem is to apply an affine transformation and obtain a shortest cycle problem. The transformation suitable for our needs is $\boldsymbol{x}' = \frac{1}{2}\mathbb{1} - \boldsymbol{x}$, which yields

$$\boldsymbol{x}(C) \le \tfrac{1}{2}|C| - 1 \qquad \Leftrightarrow \qquad \boldsymbol{x}'(C) \ge 1.$$

The separation problem can now be solved as follows: compute a shortest cycle in H_i with respect to the weights $\frac{1}{2}\mathbb{1} - \boldsymbol{x}^*$. If its weight is at most 1, this cycle yields a violated cycle inequality, otherwise no such cycle exists.

However, the weights can be negative and we have to rule out negative cycles in order to apply polynomial time methods from the literature; that is, we want the weights to be *conservative*.

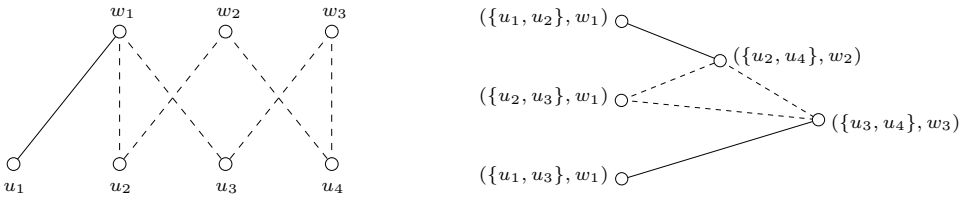


FIG. 5. *Example of the construction in section* 5.2.2. Left*: original graph* G. Right: *constructed graph* G'. *The* 6*-cycle on the left corresponds to the* 3*-cycle on the right (both shown with dashed lines).*

LEMMA 5.3. *There exists no cycle of negative weight in* H_i *with respect to* $\frac{1}{2}\mathbb{1}-\boldsymbol{x}^*$, *for* $0 \le i \le d-1$.

Proof. Let $C = (a_1, \dots, a_k)$ be a cycle in H_i and let $F_1, \dots, F_k$ be the faces that are visited by C. Recall that $\boldsymbol{x}^*$ satisfies the matching inequalities. We obtain

$$\sum_{j=1}^{k} \sum_{a \in \delta(F_i) \cap C} x^*_a \;=\; 2 \sum_{a \in C} x^*_a \;=\; 2\,\boldsymbol{x}^*(C), \tag{5.5}$$

since each edge weight is counted twice in the first term. Applying the matching inequalities (5.1) on the left-hand side yields that $\boldsymbol{x}^*(C) \le \frac{1}{2}k = \frac{1}{2}|C|$. Hence, the weight of C with respect to $\frac{1}{2}\mathbb{1} - \boldsymbol{x}^*$ can be bounded as follows:

$$\sum_{a \in C} \left(\tfrac{1}{2} - x^*_a\right) = \tfrac{1}{2}|C| - \boldsymbol{x}^*(C) \ge 0,$$

which proves the lemma. ∎

We have now reduced the separation problem to finding a shortest cycle in a weighted undirected graph $G = (V, E)$ without negative cycles.

By using T-join techniques, one can compute a shortest path in an undirected graph without negative cycles in $\mathcal{O}(n_i(m_i + n_i \log n_i))$ time, where in this formula $n_i = |\mathcal{F}^i|$ and $m_i = |A_i|$; see Schrijver [29, Chapter 29]. It follows that a shortest cycle can be computed in $\mathcal{O}(m_i n_i(m_i + n_i \log n_i))$ time. Since $|A_i| \le (i+2)n_i$, this leads to an $\mathcal{O}\big((d+1)^2 n^3 + (d+1)n^3 \log n\big)$ overall algorithm, where $n := |\mathcal{F}|$ is the number of faces and d is the dimension of the complex.

5.2.2. Transforming the graph. Another method for the separation problem of cycle inequalities, which is easier to implement, works as follows.

Let $G = (U \,\dot\cup\, W, E)$ be a bipartite graph, e.g., $G = H_i$ ($i \in \{0, \dots, d-1\}$), the ith level of the Hasse diagram. Let $\ell : E \to \mathbb{R}_{\ge 0}$ be a length function for the edges of G. In the following we write $\ell(u, v) = \ell(v, u)$ for the length $\ell(\{u, v\})$.

We construct a graph $G' = (V', E')$ and lengths $\ell' : E' \to \mathbb{R}_{\ge 0}$ as follows; see Figure 5 for an example. The set of nodes of G' is

$$\big\{(\{u, u'\}, w) \,:\, u, u' \in U,\ u \ne u',\ w \in W,\ \{u, w\} \in E,\ \{u', w\} \in E\big\}.$$

Hence, G' has a node for each path with two edges in G. There is an edge between two nodes $(\{u_1, u_1'\}, w_1)$ and $(\{u_2, u_2'\}, w_2)$ if

$$|\{u_1, u_1'\} \cap \{u_2, u_2'\}| = 1 \quad \text{and} \quad w_1 \ne w_2.$$

The length of such an edge e' is defined by

$$\ell'(e') = \tfrac{1}{2}\big(\ell(u_1, w_1) + \ell(u_1', w_1) + \ell(u_2, w_2) + \ell(u_2', w_2)\big).$$

Hence, G' contains an edge for each path with four edges in G and its length is the length of this path divided by 2. We now consider the relation of cycles in G and G'.

LEMMA 5.4. *$C = (u_0, w_0, u_1, w_1, \ldots, w_{k-1}, u_1)$ is a cycle in G with $k > 1$ of length $\ell(C)$ if and only if*

$$C' = \big((\{u_0, u_1\}, w_0), (\{u_1, u_2\}, w_1), \ldots, (\{u_{k-1}, u_1\}, w_{k-1}), (\{u_0, u_1\}, w_0)\big)$$

is a cycle in G' with $\ell'(C') = \ell(C)$.

We omit the straightforward proof.

The previous lemma does not cover cycles in G of length four. These do not occur for the case of $G = H_i$, since H_i is a level in the Hasse diagram of a *simplicial* complex. Moreover, cycles of length four can readily be detected in the construction of G' and handled accordingly (there is only a polynomial number of them).

To solve our separation problem, let $G = H_i$, $i \in \{0, \ldots, d-1\}$, and $\ell(e) = x_e^*$ for $e \in G$. Then we have $\ell'(e') \in [0, 1]$ for each $e' \in E'$, because of the matching inequalities. We now set $\tilde{\ell}(e') = 1 - \ell'(e')$ for $e' \in G'$ and hence $\tilde{\ell}(e') \in [0, 1]$. Let C be a cycle in G with at least six edges and C' be the corresponding cycle in G'. Note that $|C'| = \frac{1}{2}|C|$. We then have the following:

$$\begin{aligned}
&\tilde{\ell}(C') = \sum_{e' \in C'} \tilde{\ell}(e') = \sum_{e' \in C'} (1 - \ell'(e')) < 1 \\
\Leftrightarrow \quad &\sum_{e' \in C'} \ell'(e') > |C'| - 1 \\
\Leftrightarrow \quad &\ell'(C') > |C'| - 1 \\
\Leftrightarrow \quad &\ell(C) > \tfrac{1}{2}|C| - 1 \qquad \text{(by Lemma 5.4).}
\end{aligned}$$

Hence, C violates the cycle inequality (5.2) if and only if $\tilde{\ell}(C') < 1$. Since $\tilde{\ell}(e') \geq 0$, we can use the Floyd–Warshall algorithm to solve the separation problem in time $\mathcal{O}\big(|V'|^3\big)$; see Korte and Vygen [22].

If $G = H_i$ and W is the part arising from the higher dimensional faces, we have $|V'| = \binom{i+2}{2}|W| = \binom{i+2}{2} f_{i+1}$. This leads to an $\mathcal{O}\big((d+1)^6 n^3\big)$ algorithm for separating cycle inequalities, which is roughly as fast as the method discussed in section 5.2.1, but much easier to implement.

6. Computational results. In this section we report on computational experience with a branch-and-cut algorithm along the lines of section 5. The C++ implementation uses the framework SCIP (Solving Constraint Integer Programs) by Achterberg; see [1]. It furthermore builds on `polymake`; see [11, 12]. As an LP solver we used CPLEX 9.0.

As the basis of our implementation we take the formulation of MAXMM in section 5. Matching inequalities (5.1) and Betti inequalities (5.4) (together with variable bounds) form the initial LP. The computation of the simplicial homology from which the Betti numbers are computed is very fast, because the examples are small; its running time is not included in the following. Cycle inequalities (5.2) are separated as described in section 5.2.2. Additionally, Gomory cuts are added. As a branching rule we use *reliability branching* implemented in SCIP, a variable branching rule introduced by Achterberg, Koch, and Martin [2].

TABLE 1

Computational results of the branch-and-cut algorithm with separating cycle inequalities and Gomory cuts.

name	n	m	d	nodes	depth	time	β	c
solid_2_torus	24	42	2	1	0	0.00	2	2
simon2	31	60	2	1	0	0.00	1	1
projective ($\mathbb{RP}_2$)	31	60	2	1	0	0.01	3	3
bjorner	32	63	2	1	0	0.05	2	2
nonextend	39	77	2	6	5	0.16	1	1
simon	41	82	2	1	0	0.18	1	1
dunce	49	99	2	385	10	2.62	1	3
c-ns3	63	128	2	349	10	3.47	1	3
c-ns	75	152	2	28	10	1.95	1	3
c-ns2	79	159	2	14	7	1.11	1	1
ziegler	119	310	3	1	0	0.01	1	1
gruenbaum	167	434	3	1	0	25.24	1	1
lockeberg	216	600	3	1	0	36.25	2	2
rudin	215	578	3	77	30	103.78	1	1
mani-walkup-D	392	1112	3	111	23	512.81	2	2
mani-walkup-C	464	1312	3	135	83	1658.02	2	2
MNSB	103	267	3	12	10	73.39	1	1
MNSS	250	698	3	292	110	750.36	2	2
CP2	255	864	4	230	80	558.14	3	3

We implemented the following primal heuristic. First a simple greedy algorithm is run: We start with the empty matching $M = \varnothing$. We add arcs of the Hasse diagram to M in the order of decreasing value of the current LP solution as long as M stays an acyclic matching (which can easily be tested). Then the outcome is iteratively improved by a method described in Forman [8]; one searches for a unique path between two critical faces in $H(M)$. Such a path is alternating with respect to M. Then M can be augmented along the path (the new matching is the symmetric difference of M and the path). As is easily seen, this generates an acyclic matching, because the path is unique. This heuristic turns out to be extremely successful; see below.

We tested the implementation on a set of simplicial complexes collected by Hachimori; see [15] for more details. This test set was also used by Lewiner, Lopes, and Tavares [24]. Additionally, we considered the following complexes: `CP2` (complex projective plane), `CP2+CP2` (connected sum of `CP2` with itself), `MNSB` (vertex minimal nonshellable ball), and `MNSS` (nonshellable sphere with the fewest number of vertices known). The last two examples are due to Lutz [25, 26].

All computational experiments were run on a 3 GHz Pentium machine running Linux. In the tables of computational results, n denotes the number of faces, m the number of arcs in the Hasse diagram (= number of variables), d the dimension, *nodes* the number of nodes in the branch-and-bound tree, *depth* the maximal depth in the tree, *time* the computation time in seconds, β the lower bound obtained by adding all Betti inequalities (5.4), and c the number of critical faces in the optimal solution.

Our implementation could not solve the larger problems of Hachimori's collection in reasonable time: `bing`, `knot`, `poincare`, `nonpl_sphere`, and `nc_sphere`. In fact, for `poincare` we ran our code in different settings, each for about a week, without success. Table 1 shows the results of a computation where we separate cycle inequalities and Gomory cuts and run the heuristic every 10th level. At most seven separation rounds of cycle inequalities were performed at a node. We do not report results on the problems by Moriyama and Takeuchi in Hachimori's collection—they all could be solved within a second. The version with cut separation could not solve `CP2+CP2` within 90 minutes.

TABLE 2
Computational results of the branch-and-cut algorithm without separation.

name	n	m	d	nodes	depth	time	β	c
solid_2_torus	24	42	2	1	0	0.00	2	2
simon2	31	60	2	1	0	0.01	1	1
projective ($\mathbb{RP}_2$)	31	60	2	1	0	0.00	3	3
bjorner	32	63	2	1	0	0.01	2	2
nonextend	39	77	2	3	2	0.02	1	1
simon	41	82	2	4	3	0.02	1	1
dunce	49	99	2	168367	42	145.60	1	3
c-ns3	63	128	2	3665581	53	3940.40	1	3
c-ns	75	152	2	16625713	58	19359.69	1	3
c-ns2	79	159	2	4	3	0.03	1	1
ziegler	119	310	3	1	0	0.01	1	1
gruenbaum	167	434	3	21	20	0.68	1	1
lockeberg	216	600	3	1	0	0.05	2	2
rudin	215	578	3	81	80	3.18	1	1
mani-walkup-D	392	1112	3	107	100	2.00	2	2
mani-walkup-C	464	1312	3	1498	456	30.54	2	2
MNSB	103	267	3	1	0	0.01	1	1
MNSS	250	698	3	163	126	4.63	2	2
CP2	255	864	4	198	190	4.77	3	3
CP2+CP2	460	1592	4	5178	534	110.21	4	4

For most problems the bound obtained by adding Betti inequalities (5.4), as indicated in column "β," is tight. This means that the algorithm is done once an optimal solution is found. This usually happens very fast and shows that the heuristic is efficient. In fact, there are only three problems for which the bound is not tight and could be solved by our algorithm (`dunce`, `c-ns`, and `c-ns3`). These three problems are solved easily by the version with cut separation. In our problem set there exists no hard but still solvable problem with a "Betti bound" which is not sharp. We therefore cannot estimate the limits of our implementation for these cases (`poincare` is the next larger problem of this kind with 1112 variables, but we could not solve it).

The tractability of problems with a tight "Betti bound" is supported by the results obtained by running the implementation without any separation; see Table 2. Only integer solutions are checked whether they are acyclic and the heuristic is run every 10th level. This essentially is a test of the performance of the primal heuristic. Indeed, all problems with tight "Betti bound" were solved within a few seconds (`CP2+CP2` and `mani-walkup-C` being the exception, but could be solved within two minutes). The results for the problems `c-ns`, `c-ns3`, and `dunce` show that the cycle inequalities and Gomory cuts are very effective in reducing the number of nodes in the tree and the computing time for problems where the "Betti bound" is not sharp.

Summarizing, we can say that our implementation can solve large instances with up to about 1500 variables if the bounds from the Betti numbers are tight and small instances with up to about 150 variables if the bounds are not tight. In all the instances computed so far, the topology of the spaces involved was known. In the future, we plan to apply our techniques to other cases.

Acknowledgments. We are indebted to Tobias Achterberg for his support of the implementation. We also thank both referees for their helpful comments.

REFERENCES

[1] T. ACHTERBERG, *SCIP—A framework to integrate constraint and mixed integer programming*, ZIB-Report 04-19, Zuse Institute Berlin, Berlin, Germany, 2004.

[2] T. Achterberg, T. Koch, and A. Martin, *Branching rules revisited*, Oper. Res. Lett., 33 (2005), pp. 42–54.
[3] E. Babson, A. Björner, S. Linusson, J. Shareshian, and V. Welker, *Complexes of not i-connected graphs*, Topology, 38 (1999), pp. 271–299.
[4] E. Batzies and V. Welker, *Discrete Morse theory for cellular resolutions*, J. Reine Angew. Math., 543 (2002), pp. 147–168.
[5] M. K. Chari, *On discrete Morse functions and combinatorial decompositions*, Discrete Math., 217 (2000), pp. 101–113.
[6] M. K. Chari and M. Joswig, *Complexes of discrete Morse functions*, Discrete Math., 302 (2005), pp. 39–51.
[7] Ö. Eğecioğlu and T. F. Gonzalez, *A computationally intractable problem on simplicial complexes*, Comput. Geom., 6 (1996), pp. 85–98.
[8] R. Forman, *Morse theory for cell complexes*, Adv. Math., 134 (1998), pp. 90–145.
[9] R. Forman, *Morse theory and evasiveness*, Combinatorica, 20 (2000), pp. 489–504.
[10] R. Forman, *A user's guide to discrete Morse theory*, Sém. Lothar. Combin., 48 (2002), pp. Art. B48c, 35 pp.
[11] E. Gawrilow and M. Joswig, polymake: *a framework for analyzing convex polytopes*, in Polytopes—Combinatorics and Computation, G. Kalai and G. M. Ziegler, eds., DMV Sem. 29, Birkhäuser, Basel, Switzerland, 2000, pp. 43–73.
[12] E. Gawrilow and M. Joswig, polymake: *Version* 2.1.0, http://www.math.tu-berlin.de/polymake, 2004. With contributions by T. Schröder and N. Witte.
[13] M. Grötschel, M. Jünger, and G. Reinelt, *A cutting plane algorithm for the linear ordering problem*, Oper. Res., 32 (1984), pp. 1195–1220.
[14] M. Grötschel, M. Jünger, and G. Reinelt, *On the acyclic subgraph polytope*, Math. Programing, 33 (1985), pp. 28–42.
[15] M. Hachimori, *Simplicial complex library*. Available online from http://infoshako.sk.tsukuba.ac.jp/~hachi/math/library/index_eng.html, 2001.
[16] P. L. Hammer, E. L. Johnson, and U. N. Peled, *Facets of regular* 0-1 *polytopes*, Math. Programing, 8 (1975), pp. 179–206.
[17] P. Hersh, *On optimizing discrete Morse functions*, Adv. in Appl. Math., 35 (2005), pp. 294–322.
[18] C. S. Iliopoulos, *Worst-case complexity bounds on algorithms for computing the canonical structure of finite Abelian groups and the Hermite and Smith normal forms of an integer matrix*, SIAM J. Comput., 18 (1989), pp. 658–669.
[19] J. Jonsson, *On the topology of simplicial complexes related to* 3*-connected and Hamiltonian graphs*, J. Combin. Theory Ser. A, 104 (2003), pp. 169–199.
[20] M. Joswig, *Computing invariants of simplicial manifolds*, preprint. Available online from math.AT/0401176, 2004.
[21] M. Jünger, *Polyhedral Combinatorics and the Acyclic Subdigraph Problem*, Research and Exposition in Mathematics, 7, Heldermann Verlag, Berlin, 1985.
[22] B. Korte and J. Vygen, *Combinatorial optimization. Theory and algorithms*, 2nd ed., Algorithms and Combinatorics, 21, Springer-Verlag, Berlin, 2002.
[23] T. Lewiner, H. Lopes, and G. Tavares, *Optimal discrete Morse functions for* 2*-manifolds*, Comput. Geom., 26 (2003), pp. 221–233.
[24] T. Lewiner, H. Lopes, and G. Tavares, *Towards optimality in discrete Morse theory*, Experiment Math., 12 (2003), pp. 271–285.
[25] F. H. Lutz, *Small examples of nonconstructible simplicial balls and spheres*, SIAM J. Discrete Math, 18 (2004), pp. 103–109.
[26] F. H. Lutz, *A vertex-minimal nonshellable simplicial* 3*-ball with* 9 *vertices and* 18 *facets*, Electronic Geometry Models, (2004). Available online from www.eg-models.de.
[27] J. R. Munkres, *Elements of Algebraic Topology*, Addison-Wesley, Menlo Park, CA, 1984.
[28] G. Reinelt, *The Linear Ordering Problem: Algorithms and Applications*, Research and Exposition in Mathematics, 8, Heldermann Verlag, Berlin, 1985.
[29] A. Schrijver, *Combinatorial Optimization: Polyhedra and Efficiency*, Algorithms and Combinatorics, 24, Springer-Verlag, Berlin, 2003.
[30] J. Shareshian, *Discrete Morse theory for complexes of* 2*-connected graphs*, Topology, 40 (2001), pp. 681–701.
[31] D. M. Warme, P. Winter, and M. Zachariasen, *Exact solutions to large-scale plane steiner tree problems*, in Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 1999, pp. 979–980.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 26–41

RANDOMIZED PURSUIT-EVASION WITH LOCAL VISIBILITY*

VOLKAN ISLER†, SAMPATH KANNAN‡, AND SANJEEV KHANNA‡

Abstract. We study the following pursuit-evasion game: One or more hunters are seeking to capture an evading rabbit on a graph. At each round, the rabbit tries to gather information about the location of the hunters but it can see them only if they are located on adjacent nodes. We show that two hunters suffice for catching rabbits with such local visibility with high probability. We distinguish between reactive rabbits who move only when a hunter is visible and general rabbits who can employ more sophisticated strategies. We present polynomial time algorithms that decide whether a graph G is hunter-win, that is, if a single hunter can capture a rabbit of either kind on G.

Key words. pursuit-evasion games, local information, path planning, visibility

AMS subject classifications. 49N75, 91A43

DOI. 10.1137/S0895480104442169

1. Introduction.

Pursuit-evasion games are problems of fundamental interest in many diverse fields such as computer science, operations research, game theory, and control theory. The goal of a pursuit-evasion game is to find a strategy for a pursuer trying to catch an evader who, in turn, tries to avoid capture indefinitely. There are many different variations of pursuit evasion games based on the following:

- *Environment where the game is played*: Examples include plane, grid, and graph.
- *Information available to the players*: Do they know each others' positions all the time? Does the pursuer know the evader's strategy?
- *Controllability of the players' motion*: Is there a bound on their speed? Can they turn with arbitrary angles?
- *Meaning of capture*: In some games, the pursuer captures the evader if the distance between them is less than a threshold. In other games, the pursuers must see or surround the evader in order to capture it.

Earlier studies of pursuit-evasion were motivated by control tasks such as intercepting missiles [4]. The problem is addressed in the robotics community for its applications in collision avoidance, search and rescue, and air-traffic control [10, 9]. In these models typically the motion of the evader is modeled by a stochastic process. However, recently there has been increasing interest in modeling games where the evader is more "intelligent" and has certain sensing capabilities [19]. Pursuit-evasion games on graphs [18, 16, 13, 12, 6, 1] have been studied not only for their applications in network security and protocol design (e.g., [3, 11]) but also for their relations to fundamental properties of graphs such as vertex separation [7]. A remark about

*Received by the editors March 17, 2004; accepted for publication (in revised form) August 5, 2005; published electronically February 15, 2006. A preliminary version of this paper appeared in *Proceedings of the ACM–SIAM Symposium on Discrete Algorithms* (SODA04).
http://www.siam.org/journals/sidma/20-1/44216.html

†Corresponding author. Department of Computer Science, Rensselaer Polytechnic Institute, 110 Eighth Street, Lally 205, Troy, NY 12180-3590 (isler@cs.rpi.edu). The work of this author was supported in part by NSF-IIS-0083209, NSF-IIS-0121293, MURI DAAH-19-02-1-03-83, and a grant from Rensselaer Polytechnic Institute.

‡Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104 (kannan@cis.upenn.edu, sanjeev@cis.upenn.edu). The work of the second author was supported in part by NSF grant CCR0105337. The work of the third author was supported by an Alfred P. Sloan Research Fellowship and by an NSF Career Award CCR-0093117.

the terminology is in order. In the literature, the names pursuer-evader, cop-robber, monster-princess, hunter-rabbit, and sheriff-thief have been used somewhat synonymously. We adopt the hunter-rabbit term for it emphasizes the discrete nature of the game [5, 1].

In this paper, we address a different aspect of the problem that has not received much attention so far. We study the relationship between the information available to the rabbit and the conditions to capture it. The basic model of our game is as follows: The players are located on the nodes of a graph. At every time step, they move to nodes in their neighborhoods (which include the current node) simultaneously. We say a rabbit is *caught* or *captured* if at the beginning of a time step it occupies the same node as a hunter. We associate the information available to the rabbit with its visibility. If the rabbit has complete information about the location of the hunter(s) during the entire game, we say the rabbit has *full visibility*. On the contrary, if the rabbit has no information about the hunters, then we say it has *no visibility*.

In our present work, we study the game when the rabbit has *local visibility*. That is, it can only see the nodes that are adjacent to its current location. When the hunter is located at an adjacent node, the rabbit has complete information about his location. However, if the hunter is not visible, then the rabbit must infer the hunter's location based on the time and location of their last encounter. Note that this model is different from the "visibility-based pursuit-evasion" work [9, 17], where the goal is to eventually "see" an evader which has complete visibility and unbounded speed.

Recently, Adler et al. studied the game when the rabbit has no visibility [1]. They showed that a single hunter can catch the rabbit on any (connected) graph. The full visibility version has also been studied [16, 6]. It is known that under the full visibility model, the class of graphs on which a single hunter suffices is the class of *dismantlable* graphs. The number of hunters necessary to capture the rabbit on a graph G is known as the cop (hunter) number of G. It is known that [2] the cop number of planar graphs is at most 3 but the cop number of general graphs is still an open question [15, 8].

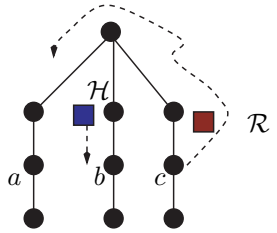


FIG. 1. *On this graph, the hunter cannot capture the rabbit using a deterministic strategy.*

An interesting aspect of our game is that on most graphs the rabbit cannot be captured using a deterministic strategy. A simple example is illustrated in Figure 1. Suppose that, on this graph, the hunter has a deterministic strategy of visiting the labeled vertices in the order a, b, c. Then, we can design a rabbit strategy that waits until the hunter arrives at b and escapes to a. Afterward, while the hunter is visiting c, the rabbit escapes to b and it is easy to see that by repeating similar moves, the rabbit can always avoid the hunter. However, on this graph there is a simple randomized strategy for the hunter: Pick one of the leaves at random and visit that leaf!

Therefore, we will focus on randomized strategies. The previous body of work for the full visibility case [16, 6, 15, 8, 2] derandomized the game by forcing the players to take turns moving, rabbit followed by hunter at each step. However, when the players

move simultaneously, the game is not well defined for deterministic strategies even if the players have full visibility: Suppose the game is played on a complete graph. In this case it is easy to see that a single hunter can catch the rabbit simply by guessing its location in the next turn. However, if the hunter's strategy is deterministic, knowing it, the rabbit would never get caught. Similarly, the hunter could always catch the rabbit in a single move if he knew its strategy.

Our results and techniques. Our main result is an algorithmic characterization for the local visibility case. We show that two hunters always suffice on general graphs and present a polynomial time procedure that decides whether a single hunter is sufficient to capture the rabbit on an input graph G. In order to obtain an efficient decision procedure, we establish that the uncertainty in the rabbit's knowledge of the hunter's location satisfies an interesting monotonicity property. This monotonicity property turns out to be crucial for obtaining a polynomial time characterization.

In the winning strategy for two hunters, a central component is to have one hunter mainly focus on keeping the rabbit on the move. This motivated us to study a natural class of *reactive* rabbit strategies, where the rabbit moves only when the hunter is in its sight. We show that the class of hunter-win graphs (i.e., graphs on which a single hunter suffices) against general rabbits is strictly smaller than the class of hunter-win graphs against reactive rabbits. We present a characterization algorithm for reactive rabbits as well.

The characterization algorithms mark pairs of vertices according to certain rules, where the pairs correspond to players' positions. To understand the corresponding hunter strategies on hunter-win graphs, we first present a hunter strategy for the full visibility case. Next, we show that omitting one of the rules from the characterization algorithms yields an algorithm that recognizes graphs that are hunter-win against rabbits with full visibility. Using these two results, we show how the hunter exploits the local visibility if the game is played on a graph G such that on G, the hunter can win against a rabbit with local visibility but not against a rabbit with full visibility.

We note that when the rabbit's visibility is extended to distance 2, there exist graphs for which $\tilde{\Omega}(\sqrt{n})$ hunters are necessary.

Organization of the paper. The paper is organized as follows: In section 2, we review necessary concepts that will be used throughout the paper. In section 3, we present a winning strategy for two hunters on general graphs. Next, we study the graphs on which a single hunter suffices, both for reactive (section 4.1) and general (section 4.2) rabbits. Section 5 is dedicated to the study of hunter strategies on hunter-win graphs. A gap example distinguishing the power of the two types of rabbit strategies is also presented in section 5. We conclude the paper with a discussion on extensions of our work.

2. Preliminaries. Throughout the paper, we use the following notation for the neighborhood of vertex v: $N(v)$ denotes the set of vertices that are adjacent to v and we always assume that $v \in N(v)$. $N^i(v)$ is defined as $\cup_{u \in N^{i-1}(v)} N(u)$. Unless otherwise stated, n denotes the number of vertices.

The game we study is formally defined as follows: It is played in rounds. In the beginning of a round, suppose a player (either a hunter or a rabbit) is located at vertex v. First, the player checks $N(v)$ and if there is another player located at a vertex $u \in N(v)$, this information is revealed to the player. In this case we say the two players *see* each other. Next, all the players make a decision about where to move and choose a vertex in their neighborhoods. At the end of the round, all players move to their chosen vertex simultaneously. A hunter *catches* the rabbit if they are located

on the same vertex.

A *reactive rabbit strategy* is a rabbit strategy where the rabbit is not allowed to move from a vertex v unless the hunter is in $N(v)$. A rabbit strategy is *general* (sometimes called *nonreactive*) if it is not forced to be reactive. In other words, the rabbit can move even if the hunter is not visible. A graph G is *hunter-win against reactive rabbits* if there exists a hunter strategy that catches any reactive rabbit on G with nonzero probability for all possible starting configurations. A graph that is *hunter-win against general rabbits* is defined similarly.

Configuration versus state. For a single hunter game, a *configuration* refers to an ordered pair (h, r) which corresponds to the locations of the hunter and the rabbit, respectively. Note that this information may not be available to the rabbit at all times due to its local visibility. A configuration (h, r) is *adjacent* if $h \in N(r)$. We use the notation $\langle H, r\rangle$ to denote the *state* of the game where r is the location of the rabbit and H corresponds to the set of vertices where the hunter can possibly be located (based on the information available to the rabbit). For the full visibility case, if the current configuration is (h, r), the state is $\langle\{h\}, r\rangle$. For the zero visibility case, the state is either $\langle G - \{r\}, r\rangle$ or $\langle\{r\}, r\rangle$. For the local visibility case that we study, state has a more complex structure, and it evolves over time even when neither the hunter nor the rabbit is in motion.

Suppose u and v are two nodes of a graph G such that $N(u) \subseteq N(v)$. Then, the operation of deleting u from G is called a *folding* of G and we say u *folds onto* v. A graph is called *dismantlable* if there is a sequence of folds reducing it to a single vertex. We say u *eventually folds onto* v if there is a sequence $u_0 = u, u_1, \ldots, u_k = v$ such that u_i folds onto u_{i+1}, $0 \le i < k$. Let G be a dismantlable graph and ψ be a folding sequence reducing G to a single vertex v. We can visualize ψ as a tree T whose vertices are the vertices of G such that when rooted at v every vertex in T is folded onto its parent.

If a graph G is not dismantlable, this means that after a sequence of foldings ψ it reduces to a graph H which cannot be folded any further. We refer to the graph H as the *residual graph* of G, or just the *residual*, if G can be inferred from the context. It is known that the residual is unique up to isomorphism [6]. We can visualize the folding process for nondismantlable graphs as a forest of trees T_h hanging from each vertex $h \in H$ (see Figure 3). T_h is composed of vertices that eventually fold onto h. We define $\psi(u) = w$ if and only if $u \in T_w$, $w \in H$. We note that the tree representation depends on the folding sequence ψ and in general it is not unique.

3. A winning strategy with two hunters. In this section, we present a strategy with two hunters that catches the rabbit on any graph. In general, a single hunter cannot always capture the rabbit. This can be seen by considering a cycle of length at least 4 as the input graph: The rabbit's strategy is to wait until the hunter becomes visible and move to its neighbor which does not contain the hunter. This strategy guarantees that it will never get caught.

The strategy of the two hunters is divided into epochs that are comprised of two phases. An epoch starts with the hunters located at a predetermined vertex. The first phase starts at time $t = 1$.

In *Phase One*, two hunters move together and their goal is to see the rabbit. To achieve this, the hunters generate a random vertex label $v \in \{1 \ldots n\}$ and move together to v. Afterward, they wait at v until either $(t \bmod n) = 0$ or the rabbit becomes visible. If the rabbit becomes visible at any time, the first phase is over and the second phase starts. Otherwise, the hunters repeat the same process by generating

a new label v.

We claim that the first phase lasts only $n^2 \log n$ steps with high probability. To see this, let $r_1, r_2, \ldots$ be the location of the rabbit at times $n, 2n, 3n, \ldots$. Suppose the hunters have not seen the rabbit until time $i \times n$. At that time, the probability that they generate a label in $N(r_{i+1})$ is at least $\frac{1}{n}$. Since they generate a label after every n steps, the first phase will be over in $n^2 \log n$ steps with high probability.

In *Phase Two*, the hunters try to catch the rabbit as follows: Suppose the second phase starts at time $t = t_0$ and let $t_i = t_0 + i$. At that time both hunters H_1 and H_2 are at vertex h and the rabbit is at vertex r, with $r \in N(h)$. For the rest of the second phase, let r_i denote the position of the rabbit at time $t = t_i$ and let us define $r_0 = h$.

The strategy of H_1 is as follows: At time $t = t_i$, he is located at r_{i-1}. With probability $p_1 = \frac{1}{n^2}$, he attacks the rabbit by generating a random neighbor of r_{i-1} and going there in the next step. With probability $1 - p_1$, he chases the rabbit by going to r_i in the next step. The second phase ends with failure if H_1 attacks and misses the rabbit.

The strategy of H_2 is based on the following observation: If H_1 chases the rabbit for more than n steps, the rabbit must revisit a vertex by the pigeonhole principle. Let u be the first vertex revisited and suppose that at time t_r, the rabbit visits a vertex $v \in N(u)$ for the first time before revisiting u. The goal of H_2 is to enter v at the same time as the rabbit. To achieve this, first he guesses u, v, and t_r. In order to reach u, he chases H_1 by moving to his location in the previous time step until u. Afterward, H_2 waits until time $t = t_r - 1$ and goes to v from u. We say H_2 is in *chasing mode* if he is following H_1 and he is in *attacking mode* after he arrives at u. The second phase ends with failure if H_2 misses the rabbit when it arrives at v. To summarize, at time $t = t_0$, the hunters are at r_0 and the rabbit is at r_1. When the hunters are chasing, the locations of the rabbit H_1 and H_2 at time t_i are r_i, r_{i-1}, r_{i-2}, respectively. The phase ends when either hunter attacks. If no hunter attacks within n^2 steps, they end the phase and move to the predetermined vertex to start a new epoch.

Next, we state the crucial property of the strategy of the hunters.

LEMMA 1. *During Phase Two, the rabbit cannot distinguish between the modes of hunter H_2.*

Proof. If the attacking mode starts at time $t = t_1$, the location of H_2 is the same for both modes. If it starts afterward, we show that if the rabbit sees H_2, it will get caught with nonzero probability.

Suppose the rabbit sees H_2 at time $t = t_2$, which implies $r_2 \in N(r_0)$. In this case, with probability at least $\frac{p_1}{n}$, H_1 can decide to attack from r_0 to r_2 at time $t = t_1$ and catch the rabbit.

Next, suppose the rabbit sees H_2 at time $t > t_2$. If H_2 was in chasing mode at that time, the fact that the rabbit sees H_2 implies $r_i \in N(r_{i-2})$. In this case as well, H_1 could decide to attack in the previous step and catch the rabbit with probability $\frac{p_1}{n}$. Therefore H_2 must be invisible to the rabbit during the chasing mode. But, H_2 will also be invisible in the attacking mode because as soon as the rabbit enters a vertex v where it can see H_2, H_2 can catch it by guessing v and the arrival time correctly.

Therefore in order to avoid getting caught, the rabbit must avoid seeing H_2. But then the information available to the rabbit will be the same, no matter which mode H_2 is in: H_2 is out of its sight since the beginning of the second phase. ∎

LEMMA 2. *During Phase Two, the hunters succeed with nonzero probability.*

Proof. As discussed previously, after the start of the second phase, the rabbit must revisit a vertex u at time $k \leq n$. If the rabbit does not see H_2 until $t = k$, H_2 can catch it with probability $\frac{1}{n^3}$ at least by guessing $t_r, u, v \leq n$. Note that H_1 will still be chasing the rabbit with probability at least $1 - \frac{k}{n^2} \geq 1 - \frac{1}{n}$. On the other hand, if the rabbit sees H_2, it is caught with probability at least $\frac{1}{n^3} = \min\{\frac{p_1}{n}, \frac{1}{n^3}\}$, by Lemma 1. □

The length of an epoch is $O(n^2 \log n)$: Phase One lasts $O(n^2 \log n)$ time with high probability and Phase Two lasts $\Theta(n^2)$ steps. We have established that in Phase Two, the rabbit is caught with probability at least $\frac{1}{n^3}$. Therefore after $n^3 \log n$ epochs, each of which last $O(n^2 \log n)$ steps at most, the rabbit will be caught, yielding our main result.

THEOREM 3. *Two hunters can catch a rabbit with local visibility on any graph with high probability.*

4. Hunter-win graphs. In this section, we start the study of graphs on which a single hunter suffices. An interesting feature of the strategy of two hunters is that one hunter makes the rabbit move constantly and therefore forces it into making mistakes. This suggests that moving when a hunter is not visible may be a disadvantage for the rabbit.

To study this phenomenon we introduce reactive strategies where the rabbit moves only when the hunter is visible and ask the question of whether the class of hunter-win graphs against reactive graphs is equivalent to the class of hunter-win graphs against general rabbits. The answer turns out to be negative.

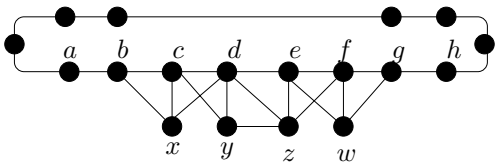


FIG. 2. *This graph is hunter-win against reactive rabbits but not against general rabbits.*

The graph in Figure 2 is hunter-win against reactive rabbits. The input graph consists of a cycle and the gadget shown in the figure. The hunter's strategy is to drive the rabbit into the gadget, by chasing it along the cycle. Once the rabbit is in the gadget, the hunter drives the rabbit to a vertex such that he can reach another vertex (without being seen) whose neighborhood dominates the rabbit's neighborhood. Next, we present the details of the hunter's strategy. In the following, without loss of generality, we assume that the hunter knows the rabbit's next move.

In order to capture the (reactive) rabbit, the hunter first chases it counterclockwise until the rabbit is at b and the hunter is at a. It can be easily verified that the rabbit cannot avoid reaching b without being captured.

If the rabbit moves to x from b, the hunter travels clockwise, arrives at c via y, and wins the game (note that the rabbit, being reactive, will not move in the meantime). Otherwise, if the rabbit moves to c, the hunter moves to b. In the next move, if the rabbit moves to y from c, the hunter travels clockwise, arrives at d through e, and wins the game. If the rabbit moves to d from c, then the hunter moves to c.

From d (while the hunter is at c), the rabbit has two options (it will be captured if it goes to y). If it moves to e from d, the hunter goes to y and then to z. The rabbit must then move to w to avoid capture. In this case the hunter goes to f and wins the

game. Otherwise, if the rabbit moves to z from d, the hunter travels clockwise again and arrives at e through g and w. From z the rabbit can only go to y, in which case the hunter moves to d from e and wins the game.

Therefore, no matter which strategy it chooses, the hunter can capture a reactive rabbit. However, once it arrives at b, a general rabbit can keep moving in the opposite direction of a until it leaves the gadget. If the length of the cycle is greater than 14, the hunter cannot reach the other entrance of the gadget before the rabbit and therefore a general rabbit is safe on this graph.

4.1. Characterization of hunter-win graphs against reactive rabbits. In this section, we describe an algorithm that recognizes hunter-win graphs against reactive rabbits. The algorithm marks configurations (h, r) according to the following rules.

Algorithm Mark-Reactive:
Mark all configurations (v, v) for every vertex v. (Initialization)
Repeat
Mark (h, r) if for all $r' \in N(r)$ there exists a vertex $h' \in N(h)$ with (h', r') marked. *(Stride Rule)*
For all (h', r) that are marked, mark (h, r) for all $h \in N(h') \setminus N(r)$. *(Stealth Rule)*
Until no further marking is possible.

Next, we prove the *soundness* (if all configurations are marked, then the graph is hunter-win) and *completeness* (if the graph is hunter-win, then all configurations will be marked) properties of the marking algorithm.

Soundness. The proof is by induction on the round k in which a configuration is marked.

When $k = 1$ only the configurations (v, v) are marked and the hunter trivially wins the game in these configurations.

Suppose the configurations marked in the first k rounds are sound and consider the configuration (h, r) marked during step $k + 1$. If (h, r) was marked using the Stride Rule, during the execution of the game, the hunter can force a configuration marked during the kth step with nonzero probability. Hence these configurations are sound. If, on the other hand, the configuration (h, r) is marked by the Stealth Rule, we observe that the rabbit will remain at vertex r since the hunter is out of its sight and hence the hunter can reach the configuration (h', r) which has been marked during the previous steps. Therefore the Stealth Rule is also sound by the inductive hypothesis.

Completeness. Clearly, if the rabbit is captured the game ends at a marked configuration. Otherwise, we show that the rabbit can always stay in an unmarked configuration and hence never get caught. Suppose there is an unmarked configuration (h, r) and the hunter and the rabbit are at vertices h and r, respectively. There are two cases: If $h \in N(r)$, the rabbit must have a move to a vertex r' such that there exists no $h' \in N(h)$ with (h', r') marked. Otherwise (h, r) would be marked by the Stride Rule. On the other hand, if $h \notin N(r)$, no matter which vertex h' the hunter moves, (h', r) is unmarked. Otherwise (h, r) would be marked by the Stealth Rule.

We can now state the result of this section which follows from the soundness and completeness of the marking algorithm.

THEOREM 4. *A graph G is hunter-win against reactive rabbits if and only if the algorithm Mark-Reactive marks all configurations.*

4.2. Characterization of hunter-win graphs against general rabbits. For reactive rabbits, it is easy to see that on a hunter-win graph every rabbit walk can be intercepted (i.e., the rabbit gets caught) by the hunter in $O(n^3)$ steps. However, it is far from being clear that such a polynomial length intercepting walk (i.e., a *witness*) exists for nonreactive rabbits. The difficulty is that at any point in time, the rabbit can infer a subset $H \subseteq V$ of possible hunter locations and plan its motion accordingly. This suggests that the state of the game may require specifying arbitrary subsets of vertices, potentially leading to exponential witnesses. Fortunately, we can establish a monotonicity property to establish once again polynomial size witnesses.

Let $\langle H, r\rangle$ be the state of the game where H is the set of possible hunter locations when the rabbit is at r. When the rabbit and the hunter are at adjacent vertices r and h, respectively, the rabbit knows the hunter's position with certainty and therefore $H = \{h\}$. Now suppose the game starts at configuration (h, r).

PROPOSITION 5. *The hunter can reach an adjacent configuration from any starting configuration* (h, r).

The proof of Proposition 5 is implicit in the strategy presented in section 3. During Phase One, the two hunters act as one and we showed that their strategy ensures that the hunters and the rabbit will end up in adjacent vertices in n steps with nonzero probability. This means that, no matter which path rabbit takes, there exists a hunter path of length at most n that leads to an adjacent configuration.

PROPOSITION 6. *A graph G is hunter-win if and only if the hunter wins starting from any adjacent configuration.*

Proof. If the graph is hunter-win, the hunter must win from all starting configurations including the adjacent ones. Conversely, if the hunter can win from any adjacent configuration, then starting from any configuration he can reach an adjacent configuration by Proposition 5 and win the game from here on. ∎

Therefore, by Proposition 6, on a hunter-win graph, we can assume that the game starts from an initial configuration where the players see each other. In addition, without loss of generality, we assume that the rabbit moves so as to maximize the time taken for capture and the hunter moves so as to minimize it.

We can view any hunter-win game as a sequence of rounds $R_1, \ldots, R_p$ where each round starts with the players located at adjacent vertices. Hence, the rabbit has full knowledge of the hunter's position. Clearly, there are at most n^2 rounds and the rounds do not repeat.

LEMMA 7. *For the optimal hunter strategy, the length of each round is $O(n^2)$.*

Proof. Partition the round into segments of length $n + 1$ each. The rabbit must revisit a vertex r within the same segment. Let $\langle H_1, r_1\rangle$ and $\langle H_2, r_2\rangle$ be the state of the game during the first and second visits. First, we show that $H_1 \subseteq H_2$. This is because, between r_1 and r_2, the rabbit cannot visit any vertex u with $u \in N(h)$, $h \in H_1$. If the hunter is at h, the rabbit would be captured. Next, if $H_1 = H_2$, then the part of the hunter strategy between r_1 and r_2 is redundant and hence the hunter can shorten the game. Therefore as the rabbit keeps visiting the same vertex, its uncertainty is monotonically increasing and after at most n revisits the state of the game becomes $\langle G - N(r), r\rangle$. In this case, either the rabbit gets caught if it moves or the hunter reveals himself, ending the round. Since the rabbit has to revisit a vertex every n steps and there are at most n revisits, the lemma follows. ∎

Since the length of a round is $O(n^2)$ and there are n^2 rounds, we conclude that the total length of a hunter-win game is $O(n^4)$.

Our characterization algorithm for general rabbits is based on the existence of such a polynomial size witness. *We will mark only adjacent configurations*: if the

adjacent configurations are all marked, by Proposition 6 the hunter wins from all starting configurations. A general rabbit can move even if the hunter is not visible. In order to capture this capability we need to generalize the stealth moves, described next.

4.2.1. Stealth moves. A *k-stealth move* from configuration (h, r) with $h \in N(r)$ to a marked configuration (h', r') is defined as follows: For every rabbit path $P_r = \{r, r_1, \ldots, r_k = r'\}$ of length k, the hunter has a path $P_h = \{h, h_1, \ldots, h_k = h'\}$ such that $h_i \notin N(r_i)$ for $i = 1, \ldots, k-1$, $h_k \in N(r_k)$, and (h_k, r_k) is marked. We refer to P_h as the *stealth path* corresponding to P_r. A configuration (h, r) is marked by the *Stealth Rule* if for all $r' \in N^k(r)$, there exists a k-stealth move to a marked configuration (h', r'). Note that the Stealth Rule for $k = 1$ subsumes the Stride Rule.

LEMMA 8. *The markings corresponding to stealth moves are sound.*

Proof. Suppose all previously marked adjacent configurations are sound and consider the next adjacent configuration (h, r) marked by a stealth move of length k. At time $t = 0$ the rabbit is located at r. Since we mark only the adjacent configurations, the state of the game is $\langle \{h\}, r \rangle$. Take any rabbit path of length k, and suppose at time $t = i$ the rabbit is at vertex r_i. Let $r'_1, \ldots, r'_p$ be the vertices accessible from r_i in the remaining $k - i$ steps and $P_1, \ldots P_p$ be the corresponding stealth paths such that at the end of k steps, P_j ends at vertex h'_j and (h'_j, r'_j) is marked. Let E_j be the event that the hunter has chosen path P_j, $j = 1, \ldots, p$, and let h_j be the jth vertex on P_j. The claim follows from the observation that no matter which path P_j the hunter chooses, the information available to the rabbit is the same—namely, the hunter was not visible for the last i steps. Therefore the state of the game is $\langle H, r \rangle$ where $\{h_j | 1 \leq j \leq p\} \subseteq H$. Since the rabbit cannot distinguish between the events E_j, no matter which final destination r'_j it chooses, the hunter can be at the corresponding vertex h_j and arrive at the already marked configuration (h'_j, r'_j). ∎

The stealth moves starting from configuration (h, r) and ending at configuration (h', r') can be computed efficiently by dynamic programming.

We will need an intermediate look-up table T, with $T[h, r, h', r', k] =$ TRUE if and only if for any rabbit path $\{r, r_1, \ldots, r_k = r'\}$ of length k there is a stealth path of length k that starts from h and ends at h'.

The entries of the table T are filled as follows:

(i) $T[h, r, h', r', 0] =$ TRUE if and only if $h = h'$, $r = r'$, and $h' \in N(r')$.

(ii) $T[h, r, h', r', 1] =$ TRUE if and only if $h' \in N(h)$, $r' \in N(r)$, and $h' \in N(r')$.

(iii) $T[h, r, h', r', k+1] =$ TRUE if and only if for all $u \in N(r)$ there is a vertex $v \in N(h) \setminus N(u)$ with $T[v, u, S, h', r', k] =$ TRUE for $1 \leq k \leq n^2$.

We now present a marking algorithm that uses the look-up table T to compute the stealth moves.

Algorithm Mark-General:
Mark all configurations (v, v) for every vertex v. (Initialization)
Repeat

For all configurations (h, r) with $h \in N(r)$, mark (h, r) if there exists an index $k \leq n^2$ such that for all $r' \in N^k(r)$ there exists a vertex h' with $T[h, r, h', r', k] =$ TRUE and (h', r') is marked. *(Stealth Rule)*
Until no further marking is possible.

LEMMA 9. *If the graph is hunter-win, then the marking algorithm Mark-General will mark all adjacent configurations.*

Proof. Let (h, r) be an adjacent configuration left unmarked after the execution of algorithm Mark-General. We claim that the rabbit can get to an adjacent configuration (h', r') that is unmarked. Suppose not. This means that for any rabbit path $r, r_1, r_2, \ldots, r_k$ there is a hunter path $h, h_1, h_2, \ldots, h_k$ with $h_k \in N(r_k)$ and (h_k, r_k) is marked. By Lemma 7, we have $k \leq n^2$. This implies that (h, r) would be marked by the Stealth Rule, which gives us the desired contradiction.

Therefore, starting from any unmarked adjacent configuration (h, r), the rabbit can reach another unmarked adjacent configuration. This means that the rabbit will never get caught, since a capture implies that the game enters the configuration (v, v) for some vertex v which is a marked adjacent configuration. ∎

THEOREM 10. *A graph G is hunter-win against general rabbits if and only if the algorithm Mark-General marks all adjacent configurations.*

Proof. If all the configurations are marked, G is hunter-win due to the fact that the Stealth Rule is sound (Lemma 8). Conversely, if there is an unmarked configuration, the rabbit is never caught by Lemma 9. ∎

5. Complete visibility and dismantlable graphs. When the rabbit has full visibility, the Stealth Rule does not make sense. In fact, we will show that the Stride Rule against reactive rabbits is sound and complete against rabbits with full visibility.

Algorithm Mark-FullVisibility:
Mark all configurations (v, v) for every vertex v.
Repeat
Mark (h, r) if for all $r' \in N(r)$ there exists a vertex $h' \in N(h)$ with (h', r') marked. *(Stride Rule)*
Until no further marking is possible.

It turns out that the algorithm Mark-FullVisibility recognizes hunter-win graphs against rabbits with full visibility.

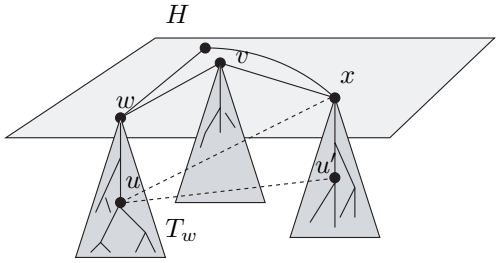


FIG. 3. *Visualization of the folding procedure for a nondismantlable graph. The vertices w, v, and x are in the residual H. Since there is no edge from w to x, the edges shown with dashed lines cannot exist.*

We will need the following property of nondismantlable graphs.

PROPOSITION 11. *Let G be a nondismantlable graph, ψ be a folding sequence, and H be the residual. Let x and w be two distinct vertices in H and T_x and T_w be the corresponding folding trees (see Figure 3). If there exists a vertex $u \in T_w$ that is adjacent to a vertex $u' \in T_x$, then $x \in N(w)$.*

Proof. Without loss of generality, suppose u was folded before u'. This implies that the parent of u must be adjacent to u'. We replace u with its parent and continue

this process of propagating the edge between u and u', which must eventually reach the roots w and x of the corresponding trees. ∎

THEOREM 12. *The algorithm Mark-FullVisibility marks all configurations if and only if the input graph is dismantlable.*

Proof. Suppose the input graph G is dismantlable. We can prove that all configurations will be marked by induction on the order of G. Since G is dismantlable, it must have two vertices u and v with $N(u) \subseteq N(v)$. Let $G' = G - \{u\}$ and run algorithm Mark-FullVisibility on G'. Suppose, inductively, that all configurations in G' are marked. Consider the marking algorithm for G which marks (u, u) first and simulates the marking algorithm on G' afterward. In addition, whenever (x, v) is marked for a vertex $x \in G'$, we also mark (x, u). This is possible since that (x, v) is marked implies that for all $v' \in N(v)$, there exists a vertex $x' \in N(x)$ with (x', v') marked and $N(u) \subseteq N(v)$. Next, we show that all the configurations (x, y) in G' will also get marked in G. Suppose there exists a configuration (x, y) that is marked in G' but not in G. Consider the first such configuration that is discovered in the marking of G. It must be that $u \in N(y)$ and that for all $x' \in N(x)$, (x', u) is not marked at this point. Also, $v \in N(y)$ since $N(u) \subseteq N(v)$. Now using the fact that (x, y) gets marked at this stage in G', we know that there exists $x'' \in N(x)$ such that (x'', v) is already marked. But then (x'', u) must also be marked at this point according to the modified marking rule. A contradiction! Thus, any (x, y) marked in G' will also be marked in G. It follows that for any x such that (x, v) is marked in G', we can mark (x, u) in G. It is easy to see that for any x, the configuration (u, x) will also be marked in G since u is adjacent to v and, by the argument above, for all $x' \in N(x)$, (v, x') is marked.

Now suppose the input graph is not dismantlable. Let ψ be a sequence of folds reducing G to a residual graph H. For any two vertices $u \in G$ and $v \in H$, we claim that (u, v) is unmarked if $\psi(u) \neq v$. Suppose this is not true and let (u, v) be the first marked configuration such that $\psi(u) \neq v$ (Figure 3). Let $w = \psi(u)$, $w \neq v$. Note that v must have a neighbor x such that $x \notin N(w)$; otherwise, v would fold onto w. When (u, v) gets marked, there must be a vertex $u' \in N(u)$ such that (u', x) is marked. If $\psi(u') = x$, this would imply $x \in N(w)$ by Proposition 11. So it must be the case that $\psi(u') \neq x$. But then, the fact that (u', x) is marked contradicts the fact that (u, v) is the first configuration marked with $\psi(u) \neq v$. Therefore, we conclude that if the graph is not dismantlable, the marking process will not mark all configurations. ∎

As stated earlier, it has been shown that the class of graphs that are hunter-win against rabbits with full visibility are precisely the class of dismantlable graphs [6]. Therefore we obtain the following corollary.

COROLLARY 13. *A graph G is hunter-win against rabbits with full visibility if and only if the algorithm Mark-FullVisibility marks all configurations.*

We know that there are nondismantlable graphs that are hunter-win against rabbits with local visibility. An example is shown in Figure 4. The labels on the vertices indicate their folding order: First, vertex 1 folds onto vertex 2; afterward, vertex 2 folds onto vertex 9, etc. After folding vertices 1 to 8, vertices 9 to 12 cannot be folded, leaving a four-cycle as the residual. Therefore this graph is not dismantlable and consequently it is not hunter-win against rabbits with full visibility. To see that the hunter wins against rabbits with local visibility, let us define the mapping $p : V \to V$, where V is the set of vertices. For $v \in V$ with $1 \leq v \leq 8$, $p(v)$ is the vertex which v folds onto. We define $p(9) = 2$, $p(10) = 8$, $p(11) = 6$, and $p(12) = 4$. The first observation is that the hunter wins the game if he can force the rabbit to go to vertex 1 while he is at vertex 2. Next, we observe that if the rabbit is at vertex $v \neq 1$ and the

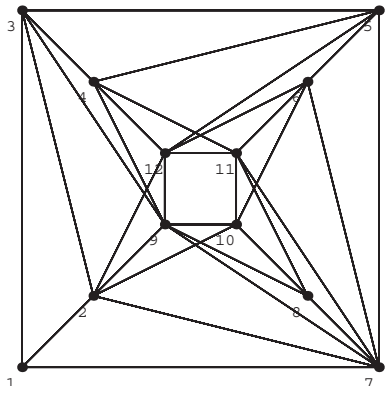


FIG. 4. *This graph is hunter-win against rabbits with local visibility. However, a rabbit with full visibility never gets caught.*

hunter is at $p(v)$, the rabbit must move to a lower numbered vertex. Now suppose the rabbit is reactive. In this case, it can be verified that for any rabbit location r and for any hunter location $h \notin N(r)$, the hunter has a path to $p(r)$ that does not enter $N(r)$. Therefore, by visiting $p(r)$ repeatedly the hunter can force a reactive rabbit to eventually move to vertex 1 and win the game afterward.

Hence, the rabbit must have a nonreactive strategy, meaning that it must move when the hunter is not visible. Consider the first time this happens: Suppose the hunter and the rabbit are at vertices h and r with $h \in N(r)$ and the rabbit takes the path $r \to r' \to r''$ such that the hunter is not visible from r'. It can be shown, by enumeration, that for any such vertices h, r, r', and r'', the hunter has a path $h \to h' \to r''$ that captures the rabbit. Therefore the rabbit cannot have a nonreactive strategy either and the graph is hunter-win against both types of rabbits.

We conclude this section with an interpretation of Theorem 12: If G is a graph that is hunter-win against rabbits with local visibility but not against rabbits with full visibility, the hunter captures the rabbit with local visibility using the stealth moves.

5.1. Hunter strategy for dismantlable graphs. Given a folding tree T rooted at vertex v, consider the vertex r where the rabbit is located. We say the hunter is an *ancestor* of the rabbit if he is located on the path from r to v. Suppose the vertices of T are ordered by their deletion times. The hunter strategy is based on the following two lemmas.

LEMMA 14. *The hunter can always maintain ancestry.*

Proof. Suppose the hunter is at vertex h and is an ancestor of the rabbit who is located at vertex r. Let r' be the rabbit's location in the next round. If h is a common ancestor of r and r' on the folding tree T, then the lemma is trivially true. Otherwise, since h is an ancestor of r and (r, r') is an edge, using basic properties of foldings it can be shown that h is adjacent to a vertex on the path that connects r' to the root of T. We show that there is always such a vertex h' with $h' \geq r'$ by a case analysis on r' (see Figure 5). Suppose for contradiction $h' < r'$. We will show that h must be adjacent to r' thus allowing the hunter to catch the rabbit in one step.

Case $(h > r' > r)$. In this case all the ancestors of h' deleted before h (including r') must have edges to h.

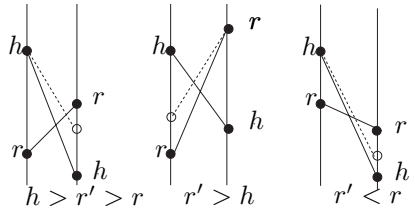


FIG. 5. *The hunter can always stay above the rabbit. The height of a vertex is proportional to its label.*

Case ($r' > h$). All the ancestors of r deleted before r' (including h) must have an edge to r'.

Case ($r' < r$). All the ancestors of h' deleted before r (including r') must have an edge to h. □

In fact, not only can the hunter maintain ancestry, but he can also reduce his height in the tree gradually and therefore get closer and closer to the rabbit.

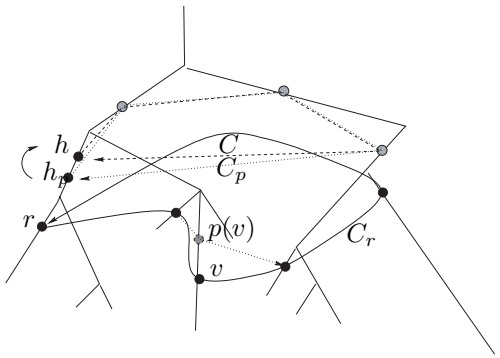


FIG. 6. *The hunter can make progress every time the rabbit revisits a vertex.*

LEMMA 15. *Every time the rabbit revisits a vertex, the hunter can reduce its height in the tree while maintaining ancestry.*

Proof. Fix any rabbit cycle C_r and let v be the vertex with the lowest label on this cycle and $p(v)$ be its parent (see Figure 6). Since v was deleted first, $p(v)$ must have edges to the neighbors of v on the cycle, so we can make a new cycle by replacing v with $p(v)$. We continue this process until the cycle reaches h, the location of the hunter (this must happen since the hunter is an ancestor at all times). Let us call this cycle C. Let C_p be the cycle just before C which contains h's child h_p, instead of h. Consider the path $P = \{h\} \cup (C \cap C_p) \cup \{h_p\}$. If the rabbit follows the cycle C_r, the hunter can follow the path P and end up at h_p which is lower than h. □

We are now ready to present the hunter strategy on a dismantlable graph G. First, the hunter builds the folding tree T for any folding sequence ψ. Afterward, he simply guesses the vertex the rabbit will jump to and jumps to the lowest possible ancestor of this vertex (see Figure 6). By Lemma 14 he can always remain an ancestor of the rabbit. Further, he can reduce his height in T every time the rabbit revisits a vertex (Lemma 15). Since the tree has a finite height, he can eventually catch the rabbit.

5.2. Extension to nondismantlable graphs. For nondismantlable graphs, we can extend the notion of ancestry as follows. Suppose the rabbit is at r and the hunter is at h. We say the hunter is an ancestor of the rabbit if there is a folding of the vertices such that in the corresponding forest representation, h is located on the path from r to the root of the tree that contains r. Once the hunter establishes ancestry, it is easy to see that Lemmas 14 and 15 still hold—both for reactive and general rabbits. Therefore the hunter can win the game afterward. Note that the hunter can trivially establish ancestry on dismantlable graphs.

In addition, if we define each vertex as its trivial parent, it is clear that the rabbit wins the game if the hunter can never become an ancestor. Therefore the class of hunter-win graphs is precisely the class of graphs on which the hunter can become an ancestor. One can view the stealth moves as giving the hunter the power to become an ancestor on nondismantlable but hunter-win graphs such as the one in Figure 4.

6. Extending the rabbit's visibility. Let us define rabbits with *i-visibility* as the rabbits who can see all vertices within distance i. It is known that one hunter always suffices to catch rabbits with 0-visibility [1]. In this paper, we studied rabbits with 1-visibility and established that two hunters always suffice to catch such rabbits. A natural question is how many hunters suffice when the rabbit has i-visibility.

Surprisingly, the number of hunters required for 2-visibility is unbounded: Consider the random bipartite graph $G = (U, V, E)$ with $|U| = |V| = n$ and each edge (u, v) is added with probability $1/\sqrt{n}$.

For an arbitrary vertex u, let x_i be the 0/1 random variable, which takes the value 1 if and only if $(u, i) \in E$. The size of $N(u)$ then becomes a random variable $X = \sum_i x_i$ with the expected value of $E[X] = n \cdot \frac{1}{\sqrt{n}} = \sqrt{n}$.

Using the Chernoff bound (see [14, p. 70]) with $\delta = 0.5$,

$$\Pr[X < (1-\delta)E[X]] < \exp(-E[X]\delta 2/2) = \exp(-\sqrt{n}/8). \tag{1}$$

Let E_1 be the event that a vertex has neighborhood of size less than $\sqrt{n}/2$. Using the union bound and (1), the probability of E_1 is at most $\frac{n}{\exp(\sqrt{n}/8)}$.

Let us also define the random variable y_i which takes the value 1 if and only if $(u, i) \in E$ and $(v, i) \in E$. Here, $v \neq u$ is an arbitrary vertex. Let $Y = \sum_i y_i$ be the size of the common neighborhood $N(u) \cap N(v)$ with $E[Y] = n \cdot \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}} = 1$.

To bound the value of Y, we use the equation (see [14, p. 71])

$$\Pr[Y > (1+\delta)E[Y]] < 2^{-(1+\delta)E[Y]} = \frac{1}{n^3}, \tag{2}$$

where δ is chosen such that $(1 + \delta) = 3\log(n)$.

Let E_2 be the event that no two vertices have a common neighborhood of size greater than $3\log(n)$. Summing (2) over all pairs of vertices and using the union bound, we get that the probability of E_2 is at most $\frac{1}{n}$.

The probability that neither of the events, E_1 and E_2, happen is at least

$$p = 1 - \frac{n}{e^{\sqrt{n}/8}} - \frac{1}{n}. \tag{3}$$

Since p becomes nonzero as n grows large, this means that for any (large) n, there exists a graph G^* where every vertex has at least $\frac{\sqrt{n}}{2}$ neighbors and the common neighborhood of any two vertices has size at most $3\log n$.

Now suppose a rabbit with 2-visibility is evading G^*. Note that the rabbit can see the hunters all the time. Without loss of generality, suppose the rabbit is located at a vertex $u \in U$. We can also assume that all the hunters are located in U without any decrease in their power. It easy to see that, on G^*, the number of hunters required is at least $(\frac{\sqrt{n}}{2})/(3 \log n) = \tilde{\Omega}(\sqrt{n})$. Otherwise the rabbit will always have a safe vertex not accessible by the hunters.

7. Concluding remarks. In this paper, we have studied a pursuit-evasion game where the players have only local visibility. We showed that two hunters can catch the rabbit with high probability on any graph. In addition, we presented an algorithmic characterization of graphs on which a single hunter suffices for capture. To the best of our knowledge, this is the only pursuit-evasion game in the literature where the pursuers' strategy explicitly exploits the local visibility of the evader.

An important aspect of the game is the time required to catch the rabbit. For 0-visibility, one hunter succeeds in time $O(n \log n)$ [1]. For 1-visibility we showed that two hunters succeed in $\tilde{O}(n^5)$ time. However, it is not clear whether a single hunter can catch a rabbit on a hunter-win graph in polynomial time. We leave this as a direction for future work.

Acknowledgment. The authors would like to thank Sudipto Guha for several useful discussions.

REFERENCES

[1] M. Adler, H. Räcke, N. Sivadasan, C. Sohler, and B. Vöcking, *Randomized pursuit-evasion in graphs*, in Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP), Málaga, Spain, 2002, pp. 901–912.

[2] M. Aigner and M. Fromme, *A game of cops and robbers*, Discrete Appl. Math., 8 (1984), pp. 1–12.

[3] T. Basar and P. R. Kumar, *On worst case design strategies*, Comput. Math. Appl., 13 (1987), pp. 239–245.

[4] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory*, 2nd ed., Classics Appl. Math. 23, SIAM, Philadelphia, 1998.

[5] P. Bernhard, A.-L. Colomb, and G. P. Papavassilopoulos, *Rabbit and hunter game: Two discrete stochastic formulations*, Comput. Math. Appl., 13 (1987), pp. 205–225.

[6] G. R. Brightwell and P. Winkler, *Gibbs measures and dismantlable graphs*, J. Combin. Theory Ser. B, 78 (2000), pp. 141–166.

[7] J. A. Ellis, I. H. Sudborough, and J. S. Turner, *The vertex separation and search number of a graph*, Inform. and Comput., 113 (1994), pp. 50–79.

[8] S. Fitzpatrick and R. Nowakowski, *Copnumber of graphs with strong isometric dimension two*, Ars Combin., 59 (2001), pp. 65–73.

[9] L. J. Guibas, J.-C. Latombe, S. M. LaValle, D. Lin, and R. Motwani, *A visibility-based pursuit-evasion problem*, Internat. J. Comput. Geom. Appl., 9 (1999), pp. 471–493.

[10] J. P. Hespanha, G. J. Pappas, and M. Prandini, *Greedy control for hybrid pursuit-evasion games*, in Proceedings of the European Control Conference, Porto, Portugal, 2001, pp. 2621–2626.

[11] I. Chatzigiannakis, S. Nikoletseas, and P. Spirakis, *An efficient communication strategy for ad-hoc mobile networks*, in Proceedings of the 15th Symposium on Distributed Computing (DISC'2001), University of Lisbon, Lisbon, Portugal, 2001, pp. 285–299.

[12] A. S. LaPaugh, *Recontamination does not help to search a graph*, J. ACM, 40 (1993), pp. 224–245.

[13] N. Megiddo, S. L. Hakimi, M. R. Garey, D. S. Johnson, and C. H. Papadimitriou, *The complexity of searching a graph*, J. ACM, 35 (1988), pp. 18–44.

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

[15] S. Neufeld and R. Nowakowski, *A game of cops and robbers played on products of graphs*, Discrete Math., 186 (1998), pp. 253–268.

[16] R. Nowakawski and P. Winkler, *Vertex-to-vertex pursuit in a graph*, Discrete Math., 43 (1983), pp. 235–239.

[17] S.-M. Park, J.-H. Lee, and K.-Y. Chwa, *Visibility-based pursuit-evasion in a polygonal region by a searcher*, in Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP), Lecture Notes in Comput. Sci. 2076, Springer-Verlag, New York, 2001, pp. 456–468.

[18] T. D. Parsons, *Pursuit evasion in a graph*, in Theory and Application of Graphs, Y. Alavi and D. R. Lick, eds., Springer-Verlag, New York, 1976, pp. 426–441.

[19] R. Vidal, O. Shakernia, J. Kim, D. Shim, and S. Sastry, *Probabilistic pursuit-evasion games: Theory, implementation and experimental evaluation*, IEEE Trans. Robotics and Automation, 18 (2002), pp. 662–669.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 42–48

ODD HOLE RECOGNITION IN GRAPHS OF BOUNDED CLIQUE SIZE*

MICHELE CONFORTI†, GÉRARD CORNUÉJOLS‡, XINMING LIU§, KRISTINA VUŠKOVIĆ¶, AND GIACOMO ZAMBELLI‖

Abstract. In a graph G, an odd hole is an induced odd cycle of length at least 5. A clique of G is a set of pairwise adjacent vertices. In this paper we consider the class $\mathcal{C}_k$ of graphs whose cliques have a size bounded by a constant k. Given a graph G in $\mathcal{C}_k$, we show how to recognize in polynomial time whether G contains an odd hole.

Key words. odd hole, recognition algorithm, cleaning, decomposition

AMS subject classification. 05C17

DOI. 10.1137/S089548010444540X

1. Introduction. A *hole* is a graph induced by a cycle of length at least 4. A hole is *odd* if it contains an odd number of vertices. Otherwise, it is even. Graph G *contains* graph H if H is isomorphic to an induced subgraph of G. Chudnovsky, Cornuéjols, Liu, Seymour, and Vušković recently proved that it is polynomial to test whether a graph contains an odd hole or its complement [2]. However, it is still an open problem to test whether a graph contains an odd hole. Bienstock [1] proved that it is NP-complete to test whether a graph contains an odd hole passing through a specific vertex. A *clique* is a set of pairwise adjacent vertices. The *clique number* of a graph is the size of its largest clique. In this paper, we show that it is polynomial to test whether a graph of bounded clique number contains an odd hole.

We use the same general strategy as in [2]. Let H be an odd hole in a graph G. We say that $u \in V(G) \setminus V(H)$ is *H-minor* if its neighbors in H lie in some 2-edge path of H. In particular, u is H-minor if u has no neighbor in H. A vertex $u \in V(G) \setminus V(H)$ is *H-major* if it is not H-minor. We say that H is *clean* if G contains no H-major vertex. A graph G is *clean* if either it is odd-hole-free or it contains a clean shortest odd hole. As in [2] our approach for testing whether a graph G of bounded clique number contains an odd hole consists of two steps:

(i) constructing in polynomial time a clean graph G' that contains an odd hole if and only if G does, or in some cases identifying an odd hole of G, and
(ii) checking whether the clean graph G' contains an odd hole.

*Received by the editors August 1, 2004; accepted for publication (in revised form) August 26, 2005; published electronically February 15, 2006.
http://www.siam.org/journals/sidma/20-1/44540.html

†Dipartimento di Matematica Pura ed Applicata, Università di Padova, Via Belzoni 7, 35131 Padova, Italy (conforti@math.unipd.it).

‡Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213 and LIF, Faculté des Sciences de Luminy, 13288 Marseille, France (gc0v@andrew.cmu.edu). The work of this author was supported by NSF grant DMI-0352885 and ONR grant N00014-03-1-0133.

§Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213 (xinming.liu@gmail.com).

¶School of Computing, University of Leeds, Leeds LS2 9JT, UK (vuskovi@comp.leeds.ac.uk). The work of this author was supported by EPSRC grant GR/R35629/01.

‖Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada (gzambell@math.uwaterloo.ca).

For step (ii), we can use the polynomial algorithms in [2]. The main result of this paper is a polynomial algorithm for step (i). Step (i) is called *cleaning* the graph G.

1.1. Notation. For a graph G and a set B of vertices of G, we denote by $G(B)$ the subgraph of G induced by the vertex set B. For a vertex v, $N(v)$ denotes the set of vertices adjacent to v.

A *pyramid* $\Pi(xyz;u)$ is a graph induced by three paths $P_1 = x,\dots,u$, $P_2 = y,\dots,u$, and $P_3 = z,\dots,u$ having no common or adjacent intermediate vertices, such that at most one of the paths is of length 1 and the vertex set $\{x,y,z\}$ induces a clique of size 3. Note that every two of the paths P_1,P_2,P_3 induce a hole. Since two of the three paths must have the same parity, one of these holes is odd. Therefore, every pyramid contains an odd hole.

A *wheel*, denoted by (H,x), is a graph induced by a hole H and a vertex $x \notin V(H)$ having at least three neighbors in H, say, $x_1,\dots,x_n$. Vertex x is the *center* of the wheel. A subpath of H connecting x_i and x_j is a *sector* if it contains no intermediate vertex x_l, $l \in \{1,\dots,n\}$. A *short sector* is a sector of length 1, and a *long sector* is a sector of length at least 2. A wheel is *odd* if it contains an odd number of short sectors and *even* otherwise. Each of the long sectors together with vertex x induces a hole. If each of these holes is even and the wheel (H,v) is odd, then H is an odd hole, since the wheel (H,x) contains an odd number of short sectors. Therefore, every odd wheel contains an odd hole.

In a graph G, a *jewel* is a sequence $v_1,\dots,v_5$, P such that $v_1,\dots,v_5$ are distinct vertices, v_1v_2, v_2v_3, v_3v_4, v_4v_5, v_5v_1 are edges, v_1v_3, v_2v_4, v_1v_4 are nonedges, and P is a path of G between v_1 and v_4 such that v_2, v_3, v_5 have no neighbors in $V(P)\setminus\{v_1,v_4\}$. Clearly a jewel contains either an odd wheel or a 5-hole, so if there is a jewel in a graph G, then there is an odd hole in G.

Chudnovsky and Seymour found an $O(|V(G)|^9)$ algorithm to test whether a graph G contains a pyramid and an $O(|V(G)|^6)$ algorithm to test whether a graph G contains a jewel (see [2]).

2. Cleaning. In this section, we show how to clean a graph G of bounded clique number. That is, we perform step (i) above. The cleaning algorithm produces a polynomial family of induced subgraphs of G such that if G contains a shortest odd hole H^*, then one of the graphs produced by the cleaning algorithm, say, G', contains H^* and H^* is clean in G'.

Roughly speaking, this is accomplished by showing that there exists a set X of vertices of H^*, whose size depends only on the clique number, such that every major vertex for H^* has a neighbor in X. Since the set Y of vertices of H^* with neighbors in X has at most $2|X|$ elements, we may enumerate all possible choices for X and Y, and for each choice of X and Y add to the family the graph obtained by removing the vertices of $V(G)\setminus Y$ that have a neighbor in X.

2.1. Vertices with at most three neighbors in H^*.

LEMMA 1. *Let H^* be a shortest odd hole in G. Suppose that G does not contain a pyramid. If a vertex $u \notin V(H^*)$ has a neighbor but no more than three neighbors in H^*, then u is H^*-minor.*

Proof. If u has one neighbor in H^*, then u is H^*-minor. Now suppose that u has two neighbors in H^*, say, u_1 and u_2. Let P_1 and P_2 be the two u_1u_2-subpaths of H^*. Since H^* is odd, P_1 and P_2 have different parity; say, P_1 is odd. If P_1 is of length 1, then u is H^*-minor. Otherwise, $V(P_1)\cup\{u\}$ induces an odd hole. Since this hole cannot be shorter than H^*, P_2 is of length 2, and hence u is H^*-minor.

Now assume that u has three neighbors in H^*, and let P_1, P_2, and P_3 be the three sectors of the wheel (H^*, u). If exactly one of the sectors is short, then $V(H^*) \cup \{u\}$ induces a pyramid. If two of the sectors are short, then u is H^*-minor. Finally suppose that all three sectors are long. Since H^* is odd, at least one of the sectors, say, P_1, is odd. Then $V(P_1) \cup \{u\}$ induces an odd hole shorter than H^*, which is a contradiction. □

2.2. Vertices with more than three neighbors in H^*. Let H^* be a shortest odd hole in G. Let $S(H^*)$ be the set of H^*-major vertices that have four or more neighbors in H^*. Note that, for any $u \in S(H^*)$, every long sector of the wheel (H^*, u) is of even length since H^* is a shortest odd hole of G; hence, (H^*, u) contains an odd number of short sectors.

Let $S \subseteq V(G)$. We say that vertex $x \in V(G) \setminus S$ is *S-complete* if x is adjacent to every vertex in S. We say that an edge xy is *S-complete* if both vertices x and y are S-complete.

LEMMA 2. *Let H^* be a shortest odd hole in G. Suppose that G does not contain a jewel. If $u, v \in S(H^*)$ are not adjacent, then an odd number of edges of H^* are $\{u, v\}$-complete.*

Proof. Let u and v be nonadjacent vertices of $S(H^*)$. Suppose that an even number of edges of H^* are $\{u, v\}$-complete. Then some long sector P of the wheel (H^*, u) contains an odd number of short sectors of (H^*, v). Let u_1 and u_2 be the endvertices of P. P has even length. Let P' be the subpath of H^* induced by $(V(H^*) \setminus V(P)) \cup \{u_1, u_2\}$. P' has odd length. Note that P' must be of length at least 4, since otherwise (H^*, u) is a jewel, which is a contradiction. If P contains three or more neighbors of v, then the vertex set $V(P) \cup \{u, v\}$ induces an odd wheel with center v, and hence contains an odd hole shorter than H^*, contradicting our choice of H^*. Otherwise, let v_1 and v_2 be the two neighbors of v in P. Vertex v cannot have exactly four neighbors in H^*, say, v_1, v_2, v_3, v_4, such that both v_3u_1 and v_4u_2 are edges, because otherwise the vertex set $(V(H^*) \setminus V(P)) \cup \{v\}$ induces a shorter odd hole than H^*, since P is even and P' is of length at least 4. Therefore, there exist vertices $u_3, v_3 \in V(H^*) \setminus V(P)$, the neighbors of u and v, respectively, such that u and v have no other neighbors on u_3v_3-subpath of H^* (call it Q) and vertices u_3 and v_3 are not adjacent to u_1 or u_2. But now the vertex set $V(Q) \cup V(P) \cup \{u, v\}$ induces a pyramid $\Pi(v_1v_2v; u)$, and hence contains an odd hole shorter than H^*, contradicting our choice of H^*. □

The following, which is an easy consequence of Lemma 2, will be used in several places.

LEMMA 3. *Let H^* be a shortest odd hole in G, P be a subpath of H^* such that $|V(H^*) \setminus V(P)| \geq 3$, and x, y be two nonadjacent vertices in $S(H^*)$. Assume that no ends of P are $\{x, y\}$-complete and there is no $\{x, y\}$-complete edge in P. Then there exists an $\{x, y\}$-complete vertex in H^* with no neighbor in P.*

Proof. By Lemma 2, there exists an $\{x, y\}$-complete edge e in H^*. One of the two endvertices of e has the desired property. □

LEMMA 4. *Suppose that G does not contain a jewel. If $A \subseteq S(H^*)$ is a stable set, then an odd number of edges of H^* are A-complete.*

Proof. Let $A \subseteq S(H^*)$ be a stable set and suppose that an even number of edges of H^* are A-complete. Let A' be a smallest subset of A with the property that an even number of edges of H^* are A'-complete. Note that by Lemma 2, $|A'| \geq 3$. Let $s_1, \ldots, s_m$ be the vertices of H^* adjacent to at least one vertex in A', encountered in that order when traversing H^* clockwise. For $i \in [m]$, let S_i be the s_is_{i+1}-subpath

of H^* (indices taken modulo m) that does not contain any intermediate vertex s_j, $j \in [m]$.

Claim. For every $i \in [m]$, either S_i is an edge whose endvertices are both adjacent to some vertex $x \in A$, or S_i has even length.

Proof of claim. If there is a vertex $x \in A'$ adjacent to both s_i and s_{i+1}, then S_i is a sector of the wheel (H^*, x) and hence the result holds. Otherwise, let x_1 and x_2 be vertices of A' such that x_1 is adjacent to s_i and x_2 is adjacent to s_{i+1}. By Lemma 3 there exists an $\{x_1, x_2\}$-complete vertex u in H^* with no neighbor in S_i. Then the vertex set $V(S_i) \cup \{x_1, x_2, u\}$ induces a hole. Since both x_1 and x_2 have at least four neighbors in H^*, this hole is shorter than H^*, so it must be even; hence S_i is of even length. This completes the proof of the claim. □

For $C \subseteq A'$, let δ_C denote the number of edges of H^* that are C-complete. Let δ be the number of paths in $S_1, \ldots, S_m$ of length 1. Then

$$\delta = \sum_{i=1}^{|A'|} (-1)^{i+1} \sum_{C \subseteq A', |C|=i} \delta_C.$$

By the choice of A', for every $C \subseteq A'$ such that $C \neq A'$, δ_C is odd. Hence the parity of δ is equal to the parity of

$$\sum_{i=1}^{|A'|-1} \left(\begin{array}{c} |A'| \\ i \end{array} \right) + \delta_{A'},$$

which is itself equal to the parity of $\delta_{A'}$ since

$$\sum_{i=1}^{|A'|-1} \left(\begin{array}{c} |A'| \\ i \end{array} \right) = 2^{|A'|} - 2.$$

By the claim and because H^* is odd, δ is odd. Hence $\delta_{A'}$ must be odd as well, contradicting the choice of A'. □

THEOREM 5. *Suppose that G does not contain a jewel. Let A be a stable set of $S(H^*)$ and let x_1x_2 be an edge of H^* such that every vertex of A is adjacent to both x_1 and x_2 (such an edge exists by Lemma 4). Let B be the set of vertices of $S(H^*)$ that have no neighbor in $\{x_1, x_2\}$ and have both a neighbor and a nonneighbor in A. Then there exists an edge y_1y_2 of H^* such that y_1 is A-complete and every vertex of B has a neighbor in $\{y_1, y_2\}$.*

Proof. If $B = \emptyset$ then the result is trivially true, so we may assume that $B \neq \emptyset$. Since every vertex of B is major, this implies that H^* is of length greater than 5.

CLAIM 1. *For every $u \in B$, an edge of H^* is $(A \cup \{u\})$-complete.*

Proof of Claim 1. Let A_1 be the neighbors of u in A and $A_2 = A \setminus A_1$. By Lemma 4, there is an edge u_1u_2 of H^* such that every vertex of $A_2 \cup \{u\}$ is adjacent to both u_1 and u_2. Since u has no neighbor in $\{x_1, x_2\}$, every vertex of A_1 must be adjacent to both u_1 and u_2, or else there is a 5-hole. This completes the proof of Claim 1. □

CLAIM 2. *If X is a stable set of B, then there exists an edge z_1z_2 of H^* such that z_1 is A-complete and every vertex of X has a neighbor in $\{z_1, z_2\}$.*

Proof of Claim 2. We consider the following two cases.

Case 1. There is a vertex in A that is not adjacent to any vertex in X.

Let $A_1 \subseteq A$ be such that $A_1 \cup X$ is a maximal stable set. By Lemma 4, an edge of H^* is $(A_1 \cup X)$-complete—say, u_1u_2. Let $w \in A \setminus A_1$. Note that w is adjacent to some

$x \in X$. If w is not adjacent to u_1 or u_2, then there is a 5-hole in the graph induced by $\{x, y, w, u_1, u_2, x_1, x_2\}$, where $y \in A_1$. So every vertex of $A \setminus A_1$ is adjacent to both u_1 and u_2.

Case 2. Every vertex of A is adjacent to some vertex in X.

By Claim 1 and Case 1, we may assume w.l.o.g. that $|X| > 1$ and for every proper subset of X the result holds. Let $w \in A$ be such that $|N(w) \cap X|$ is minimum. Let $Z = N(w) \cap X$. Since every vertex of X has a nonneighbor in A and $|Z|$ is minimum, $|Z| < |X|$. By our assumption, there exists an edge y_1y_2 of H^* such that y_1 is A-complete and every vertex of $X \setminus Z$ has a neighbor in $\{y_1, y_2\}$. By Lemma 4 an edge of H^* is X-complete—say, edge y_3y_4.

We may assume that vertices y_1, y_2, y_3, y_4 are all distinct and y_1y_3 and y_1y_4 are not edges, since otherwise the result trivially holds. Also w.l.o.g. y_2y_4 is not an edge.

Suppose that wy_4 is not an edge. We may assume that some $z \in Z$ is not adjacent to y_1, since otherwise the edge y_1y_2 satisfies the claim. If some $v \in X \setminus Z$ is adjacent to y_1, then $\{y_1, v, w, z, y_4\}$ induces a 5-hole. So for every $v \in X \setminus Z$, vy_1 is not an edge, and hence vy_2 is an edge. If w is adjacent to y_2, then $\{y_2, w, v, z, y_4\}$ induces a 5-hole. So w is not adjacent to y_2. By Lemma 3, there is a vertex u of H^* adjacent to both v and w, but with no neighbor in $\{y_1, y_2\}$. Then $\{y_1, y_2, u, v, w\}$ induces a 5-hole.

Therefore wy_4 is an edge. We now show that y_4 is A-complete. Let $w' \in A$ and assume $w'y_4$ is not an edge. By the choice of w and by the above argument, there is a vertex $v \in X \setminus Z$ adjacent to w'. But then the graph induced by $\{w, w', x_1, x_2, v, y_4\}$ contains a 5-hole. This completes the proof of Claim 2. □

Claim 3. *For every edge v_1v_2 in $G(B)$, there exists $v \in A$ that is adjacent to neither v_1 nor v_2.*

Proof of Claim 3. Let A_1 be the set of neighbors of v_1 in A, and $A_2 = A \setminus A_1$. Suppose the claim does not hold. Then v_2 is universal for A_2. Let w_1 be a vertex of A_1 that v_2 is not adjacent to. Then $v_1, v_2, w_2, x_2, w_1, v_1$, where $w_2 \in A_2$, is a 5-hole. This completes the proof of Claim 3. □

By Claim 1, we may assume that for every proper subset B' of B, the statement holds. By Claim 2 we may assume that B is not a stable set. Let v_1v_2 be an edge of $G(B)$. By Claim 3, let v be a vertex of A that is adjacent to neither v_1 nor v_2. Let y_1y_2 be an edge of H^* such that y_1 is A-complete and all vertices of $B \setminus v_2$ have a neighbor in $\{y_1, y_2\}$. Let y_3y_4 be an edge of H^* such that y_3 is A-complete and all vertices of $B \setminus v_1$ have a neighbor in $\{y_3, y_4\}$. Then the theorem follows from the following claim.

Claim 4. *v_1 has a neighbor in $\{y_3, y_4\}$, or v_2 has a neighbor in $\{y_1, y_2\}$.*

Proof of Claim 4. Suppose the claim does not hold. v_1 has no neighbor in $\{y_3, y_4\}$ and v_2 has no neighbor in $\{y_1, y_2\}$.

If a vertex of $\{y_1, y_2\}$ coincides with a vertex of $\{y_3, y_4\}$, then $\{y_1, y_2, y_3, y_4, v_1, v_2\}$ induces a 5-hole. Therefore, vertices y_1, y_2, y_3, y_4 are all distinct.

We now show that v and v_1 must have a common neighbor in $\{y_1, y_2\}$. Assume not. Then vy_1 and v_1y_2 are edges, and vy_2 and v_1y_1 are not. By Lemma 3, there is a vertex u of H^* that is $\{v, v_1\}$-complete but has no neighbor in $\{y_1, y_2\}$. Then $\{y_1, y_2, v, v_1, u\}$ induces a 5-hole. Therefore, v and v_1 have a common neighbor y in $\{y_1, y_2\}$, and similarly v and v_2 have a common neighbor y' in $\{y_3, y_4\}$. If yy' is not an edge, then $\{y, y', v, v_1, v_2\}$ induces a 5-hole. Therefore, yy' is an edge.

Let a, y, y', b be the subpath of H^* induced by $\{y_1, y_2, y_3, y_4\}$. Then vy, vy', v_1y, v_2y' are edges and v_2a, v_2y, v_1y', v_1b are not.

Let z_2 be the neighbor of v_2 in H^* that is closest to a in $H^* \setminus \{y, y'\}$. Note that $z_2 \neq b$ since v_2 is a major vertex. Let P_2 be the az_2-subpath of H^* that does not contain y.

Suppose v does not have a neighbor in P_2. By Lemma 3, some vertex u of H^* is $\{v, v_2\}$-complete and has no neighbor in P_2. Note that $u \neq b$ since b is not $\{v, v_2\}$-complete. But then $P_2 \cup \{y, y', v, v_2, u\}$ induces a pyramid $\Pi(vyy', v_2)$, and hence there is an odd hole shorter than H^*, which is a contradiction. Therefore v must have a neighbor in P_2.

We now show that a is the unique neighbor of v in P_2. Let v' be the neighbor of v in P_2 that is closest to z_2. Assume that $v' \neq a$. Let P' be the $v'z_2$-subpath of P_2. If v_1 has no neighbor in P', then the graph induced by $S = P' \cup \{y, y', v, v_1, v_2\}$ is a pyramid $\Pi(vyy', v_2)$; hence there is an odd hole shorter than H^*. If v_1 has a neighbor in $P' \setminus z_2$, then the graph induced by S contains a pyramid $\Pi(vyy', v_1)$; hence there is an odd hole shorter than H^*. So v_1 is adjacent to z_2. If the graph induced by $P_2 \cup \{y, y', v_1, v_2\}$ is an odd wheel with center v_1, there is an odd hole shorter than H^*. Hence v_1 must have a neighbor in $P_2 \setminus P'$. If v_1 has a neighbor z in P_2 that lies strictly between a and v', then there is a path Q from v to v_1 with interior in z, P_2, v'. But then $Q \cup \{y, y', v_2\}$ induces a pyramid $\Pi(vyy', v_1)$, which contains an odd hole shorter than H^*. Therefore a and z_2 are the only neighbors of v_1 in P_2. Then v is not adjacent to a for otherwise a, v, y', v_2, v_1, a is an odd hole. Let v'' be the neighbor of v closest to a in P_2. Note that $v'' \neq z_2$ since otherwise $P_2 \cup \{y, y', v_2, v\}$ induces an odd wheel with center v; hence there is an odd hole shorter than H^*. Let P'' denote the av''-subpath of P_2. By Lemma 3, some vertex u of H^* is $\{v, v_1\}$-complete and has no neighbor in P''. But then the graph induced by $P'' \cup \{y, v, v_1, u\}$ is a pyramid $\Pi(ayv_1, v)$; hence there is an odd hole shorter than H^*. Therefore a is the unique neighbor of v in P_2.

Then v_1 is not adjacent to a for otherwise a, v, y', v_2, v_1, a is an odd hole. Suppose v_1 has a neighbor in P_2. By Lemma 3, there exists a vertex u of H^* adjacent to both v and v_1, but with no neighbor in P_2. Then the graph induced by $P_2 \cup \{y, v, v_1, u\}$ contains a pyramid $\Pi(ayv, v_1)$; hence there is an odd hole shorter than H^*. Therefore, v_1 has no neighbor in P_2.

Let z_1 be the neighbor of v_1 in H^* that is closest to b in $H^* \setminus \{y, y'\}$. Let P_1 be the bz_1-subpath of H^* that does not contain y. By symmetry, b is the unique neighbor of v in P_1 and v_2 has no neighbor in P_1. Since P_2, a, y, y' is a sector of wheel (H^*, v_2), P_2 must be even, and similarly P_1 is even. Note that z_1z_2 is not an edge since H^* and the path a, y, y', b have odd length and P_1, P_2 have even length. But then $P_1 \cup P_2 \cup \{v, v_1, v_2\}$ induces an odd hole shorter than H^*, which is a contradiction. □

2.3. Cleaning algorithm. In this section, we present our cleaning algorithm for the class of graphs of bounded clique number. The running time depends on the clique number.

Input: A graph G of bounded clique number k.

Output: Either an odd hole or a family $\mathcal{F}$ of induced subgraphs of G that satisfies the following properties:

(1) G contains an odd hole if and only if some graph of $\mathcal{F}$ contains a clean shortest odd hole.
(2) $|\mathcal{F}|$ is $O(|V(G)|^{8k})$.

Step 1. Check whether G contains a jewel or a pyramid (by algorithms in [2]). If it does, output an odd hole and stop. Otherwise, set $\mathcal{F}_1 = \{G\}$ and $\mathcal{F}_2 = \emptyset$.

Step 2. Repeat the following k times. For each graph $F \in \mathcal{F}_1$ and every (P_1, P_2) where $P_1 = x_0, x_1, x_2, x_3$ and $P_2 = y_0, y_1, y_2, y_3$ are two induced paths of F, add to $\mathcal{F}_2$ the graph obtained from F by removing the vertex set $(N(x_1) \cup N(x_2) \cup N(y_1) \cup N(y_2)) \setminus (V(P_1) \cup V(P_2))$. Set $\mathcal{F}_1 = \mathcal{F}_2$ and $\mathcal{F}_2 = \emptyset$.

Step 3. Set $\mathcal{F} = \mathcal{F}_1$.

THEOREM 6. *This algorithm produces the desired output, and its running time is* $O(|V(G)|^{8k})$.

Proof. Suppose that the algorithm does not output an odd hole. Suppose G contains a shortest odd hole H^*. By Step 1 G contains no jewel and no pyramid. Now we show how Step 2 generates a graph in $\mathcal{F}_1$ that contains H^* and H^* is clean in it.

By Lemma 1, $S(H^*)$ is the set of all H^*-major vertices. Let A be a maximal stable set of $S(H^*)$. We follow the notation in Theorem 5. Let $P_1 = x_0, x_1, x_2, x_3$ and $P_2 = y_0, y_1, y_2, y_3$ such that x_1x_2 and y_1y_2 satisfy the conditions stated in Theorem 5. Let $S'(H^*)$ denote the set of vertices of $S(H^*)$ that have no neighbor in $\{x_1, x_2\}$ and are A-complete. Let G' be the graph obtained from G by removing $(N(x_1) \cup N(x_2) \cup N(y_1) \cup N(y_2)) \setminus (V(P_1) \cup V(P_2))$. Then G' contains H^* and the set of major vertices for H^* in G' is contained in $S'(H^*)$. The clique number of the graph induced by $S'(H^*)$ is one less than the clique number of the graph induced by $S(H^*)$. Hence, by the fact that the clique number of G is bounded by k, Theorem 5 implies that, when the k iterations of Step 2 are completed, some graph $F \in \mathcal{F}_1$ contains H^* and H^* is clean in F. Hence (1) holds.

$O(|V(G)|^{8k})$ graphs are created in Step 2. Hence, (2) holds. The running time of Step 1 is $O(|V(G)|^9)$ as discussed in [2]. The running time of Step 2 is $O(|V(G)|^{8k})$. Therefore, the overall running time is $O(|V(G)|^{8k})$. □

In [2] a polynomial time algorithm with following specification is obtained.

Input: A clean graph G.

Output: ODD-HOLE-FREE when G is odd-hole-free, and NOT ODD-HOLE-FREE otherwise.

The above two algorithms imply that it is polynomial to test whether a graph of bounded clique number contains an odd hole.

REFERENCES

[1] D. BIENSTOCK, *On complexity of testing for odd holes and induced odd paths*, Discrete Math., 90 (1991), pp. 85–92.

[2] M. CHUDNOVSKY, G. CORNUÉJOLS, X. LIU, P. SEYMOUR, AND K. VUŠKOVIĆ, *Recognizing Berge graphs*, Combinatorica, 25 (2005), pp. 143–186.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 49–54

AVOIDING PATTERNS IN MATRICES VIA A SMALL NUMBER OF CHANGES*

MARIA AXENOVICH† AND RYAN MARTIN†

Abstract. Let $\mathcal{A} = \{A_1, \dots, A_r\}$ be a partition of a set $\{1, \dots, m\} \times \{1, \dots, n\}$ into r nonempty subsets, and let $A = (a_{ij})$ be an $m \times n$ matrix. We say that A has a pattern $\mathcal{A}$ provided that $a_{ij} = a_{i'j'}$ if and only if $(i,j), (i',j') \in A_t$ for some $t \in \{1, \dots, r\}$. In this note we study the following function f defined on the set of all $m \times n$ matrices M with s distinct entries: $f(M; \mathcal{A})$ is the smallest number of positions where the entries of M need to be changed such that the resulting matrix does not have any submatrix with pattern $\mathcal{A}$. We give an asymptotically tight value for

$$f(m, n; s, \mathcal{A}) = \max\{f(M; \mathcal{A}) : M \text{ is an } m \times n \text{ matrix with at most } s \text{ distinct entries}\}.$$

Key words. forbidden patterns, editing, graph editing, editing distance, coloring

AMS subject classifications. 15A99, 05C15, 05B20, 05C50, 05C80

DOI. 10.1137/S0895480104445150

1. Introduction.

The problem of studying the properties of matrices that avoid certain submatrices or patterns is a classical and well-studied problem in combinatorics. It is investigated from a matrix point of view as well as in an equivalent formulation of forbidden subgraphs of bipartite graphs; see [1], [7], [4], [12]. Most of the previous research is devoted to extremal and structural problems of matrices with no forbidden submatrices. There are only a few results studying efficient modifications of matrices or graphs such that the resulting structure satisfies certain properties—for example, [5] and [6]. In this paper, we apply powerful graph theoretic techniques to study the distance properties between certain classes of matrices. Our main goal is to investigate the number of positions where the entry-changes need to be performed on a given matrix such that the resulting matrix does not have a fixed subpattern. Although this problem is of independent theoretical interest, it has multiple applications in computational biology such as in the compatibility of evolutionary trees and in studying metabolic networks; see [3], [13].

For positive integers m, n, s, with $s \le mn$, let $\mathcal{M}(m, n; s)$ denote the set of all $m \times n$ matrices with a fixed number, s, of distinct entries. Let $[m] \stackrel{\text{def}}{=} \{1, \dots, m\}$. Let $\mathcal{A} = \{A_1, \dots, A_r\}$ be a partition of pairs from $[m] \times [n]$ into r nonempty classes. An $m \times n$ matrix $A = (a_{ij})$ is said to have a *pattern* $\mathcal{A}$ provided that $a_{ij} = a_{i'j'}$ if and only if $(i,j), (i',j') \in A_t$ for some $t \in \{1, \dots, r\}$. It follows, in particular, that two $m \times n$ matrices A and B with sets of distinct entries $S(A)$ and $S(B)$, respectively, have the same pattern if there is a bijection $g : S(A) \to S(B)$ such that $B(i,j) = g(A(i,j))$ for all $1 \le i \le m$ and all $1 \le j \le n$.

Example 1. Matrices A and B have the same pattern with a corresponding bijection g; matrices A and B' have different patterns:

$$A = \begin{pmatrix} 1 & 4 & 3 \\ 1 & 1 & 4 \end{pmatrix}, \qquad B = \begin{pmatrix} 5 & 1 & 2 \\ 5 & 5 & 1 \end{pmatrix}, \qquad B' = \begin{pmatrix} 5 & 1 & 2 \\ 0 & 5 & 1 \end{pmatrix}.$$

*Received by the editors July 14, 2004; accepted for publication (in revised form) August 13, 2005; published electronically February 15, 2006.
http://www.siam.org/journals/sidma/20-1/44515.html
†Department of Mathematics, Iowa State University, Ames, IA 50011 (axenovic@math.iastate.edu, rymartin@iastate.edu).

In this case, $g(1) = 5$, $g(4) = 1$, $g(3) = 2$.

A $k \times \ell$ matrix B is a *submatrix* of an $m \times n$ matrix A if there are nonempty subsets $\{i_1, \dots, i_k\}$ and $\{j_1, \dots, j_\ell\}$ of distinct indices with $\{i_1, \dots, i_k\} \subseteq [m]$, $\{j_1, \dots, j_\ell\} \subseteq [n]$ such that $B(\alpha, \beta) = A(i_\alpha, j_\beta)$, $1 \le \alpha \le k$, $1 \le \beta \le l$. If, for a matrix M', there is a submatrix M with pattern $\mathcal{A}$, then we say that M' has a *subpattern* $\mathcal{A}$.

DEFINITION 1. *For a pattern $\mathcal{A}$ and positive integers m, n, s, we define* $\mathrm{Forb}(m, n; s, \mathcal{A})$ *to be the set of all $m \times n$ matrices with at most s distinct entries and not containing subpattern $\mathcal{A}$.*

Example 2. Let $\mathcal{A} = \{\{(1,1),(1,2),(2,1)\},\{(2,2)\}\}$. The set $\mathrm{Forb}(m, n; 2, \mathcal{A})$ consists of all $m \times n$ matrices which have at most two distinct entries and contain no submatrix of the form $\left(\begin{smallmatrix} x & x \\ x & y \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} y & x \\ x & x \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} x & x \\ y & x \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} x & y \\ x & x \end{smallmatrix}\right)$, $x \neq y$. In particular, $\mathrm{Forb}(m, n; 2, \mathcal{A})$ consists of $m \times n$ matrices with all entries equal and all $m \times n$ matrices with two distinct entries such that each row has all equal entries.

Next we define the distance between two matrices and between classes of matrices. For two matrices A and B of the same dimensions, we say that $\mathrm{Dist}(A, B)$ is the number of positions in which A and B differ; i.e., it is the matrix Hamming distance. For a class of matrices $\mathcal{F}$ and a matrix A, all of the same dimensions, we denote $\mathrm{Dist}(A, \mathcal{F}) = \min\{\mathrm{Dist}(A, F) : F \in \mathcal{F}\}$. Finally,

$$f(m, n; s, \mathcal{A}) = \max\{\mathrm{Dist}(A, \mathcal{F}) : A \in \mathcal{M}(m, n; s), \mathcal{F} = \mathrm{Forb}(m, n; s, \mathcal{A})\}.$$

This function corresponds to the minimum number of positions on which the entries need to be changed in any $m \times n$ matrix with at most s distinct entries in order to eliminate all subpatterns $\mathcal{A}$. This problem is also called an *editing distance problem*, since we consider the minimum number of editing operations on a matrix, where each editing operation is a change of an entry in some position.

Note that $\mathrm{Forb}(m, n; s, \mathcal{A})$ might be an empty set of matrices for some patterns $\mathcal{A}$. For example, let s be fixed, and let $\mathcal{A}$ be a pattern having exactly one set, i.e., a pattern corresponding to matrices with all entries being equal. We call such a pattern a *trivial pattern*. If m and n are large, then there is no $m \times n$ matrix with a fixed number of distinct entries avoiding pattern $\mathcal{A}$. This follows from the finiteness of the bipartite Ramsey number; see [8]. On the other hand, when a pattern $\mathcal{A}$ has at least two distinct entries, then the class $\mathrm{Forb}(m, n; s, \mathcal{A})$ is nonempty since it contains all $m \times n$ matrices with a trivial pattern. Our main result is the following.

THEOREM 1.1. *Let s, r be positive integers, $s \ge r$. Let b_1, b_2 be positive constants such that $b_1 \le m/n \le b_2$. Let $\mathcal{A}$ be a nontrivial pattern with r distinct entries; then*

$$f(m, n; s, \mathcal{A}) = (1 + o(1)) \left(\frac{s - r + 1}{s} \right) mn.$$

We shall prove these results using graph-theoretic formulations. A graph $H = (V, E)$ is bipartite if its vertex set can be partitioned such that $V = X \cup Y$, $X \cap Y = \emptyset$, and its edge set E is a subset of $X \times Y$. If $m = |X|$, $n = |Y|$, and $E = X \times Y$, then this graph is denoted $K_{m,n}$ and called a complete bipartite graph. Now, we can introduce a pattern on the edges of a complete bipartite graph as a partition of the edges in exactly the same manner as above. Let $\mathcal{A} = \{A_1, \dots, A_r\}$ such that $E = A_1 \cup \dots \cup A_r$ and A_i's are nonempty and pairwise disjoint. Then $\mathcal{A}$ is called a pattern on E. Now, let c be a coloring of edges of $K_{m,n}$. We say that c has a *pattern* $\mathcal{A}$ if it satisfies the property that $c(e) = c(e')$ if and only if $e, e' \in A_i$ for some $i = 1, \dots, r$. If c is an edge-coloring of a graph G, we say that a coloring c' of a graph G' occurs in G under coloring c if there is a subgraph H of G isomorphic to G' such that the

coloring c restricted to H coincides with the coloring c' of G'. Similar to the case with matrices, for a color pattern $\mathcal{A}$ defined on the edges of a graph G', we say that G has a subpattern $\mathcal{A}$ if there is an occurrence of a subgraph H in G such that H is isomorphic to G' and the coloring c restricted to H has a pattern $\mathcal{A}$.

For two edge-colorings c and c' of a graph G, we say that the *edit distance* between c and c' on G is the smallest number of edge-recolorings in G colored under c needed to obtain c'. For a given pattern $\mathcal{A}$ on edges of a complete bipartite graph, and an edge-colored $K_{m,n}$ with coloring c, let $F(m,n;c,\mathcal{A})$ be the smallest number of edge-recolorings of $K_{m,n}$ colored by c such that the resulting coloring does not contain a subpattern $\mathcal{A}$. Define

$$F(m,n;s,\mathcal{A}) := \max\{F(m,n;c,\mathcal{A}) : c \text{ uses } s \text{ colors}\}.$$

Observation. There is a bijection g between all $m \times n$ matrices with s distinct entries and all edge-colorings of $K_{m,n}$ using s colors. Indeed, this bijection can be defined as $g(M(i,j)) = c(\{i,j\})$, $i \in \{1,\dots,m\}$, $j \in \{1,\dots,n\}$, where $c(\{i,j\})$ is the color of an edge $\{i,j\}$ and $M(i,j)$ is the (i,j)th entry of the matrix. Moreover, a matrix M does not have subpattern $\mathcal{A}$ if and only if a coloring $g(M)$ does not have a subpattern $\mathcal{A}$.

For all other graph-theoretic terminology, we refer the reader to [14]. Our main theorem is proven in terms of graph colorings.

THEOREM 1.2. *Let ϵ, $0 < \epsilon < 1$, be fixed, and let m', n', s, r, $s \geq r$, be fixed as well. Let $m+n$ be sufficiently large and let $\mathcal{A}$ be a pattern on $K_{m',n'}$ with r colors. Then,*

$$\left(1 - \epsilon\left(5s + 2 + (s+1)\left(\frac{m}{n} + \frac{n}{m}\right)\right)\right)\left(\frac{s-r+1}{s}\right)mn \leq F(m,n;s,\mathcal{A}) \leq \left(\frac{s-r+1}{s}\right)mn.$$

Observe that now Theorem 1.1 is an immediate corollary of Theorem 1.2 which we prove in section 3. Section 2 describes the techniques that we use in the proof.

2. Main tools. For two disjoint sets of vertices X and Y, we shall refer to a pair (X,Y) as a complete bipartite graph with partite sets X and Y. We denote its edges by $E(X,Y)$. Let $c : E(X,Y) \to \{1,\dots,s\}$ be an edge-coloring of a pair (X,Y). For each color $\nu \in \{1,\dots,s\}$ and any two subsets $X' \subseteq X$, $Y' \subseteq Y$, we denote by $E_\nu(X',Y')$ the set of edges of color ν in a pair (X',Y'). Then $d_\nu(X',Y')$ is the *density of a color ν* in the subgraph induced by X' and Y', defined as follows:

$$d_\nu(X',Y') = \frac{|E_\nu(X',Y')|}{|X'||Y'|}.$$

For $x \in X \cup Y$, we define $N_\nu(x)$ to be the set of all vertices joined to x by edges of color ν. We say that a pair (X,Y) is *ϵ-regular in color ν* if for every $X' \subseteq X$ and $Y' \subseteq Y$ with sizes $|X'| \geq \epsilon|X|$, $|Y'| \geq \epsilon|Y|$, we have

$$|d_\nu(X,Y) - d_\nu(X',Y')| < \epsilon. \tag{2.1}$$

Lemma 2.1 is based on the so-called many-color regularity lemma of Szemerédi (see [10]) and is an implication of the refinement argument, i.e., Theorem 8.4 in [11].

LEMMA 2.1 (bipartite many-color regularity lemma [11]). *For any $\epsilon > 0$ and integers s, m_0 there exists M, a positive integer, such that if the edges of a pair (X, Y) are colored with $1, \dots, s$, then the vertex set $X \cup Y$ can be partitioned into sets $V_0, V_1, \dots, V_k$ for some k, $m_0 \le k \le M$, so that $|V_0| < \epsilon(|X| + |Y|)$, and $|V_i| = |V_j|$ for $i, j \in \{1, \dots, k\}$, and all but at most ϵk^2 pairs (V_i, V_j) are ϵ-regular in color ν for each $\nu = 1, \dots, s$, and either $V_i \subseteq X$ or $V_i \subseteq Y$ for $i = 1, \dots, k$.*

In addition, we need to prove a multicolor version of the so-called intersection property, which is stated in [11] and revised in [2].

FACT 2.2 (many-color intersection property). *Let $\epsilon > 0$ and $\delta > 0$ be fixed and r and ℓ be positive integers. Let (A, B) be a pair with edges colored such that color ν is ϵ-regular with density d_ν, $d_\nu \ge \delta$ for $\nu = 1, \dots, r$. Let $Y \subset B$. Assume that $(\delta - \epsilon)^{\ell-1}|Y| > \epsilon|B|$. Let k_ν for $\nu = 1, \dots, r$ be a positive integer such that $\sum_{\nu=1}^{r} k_\nu = \ell$ and let any vector $\mathbf{a} \in A^\ell$ be indexed such that*

$$\mathbf{a} = \left(a_{[1,1]}, \dots, a_{[1,k_1]}, a_{[2,1]}, \dots, a_{[r-1,k_{r-1}]}, a_{[r,1]}, \dots, a_{[r,k_r]}\right).$$

Then,

$$\#\left\{\mathbf{a} \in A^\ell : \left|Y \cap \bigcap_{\nu=1}^{r} \bigcap_{i=1}^{k_\nu} N_\nu(a_{[\nu,i]})\right| < \prod_{\nu=1}^{r} (d_\nu - \epsilon)^{k_\nu} |Y|\right\} \le \ell\epsilon|A|^\ell. \tag{2.2}$$

The proof of Fact 2.2 is a standard argument which follows by induction on ℓ.

COROLLARY 2.3. *Let $\epsilon > 0$ and $\delta > 0$ be fixed and r and ℓ be positive integers. Let c be an edge-coloring of a pair (A, B) with at least r colors from $\{1, \dots, r, \dots\}$ such that color ν is ϵ-regular with density d_ν, $d_\nu \ge \delta$, for $\nu = 1, \dots, r$. Let us be given that $(\delta - \epsilon)^{\ell-1} > \epsilon$, $2r^\ell \ell\epsilon < 1$, and $(\delta - \epsilon)^\ell |B| \ge \ell$. Then any edge-coloring of $K_{\ell,\ell}$ with colors from $\{1, \dots, r\}$ will occur as a subcoloring of c.*

3. Proof of Theorem 1.2.

3.1. Upper bound. We shall show that for any s-edge-coloring of a complete bipartite graph with vertex class of sizes m and n, there are at most $\left(\frac{s-r+1}{s}\right) mn$ editing operations sufficient to destroy a fixed color pattern with r colors.

Let $\mathcal{A}$ be a color pattern with r sets defined on a complete bipartite graph G and let c be an edge-coloring of $K_{m,n}$ with s colors. Without loss of generality, let 1 be the color of the largest color class in c. We shall recolor the $s - r + 1$ smallest color classes of c so that their new color is 1. The resulting coloring will use only $r - 1$ colors and thus will not contain a forbidden pattern. The $s - r + 1$ smallest color classes account for at most $(1 - (r-1)/s)\, mn$ edges. Thus,

$$F(n, m; s, \mathcal{A}) \le \left(\frac{s - r + 1}{s}\right) mn.$$

3.2. Lower bound. To establish the lower bound, we show that there is a coloring of the given complete bipartite graph requiring many edit-operations to destroy a forbidden pattern. We begin with a claim that gives us a coloring which is highly regular.

CLAIM 1. *Let s be a positive integer, and $0 < \epsilon < 1/2$. There is an integer M such that if $|X| \ge M$ and $|Y| \ge M$, then there is an edge-coloring c of a complete bipartite graph $G = X \times Y$, with colors $1, 2, \dots, s$, satisfying the following property: If $X' \subseteq X$ and $Y' \subseteq Y$, such that $|X'|, |Y'| > (|X| + |Y|)(1 - \epsilon)/M$, then $d_\nu(X', Y') \in (1/s - \epsilon, 1/s + \epsilon)$, $\nu = 1, \dots, s$.*

Claim 1 follows from standard applications of the Chernoff bound (see [9, Chapter 2]).

Fix $\epsilon > 0$, let c' be a coloring of the pair (X, Y) $|X| = m, |Y| = n$, of minimum edit distance from c with the property that c' contains no subpattern $\mathcal{A}$. Apply Lemma 2.1 with parameters ϵ, s, and $m_0 = 1$ to the coloring c'. Let M be the constant given by Lemma 2.1 and the partition having all the nonleftover sets being enumerated as $X_1, \ldots, X_p, Y_1, \ldots, Y_q$ with $|X_i| = |Y_j| = Q$ and $X_i \subseteq X$, $Y_j \subseteq Y$ for $1 \le i \le p$, $1 \le j \le q$. We call a pair (X_i, Y_j) a *good* pair if it is ϵ-regular in each color $\nu \in \{1, 2, \ldots, s\}$ in coloring c'. We have that there are at most $s\epsilon(p+q)^2$ pairs which are not good. Moreover, for each good pair (X_i, Y_j) there are at most $r-1$ colors such that the density of those classes in coloring c' is at least $\delta = 2\epsilon$. Otherwise, Corollary 2.3 would imply that pattern $\mathcal{A}$ appears in c', which is a contradiction. Therefore, for a good pair (X_i, Y_j), there are at least $(s-r+1)\left(\frac{1}{s} - 3\epsilon\right)Q^2$ edit-operations needed to obtain coloring c' from the coloring c. The regularity lemma gives that $m \ge pQ \ge m - \epsilon(m+n)$ and $n \ge qQ \ge n - \epsilon(m+n)$. Therefore, the total number of recolored edges is at least

$$
\begin{aligned}
&(s-r+1)\left(\frac{1}{s} - 3\epsilon\right)Q^2\left(pq - s\epsilon(p+q)^2\right) \\
&\ge \left(\frac{s-r+1}{s}\right)(1-3s\epsilon)\left(pQqQ - s\epsilon(pQ+qQ)^2\right) \\
&\ge \left(\frac{s-r+1}{s}\right)(1-3s\epsilon)\left((m-\epsilon(m+n))(n-\epsilon(m+n)) - s\epsilon(m+n)^2\right) \\
&\ge \left(\frac{s-r+1}{s}\right)mn\left(1-\epsilon\left(5s+2+(s+1)\left(\frac{m}{n}+\frac{n}{m}\right)\right)\right). \qquad \square
\end{aligned}
$$

Remark. It should be noted that, although we prove theorems for submatrices, our results easily follow for other patterns. Suppose we wish to forbid patterns of the form $\left(\begin{smallmatrix}1 & 2\\ 1 & *\end{smallmatrix}\right)$, where the $*$ represents any entry, either a repeated 1 or 2 or a new entry 3. Our result depends only on the number of distinct entries in the pattern, so the (asymptotic) number of changes necessary and sufficient to forbid this pattern is the same as the number of changes needed to forbid $\left(\begin{smallmatrix}1 & 2\\ 1 & 1\end{smallmatrix}\right)$ or $\left(\begin{smallmatrix}1 & 2\\ 1 & 2\end{smallmatrix}\right)$ (that is, $(1 + o(1))\left(\frac{s-1}{s}\right)mn$) but fewer than to forbid $\left(\begin{smallmatrix}1 & 2\\ 1 & 3\end{smallmatrix}\right)$ (that is, $(1 + o(1))\left(\frac{s-2}{s}\right)mn$).

Acknowledgment. We are indebted to anonymous referees whose careful reading and friendly suggestions helped to significantly improve the presentation of the results.

REFERENCES

[1] R. Anstee, *General forbidden configuration theorems,* J. Combin. Theory Ser. A, 40 (1985), pp. 108–124.

[2] M. Axenovich, A. Kézdy, and R. Martin, *On editing distance in graphs*, J. Graph Theory, submitted.

[3] D. Chen, O. Eulenstein, D. Fernández-Baca, and M. Sanderson, *Flipping: A supertree construction method*, in Bioconsensus, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 61, AMS, Providence, RI, 2003, pp. 135–160.

[4] V. Deĭneko, R. Rudolf, and G. J. Woeginger, *A general approach to avoiding two by two submatrices*, Computing, 52 (1994), pp. 371–388.

[5] P. Erdős, A. Gyárfás, and M. Ruszinkó, *How to decrease the diameter of triangle-free graphs*, Combinatorica, 18 (1998), pp. 493–501.

[6] P. Erdős, E. Győri, and M. Simonovits, *How many edges should be deleted to make a triangle-free graph bipartite*? in Sets, Graphs and Numbers (Budapest, 1991), Colloq. Math. Soc. János Bolyai 60, North–Holland, Amsterdam, 1992, pp. 239–263.

[7] Z. Füredi, *Turán type problems*, in Surveys in Combinatorics, London Math. Soc. Lecture Note Ser. 166, Cambridge University Press, Cambridge, UK, 1991, pp. 253–300.

[8] R. L. Graham, B. L. Rothschild, and J. H. Spencer, *Ramsey Theory*, 2nd ed., Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley and Sons, Inc., New York, 1990.

[9] S. Janson, T. Łuczak, and A. Ruciński, *Random Graphs*, Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000.

[10] J. Komlós, A. Shokoufandeh, M. Simonovits, and E. Szemerédi, *The regularity lemma and its applications in graph theory*, in Theoretical Aspects of Computer Science (Tehran, 2000), Lecture Notes in Comput. Sci. 2292, Springer-Verlag, Berlin, 2002, pp. 84–112.

[11] J. Komlós and M. Simonovits, *Szemerédi's regularity lemma and its applications in graph theory*, in Combinatorics, Paul Erdős is Eighty, Vol. 2 (Keszthely, 1993), Bolyai Soc. Math. Stud. 2, János Bolyai Math. Soc., Budapest, 1996, pp. 295–352.

[12] H. Prömel and A. Steger, *Excluding induced subgraphs*. II. *Extremal graphs*, Discrete Appl. Math., 44 (1993), pp. 283–294.

[13] G. Stephanopoulos, A. Aristidou, and J. Nielsen, *Metabolic Engineering: Principles and Methodologies*, Academic Press, San Diego, 1998.

[14] D. West, *Introduction to Graph Theory*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 2001.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 55–61

A THEOREM ABOUT A CONTRACTIBLE AND LIGHT EDGE*

ZDENĚK DVOŘÁK† AND RISTE ŠKREKOVSKI‡

Abstract. In 1955 Kotzig [A. Kotzig, *Math. Slovaca*, 5 (1955), pp. 111–113] proved that every planar 3-connected graph contains an edge such that the sum of degrees of its end-vertices is at most 13. Moreover, if the graph does not contain 3-vertices, then this sum is at most 11. Such an edge is called light. The well-known result of Steinitz [E. Steinitz, *Enzykl. Math. Wiss.*, 3 (1922), pp. 1–139] that the 3-connected planar graphs are precisely the skeletons of 3-polytopes gives an additional trump to Kotzig's theorem. On the other hand, in 1961, Tutte [W. T. Tutte, *Indag. Math.*, 23 (1961), pp. 441–455] proved that every 3-connected graph, distinct from K_4, contains a contractible edge. In this paper, we strengthen Kotzig's theorem by showing that every 3-connected planar graph distinct from K_4 contains an edge that is both light and contractible. A consequence is that every 3-polytope can be constructed from tetrahedron by a sequence of splittings of vertices of degree at most 11.

Key words. light graph theory, contractible edges, planar graphs

AMS subject classifications. 05C10, 05C40

DOI. 10.1137/05062189X

1. Light edges.

Throughout this paper, we consider 3-connected planar graphs without loops and multiple edges. The *weight* of an edge is the sum of the degrees of its end-vertices. It is well known that every planar graph contains a vertex of degree at most 5. Kotzig [5] proved a similar result on edges.

THEOREM 1 (Kotzig). *Every 3-connected planar graph G contains an edge of weight at most 13. Moreover, if G has minimum degree at least 4, then G contains an edge of weight at most 11.*

An edge of a 3-connected planar graph is called *light* if it satisfies the requirements of the above theorem. In particular, if the graph has minimum degree at least 4, then an edge is light only if it is of weight at most 11.

The bounds of 13 and 11 from Kotzig's theorem are the best possible in the sense that there exists a planar 3-connected graph G_1 such that each edge of G_1 has weight at least 13, and that there exists a planar 3-connected graph G_2 of minimum degree 4 such that each edge of G_2 has weight at least 11. As for G_1, consider a copy of icosahedron and insert into each face a vertex and connect it with the three vertices of the face. As for G_2, consider any fulleren where no two vertices of degree 5 are adjacent.

The well-known theorem of Steinitz [9, 10] states that the 3-connected planar graphs are precisely the skeletons of the 3-dimensional polytopes. This gives an additional importance to Theorem 1.

Kotzig's Theorem has been generalized in many directions. It served as a starting point for looking for other subgraphs of small weight in plane graphs. This subject

*Received by the editors January 3, 2005; accepted for publication (in revised form) October 4, 2005; published electronically February 15, 2006.

http://www.siam.org/journals/sidma/20-1/62189.html

†Institute for Theoretical Computer Science (ITI), Charles University, Malostranské náměstí 25, Praha, Czech Republic (rakdver@kam.mff.cuni.cz). The work of this author was supported in part by project LN00A056 of the Czech Ministry of Education.

‡Department of Mathematics, University of Ljubljana, Jadranska 19, 1111 Ljubljana, Slovenia (skreko@fmf.uni-lj.si). The work of this author was supported in part by Ministry of Science and Technology of Slovenia, research project Z1-3129.

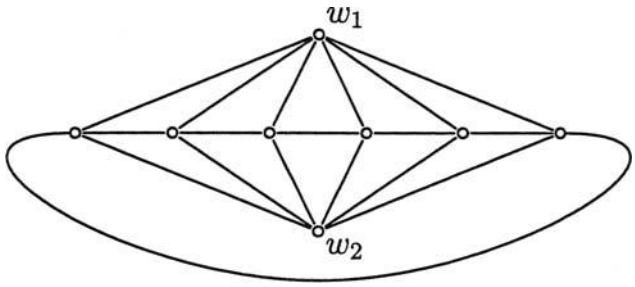


FIG. 1. *A double wheel.*

later developed into light graph theory: let $\mathcal{H}$ be a family of graphs, and let H be a connected graph such that infinitely many members of $\mathcal{H}$ contain a subgraph isomorphic to H. Let $\mathcal{H}_H$ be the subfamily of graphs in $\mathcal{H}$ that contain H as a subgraph. We say that H is a *light* graph in the family $\mathcal{H}$ if there exists a constant c such that each graph $G \in \mathcal{H}_H$ contains a subgraph $K \cong H$ with $d_G(v) \leq c$ for every vertex $v \in K$. Let us mention a few results from light graph theory: Fabrici and Jendrol' [2] proved that only the paths are light in the family of all 3-connected plane graphs; the same holds also for the family of all 3-connected plane graphs of minimum degree 4 (see [3]). A survey on light graphs in various families of plane, projective plane and general graphs can be found in the paper by Jendrol' and Voss [4].

1.1. Light edge avoiding prescribed triangle. In this section, we prove the existence of a light edge which avoids vertices of a prescribed triangular face.

LEMMA 1. *Let $G \neq K_4$ be a plane 3-connected graph with the outer-face $O = x_1x_2x_3$ of length 3. Let δ' be the minimum degree of the vertices of G that are distinct from x_1, x_2 and x_3. Let d be 13 if $\delta' = 3$, and 11 otherwise. The graph G then contains an edge of weight at most d which is not incident with x_1, x_2 and x_3.*

Proof. Suppose that the statement of the lemma is false and G is a counterexample on n vertices. Obviously, $n \geq 5$. In addition, we assume that G has maximum number of edges among all such graphs.

We claim that *every face incident with x_1, x_2, or x_3 is a triangle.* Otherwise, we may assume that x_1 lies on a face f' of length ≥ 4. Hence, we can insert an edge between x_1 and a vertex of f' which is not adjacent to x_1. This is always possible since G is 3-connected. Let G' be the resulting graph. Notice that if G is a graph of minimum degree ≥ 4, then G' also has minimum degree ≥ 4. Hence, G' is a counterexample to the lemma with the same number of vertices but it has more edges than G, a contradiction.

By the above-mentioned claim, it easily follows that at most one of x_1, x_2 and x_3 is a vertex of degree 3. Thus, we may assume that $d(x_1) \geq 4$ and $d(x_2) \geq 4$. Notice that $d(x_3) \geq 3$ since G is 3-connected.

Next, consider the double wheel W of order 8 as depicted in Figure 1. Let w_1 and w_2 be the two 6-vertices of W. We construct a planar graph W_G by gluing a copy of G in each face of W in such a way that the vertex x_3 of the copy is identified with either w_1 or w_2. It follows from the assumption on the degrees of vertices x_1, x_2 and x_3 in G that each vertex of W has degree ≥ 12 in W_G. It is easy to see that if two 3-cycles of two 3-connected graphs are identified, the resulting graph is also 3-connected. This implies that W_G is 3-connected.

By Kotzig's Theorem, the graph W_G contains a light edge e_w. This edge is not incident with any vertex of the copy of W, since all these vertices are of degree ≥ 12.

Hence, e_w corresponds to an edge e of G which is not incident with x_1, x_2 and x_3. Notice that if $\delta' \geq 4$, then W_G has minimum degree ≥ 4 and thus the weight of e_w is at most 11. This implies that the weight of e satisfies requirements of the lemma. □

2. Contractible edges. A subset S of vertices of a connected graph G is a *cut*, if the graph $G - S$ is disconnected and S is a minimal set with this property. If S is of size k, then it is called a *k-cut*. A graph G is *k-connected* if it has at least $k+1$ vertices and it has no cuts of size $< k$.

Let $e = ab$ be an edge of a 3-connected graph G, and let G/e be the graph obtained by identifying the vertices a and b into a new vertex w, and by removing the arising loop and multiple edges (in order to obtain a simple graph). We say that G/e is obtained from G by contracting the edge e. Similarly, we say that G is obtained from G/e by splitting w. If G/e is a 3-connected graph, then we say that the edge e is *contractible*. If e is not contractible, we say it is *noncontractible*. It is easy to see that e is noncontractible if and only if G has a 3-cut S such that $\{a, b\} \subseteq S$.

Tutte [11] proved that every 3-connected graph, that is distinct from K_4, contains a contractible edge, and as a consequence, Theorem 2 follows.

THEOREM 2 (Tutte). *A graph G is 3-connected if and only if there exists a sequence $G_0, \ldots, G_n$ of graphs with the following properties:*

(a) *$G_0 = K_4$, $G_n = G$, and*
(b) *G_{i+1} has an edge xy with $d(x), d(y) \geq 3$ and $G_i = G_{i+1}/xy$, for every $i < n$.*

In fact, every 3-connected graph on ≥ 5 vertices has more than just one contractible edge. See the survey of Kriesell [6] for more results of this kind.

Notice that if G is a 3-connected planar graph and S is a 3-cut, then $G - S$ comprises of precisely two components: there cannot be more than two, otherwise we obtain a subdivision of $K_{3,3}$ in G. Let these two components be denoted by $G_1(S)$ and $G_2(S)$. Let $G_i^*(S)$ be the subgraph of G induced by $V(G_i(S)) \cup S$. In particular, $G_i(S) = G_i^*(S) - S$ for $i \in \{1, 2\}$. Observe that if $x, y \in S$ are nonadjacent, then there exists precisely one face incident with both of them. When the graph G is clear from the context, its face which contains the vertices x and y is denoted by $f_{x,y}$.

A triangle $v_1v_2v_3$ of a graph is called *separating* if $\{v_1, v_2, v_3\}$ is a cut. If $v_1v_2v_3$ is a separating triangle of G, then each of the edges v_1v_2, v_1v_3 and v_2v_3 is obviously noncontractible. On the other hand, it is not necessarily true that every noncontractible edge of G belongs to a separating triangle. However, in this section we show that unless G contains a light contractible edge, we may extend G to a supergraph that satisfies this condition by adding new noncontractible edges and without creating any new contractible edges; see Lemma 8.

The proofs of the following three folklore lemmas can be found in [1, 7, 8]:

LEMMA 2. *Let G be a 3-connected graph of order at least five. Suppose x is a 3-vertex of G whose neighbors are a, b and c. If ab is an edge of G, then xc is contractible.*

If H is a subgraph of G, then we denote by G/H the graph constructed from G by contracting all edges of H.

LEMMA 3. *Let x be a 3-vertex of a 3-connected graph $G \neq K_4$. If xa and xb are two noncontractible edges of G, then a and b are adjacent vertices of degree 3. Moreover, $G^* = G/axb$ is 3-connected.*

LEMMA 4. *Let G be a 3-connected graph and let $C = x_1x_2x_3$ be a 3-cycle of G with all vertices of degree 3. An edge e of G/C is contractible if and only if its corresponding edge e in G is contractible.*

We are now ready to prove the following lemma on minimal 3-connected graphs without a light contractible edge.

LEMMA 5. *If $G \neq K_4$ is a 3-connected planar graph with the smallest possible number $n \geq 5$ of vertices such that every light edge of G is noncontractible, then G does not contain a 3-cycle whose vertices are all of degree 3.*

Proof. First, suppose that $n < 7$. Hence, the degree of each vertex of G is at most 5, and thus each edge of G is light. Since every 3-connected graph of order at least 5 contains a contractible edge, the graph G contains a light contractible edge.

Let us now assume that $n \geq 7$. Suppose that $C = x_1x_2x_3$ is a 3-cycle of G such that all vertices of C are of degree 3. Let y_i be the neighbor of x_i that does not belong to C. Note that the vertices y_1, y_2 and y_3 are mutually distinct, since G is 3-connected and $G \neq K_4$. Let $G^* = G/C$ and let w be the vertex of G^* into which C is contracted. By Lemma 3, the graph G^* is 3-connected. Hence, w is a 3-vertex whose neighbors are y_1, y_2 and y_3. Also notice that each edge e^* of G^* has the same weight as the corresponding edge e of G. Lemma 4 claims that e^* is contractible in G^* if and only if e is contractible in G. This implies that every light edge of G^* is noncontractible. Since G^* has at least five vertices, it contradicts the minimality of G. □

The following two lemmas describe the structure of a graph containing a noncontractible edge xy that becomes contractible after a new edge bc is inserted in the graph.

LEMMA 6. *Let G be a planar 3-connected graph, xy a noncontractible edge of G, and b and c two nonadjacent vertices of G that lie on a common face. Suppose that xy is contractible in $G \cup \{bc\}$. If a vertex z is contained in a 3-cut $S = \{x, y, z\}$ of G, then the following four claims hold:*

(i) *b and c are distinct from x, y and z, and they belong to distinct components of $G - S$,*
(ii) *z belongs to $f_{b,c}$, and precisely one of x and y belongs to $f_{b,c}$ (let this vertex be denoted by w),*
(iii) *$f_{b,c} = w \cdots b \cdots z \cdots c \cdots w$, and*
(iv) *w and z are nonadjacent.*

Proof. Since xy is contractible in $G \cup \{bc\}$ but not in G, it follows that b and c belong to distinct components of $G - S$. Therefore, the vertices b and c are distinct from x, y and z.

Since S is a cut and the edge bc connects the two components of $G - S$, it follows that b, c and z belong to a common face. Moreover, one of x and y lies on the same face as well (but not both since no face may contain all three vertices of a 3-cut of G). The order of the vertices w, z, b and c that appear around the face must be as described in the claim, because b and c belong to distinct components of $G - S$. Since G is 3-connected, it follows that w and z are nonadjacent. □

LEMMA 7. *Let ab and xy be two noncontractible edges and let $S_1 = \{a, b, c\}$ and $S_2 = \{x, y, z\}$ be two 3-cuts of G. If the edge xy is contractible in $G \cup \{bc\}$, then the following two claims hold:*

(i) *If $a \notin \{x, y\}$, then c is a 3-vertex with $N(c) = \{z, x, y\}$ and cxy is a 3-face.*
(ii) *If $a = x$, then y is a 3-vertex with $N(y) = \{a, b, c\}$ and aby is a 3-face.*

Proof. First notice that $G \cup \{bc\}$ is a planar graph, since the vertices b and c lie on a common face in G. Also notice that b and c are nonadjacent in G. By Lemma 6, the vertices b and c belong to different components of $G - S_2$, and they are distinct from x, y and z. By the same lemma, without loss of generality, we may assume that y and z are nonadjacent and that they lie on the same face with b and c (i.e.,

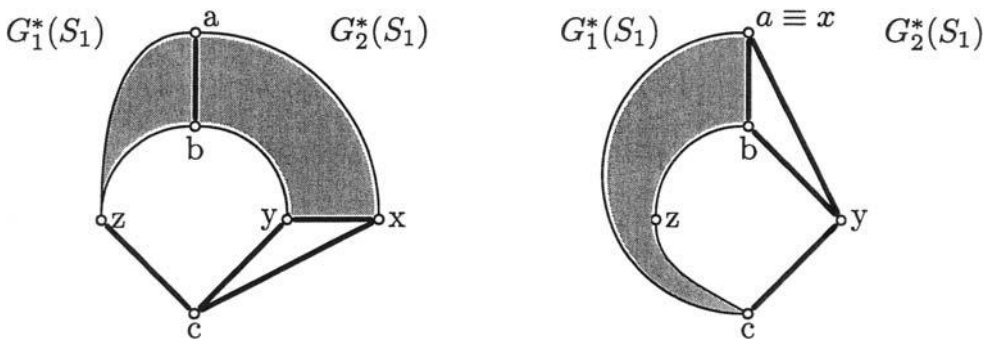


FIG. 2. *Configurations in Lemma* 7.

$f_{b,c} = f_{y,z}$) and $f_{y,z} = y \cdots c \cdots z \cdots b \cdots y$. We may also assume that z is a vertex of $G_1^* = G_1^*(S_1)$ and x, y are vertices of $G_2^* = G_2^*(S_1)$. Consider now the claims of this lemma separately and see Figure 2 for illustration:

(i) Observe that z is a cut-vertex in G_1^* which separates a and b from c; otherwise we can infer that S_2 is not a cut of G. Since G is 3-connected it follows that c is adjacent only to z in G_1^*.
Similarly one can show that $\{x, y\}$ is a cut in G_2^* which also separates a and b from c. To show the minimality of $\{x, y\}$ observe that if x or y is a vertex-cut in G_2^*, then $\{x, z\}$ or $\{y, z\}$ is a 2-cut in G, respectively.
If there is a vertex adjacent to c in G_2^* which is distinct from x and y, then $\{c, x, y\}$ is a 3-cut in $G \cup \{bc\}$ but this contradicts the assumption that xy is a contractible edge in $G \cup \{bc\}$. Since $\{x, y\}$ is a cut in G_2^*, both x and y are adjacent to c. Thus, x, y are the only neighbors of c in G_2^*. This implies that cxy is a 3-face and $N(c) = \{z, x, y\}$.

(ii) Since $\{x, y, b\}$ is not a 3-cut in $G \cup \{bc\}$, we infer that aby is a 3-face. Similarly, since $\{x, y, c\}$ is not a 3-cut in $G \cup \{bc\}$, it follows that cy is an edge of G, and hence $N(y) = \{a, b, c\}$. □

We are now ready to show that in a maximal graph which does not contain a light contractible edge, every noncontractible edge belongs to a separating 3-cycle.

LEMMA 8. *Suppose that there exists a planar graph on $n \geq 5$ vertices such that each of its light edges is noncontractible. If G is such a graph with n vertices with maximum number of edges, then every noncontractible edge of G belongs to a separating* 3*-cycle.*

Proof. Suppose that the claim is false and G is a counterexample with minimum number of vertices $n \geq 5$. Let ab be a noncontractible edge which does not belong to a separating 3-cycle and let $S = \{a, b, c\}$ be a 3-cut of G. Without loss of generality, we may assume that b and c are nonadjacent.

Consider the graph $G \cup \{bc\}$. By the maximality of $|E(G)|$, the graph $G \cup \{bc\}$ contains a light contractible edge xy. Obviously the edge xy is distinct from bc, since bc is noncontractible. The edge xy is light in G as well, thus it must be noncontractible in G. Let $\{x, y, z\}$ be a 3-cut of G. We may assume that $x, y \in V(G_2^*(S))$ and $z \in V(G_1^*(S))$. By Lemma 6, we may assume that b, y, c and z belong to a common face. Consider now the following two cases and see Figure 3 for illustration:

Case 1: $a \notin \{x, y\}$. By Lemma 7(i), we may assume that c is a 3-vertex with neighbors x, y and z. The maximality of G implies that the graph $G \cup \{yz\}$ must contain a light contractible edge $e = a'b'$. Notice that this edge is noncontractible

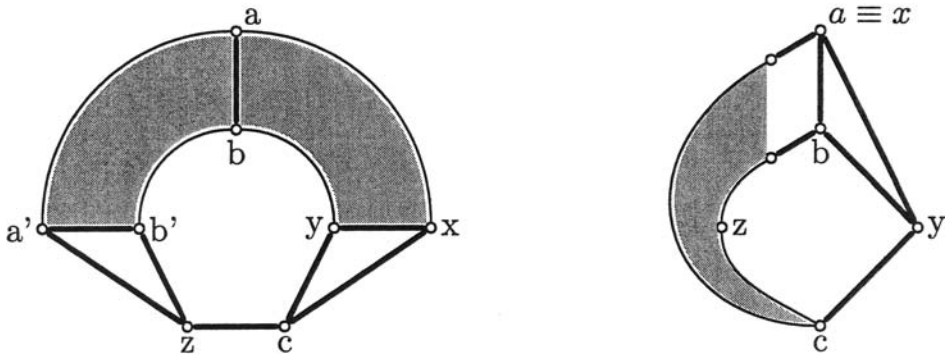


FIG. 3. *Configurations in Lemma* 8.

in G. By Lemma 6 one of the end-vertices of e must be incident with $f_{y,z}$, say b'. Observe that the only 3-cut that shows noncontractibility of e is $\{a', b', c\}$. If e belongs to $G_2^*(S)$, then $\{a', b', z\}$ is a cut of $G \cup \{yz\}$ which separates a or b from y, x or c, and $a'b'$ would be noncontractible in $G \cup \{yz\}$. Therefore, we may assume that e belongs to $G_1^*(S)$. See the left graph of Figure 3. In particular, the edges $a'b'$ and xy are not incident. Hence, by Lemma 7(i), z is a 3-vertex and $za'b'$ is a 3-face of G. Finally, Lemma 2 implies that zc is a contractible edge of weight 6.

Case 2: $a \in \{x, y\}$, say $a = x$. We assume that no choice of x, y, z, a, b and c may satisfy Case 1. By Lemma 7(ii), y is a 3-vertex with neighbors a, b, c and aby is a face. By the maximality of G, the graph $G \cup \{yz\}$ contains a light contractible edge $a'b'$. The edge $a'b'$ must be noncontractible in G and distinct from ay. Excluding Case 1, the edge $a'b'$ must be incident with the edge ay. However, adding the edge yz does not affect contractibility of any edge incident with y or z; therefore, the edge $a'b'$ must be incident with a. We may assume that $a = a'$. Notice that b' is a vertex of $G_1^*(S)$ and by Lemma 7(ii), we conclude that ayb' is a 3-face and that degree of b' is 3. Hence $b' = b$ or $b' = c$. If $b' = c$, then the edge cy has weight 6 and by Lemma 2 it is contractible in G.

Now consider the case $b = b'$ and see the right graph of Figure 3. The degree of b is 3 and the edge by has weight 6. If by is a noncontractible edge, then b is a 3-vertex incident with two noncontractible edges ab and by. Lemma 3 implies that the degree of a is also 3. Hence, G contains a 3-cycle with each vertex of degree 3, but Lemma 5 excludes such a subgraph in G. We conclude that by is a contractible light edge in G. This finishes the proof. □

3. Contractible light edge. If C is a cycle of a plane graph G, then $\mathrm{Int}(C)$ denotes the subgraph of G induced by the vertices and edges of G which lie on C or in its interior. We are now ready to prove the theorem.

THEOREM 3. *Every* 3*-connected planar graph, distinct from* K_4*, contains a light and contractible edge.*

Before we proceed with the proof of the theorem, let us emphasize that this result strengthens Theorem 1, i.e., we show precisely the same bounds on the weight of contractible edges.

Proof. Suppose that the theorem is false and G is a counterexample with the minimum number of vertices $n \geq 5$. In particular, every light edge of G is noncontractible. We may also assume that G has the maximum number of edges among all such graphs of order n.

By Lemma 8, every noncontractible edge of G belongs to a separating 3-cycle. Since G is 3-connected, it follows that every vertex that belongs to a separating 3-cycle is of degree ≥ 4. Therefore, every 3-vertex is incident only to contractible edges. This implies that every 3-vertex of G is adjacent only to vertices of degree ≥ 11. In order to complete the proof, consider the following two possibilities:

First, suppose that *every separating 3-cycle C of G satisfies* $\mathrm{Int}(C) = K_4$. By Theorem 1, the graph G contains a light edge $e = uv$. This edge e does not lie on a separating 3-cycle; otherwise u and v are adjacent with a 3-vertex, and each of them is of degree ≥ 11 by the argument in the above paragraph. We conclude that e is a contractible light edge.

Now suppose that *G has a separating 3-cycle C such that* $\mathrm{Int}(C) \neq K_4$. We may additionally assume that $C = x_1x_2x_3$ is chosen so that $G' := \mathrm{Int}(C)$ has the smallest possible number of vertices. The graph G' has at least five vertices. By the choice of C, each separating 3-cycle C' of G' satisfies $\mathrm{Int}(C') = K_4$. By Lemma 1, G' contains an edge e' that is not incident with x_1, x_2 and x_3 such that e' is light in G. Applying a similar argument as in the previous paragraph, one can observe that e' is also contractible. This establishes the theorem. □

Theorems 2 and 3 imply the following result:

COROLLARY 1. *Every 3-polytope G can be constructed from tetrahedron by sequential splittings of vertices of degree at most* 11.

Acknowledgments. We would like to thank Daniel Král' for careful reading and valuable suggestions regarding the presentation of the results.

REFERENCES

[1] K. ANDO, H. ENOMOTO, AND A. SAITO, *Contractible edges in 3-connected graphs*, J. Combin. Theory Ser. B, 42 (1987), pp. 87–93.

[2] I. FABRICI AND S. JENDROL', *Subgraphs with restricted degrees of their vertices in planar 3-connected graphs*, Graphs Combin., 13 (1997), pp. 245–250.

[3] I. FABRICI, E. HEXEL, S. JENDROL', AND H. WALTHER, *On vertex-degree restricted paths in polyhedral graphs*, Discrete Math., 212 (2000), pp. 61–73.

[4] S. JENDROL' AND H.-J. VOSS, *Light subgraphs of graphs embedded in the plane and in the projective plane – a survey*, MATH-AL-02-2001, Technical University of Dresden, Dresden, Germany.

[5] A. KOTZIG, *Contribution to the theory of Eulerian polyhedra*, Math. Slovaca, 5 (1955), pp. 111–113.

[6] M. KRIESELL, *A survey on contractible edges in graphs of a prescribed vertex connectivity*, Graphs Combin., 18 (2002), pp. 1–30.

[7] K. OTA, *The number of contractible edges in 3-connected graphs*, Graphs Combin., 4 (1988), pp. 333–354.

[8] W. MCCUAIG, *Edge contractions in 3-connected graphs*, Ars Combin., 29 (1990), pp. 299–308.

[9] E. STEINITZ, *Polyeder und Raumeinteilungen*, Enzykl. Math. Wiss., Vol. 3 (Geometrie), Part 3AB12 (1922), pp. 1–139.

[10] E. STEINITZ AND H. RADEMACHER, *Vorlesungen ber die Theorie der Polyeder unter Einschluss der Elemente der Topologie*. Reprint der 1934 Auflage. Grundlehren Math. Wiss. 41, Springer-Verlag, New York, 1976.

[11] W. T. TUTTE, *A theory of 3-connected graphs*, Indag. Math., 23 (1961), pp. 441–455.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 62–78

IMPROVED BOUNDS FOR TOPOLOGICAL CLIQUES IN GRAPHS OF LARGE GIRTH*

DANIELA KÜHN† AND DERYK OSTHUS†

Abstract. We prove that every graph of minimum degree at least r and girth at least 27 contains a subdivision of K_{r+1}. This implies that the conjecture of Hajós, that every graph of chromatic number at least r contains a subdivision of K_r, is true for graphs of girth at least 27. This conjecture is known to be false in general.

Key words. subdivisions, topological minors, girth, Hajós conjecture

AMS subject classifications. 05C83, 05C35, 05C15

DOI. 10.1137/040617765

1. Introduction. It is well known that the existence of a subdivision of the complete graph K_r is forced by large but constant minimum degree: let $d(r)$ be the smallest number such that every graph G of minimum degree at least $d(r)$ contains a subdivided K_r. The existence of $d(r)$ was proved by Mader (see, e.g., [1]). Bollobás and Thomason [2, 3] as well as Komlós and Szemerédi [8] independently proved that $d(r)$ is quadratic in r. As observed by Jung [7] earlier on, complete bipartite graphs show that $d(r)$ is at least quadratic in r.

On the other hand, Mader [15] proved that the situation is rather different for graphs which have large girth, i.e., which do not contain short cycles: he showed that there is a function $g(r)$ so that every graph of minimum degree at least r and girth at least $g(r)$ contains a subdivided K_{r+1}. At first, this might seem rather surprising since the condition on the minimum degree only ensures that every vertex has sufficiently many neighbors to be a candidate for a branch vertex. Mader's bound on the $g(r)$ was linear in r. He asked about the growth of $g(r)$ and pointed out that it might even be true that $g(r) = 5$ for $r \geq 4$ (see [17]). The complete bipartite graph $K_{r,r}$ provides the lower bound $g(r) \geq 5$ for $r \geq 4$. In [9], we showed that $g(r) \leq 186$ and that $g(r) \leq 15$ for all $r \geq 435$. In this paper, we prove that $g(r) \leq 27$.

THEOREM 1. *Let $r \geq 1$ be a natural number. Every graph of minimum degree at least r and girth at least 27 contains a subdivision of K_{r+1}.*

In [12] we proved the related result that if we relax the condition of having girth at least 27 to being C_4-free, then at least we can find a subdivision of a complete graph whose order is almost linear in the minimum degree of the host graph.

Theorem 1 has an immediate application to the well-known conjecture of Hajós (see [6]), which states that every graph of chromatic number r contains a subdivision of K_r. Catlin [4] found several counterexamples to this. A little later, Erdős and Fajtlowicz [5] proved that the conjecture fails even for almost all graphs. On the other hand, since every graph of chromatic number at least r has a subgraph of minimum degree at least $r-1$, Theorem 1 shows that the conjecture does hold for all graphs whose girth is at least 27.

*Received by the editors October 27, 2004; accepted for publication (in revised form) September 13, 2005; published electronically February 21, 2006.
http://www.siam.org/journals/sidma/20-1/61776.html
†School of Mathematics, Birmingham University, Edgbaston, Birmingham B15 2TT, UK (kuehn@maths.bham.ac.uk, osthus@maths.bham.ac.uk).

COROLLARY 2. *Let $r \geq 2$ be a natural number. Every graph of chromatic number r and girth at least 27 contains a subdivision of K_r.*

Thomassen [22] asked whether the conjecture of Hajós might even be true for all triangle-free graphs. Note that there is a difference to the minimum degree condition here: while it may be that Theorem 1 remains true if one replaces the condition of "girth $\geq$ 27" by "girth $\geq$ 5," one cannot replace it by "girth $\geq$ 4." Also, no counterexamples to the conjecture of Hajós are known for $r = 5, 6$.

Based on a result of Mader [18], in [11] we proved an analogue of Theorem 1 with the minimum degree condition replaced by one on the average degree: for every $\varepsilon > 0$ there exists an integer $f(\varepsilon)$ such that for all $r \geq 2$ every graph G of average degree at least $r + \varepsilon$ and girth at least $f(\varepsilon)$ contains a subdivision of K_{r+2}. (In [18], the same result is proved with the difference that the function f also depends on r.)

Apart from the obvious question of whether the girth bound in Theorem 1 and its relatives can be improved, one could also try to strengthen the above results by asking for an *induced* subdivision. In particular, Shi [19] posed the following question.

PROBLEM 3. *Is there a function $h(r)$ such that every graph of minimum degree at least $r \geq 3$ and girth at least $h(r)$ contains a subdivision of K_{r+1} as an induced subgraph?*

This question was motivated by our earlier result [13] that for all integers $s \geq 2$ and $r \geq 3$ there exists an integer d such that every $K_{s,s}$-free graph of minimum degree at least d contains an induced subdivision of K_r. (Clearly, the condition of being $K_{s,s}$-free cannot be omitted here.) Thus the above problem has an affirmative answer if we assume that the minimum degree is sufficiently large compared to r.

The assumption of large girth also makes a major difference if one asks for ordinary minors instead of subdivisions. The first result in this direction was proved by Thomassen [21], who showed that there is a function $q(r)$ so that every graph of minimum degree at least 3 and girth at least $q(r)$ contains a K_r minor. In [10], we proved much more precise bounds on the existence of large complete minors in graphs of given minimum degree and given girth. Also, in [14] we proved that one obtains surprisingly large complete minors if one relaxes the condition of having large girth to being $K_{s,s}$-free.

This paper is organized as follows. In the next section, we give a brief outline of the proof of Theorem 1. In section 3 we then present the necessary definitions and some tools which we will need later on. Section 4 is devoted to the proof itself and is divided into two subsections: one for the easier case when we seek a subdivision of K_{r+1} for $r \geq 5$ and one for the case where we seek a subdivision of K_5. In the final section, we then very briefly discuss possible approaches for further improvements and the obstacles to these.

2. Outline of the proof. Suppose that we are given a graph G of minimum degree r and girth at least 27. First we choose a maximal set X of vertices whose pairwise distance is at least 7. We then extend these vertices into rooted induced subtrees of G such that each tree sends many edges to different other trees. (The vertices in X are the roots of these trees.) Then the minor G' of G which is obtained by contracting each of the trees into a single vertex has large minimum degree (at least $r(r-1)^3$). One could then show that G' contains an $\binom{r+1}{2}$-linked subgraph G^* (provided that r is not too small). Thus G^*, and therefore also G', contains a subdivision of a K_{r+1}. But unfortunately, this need not correspond to a subdivision of a K_{r+1} in our original graph G.

Thus we have to find a highly linked substructure G^* of G' which allows us more control on what the subdivision of K_{r+1} in G^* (and in G') looks like in order to guarantee that this subdivision will correspond to one in G. Indeed, we find a set A of vertices of G' together with their neighbors B having the property that the graph G^* obtained from $G'[A \cup B]$ by contracting an independent set F of A–B edges is highly connected (Lemma 5). At first it might not seem to be a good idea to consider a highly connected minor of G' instead of a highly connected subgraph. But the advantage of this is that at least the vertices in A are now guaranteed to "keep" all their neighbors in the connected substructure. In the case $r \geq 5$ this property can be used to find $r+1$ disjoint r-stars in $G'[A \cup B]$ (with centers in A) which correspond both to disjoint r-stars in G^* as well as subdivided r-stars in G. As G^* is highly connected and thus highly linked, we can link the leaves of the stars to obtain a subdivision of K_{r+1} in G^*. Since each star in G^* corresponds to a subdivided star in G, this subdivision of K_{r+1} in G^* will then correspond to one in G.

This strategy was first used by Mader [15] and subsequently also by us in [9]. The improvements we obtain here are partly due to a more economical construction of the stars in $G'[A \cup B]$. In particular, in order to find stars with the desired properties, the graph G' was required to have a girth which was linear in r in [15] and to have a girth of least 6 in [9], where the weaker bound of 186 on the girth of G was proved. Dropping this requirement on G' immediately leads to a significantly lower requirement on the girth of G.

The strategy described above does not work when $r = 4$ since in this case we cannot guarantee that the graph G^* is sufficiently linked. So we have to work harder here (see the beginning of section 4.2 for more details).

3. Notation and tools. We will now collect some definitions and results which we will need later. We denote the minimum degree of a graph G by $\delta(G)$. The *girth* $g(G)$ of G is the length of the shortest cycle in G. A *subdivision* of a graph G is a graph TG obtained from G by replacing the edges of G with internally disjoint paths. The *branch vertices* of TG are all those vertices that correspond to the vertices of G. An *r-star* is a star with r leaves. Given a set A of vertices of a graph G, we write $N_G(A)$ for the set of all those neighbors of vertices in A that lie outside A. Given $\ell \in \mathbb{N}$, the *ℓ-ball around a vertex x* of a graph G is the subgraph of G induced by all its vertices of distance at most ℓ from x (including x itself). We denote this subgraph by $B^\ell_G(x)$. Given integers $r \geq 3$ and $\ell \geq 1$, we define an *r-uniform tree of radius ℓ* to be the rooted tree in which all leaves have distance ℓ from the root and all other vertices have degree r. Thus an r-uniform tree of radius ℓ has precisely $r(r-1)^{\ell-1}$ leaves. Given a tree T with root x and a vertex $v \in T$, we say that a vertex $u \in T$ *lies above* v if v lies on the subpath of T which joins x to u. The *branch above v* is the subtree of T which is spanned by all the vertices lying above v.

Given $k \in \mathbb{N}$, we say that a graph G is k-linked if $|G| \geq 2k$ and for every $2k$ distinct vertices $x_1, \dots, x_k$ and $y_1, \dots, y_k$ of G there exist disjoint paths $P_1, \dots, P_k$ such that P_i joins x_i to y_i. We will make essential use of the following recent result of Thomas and Wollan [20].

THEOREM 4. *Let $k \geq 1$ be a natural number. Every $2k$-connected graph of average degree at least $10k$ is k-linked.*

The first linear bound on the necessary average degree was established by Bollobás and Thomason [2], who obtained the same result but with the $10k$ replaced by $22k$.

The following lemma is essentially due to Mader [15], who proved a slightly stronger result for triangle-free graphs. A proof of the version below is contained in [9].

LEMMA 5. *Let $c \geq 1$ be an integer and let G be a graph of minimum degree at least $2c$. Then there exist disjoint sets $A, B \subseteq V(G)$ and a set F of $|B|$ independent A–B edges such that $|A| > c > |B|$, $N_G(A) \subseteq B$ and so that the graph G^* obtained from $G[A \cup B]$ by contracting the edges in F is $\lceil c/3 \rceil$-connected.*

4. Proofs. Note that Theorem 1 is trivial for $r \leq 2$. Moreover, it is easy to check that every graph of minimum degree at least 3 contains a subdivision of K_4. (This was first observed by Dirac; see, e.g., [1, Ch. VII, Thm. 2.2].) Thus Theorem 1 also holds when $r = 3$ (even with no assumption on the girth). Pelikán showed that every graph of minimum degree at least 4 contains a subdivision of the graph obtained from K_5 by deleting one edge (see, e.g., [1, Ch. VII, Thm. 2.5]). However, as mentioned in the introduction, an additional assumption on the girth is needed to guarantee a subdivision of K_5. (Alternatively, instead of increasing the girth, one can also guarantee a subdivided K_5 by increasing the minimum degree to 6. This follows from the result of Mader [16] that $3|G| - 5$ edges force a TK_5.) As the proof of the K_5-case of Theorem 1 needs some special arguments, it will be considered separately. So let us first prove Theorem 1 for the case $r \geq 5$.

4.1. Finding a subdivision of K_{r+1} for $r \geq 5$. Let G be a graph of minimum degree $r \geq 5$ and girth at least 27. Consider a maximal set X of vertices of G such that every two vertices in X have distance at least 7 from each other. Since the girth of G is at least 27, the 3-ball $B^3_G(x)$ around any vertex $x \in X$ is a tree. Since the vertices in X have distance at least 7 from each other, all these trees must be disjoint. We now extend these trees to connected subgraphs T_x ($x \in X$) by adding first every vertex of distance 4 from X to one of the trees to which it is adjacent, then adding all vertices of distance 5 from X to one of the subgraphs constructed in the previous step and then those of distance 6. By the maximality of X, every vertex in $V(G) \setminus X$ has distance at most 6 from X and is thus contained in some T_x. Our assumption that the girth of G is at least 27 now implies that the T_x ($x \in X$) satisfy the following properties:

(i) T_x is an induced subtree of G. We will view x as the root of T_x.
(ii) Each leaf of T_x has in T_x distance at most 6 from the root x of T_x.
(iii) The 3-ball $B^3_G(x) = B^3_{T_x}(x)$ around x contains an r-uniform tree of radius 3. Thus T_x has at least $r(r-1)^2$ leaves and so there are at least $r(r-1)^3$ edges in G emanating from T_x. All these edges go to different other trees T_y, i.e., between every pair T_x and T_y of trees there exists at most one edge in G.

We now consider the graph G' obtained from G by contracting each T_x ($x \in X$) into a single vertex. For each $x \in X$, we denote by x' the vertex of G' corresponding to the (contracted) tree T_x. Note that (iii) implies that

$$\delta(G') \geq r(r-1)^3.$$

Put

$$c := 6\binom{r+1}{2} + 3(r+1). \tag{1}$$

Using that $r \geq 5$, it is easy to check that

$$\delta(G') \geq 2c. \tag{2}$$

Thus we can apply Lemma 5 to G' to obtain sets A and B and a set F of $|B|$ independent A–B edges as described there. The edges of F are called *F-edges*.

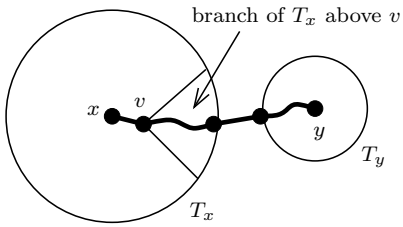


FIG. 1. *The neighbor y' of x' lies above v.*

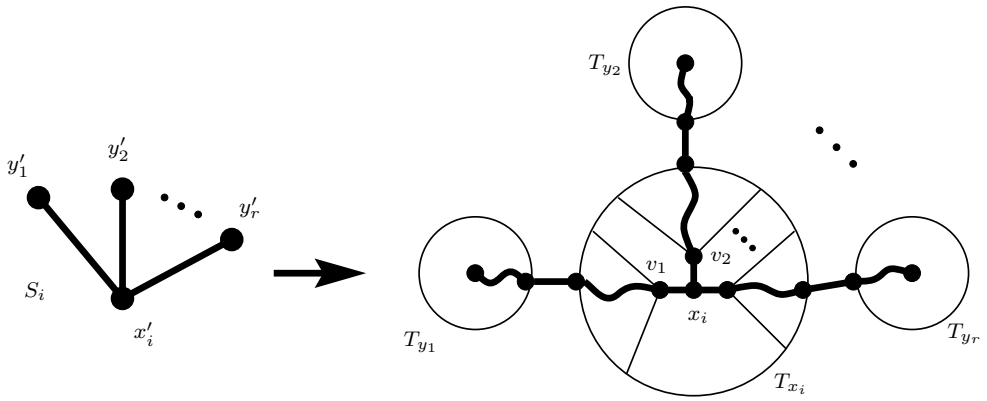


FIG. 2. *The star S_i corresponds to a subdivided star in G.*

Given a vertex $v \in T_x$ and a neighbor y' of x' in G', we say that y' *lies above* v if v lies on the subpath of T_x which joins x to the unique T_x–T_y edge of G (Figure 1). Our next aim is to find disjoint r-stars $S_1, \dots, S_{r+1}$ in $G'[A \cup B]$ such that, writing x'_i for the center of S_i, they satisfy the following properties:

(a) No edge of S_i belongs to F. No F-edge joins 2 leaves of S_i.
(b) There are no F-edges joining two different S_i.
(c) The leaves of S_i lie above different neighbors of x_i in T_{x_i}.

The vertices x_i will be the branch vertices of our subdivided K_{r+1}. Property (c) ensures that each S_i corresponds to a subdivided r-star in G (Figure 2). The stars $S_1, \dots, S_{r+1}$ can be found greedily as follows. Suppose that we have already chosen $S_1, \dots, S_{i-1}$ for some $i \le r+1$. Let W be the set consisting of all those vertices in $G'[A \cup B]$ which send an F-edge to some vertex in $V(S_1) \cup \dots \cup V(S_{i-1}) =: W'$. Since the F-edges are independent, we have $|W \cup W'| \le 2|W'| \le 2r(r+1)$. Take x'_i to be any vertex in $A \setminus (W \cup W')$. To see that such a vertex exists, we have to check that $|A| > |W \cup W'|$. But this holds since all the vertices in A have their neighbors in $A \cup B$ and thus

$$|A| \ge \delta(G') - |B| > \delta(G') - c \overset{(2)}{\ge} c \ge 2r(r+1) \ge |W \cup W'|,$$

as desired. Now we have found the center x'_i of S_i. So next we will choose its leaves. Let $v_1, \dots, v_r$ be distinct neighbors of x_i in the tree T_{x_i} (Figure 2). (So the v_j are vertices of G.) Let V_j be the set of all those neighbors of x'_i in G' which lie above v_j. Let $U_j := V_j \cap (W \cup W')$. We assume that the v_j are ordered descendingly according to the size of U_j, i.e., if $k > j$, then $|U_k| \le |U_j|$. For each v_j in turn we have to choose a neighbor y'_j of x'_i in G' such that (I) $y'_j \in V_j$, (II) $y'_j \notin U_j$, (III) $x'_i y'_j \notin F$, (IV) no

F-edge joins y'_j to any $y'_1, \dots, y'_{j-1}$ chosen previously for $v_1, \dots, v_{j-1}$ and such that (V) $y'_j \neq y'_k$ for $k < j$. Then (a)–(c) are satisfied and $y'_1, \dots, y'_r$ can be taken as leaves of S_i. We call every neighbor of x'_i in V_j which violates one of the properties (II)–(IV) a *forbidden neighbor*. We will show that there exists one such neighbor y'_j which is not forbidden. For this, first note that $V_j \cap V_k = \emptyset$ for $j \neq k$. (Indeed, if $y' \in V_j \cap V_k$, then G would contain at least two edges between T_{x_i} and T_y, contradicting (iii). In particular, this shows that (V) will automatically be satisfied.) Thus also $U_j \cap U_k = \emptyset$ for $j \neq k$. By our assumption on the ordering of the vertices $v_1, \dots, v_r$, this implies that $|U_j| \leq |W \cup W'|/j$. Since the number of neighbors which are forbidden by (III) and (IV) is clearly at most j, this implies that the total number of forbidden neighbors is at most

$$j + |U_j| \leq j + 2r(r+1)/j \leq 1 + 2r(r+1).$$

The final inequality is due to the fact that $j + 2r(r+1)/j$ is a decreasing function in j for $1 \leq j \leq r$.

On the other hand, property (iii) implies that the branch of T_{x_i} above v_j has at least $(r-1)^2$ leaves and thus sends out at least $(r-1)^3$ edges going to other trees T_y. So there are at least $(r-1)^3$ neighbors of x'_i which lie above v_j, i.e., $|V_j| \geq (r-1)^3$, which is greater than $1 + 2r(r+1)$ for $r \geq 5$. Thus there exists a neighbor y'_j of x'_i lying above v_j which is not forbidden, as desired. This proves the existence of the stars $S_1, \dots, S_{r+1}$.

We now consider the graph G^* obtained from $G'[A \cup B]$ by contracting every edge in F. Conditions (a) and (b) ensure that the images of $S_1, \dots, S_{r+1}$ in G^* are still disjoint r-stars. Our aim now is to delete the centers of these stars and then to link the leaves of them by $\binom{r+1}{2}$ disjoint paths in such a way that (after adding the centers again) we obtain a subdivision of K_{r+1} in G^* whose branch vertices are the centers. Since by condition (c) each S_i corresponds to a subdivided r-star in G, it is easy to see that this subdivision of K_{r+1} in G^* would then correspond to one in G (with branch vertices $x_1, \dots, x_{r+1}$).

Thus it suffices to show that the graph obtained from G^* by deleting the centers of the images of $S_1, \dots, S_{r+1}$ is $\binom{r+1}{2}$-linked. By Theorem 4 this is the case if the minimum degree of this graph is at least $10\binom{r+1}{2}$ and if its connectivity is at least $2\binom{r+1}{2}$. Thus is suffices to show that the minimum degree of G^* is at least $10\binom{r+1}{2} + r + 1$ and its connectivity is at least $2\binom{r+1}{2} + r + 1$. The second condition is satisfied since Lemma 5 implies that G^* is $\lceil c/3 \rceil$-connected and $c/3 = 2\binom{r+1}{2} + r + 1$. Moreover, when contracting F, each vertex $a \in A$ looses one neighbor for each F-edge to which a is either incident or which has the property that a is joined to both of its endvertices. Thus

$$\delta(G^*) \geq \delta(G') - |B| > r(r-1)^3 - c \geq 10\binom{r+1}{2} + r + 1,$$

as desired (use that $r \geq 5$).This completes the proof of Theorem 1 for the case when $r \geq 5$.

4.2. Finding a subdivision of K_5. Let G be a graph of minimum degree 4 and girth at least 27. Set

$$c := 51 = 3\left(2 \cdot \left(\binom{5}{2} - 4\right) + 5\right).$$

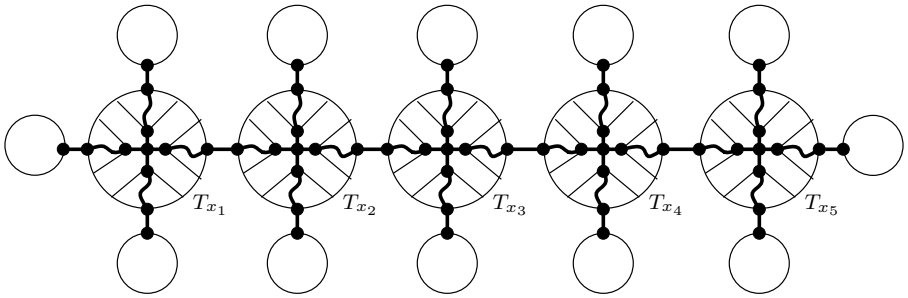


FIG. 3. *The subgraph of G corresponding to graph $H' \subseteq G'$ defined in condition (δ).*

Thus c is a little smaller than the number obtained by substituting $r = 4$ in (1). Proceed similarly as in the first part of section 4.1 to obtain trees T_x ($x \in X$), a graph G', vertex sets $A, B \subseteq V(G')$, and a set F of $|B|$ independent A–B edges. Again, denote by G^* the graph obtained from $G'[A \cup B]$ by contracting each edge in F. We now would like to choose stars $S_1, \dots, S_5$ in $G'[A \cup B]$ and continue as in section 4.1. However, c is now smaller and thus the graph obtained from G^* by deleting any five vertices (e.g., the star centers) will only be $(\binom{5}{2} - 4)$-linked in Case 1 below instead of $\binom{5}{2}$-linked as in section 4.1, and in Case 2 it will only be $(\binom{5}{2} - 5)$-linked. So we have to argue more carefully to ensure that when applying the linkedness we only have to find 6 of the 10 subdivided edges of our TK_5 in Case 1 and 5 subdivided edges in Case 2. The reason why we choose c to be smaller now is that when passing from G' to G^*, the minimum degree may decrease by $|B| \leq c - 1$, and thus we can guarantee a larger minimum degree of G^* in this way.

The idea is to distinguish two cases according to the minimum degree of G^*.

Case 1. The minimum degree of G^* is at least 65.

The strategy here is to choose the stars $S_1, \dots, S_5$ in such a way that their centers form a path of length 5. Then we can use the edges of this path to find four of the subdivided edges of our TK_5 and thus we only need the linkedness of G^* to find the six remaining edges. More formally, we wish to find stars $S_1, \dots, S_5$ in G' satisfying the following properties:

- (α) Both S_1 and S_5 are 3-stars and each of S_2, S_3, S_4 is a 2-star. All the S_i are disjoint from each other. The centers $x'_1 \dots x'_5$ of $S_1, \dots, S_5$ form a path in $G'[A]$.
- (β) No edge of S_i belongs to F. No F-edge joins 2 leaves of S_i.
- (γ) There are no F-edges joining two different S_i. In particular, none of the edges on the path $x'_1 \dots x'_5$ lies in F.
- (δ) Let H' denote the union of all the stars S_i and the path $x'_1 \dots x'_5$. (Thus every x'_i has degree 4 in H'.) The neighbors of x'_i in H' lie above different neighbors of x_i in T_{x_i} (Figure 3).

This can be achieved as follows. We first choose the path $x'_1 \dots x'_5$. This can be done greedily. Indeed, suppose that we have already found $x'_1 \dots x'_j$ for some $1 \leq j < 5$. For x'_{j+1} we can take any neighbor y' of x'_j which lies in $A \setminus \{x'_1, \dots, x'_j\}$, does not send an F-edge to any of $x'_1, \dots, x'_j$, and is such that y' and x'_{j-1} lie above different neighbors of x_j in T_{x_j}. But by (iii) there are at least $3 \cdot 3^3 = 81$ neighbors of x'_j satisfying the latter property and at most $|B| + j \leq c + 4 = 55$ of them lie in $B \cup \{x'_1, \dots, x'_j\}$. Moreover, at most $j \leq 4$ of the remaining neighbors send an F-edge to any of $x'_1, \dots, x'_j$. Thus we can find the path $x'_1 \dots x'_5$ consisting of the centers of $S_1, \dots, S_5$.

We now have to choose the leaves of the S_i. As before, this can be done greedily. It is easy to check that one can find leaves for each of $S_1, \dots, S_4$ in turn. So let us now choose the three leaves for S_5. Let v_1, v_2, v_3 be three neighbors of x_5 in T_{x_5} such that x'_4 does not lie above v_i for each $i = 1, 2, 3$. Let W be the set consisting of all those vertices in $G'[A \cup B]$ which send an F-edge to some vertex in $V(S_1) \cup \dots \cup V(S_4) \cup \{x'_5\}$. Set $W' := (V(S_1) \cup \dots \cup V(S_4)) \setminus \{x'_4\}$. Given $i \leq 3$, we say that a vertex y' is a *candidate for the ith leaf of S_5* if y' lies above v_i and if $y' \notin W \cup W'$. We are looking for three candidates, one for the first leaf, one for the second, and one for the third, such that no F-edge joins two of them. (Then these candidates can be taken as leaves of S_5. Indeed, note that the candidates are automatically distinct from x'_4 since otherwise G would contain at least two T_{x_4}–T_{x_5} edges.) There are at least 27 neighbors of x'_5 which lie above v_i. Since at most $|W \cup W'| \leq 26$ of these neighbors lie in $W \cup W'$, there is at least one candidate for the ith leaf. Moreover, this argument also shows that for at most one index i there are less than three candidates for the ith leaf. (Indeed, suppose there are at most two candidates for the first leaf, say. Then at most one of the vertices above v_2 or v_3 can be in $W \cup W'$, as G' contains no multiple edges.) Thus there must be three candidates, one for each of the three leaves, such that no F-edge joins two of them, as required.

Similarly as in section 4.1 we now consider the graph G^* obtained from $G'[A \cup B]$ by contracting each edge in F. Again, conditions (β) and (γ) ensure that the image of H' in G^* is still isomorphic to H', i.e., nothing in H' will be contracted. As in section 4.1 we now delete the images of the centers $x'_1, \dots, x'_5$ in G^* and then wish to link the images of the leaves of the S_i to obtain a subdivision of K_5 in G^* whose branch vertices are the centers. However, since the centers lie on a path of length 4 in G^* we can link adjacent branch vertices via an edge of the path. Thus we now only have to find the $\binom{5}{2} - 4 = 6$ remaining subdivided edges of our TK_5 in G^*. For this, it suffices that the graph obtained from G^* by deleting the five branch vertices is 6-linked. By Theorem 4 this is the case if the minimum degree this graph is at least 60 and if its connectivity is at least 12. This in turn holds if the minimum degree of G^* is at least 65 and its connectivity is at least 17. The first requirement holds by our assumption. The latter is satisfied since by Lemma 5 the graph G^* is $\lceil c/3 \rceil$-connected and $c/3 = 17$. Condition (δ) now implies that our subdivision of K_5 in G^* corresponds to a subdivision of K_5 in G whose branch vertices are $x_1, \dots, x_5$.

Case 2. The minimum degree of G^* is at most 64.

Recall that, as in section 4.1, when contracting the edge set F to obtain G^* from $G'[A \cup B]$, each vertex $x' \in A$ loses one neighbor for each F-edge which is incident with a or of which a sees both endvertices. Thus we obtain the lower bound

$$\delta(G^*) \geq \delta(G') - |B| > 4 \cdot 3^3 - c = 57. \tag{3}$$

This is too small for us to be able to proceed as in Case 1, since with this bound we can now only guarantee that the graph obtained by deleting five vertices of G^* is 5-linked instead of 6-linked as it was in Case 1. Our solution is to find in G' a suitable subgraph Q' consisting of a triangle and a path attached to it. The branch vertices and 5 of the 10 subdivided edges of our subdivision will be contained in the subgraph of G which corresponds to Q'. The existence of some triangle in G' is easy to show: our assumption on the minimum degree of G^* implies that there exists a triangle $x'y'z'$ in $G'[A \cup B]$ which is formed by some vertex $x' \in A$ and an F-edge $y'z'$. (In fact, some $x' \in A$ must form such a triangle with at least $108 - 64 - 1 = 43$ of the F-edges, but we will not make use of this.)

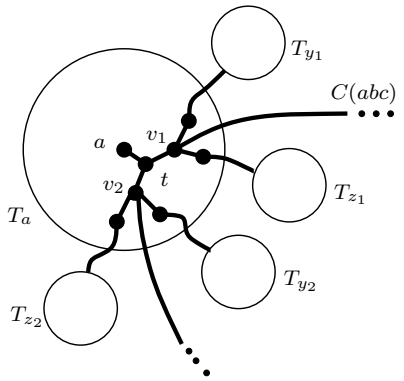


FIG. 4. *The trees corresponding to the docking points for v_1 and v_2.*

Given a triangle $u'v'w'$ in G', we denote by $C(uvw)$ its corresponding cycle in G, i.e., the unique cycle which consists of the T_u–T_v edge, the T_v–T_w edge, the T_u–T_w edge as well as the unique paths in T_u, T_v and T_w joining endvertices of these edges. The *turning point of $C(uvw)$ in T_u* is that vertex in $C(uvw) \cap T_u$ whose distance to u is minimal. The turning points in T_v and T_w are defined similarly. Consider all triangles as described in the previous paragraph, i.e., all triangles in $G'[A \cup B]$ which meet A in precisely two vertices and contain an F-edge. Among all these triangles choose $a'b'c'$ such that the distance of a to the turning point t of $C(abc)$ in T_a is minimal, where $a', c' \in A$ and $b' \in B$. We have to distinguish two cases.

Case 2.1. The distance between a and t is at most 1.

Let us start with a definition. Given $v \in T_x$, we say that distinct neighbors $y'_1, \dots, y'_i$ of x' in G' *form a v-join* if all the subpaths of T_x which join v to the unique T_x–T_{y_i} edges meet in v and are otherwise disjoint from each other. This means that if x' is the center of an i-star S in G' whose leaves form a v-join for some vertex $v \in T_x$, then S corresponds to a subdivided i-star in G whose center is v. Moreover, note that neighbors $y'_1, \dots, y'_i$ of x' form an x-join if and only if they lie above different neighbors of x in T_x.

Now let v_1 and v_2 denote the neighbors of t on $C(abc)$. The vertices t, v_1, v_2 will be three branch vertices of our subdivision of K_5. We will use the cycle $C(abc)$ to join these three branch vertices pairwise. Before this we will choose for each v_i two neighbors y'_i and z'_i of a' in G' which will serve as "docking points" when using the linkedness of G^* to join v_i to the remaining two branch vertices of the subdivided K_5 (Figure 4). (These "docking points" play the same role for v_i as the leaves of S_i do for x_i in both section 4.1 and Case 1.) The desired docking points y'_i and z'_i have to satisfy the following properties:

- No F-edge joins a' or c' to any of the y'_i or z'_i.
- There is no F-edge between any of the four vertices y'_i and z'_i.
- The vertices y'_i and z'_i are neighbors of a' in $G' - \{b', c'\}$ which lie above v_i. Moreover, y'_i, z'_i, b', c' form a v_i-join.

Note that the third condition together with the fact that either $b'a' \in F$ or $b'c' \in F$ implies that neither y'_i nor z'_i can be joined to b' by an F-edge. Let us now show that such docking points y'_1, y'_2, z'_1, and z'_2 exist. (For this, without mentioning it explicitly, we will make frequent use of the fact that every vertex sends out at most one F-edge and that G contains at most one edge between every pair T_x, T_y of trees.)

First fix two different neighbors v_i^y and v_i^z of v_i in T_a which lie above v_i and avoid $C(abc)$. Such neighbors exist since v_i is not the turning point of $C(abc)$ in T_a and thus only one of the at least three neighbors of v_i lying above v_i in T_a lies on $C(abc)$. Since v_i has distance at most 2 from the root a of T_a, the branches of T_a above v_i^y and v_i^z both send out at least three edges. (Note that the definition of v_i^y and v_i^z implies that none of these edges can go to T_b or T_c.) Thus there are at least three neighbors of a' in $G' - \{b', c'\}$ which lie above v_i^y and the same holds for v_i^z. This means that for each of y_1', y_2', z_1', z_2' there are at least three candidates which satisfy the third of the above properties. Without loss of generality we may assume that no F-edge joins a' to a candidate for z_1', y_2' or z_2' (relabel if necessary to achieve this). Moreover, we will consider only the case when no F-edge joins c' to a candidate for z_1' or z_2'. In this case we will first choose y_1' and then y_2'. The remaining cases are analogous. Since there are at least three candidates for y_1', one of them sends no F-edge to a' or c'. Denote it by y_1'. Similarly, for y_2' take any candidate which does not send an F-edge to y_1' or c'. Note that $a'y_2' \notin F$ by our assumption. Next we have to choose z_1' and z_2'. We will consider only the case when there are at least two candidates for z_2' which are joined to neither y_1' nor y_2' by an F-edge. (The case when this holds for z_1' instead of z_2' is analogous.) Take z_1' to be any candidate which is joined to neither y_1' nor y_2' by an F-edge. Finally, we can choose z_2' since by assumption there is at most one candidate for it which is joined to y_1' or y_2' by an F-edge and at most one of the remaining ≥ 2 candidates sends an F-edge to z_1'. (Recall that our earlier assumptions imply that neither a' nor c' sends an F-edge to z_1' or z_2'.)

Next we will choose one docking point for t. Let v_t^y and v_t^z be two neighbors of t in T_a which are distinct from v_1 and v_2. If $t \neq a$, we choose v_t^z to be a. Note that the branch of T_a above v_t^y sends out at least nine edges and none of them goes to T_b, T_c, T_{y_1}, T_{y_2}, T_{z_1}, or T_{z_2}. Thus there are at least nine neighbors of a' in $G' - \{b', c', y_1', y_2', z_1', z_2'\}$ which lie above v_t^y. Let y_t' be one such neighbor which sends an F-edge to none of a', c', y_1', y_2', z_1', z_2'.

For $i = 1, 2$ let S_i denote the 2-star in $G'[A \cup B]$ whose center is a' and whose leaves are y_i' and z_i'. Let S_3 be the 1-star with center a' and leaf y_t'. Thus the stars S_1, S_2, and S_3 meet in a' and are disjoint otherwise. We will now choose a 2-star S_4 and a 3-star S_5 satisfying the following properties:

- (α') S_4 and S_5 are disjoint and avoid $V(S_1) \cup V(S_2) \cup V(S_3) \cup \{b', c'\}$. The centers x_4' and x_5' of S_4 and S_5 lie in A. Moreover, either $a'x_4'x_5'$ forms a path or else there is a vertex $z_t' \in B \setminus (V(S_1 \cup \cdots \cup S_5) \cup \{b', c'\})$ such that $a'z_t'x_4'x_5'$ is a path and $z_t'x_4' \in F$.
- (β') No edge of S_4 or S_5 belongs to F. No F-edge joins two leaves of S_4 or of S_5.
- (γ') There is no F-edge joining S_4 to a vertex in $V(S_1 \cup S_2 \cup S_3 \cup S_5) \cup \{a', b', c'\}$. The analogous condition holds for S_5.
- (δ') Let P' denote the path $a'x_4'x_5'$ or $a'z_t'x_4'x_5'$ guaranteed in (α'). Let H' be the union of P', S_4, and S_5. (Thus both x_4' and x_5' have degree 4 in H'.) For each $i = 4, 5$ the 4 neighbors of x_i' in H' form an x_i-join. Moreover, the neighbor of a' on P' lies above v_t^z.

(The vertices x_4 and x_5 will be the two remaining branch vertices of our subdivision of K_5 in G.) We will only show that we can either find a path $a'x_4'x_5'$ or a path $a'z_t'x_4'x_5'$ with the desired properties. The existence of the leaves of S_4 and S_5 then follows by an argument analogous to the one in Case 1 (but with more room to spare this time). So let us consider all those at least 27 neighbors of a' that lie above v_t^z. Clearly, none of them lies in $V(S_1 \cup S_2 \cup S_3) \cup \{b', c'\}$. Moreover, at most eight of them are joined by an F-edge to a vertex in $V(S_1 \cup S_2 \cup S_3) \cup \{b', c'\}$. If one of the

remaining neighbors lies in A, we take it to be x_4'. Thus we may assume that each of the at least 19 remaining neighbors lies in B. Take z_t' to be any such neighbor which has the property that the endvertex in A of the unique F-edge incident to z_t' does not lie in $V(S_1 \cup S_2 \cup S_3) \cup \{a', b', c'\}$. Take x_4' to be this endvertex. The vertex x_5' can now be found in the same way as in Case 1.

Having chosen our stars $S_1, \dots, S_5$, we proceed similarly as in Case 1. Again, we consider the graph G^* obtained from $G'[A \cup B]$ by contracting each edge in F. The stars $S_1, \dots, S_5$ have been chosen is such a way that in the subgraph of G' induced by all the stars S_i, the path P' and the triangle $a'b'c'$ at most two edges are contracted, namely, the F-edge lying on $a'b'c'$ and the F-edge $z_t'x_4'$ (if z_t' exists). Otherwise nothing changes in this subgraph when passing over to G^*.

We will now choose our subdivision of K_5 in G. As indicated before, the branch vertices will be t, v_1, v_2, x_4, and x_5. Five of the pairs of branch vertices can be linked directly. Indeed, we can use the cycle $C(abc)$ to link t, v_1 and v_2 pairwise. Moreover, x_4 will be linked to x_5 by the path consisting of the unique T_{x_4}–T_{x_5} edge together with the paths in T_{x_4} and T_{x_5} joining x_4 and x_5 to this edge. In the case when a' and x_4' are joined by an edge, we can link t to x_4' in a similar way. If z_t' exists, i.e., if $a'z_t'x_4'$ is a path of length 2, then the path linking t to x_4 will run through T_{z_t}. We now have to find paths joining the remaining five pairs of branch vertices. Similarly as in Case 1, this will be done by using the linkedness of G^* to connect the docking points, i.e., the images of the leaves of the stars $S_1, \dots, S_5$. Each of the five paths we are looking for has to avoid the star centers as well as the image of c' in G^*. Thus we delete these four vertices. We can now link the images of the star leaves if the subgraph of G^* thus obtained is 5-linked. So by Theorem 4 we are done if the minimum degree this graph is at least 50 and if its connectivity is at least 10. This in turn holds if the minimum degree of G^* is at least 54 and its connectivity is at least 14. Inequality (3) shows that the first condition holds. The latter condition is satisfied since by Lemma 5 the graph G^* is $\lceil c/3 \rceil$-connected and $c/3 = 17$.

Case 2.2. The distance between a and t is at least 2.

Similarly as in Case 2.1, also in this case we will choose the branch vertices in such a way that three of them lie on the cycle $C(abc)$ and thus this cycle can be used to link them pairwise. The two remaining branch vertices will also be chosen in an analogous way, i.e., they will be attached to $C(abc)$ by a path. However, this time we cannot choose the first three branch vertices in $T_a \cap C(abc)$ as $C(abc)$ stays too far away from the center when passing through T_a. Instead, we will only choose two vertices in $T_a \cap C(abc)$ and one in $T_c \cap C(abc)$.

Since by assumption the distance from a to the turning point of $C(abc)$ in T_a is at least 2 and since by the choice of $a'b'c'$ the analogue also holds for the turning point of $C(abc)$ in T_c, the cycle $C(abc)$ meets both T_a and T_c in at most nine vertices (use condition (ii) to see this). Since $C(abc)$ meets T_b in at most 13 vertices and the girth of G is at least 27, it follows that $C(abc)$ has to meet each of T_a and T_c in at least five vertices. Moreover, it follows that $C(abc)$ meets at least one of T_a and T_b in at least seven vertices. Thus without loss of generality we may assume that $C(abc)$ meets T_a in at least seven and at most nine vertices and that $C(abc)$ meets T_c in at least five and at most nine vertices (otherwise relabel a' and c').

Let us now prove that there exists a vertex $c_0 \in T_c$ and neighbors y_c' and z_c' of c' in $G' - \{a', b'\}$ satisfying the following properties (the "moreover" part of the first property will not be used until Case 2.2.2):

- The vertex c_0 is either the first, the second, or the third vertex on $C(abc)$ in T_c when coming from T_a. Moreover, c_0 is not the turning point of $C(abc)$ in T_c.

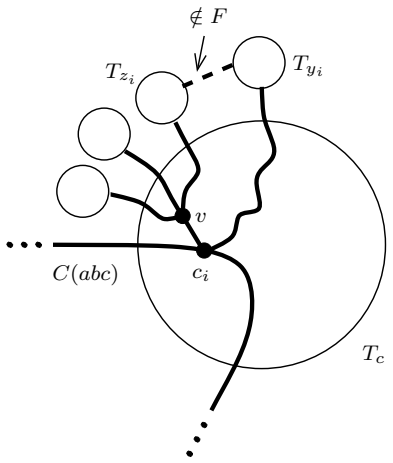


FIG. 5. *Choosing y_i' and z_i' in the case when c_i has a neighbor v in $T_c - C(abc)$.*

- The vertices y_c', z_c', a', b' form a c_0-join.
- F does not contain $y_c'z_c'$.
- Neither y_c' nor z_c' sends an F-edge to any of a', b', c'.

The vertex c_0 will be one of the branch vertices of our subdivision. Both y_c' and z_c' will serve as docking points for c_0. So let us now show that there exist such vertices c_0, y_c', and z_c'. Choose any two vertices c_1 and c_2 such that they are among the first three vertices on $C(abc)$ in T_c when coming from T_a and such that none of them equals the turning point. (These two vertices exist since $C(abc)$ meets T_c in at least five vertices.) Thus c_1 and c_2 are candidates for c_0 in the sense that they satisfy the first property above.

Before we continue, we need some more notation which generalizes the notion of "lying above." Given a subtree $T \subseteq T_x$ and a neighbor z' of x' in G', we say that z' *belongs to* T if T contains an endvertex of the unique T_x–T_z edge. Thus if $v \in T_x$, then z' lies above v if and only if z' belongs to the branch of T_x above v.

Next we show that for each $i = 1, 2$, one can find neighbors y_i' and z_i' of c' in $G' - \{a', b'\}$ such that a', b', y_i', z_i' form a c_i-join and $y_i'z_i' \notin F$. To do this, we first consider the case that all neighbors of c_i in T_c are contained in $C(abc)$. This implies that c_i sends out at least two edges to vertices in $G - (T_a \cup T_b \cup T_c)$, i.e., there are (at least) two neighbors of c' in $G' - \{a', b'\}$ which form a c_i-join together with a' and b'. Take y_i' and z_i' to be two such neighbors. Suppose that $y_i'z_i' \in F$. We may assume that $y_i' \in A$ and $z_i' \in B$ (relabel if necessary). Since the triangle $c'y_i'z_i'$ was a candidate for the choice of $a'b'c'$, it follows that the turning point of $C(cy_iz_i)$ in T_{y_i} has distance at least 2 from y_i. Thus $C(cy_iz_i)$ meets T_{y_i} in at most nine vertices. However, it meets T_{z_i} in at most 13 vertices and T_c in precisely one vertex. Thus the length of $C(cy_iz_i)$ is at most 23, contradicting the fact that the girth of G is at least 27. So we may turn to the case that c_i has a neighbor v in $T_c - C(abc)$. In the case when v is the only such neighbor, c_i has to send out an edge to $G - (T_a \cup T_b \cup T_c)$. Take y_i' to be the vertex for which the corresponding tree T_{y_i} contains the endvertex of that edge. If c_i has another neighbor w in $T_c - C(abc)$, take for y_i' any neighbor of c' which belongs to the component of $T_c - c_i$ containing w. To find z_i', consider those at least three neighbors of c' which belong to the component of $T_c - c_i$ containing v (Figure 5). At most one of them sends an F-edge to y_i'. Take z_i' to be any other such neighbor. Thus

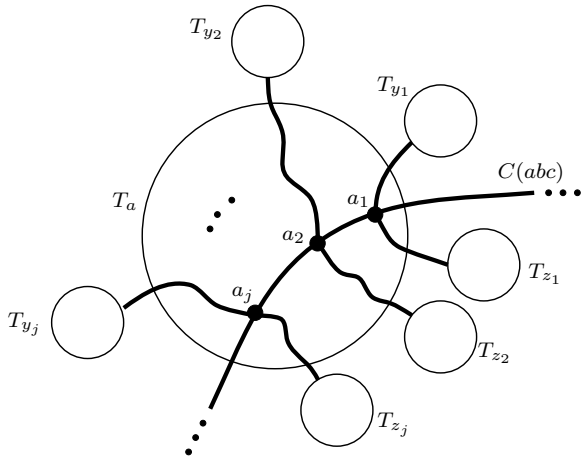


FIG. 6. *The neighbors y_i' and z_i' chosen for a_i.*

in each case we have found vertices y_i' and z_i' as required. Note that the sets $\{y_1', z_1'\}$ and $\{y_2', z_2'\}$ are disjoint. (Indeed, if, for example, $y_1' = y_2'$, then G would contain at least two T_c–T_{y_1} edges.)

Let us now show that for some i we have $a'y_i', a'z_i', c'y_i', c'z_i' \notin F$. (Note that $b'y_i', b'z_i' \notin F$ holds automatically since either $a'b' \in F$ or $b'c' \in F$ and F consists of independent edges.) If this is not the case, then, for example, we have that $a'y_1', c'y_2' \in F$. But this contradicts the fact that either a' or c' already sends an F-edge to b'. The vertices c_i, y_i', z_i' for which $a'y_i', a'z_i', c'y_i', c'z_i' \notin F$ are a good choice for c_0, y_c', z_c'.

Having found one branch vertex c_0 in $C(abc) \cap T_c$, we will choose two other branch vertices in $C(abc) \cap T_a$. So let $a_1, a_2, \dots, a_j$ denote all those vertices on $C(abc)$ which lie in T_a in the order in which they appear when coming from T_c. Thus $7 \le j \le 9$ by our assumption. For each $i = 1, \dots, j$ we choose neighbors y_i' and z_i' of a' in $G - \{b', c'\}$ which satisfy the following properties (Figure 6):

(*A*) The vertices y_i', z_i', b', c' form an a_i-join.

(*B*) F does not contain $y_i'z_i'$.

Such neighbors can be found in the same way as before. (Recall that the fact that c_0 was not a turning point was not used when showing the existence of y_c and z_c.) Among all possible choices, we choose each pair y_i', z_i' such that $|\{y_i', z_i'\} \cap \{y_c', z_c'\}|$ is minimal. Thus we try to keep the pair y_i', z_i' as disjoint as possible from y_c', z_c'. Again, the sets $\{y_i', z_i'\}$ are automatically disjoint for different indices $i \le j$.

Next we show that there exists an index $i_0 \le j$ for which we can take a_{i_0} to be the second branch vertex and both y_{i_0}' and z_{i_0}' to be docking points for a_{i_0}. So i_0 has to satisfy the following properties:

(*C*) Neither y_{i_0}' nor z_{i_0}' sends an F-edge to any of a', b', c', y_c', z_c'.

(*D*) Neither y_{i_0}' nor z_{i_0}' lies in $\{y_c', z_c'\}$.

Since either $a'b' \in F$ or $b'c' \in F$, at most one of a', b', c' sends an F-edge to a vertex in $G' - \{a', b', c'\}$. Thus altogether the vertices a', b', c', y_c', z_c' send out at most three F-edges to vertices in $G' - \{a', b', c'\}$. This implies that at most five of the indices $i \le 7$ violate (*C*) or (*D*). Hence there exists an index i_0 as required. Put $I := \{1, \dots, j\} \setminus \{i_0\}$.

Our next aim is to show that there is another a_i which we can take as a branch vertex (and such that y_i' and z_i' can serve as docking points for a_i). Thus we have to show that there exists an index $i_1 \in I$ which satisfies the following properties:

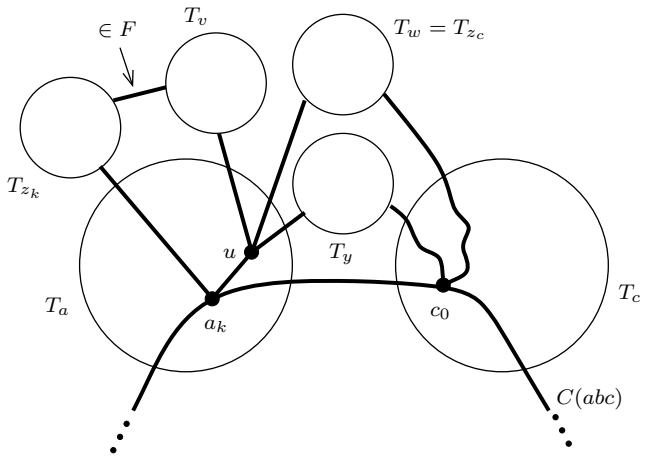


FIG. 7. *The three neighbors of a' lying above u.*

(C') Neither y'_{i_1} nor z'_{i_1} sends an F-edge to any of $a', b', c', y'_c, z'_c, y'_{i_0}, z'_{i_0}$.
(D') Neither y'_{i_1} nor z'_{i_1} lies in $\{y'_c, z'_c, y'_{i_0}, z'_{i_0}\}$.

As before, it follows that at most five indices $i \in I$ violate (C'). Since $\{y'_{i_0}, z'_{i_0}\}$ is disjoint from $\{y'_i, z'_i\}$ for each $i \in I$, at most two indices $i \in I$ violate (D'). Hence altogether at most seven of the indices $i \in I$ are forbidden. Thus i_1 exists if $j = 9$ (since then $|I| = j - 1 = 8$).

So we only need to consider the case when $7 \le j \le 8$. Moreover, we may assume that for some index $k \in I$ either y'_k or z'_k lies in $\{y'_c, z'_c\}$. Let us assume that $y'_k = y'_c =: y'$ (relabel if necessary to achieve this). Furthermore, if $j = 8$, we may assume that there is some index $\ell \in I \setminus \{k\}$ for which either y'_ℓ or z'_ℓ equals z'_c. So let us assume that $y'_\ell = z'_c$ in the case when $j = 8$. To show the existence of the index i_1, we will now prove that these assumptions lead to a contradiction. To do this, we distinguish the following two cases.

Case 2.2.1. The vertex a_k is not an endvertex of the unique T_a–T_y edge.

Let u denote the neighbor of a_k on the subpath of T_a joining a_k to the endvertex of the unique T_a–T_y edge (Figure 7). Thus there are at least three neighbors of a' which belong to the component of $T_a - a_k$ containing u. So each of these three neighbors was a candidate for y'_k when choosing the pair y'_k, z'_k. Since y'_k, z'_k was chosen to be as disjoint as possible from y'_c, z'_c, this implies there are precisely three such neighbors: one of which (namely, $y' = y'_k$) is equal to y'_c, one of which, w', say, is equal to z'_c, and the third one, v', say, sends an F-edge to z'_k. (Otherwise the pair consisting of z'_k together with one such neighbor would have been a better choice for y'_k, z'_k.) Since there are precisely three such neighbors, u has to be an endvertex of all the three edges joining T_a to each of T_v, T_w, and T_y. Moreover, the choice of y'_k, z'_k implies that a_k is an endvertex of the unique T_a–T_{z_k} edge (Figure 7). Let us consider the cycle $C(avz_k)$ corresponding to the triangle $a'v'z'_k$. This cycle meets T_a in precisely two vertices. Since the girth of G is at least 27, it meets both T_v and T_{z_k} in at least $27 - 2 - 13 = 12$ vertices. Thus the turning points of $C(avz_k)$ in T_v and T_{z_k} have to be v and z_k. But this shows that the triangle $a'v'z'_k$ would have been a better choice for $a'b'c'$, a contradiction.

Case 2.2.2. The vertex a_k is an endvertex of the unique T_a–T_y edge.

This time consider the cycle $C(acy)$ corresponding to the triangle $a'c'y'$. Let us first estimate the number of vertices in $C(acy) \cap T_c$. Note that the distance from

the root c of T_c to c_0 is as least one larger than the distance from c_0 to the turning point of $C(abc)$ in T_c (this holds since c_0 is not this turning point). Since the latter distance is at least 2, it follows that the segment of $C(acy) \cap T_c$ which joins c_0 to the endvertex of the T_c–T_y edge consists of at most four vertices. Since c_0 was one of the first three vertices on $C(abc)$ in T_c when coming from T_a, it follows that the segment of $C(acy) \cap T_c$ which joins c_0 to the endvertex of the T_a–T_c edge consists of at most three vertices. Thus $C(acy)$ contains at most six vertices in T_c (we have counted c_0 twice). Since $C(acy)$ contains at most 13 vertices in T_y and the girth of G is at least 27, it follows that $C(acy)$ meets T_a in at least 8 vertices. On the other hand, since by assumption a_k is an endvertex of the unique T_a–T_y edge, it follows that $C(acy)$ meets T_a precisely in the vertices $a_1, \dots, a_k$. Thus we may assume that $k = 8$ and therefore $j = 8$. (Recall that we had previously ruled out all possibilities for j except 7 and 8.) As shown before Case 2.2.1, this in turn implies that we may assume that $y'_\ell = z'_c =: z'$ for some index $\ell \in I \setminus \{k\}$. Thus $\ell \leq 7$. The case when a_ℓ is not an endvertex of the unique T_a–T_z edge can be dealt with as in Case 2.2.1. Thus we may assume that a_ℓ is an endvertex of the unique T_a–T_z edge. But just as for $C(acy)$, one can show that the cycle $C(acz)$ corresponding to the triangle $a'c'z'$ has to meet T_a in at least eight vertices. But this is a contradiction since it meets T_a precisely in $a_1, \dots, a_\ell$ and $\ell \leq 7$. This completes the proof of the existence of the index i_1.

So far, we have chosen three branch vertices c_0, a_{i_0}, and a_{i_1} on $C(abc)$ together with two docking points for each of them. We will now find our subdivision of K_5 similarly as in Case 2.1. As there, the remaining two branch vertices will be attached to $C(abc)$ by a path. So we have to find a suitable path in $G'[A \cup B]$ of length 2 or 3 which starts in either a' or c'. As in Case 2.1, the last two vertices on this path will be the centers of the stars S_4 and S_5. The remaining two branch vertices of our subdivision will be the roots of the two trees corresponding to the centers of S_4 and S_5. Again, S_4 will be a 2-star and S_5 a 3-star.

More precisely, we proceed as follows. First consider the case that one of the docking points $y'_c, y'_{i_0}, y'_{i_1}, z'_c, z'_{i_0}, z'_{i_1}$ lies in A. Suppose, for example, that $z'_{i_0} \in A$. (The other cases are analogous.) Then we take z'_{i_0} to be the center x'_4 of S_4. S_1 will be the 2-star in $G'[A \cup B]$ whose center is c' and whose leaves are y'_c and z'_c. S_2 will be the 2-star in $G'[A \cup B]$ whose center is a' and whose leaves are y'_{i_1} and z'_{i_1}. S_3 will be the 1-star in $G'[A \cup B]$ whose center is a' and whose leaf is y'_{i_0}.

So let us now consider the case that all the six docking points lie in B. Then at least one of them, z'_{i_0}, say, sends an F-edge to a vertex outside $\{a', c'\}$. Let x'_4 denote this vertex. Since $x'_4 \in A$, x'_4 is automatically distinct from b' and from all the docking points. Moreover, x'_4 sends neither an F-edge to any of a', b', c' nor to any docking point other than z'_{i_0}. We take x'_4 to be the center of S_4. The stars S_1, S_2, S_3 are defined similarly as before.

In both cases, the center x'_5 of the star S_5 as well as the two leaves of S_4 and the three leaves of S_5 can now be found similarly as in Case 2.1. Having chosen $S_1, \dots, S_5$, the subdivision of K_5 can also be found as in Case 2.1. This completes the proof of Case 2.2 and thus of the case when $r = 4$.

5. Further improvements: Approaches and obstacles. Below, we briefly discuss three rather natural modifications to the proof strategy which one might try out in order to improve the bound of 27 on the girth in Theorem 1.

Recall that throughout this paper, the trees which were contracted to yield the auxiliary graph G' have radius at least 3. It would of course be desirable if we could

adapt our strategy to work also for trees of radius at least 2. In this case a girth of at least 19 would now suffice to ensure that the auxiliary graph G' graph has no multiple edges. However, the obvious drawback is that the minimum degree of G' may now be smaller. In particular, in the case $r = 4$ we can only guarantee $\delta(G') \geq 36$ instead of $\delta(G') \geq 108$, so this does not seem feasible for small r.

Another approach is of course to work with trees of radius at least 3 as before but now to relax the girth requirement a little. This leads to the possibility of multiple edges between pairs of trees. One can show that this does not reduce the minimum degree by more than a constant factor (e.g., by a factor of six if one assumes a girth of at least 21 instead of a girth of 27; see Lemma 15 in [9] for the argument). However, the resulting bound on the minimum degree is again too small for our techniques to apply if r is small. (Moreover, multiple edges cause other problems, too.)

Thirdly, one might hope for an improvement on the bound of $10k$ on the necessary average degree in order to ensure k-linkedness in Theorem 4. An example in [20] shows that the best one can hope for is to replace the constant 10 by 4. But even with such an optimal bound our proof would not give a better bound on the girth. However, maybe combining this with the allowance of multiple edges and some new ideas would work.

On the other hand, one might wonder how much of our improvement from a girth of 186 in [9] to 27 here is due to the fact that we did not use Theorem [20] in [9] but used the previous bound of Bollobás and Thomason [2], where the $10k$ is replaced by $22k$: it is easy to check that our proofs would still work using the latter bound if one requires the girth to be at least 35 (and considers trees which now have radius at least 4 instead of radius at least 3).

Finally, recall that there is a lot of room to spare in the estimates in section 4.1 when r is large. However, even if we assume r to be large, the arguments in section 4.1 no longer work for trees of radius at least 2 instead of radius at least 3. (Thus they do not provide a short proof of the fact that the girth bound can be reduced to 19 instead of 27 for large r.) It is still an open question if the bound of 15 on the girth for large r [9] can be reduced.

REFERENCES

[1] B. Bollobás, *Extremal Graph Theory*, Academic Press, New York, 1978.

[2] B. Bollobás and A. Thomason, *Highly linked graphs*, Combinatorica, 16 (1996), pp. 313–320.

[3] B. Bollobás and A. Thomason, *Proof of a conjecture of Mader, Erdős and Hajnal on topological complete subgraphs*, Eur. J. Comb., 19 (1998), pp. 883–887.

[4] P. Catlin, *Hajós' graph-coloring conjecture: Variations and counterexamples*, J. Combin. Theory Ser. B, 26 (1979), pp. 268–274.

[5] P. Erdős and S. Fajtlowicz, *On the conjecture of Hajós*, Combinatorica, 1 (1981), pp. 141–143.

[6] T.R. Jensen and B. Toft, *Graph Coloring Problems*, Wiley-Interscience, New York, 1995.

[7] H. A. Jung, *Eine Verallgemeinerung des n-fachen Zusammenhangs für Graphen*, Math. Ann., 187 (1970), pp. 95–103.

[8] J. Komlós and E. Szemerédi, *Topological cliques in graphs* II, Combin. Probab. Comput., 5 (1996), pp. 79–90.

[9] D. Kühn and D. Osthus, *Topological minors in graphs of large girth*, J. Combin. Theory B, 86 (2002), pp. 364–380.

[10] D. Kühn and D. Osthus, *Minors in graphs of large girth*, Random Structures Algorithms, 22 (2003), pp. 213–225.

[11] D. Kühn and D. Osthus, *Subdivisions of K_{r+2} in graphs of average degree at least $r + \varepsilon$ and large but constant girth*, Combin. Probab. Comput., 13 (2004), pp. 361–371.

[12] D. Kühn and D. Osthus, *Large topological cliques in graphs without a 4-cycle*, Combin. Probab. Comput., 13 (2004), pp. 93–102.

[13] D. Kühn and D. Osthus, *Induced subdivisions in $K_{s,s}$-free graphs of large average degree*, Combinatorica, 24 (2004), pp. 287–304.
[14] D. Kühn and D. Osthus, *Dense minors in $K_{s,s}$-free graphs*, Combinatorica, 25 (2004), pp. 49–64.
[15] W. Mader, *Topological subgraphs in graphs of large girth*, Combinatorica, 18 (1998), pp. 405–412.
[16] W. Mader, *$3n-5$ edges do force a subdivision of K_5*, Combinatorica, 18 (1998), pp. 569–595.
[17] W. Mader, *An extremal problem for subdivisions of K_5^-*, J. Graph Theory, 30 (1999), pp. 261–276.
[18] W. Mader, *Subdivisions of a graph of maximal degree $n+1$ in graphs of average degree $n+\varepsilon$ and large girth*, Combinatorica, 21 (2001), pp. 251–265.
[19] Shi, *personal communication*, Humboldt-Universität, Berlin, 2002.
[20] R. Thomas and P. Wollan, *An improved linear edge bound for graph linkages*, Eur. J. Comb., 26 (2005), pp. 309–324.
[21] C. Thomassen, *Girth in graphs*, J. Combin. Theory B, 35 (1983), pp. 129–141.
[22] C. Thomassen, *Chromatic numbers of triangle-free graphs and their complements*, Report 1/2004 Combinatorics, Mathematisches Forschungsinstitut Oberwolfach, 2004.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 79–95

OPTIMUM SECRET SHARING SCHEME SECURE AGAINST CHEATING*

WAKAHA OGATA†, KAORU KUROSAWA‡, AND DOUGLAS R. STINSON§

Abstract. Tompa and Woll introduced a problem of cheating in (k, n) threshold secret sharing schemes. In this problem $k-1$ malicious participants aim to cheat an honest one by opening forged shares and causing the honest participant to reconstruct the wrong secret. We first derive a tight lower bound on the size of shares $|\mathcal{V}_i|$ for secret sharing schemes that protect against this type of attack: $|\mathcal{V}_i| \geq (|\mathcal{S}|-1)/\delta + 1$, where $\mathcal{V}_i$ denotes the set of shares of participant P_i, $\mathcal{S}$ denotes the set of secrets, and δ denotes the cheating probability. We next present an optimum scheme, which meets the equality of our bound, by using "difference sets." A partial converse and some extensions are also shown.

Key words. cryptography, secret sharing schemes, cheaters, balanced incomplete block design, difference sets

AMS subject classifications. 94A62, 94A60, 05B05

DOI. 10.1137/S0895480100378689

1. Introduction.

(k, n) threshold secret sharing schemes [20, 2] have been studied extensively because of their wide applications in fields such as key management and secure computation. In such a scheme, a dealer D distributes a secret s to n participants $P_1, \ldots, P_n$ in such a way that any k or more participants can recover the secret s, but any $k-1$ or fewer participants have no information on s. A piece of information given to P_i is called a share and is denoted by v_i. An important issue in secret sharing schemes is the size of shares. Let $\mathcal{V}_i$ be the set of possible shares for P_i. Let $\mathcal{S}$ be the set of possible secrets. Then it is well known that

$$|\mathcal{V}_i| \geq |\mathcal{S}| \tag{1}$$

in any (k, n) threshold scheme [13].

Tompa and Woll [23] considered the following scenario: Suppose that $k-1$ participants, say $P_1, \ldots, P_{k-1}$, want to cheat a kth participant, P_k, by opening forged shares $v'_1, \ldots, v'_{k-1}$. They succeed if the secret s' that is reconstructed from $v'_1, \ldots, v'_{k-1}$ and v_k is different from the original secret s. Tompa and Woll showed that Shamir's scheme [20] is insecure against this attack in that even a single participant can, with high probability, deceive $k-1$ honest participants. They provided a scheme that is secure against this attack, but $|\mathcal{V}_i|$ in their scheme is much larger than in (1):

$$|\mathcal{V}_i| = \left(\frac{(|\mathcal{S}|-1)(k-1)}{\epsilon} + k \right)^2, \tag{2}$$

*Received by the editors September 25, 2000; accepted for publication (in revised form) September 6, 2005; published electronically February 21, 2006. A preliminary version of this paper was presented at EUROCRYPT '96 and appeared in *Lecture Notes in Computer Science* 1070, 1996, pp. 200–211.
http://www.siam.org/journals/sidma/20-1/37868.html

†Graduate School of Innovation Management, Tokyo Institute of Technology, 2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan (wakaha@mot.titech.ac.jp).

‡Department of Computer and Information Sciences, Ibaraki University, 4-12-1 Nakanarusawa, Hitachi, Ibaraki 316-8511, Japan (kurosawa@cis.ibaraki.ac.jp).

§School of Computer Science, University of Waterloo, Waterloo N2L 3G1, ON, Canada (dstinson @uwaterloo.ca). This author's research was supported by NSERC (Canada) grant 203114-02.

where ϵ denotes the cheating probability. Carpentieri, De Santis, and Vaccaro [5] showed the following lower bound on $|\mathcal{V}_i|$ for this problem:

$$|\mathcal{V}_i| \geq \frac{|\mathcal{S}|}{\epsilon}. \tag{3}$$

There is a gap between the bounds of (2) and (3). In fact, both of them can be improved. Furthermore, in the derivation of (3) it is assumed that $k-1$ cheaters $P_1, \ldots, P_{k-1}$ somehow know the secret s before they cheat P_k. (We call this assumption the *CDV assumption*.)

In this paper, we define a (k, n, δ) *secure secret sharing scheme* without the CDV assumption. Here, δ is the maximum probability that $k-1$ participants (cheaters) succeed in cheating another participant, without knowing s. (We use a different symbol, δ rather than ϵ, to denote the cheating probability, in order to highlight the different assumptions we make.) We stress here that the optimum cheating probability δ is determined by the dealer's secret sharing algorithm and the probability distribution on the secret.

Next, we derive a tight lower bound on $|\mathcal{V}_i|$ by using basic probability arguments:

$$|\mathcal{V}_i| \geq \frac{|\mathcal{S}|-1}{\delta} + 1. \tag{4}$$

We then present an optimal scheme which meets the equality of our bound by using "difference sets." A planar difference set modulo $N = \ell(\ell-1)+1$ is a set of ℓ numbers $B = \{d_0, d_1, \ldots, d_{\ell-1}\} \subseteq \mathbb{Z}_N$ with the property that the $\ell(\ell-1)$ differences $d_i - d_j$ $(d_i \neq d_j)$, when reduced modulo N, are exactly the nonzero elements in $\mathbb{Z}_N$ in some order [15, p. 397]. It is known that there exists a planar difference set if l is a prime power [15, p. 398, Theorem 22]. Our optimal scheme is then characterized as follows. If there exists a planar difference set modulo $N = \ell(\ell-1)+1$ such that N is a prime, then there exists a (k, n) threshold secret sharing scheme with $|\mathcal{S}| = \ell$, $\delta = 1/\ell$, and $n < N$ which meets the equality of our bound (4).

Next, we prove a weak converse of the above characterization. It is known that a difference set is equivalent to a certain symmetric balanced incomplete block design (BIBD) having a certain automorphism. We prove that there exists a symmetric BIBD if there exists a (k, n) threshold secret sharing scheme which meets the bound (4). Therefore, we see that there is a tight connection between the optimal schemes and difference sets (or symmetric BIBDs).

Our optimal scheme can be generalized as follows. Let $(\Gamma, +)$ be an abelian group of order N and let $B = \{d_0, d_1, \ldots, d_{\ell-1}\} \subseteq \Gamma$. Then B is called an (N, ℓ, λ) difference set [1, p. 261] if each nonzero element x of Γ appears exactly λ times as a difference $d_i - d_j$ $(d_i \neq d_j)$. Our generalized scheme is as follows: There exists a (k, n) threshold secret sharing scheme which meets (4) such that $|\mathcal{S}| = \ell$, $\delta = \lambda/\ell$, and $n < N$ if there exists an (N, ℓ, λ) difference set B in $(\mathbb{F}_N, +)$. It is known that there exists an (N, ℓ, λ) difference set B in $(\mathbb{F}_N, +)$ if N is a prime power, $N = 4t-1$, $\ell = 2t-1$, and $\lambda = t-1$ [1, p. 264].

Finally, for the model with CDV assumption, we show a lower bound on $|\mathcal{V}_i|$ that improves (3) by using the same technique we use to derive (4). Our bound for the model with CDV assumption is as follows. If S is uniformly distributed, then

$$|\mathcal{V}_i| \geq \frac{|\mathcal{S}|-1}{\epsilon^2} + 1.$$

Note that we can prove this bound only for uniformly distributed secrets. (Actually, for nonuniformly distributed secrets, we show some counterexamples to this bound.)

1.1. Some historical remarks. Mainly, there are two concerns on secret sharing schemes with malicious players. The first is that each participant should be able to make sure that his/her share was obtained from a legitimate distribution procedure even if a dealer is dishonest. The second one is that participants should be able to make sure that the reconstructed secret is the correct one.

The problem we consider here focuses only on the second aspect. In particular, we assume that the dealer is honest. Under this assumption, a slightly different problem has been studied by other researchers. McEliece and Sarwate [16] showed that in Shamir's (k, n) threshold scheme, any group of $k + 2e$ participants which includes at most e cheaters can always correctly calculate the secret. For additional work on this problem, see [19, 24].

The problem of identifying cheaters has also been studied; see [18, 3, 4, 14]. These schemes, however, require $|\mathcal{V}_i|$ much bigger than the bound given in (4). On the other hand, in this paper, we are interested only in detecting the fact of cheating.

Verifiable secret sharing schemes (VSSs) have been well studied [7, 11, 17, 6, 18, 22, 9, 10]. Although VSS can protect against both concerns mentioned above, they generally require some computational assumption such as the discrete logarithm assumption [11, 17] or they are complicated protocols involving many interactions to distribute a secret [6, 18, 9, 11].

2. Preliminaries.

2.1. Definition of cheating. Throughout this paper, D denotes a probabilistic Turing machine called a dealer, S denotes a random variable distributed over a finite set $\mathcal{S}$, and $s \in \mathcal{S}$ is called a secret. On input $s \in \mathcal{S}$, D outputs $(v_1, \ldots, v_n)$ according to some fixed probability distribution. For $1 \le i \le n$, each participant P_i holds v_i as his share. V_i denotes the random variable induced by v_i. Let $\mathcal{V}_i = \{v_i \mid \Pr[V_i = v_i] > 0\}$. $\mathcal{V}_i$ is the set of possible shares held by P_i.

DEFINITION 2.1. *We say that D is a (k, n) threshold secret sharing scheme for S if the following two requirements hold: For any $\{i_1, \ldots, i_j\} \subseteq \{1, \ldots, n\}$ and $(v_{i_1}, \ldots, v_{i_j})$ such that* $\Pr[V_{i_1} = v_{i_1}, \ldots, V_{i_j} = v_{i_j}] > 0$,

(A1) *if $j \ge k$, then there exists a unique $s \in \mathcal{S}$ such that*

$$\Pr[S = s \mid V_{i_1} = v_{i_1}, \ldots, V_{i_j} = v_{i_j}] = 1;$$

(A2) *if $j < k$ for each $s \in \mathcal{S}$, then*

$$\Pr[S = s \mid V_{i_1} = v_{i_1}, \ldots, V_{i_j} = v_{i_j}] = \Pr[S = s].$$

In the above, a secret sharing scheme is defined based on a given probability distribution S. In contrast, most constructions of secret sharing schemes will be valid for any probability distribution defined on S.

DEFINITION 2.2. *For $v_{i_1} \in \mathcal{V}_{i_1}, \ldots, v_{i_k} \in \mathcal{V}_{i_k}$, define*

$$\mathsf{Sec}(v_{i_1}, \ldots, v_{i_k}) = \begin{cases} s & \text{if } \exists s \in \mathcal{S} \text{ s.t. } \Pr[S = s \mid V_{i_1} = v_{i_1}, \ldots, V_{i_k} = v_{i_k}] = 1, \\ \perp & \text{otherwise.} \end{cases}$$

That is, $\mathsf{Sec}(v_{i_1}, \ldots, v_{i_k})$ denotes the secret reconstructed from the k possible shares $(v_{i_1}, \ldots, v_{i_k})$ associated with $(P_{i_1}, \ldots, P_{i_k})$, respectively. The symbol $\perp$ is used to indicate when no secret can be reconstructed from the k shares. We will often aggregate the first $k-1$ arguments of Sec into a vector, by defining $\mathbf{b} = (v_{i_1}, \ldots, v_{i_{k-1}})$ and $\mathsf{Sec}(\mathbf{b}, v_{i_k}) = \mathsf{Sec}(v_{i_1}, \ldots, v_{i_k})$.

DEFINITION 2.3. *Suppose that $k-1$ cheaters $P_{i_1}, \ldots, P_{i_{k-1}}$ possess the list of shares $\mathbf{b} = (v_{i_1}, \ldots, v_{i_{k-1}})$. Let $\mathbf{b}' = (v'_{i_1}, \ldots, v'_{i_{k-1}}) \neq \mathbf{b}$ be a list of $k-1$ forged shares. Then we say that P_{i_k} is* cheated *by $\mathbf{b}'$ if*

$$\mathsf{Sec}(\mathbf{b}', v_{i_k}) \notin \{\mathsf{Sec}(\mathbf{b}, v_{i_k}), \perp\},$$

where v_{i_k} denotes the share of P_{i_k}.

3. New lower bound on $|\mathcal{V}_i|$.

3.1. Definition of secure secret sharing. In this section we derive a tight lower bound on $|\mathcal{V}_i|$ by using basic probability arguments. In deriving this bound we do not use the CDV assumption. That is, we assume that, according to the definition of a (k, n) threshold secret sharing scheme, $k-1$ cheaters have no information on s.

To define a secure secret sharing scheme clearly, we consider the following game called the "cheating game."

1. $k-1$ cheaters and the target participant are fixed. That is, we fix $i_1, \ldots, i_{k-1}$ and i_k.

2. The dealer picks $s \in \mathcal{S}$ according to distribution S, and uses D to compute shares $v_1, \ldots, v_n$ for the n participants. v_i is given to P_i for $i \in \{1, \ldots, n\}$.

3. Let $\mathbf{b} = (v_{i_1}, \ldots, v_{i_{k-1}})$. The cheaters jointly use a *probabilistic* algorithm A to compute forged shares $\mathbf{b}' = (v'_{i_1}, \ldots, v'_{i_{k-1}})$ from $\mathbf{b}$.

4. The cheaters open the forged shares $\mathbf{b}'$. If P_{i_k} is cheated by $\mathbf{b}'$ (as defined above), then we say that the cheaters win the cheating game.

In order to analyze cheating probabilities, we define some useful notation. First, define

$$\gamma(\mathbf{b}, \mathbf{b}', x) = \begin{cases} 1 & \text{if } \mathsf{Sec}(\mathbf{b}', x) \notin \{\mathsf{Sec}(\mathbf{b}, x), \perp\}, \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

and

$$\gamma(\mathbf{b}, \mathbf{b}') = \sum_x \left(\gamma(\mathbf{b}, \mathbf{b}', x)\ \Pr[V_{i_k} = x \mid \mathbf{b}]\right). \tag{6}$$

The value $\gamma(\mathbf{b}, \mathbf{b}')$ is the probability that the cheaters win if they change $\mathbf{b}$ to $\mathbf{b}'$.

A *cheating strategy* C defines conditional probabilities $\Pr[\mathbf{b}' \mid \mathbf{b}]$ for every $\mathbf{b}$ such that $\Pr[\mathbf{b}] > 0$. The success of the cheating strategy C is computed to be

$$\mathsf{Succ}(\mathsf{C}) = \sum_{\mathbf{b}} \left(\Pr[\mathbf{b}] \sum_{\mathbf{b}'} \left(\Pr[\mathbf{b}' \mid \mathbf{b}]\, \gamma(\mathbf{b}, \mathbf{b}')\right)\right). \tag{7}$$

Note that the probabilities $\Pr[\mathbf{b}]$ are determined by the dealer's secret sharing algorithm, while the probabilities $\Pr[\mathbf{b}' \mid \mathbf{b}]$ are chosen by the cheaters.

For future use, we record some equivalent formulations of $\mathsf{Succ}(\mathsf{C})$. These are obtained by substituting (6) into (7) and interchanging the order of summation:

$$\mathsf{Succ}(\mathsf{C}) = \sum_{\mathbf{b}} \left(\Pr[\mathbf{b}] \sum_{\mathbf{b}'} \left(\Pr[\mathbf{b}' \mid \mathbf{b}] \sum_x \left(\Pr[x \mid \mathbf{b}]\, \gamma(\mathbf{b}, \mathbf{b}', x)\right)\right)\right) \tag{8}$$

$$= \sum_{\mathbf{b}} \left(\Pr[\mathbf{b}] \sum_x \left(\Pr[x \mid \mathbf{b}] \sum_{\mathbf{b}'} \left(\Pr[\mathbf{b}' \mid \mathbf{b}]\, \gamma(\mathbf{b}, \mathbf{b}', x)\right)\right)\right). \tag{9}$$

We define the *maximum average cheating probability* to be the maximum value of $\mathsf{Succ}(\mathsf{C})$ over all cheating strategies C. A (k,n) threshold secret sharing scheme is called a (k,n,δ) *secure secret sharing scheme* if the maximum average cheating probability is at most δ for any $k-1$ cheaters $P_{i_1},\dots,P_{i_{k-1}}$ and any target P_{i_k}.

3.2. New lower bound on $|\mathcal{V}_i|$. As before, we fix $k-1$ cheaters $P_{i_1},\dots,P_{i_{k-1}}$ and a target P_{i_k}. Here we consider a simple cheating strategy in order to prove a lower bound on the number of possible shares. Let $\mathbf{b}=(v_{i_1},\dots,v_{i_{k-1}})$ be the shares held by the cheaters. We consider a strategy C_0 where P_{i_1} opens a forged share $v'_{i_1}\neq v_{i_1}$, and the other cheaters $P_{i_2},\dots,P_{i_{k-1}}$ open their shares $v_{i_2},\dots,v_{i_{k-1}}$ honestly. Suppose that P_{i_1} chooses $v'_{i_1}\neq v_{i_1}$ uniformly at random. More precisely,

$$\Pr[\mathbf{b}'=(v'_{i_1},v_{i_2},\dots,v_{i_{k-1}})\mid\mathbf{b}]=\begin{cases}\frac{1}{|\mathcal{V}_{i_1}|-1} & \text{if } v'_{i_1}\neq v_{i_1},\\ 0 & \text{if } v'_{i_1}=v_{i_1}.\end{cases}$$

Furthermore,

$$\Pr[\mathbf{b}'=(v'_{i_1},v'_{i_2},\dots,v'_{i_{k-1}})\mid\mathbf{b}]=0 \quad\text{if}\quad (v'_{i_2},\dots,v'_{i_{k-1}})\neq(v_{i_2},\dots,v_{i_{k-1}}).$$

LEMMA 3.1. *Suppose that* $\Pr[\mathbf{b}]>0$ *and* $\Pr[x\mid\mathbf{b}]>0$. *Then it holds that*

$$\sum_{\mathbf{b}'}(\Pr[\mathbf{b}'\mid\mathbf{b}]\,\gamma(\mathbf{b},\mathbf{b}',x))\geq\frac{|\mathcal{S}|-1}{|\mathcal{V}_{i_1}|-1}. \tag{10}$$

Proof. Let $\mathbf{b}=(v_{i_1},v_{i_2},\dots,v_{i_{k-1}})$ and let $s=\mathsf{Sec}(\mathbf{b},x)$. Observe that $s\in\mathcal{S}$ because $(\mathbf{b},x)$ is a distribution of shares to k participants that occurs with positive probability.

Now, for every $s'\in\mathcal{S}$, $s'\neq s$, there exists at least one possible share for P_{i_1}, namely $v_{s'}\in\mathcal{V}_{i_1}$, such that $\mathsf{Sec}(v_{s'},v_{i_2},\dots,v_{i_{k-1}},x)=s'$. This is because the $k-1$ shares $v_{i_2},\dots,v_{i_{k-1}},x$ yield no information on the value of the secret.

Therefore the $|\mathcal{S}|-1$ vectors $(v_{s'},v_{i_2},\dots,v_{i_{k-1}})$ ($s'\in\mathcal{S}$, $s'\neq s$) are such that

$$\gamma(\mathbf{b},(v_{s'},v_{i_2},\dots,v_{i_{k-1}}),x)=1.$$

There are $|\mathcal{V}_{i_1}|-1$ possible vectors $\mathbf{b}'$ considered in the given strategy, each of which is chosen with probability $1/(|\mathcal{V}_{i_1}|-1)$. Therefore the desired result follows. ∎

THEOREM 3.2. *The cheating strategy* C_0 *(described above) has*

$$\mathsf{Succ}(\mathsf{C}_0)\geq\frac{|\mathcal{S}|-1}{|\mathcal{V}_{i_1}|-1}$$

for any i_1.

Proof. We use (9) and Lemma 3.1:

$$\begin{aligned}\mathsf{Succ}(\mathsf{C}_0)&=\sum_{\mathbf{b}}\left(\Pr[\mathbf{b}]\sum_x\left(\Pr[x\mid\mathbf{b}]\sum_{\mathbf{b}'}(\Pr[\mathbf{b}'\mid\mathbf{b}]\,\gamma(\mathbf{b},\mathbf{b}',x))\right)\right)\\&\geq\sum_{\mathbf{b}}\left(\Pr[\mathbf{b}]\sum_x\left(\Pr[x\mid\mathbf{b}]\times\frac{|\mathcal{S}|-1}{|\mathcal{V}_{i_1}|-1}\right)\right)\\&=\frac{|\mathcal{S}|-1}{|\mathcal{V}_{i_1}|-1}.\end{aligned}$$

∎

Observe that the above bound holds for any distribution on S. Now our lower bound on $|\mathcal{V}_i|$ is an immediate consequence of Theorem 3.2.

COROLLARY 3.3. *In a (k, n, δ) secure secret sharing scheme,*

$$|\mathcal{V}_i| \geq \frac{|\mathcal{S}| - 1}{\delta} + 1 \tag{11}$$

for any i.

3.3. Generalization. Our bound (11) holds for general secret sharing schemes with monotone access structures. Let $\mathcal{P} = \{P_1, \ldots, P_n\}$. Let Γ be a collection of subsets of $\mathcal{P}$ that is *monotone*: $A \in \Gamma$, and $B \subseteq A$ implies that $B \in \Gamma$. We say that a secret sharing scheme has *access structure* Γ if A can determine the secret s for all $A \in \Gamma$ and A has no information on s for all $A \subseteq \mathcal{P}$, $A \notin \Gamma$. It is known that there exists such a secret sharing scheme for any monotone access structure; see [12].

Our definition of (k, n, δ) secure secret sharing schemes can be naturally generalized to secret sharing schemes with monotone access structures. We call such a scheme a (Γ, δ) *secret sharing scheme.*

Now suppose that there exists a (Γ, δ) secret sharing scheme such that $\min\{|A| : A \in \Gamma\} \geq 2$. It is easy to see that this implies that there exists a $(2, 2, \delta)$ secure secret sharing scheme. Moreover, for any $P_i \in \mathcal{P}$, there is a $(2, 2, \delta)$ secure secret sharing scheme in which P_i is a participant. Hence (11) holds for any $P_i \in \mathcal{P}$.

4. Optimum (k, n, δ) secure scheme. In this section, we show an optimum scheme which meets the equality of Corollary 3.3 by using "difference sets."

4.1. Difference set.

DEFINITION 4.1 (see [15, p. 397]). *A* planar difference set *modulo $N = \ell(\ell-1)+1$ is a set $B = \{d_0, d_1, \ldots, d_{\ell-1}\} \subseteq \mathbb{Z}_N$ with the property that the $\ell(\ell-1)$ differences $d_i - d_j$ $(d_i \neq d_j)$, when reduced modulo N, are exactly the numbers $1, 2, \ldots, N-1$ in some order.*

Example 4.1 (see [15, p. 398]). $\{d_0 = 0, d_1 = 1, d_2 = 3\}$ is a planar difference set modulo 7 with $\ell = 3$. Indeed, the differences modulo 7 are

$$1-0=1, \quad 3-0=3, \quad 3-1=2, \quad 0-1=6, \quad 0-3=4, \quad 1-3=5.$$

PROPOSITION 4.1 (see [15, Theorem 22, p. 398]). *Let Π be a projective plane $PG(2, q)$. A point in Π can be represented as $(\beta_1, \beta_2, \beta_3) \in (\mathbb{F}_q)^3$, or $\alpha^i \in \mathbb{F}_{q^3}$ for some i, where α is a generator of $\mathbb{F}_{q^3}$. If $\ell = q+1$ points $\alpha^{d_0}, \ldots, \alpha^{d_{\ell-1}}$ are the points on a line in Π, then $\{d_0, \ldots, d_{\ell-1}\}$ is a planar difference set modulo $q^2 + q + 1$.*

Definition 4.1 is generalized as follows.

DEFINITION 4.2 (see [1, p. 261]). *Let $(\Gamma, +)$ be an abelian group of order N. B is called an (N, ℓ, λ) difference set if it satisfies the following:*

(i) *$B \subset \Gamma$ and $|B| = \ell$.*

(ii) *The list of differences $d - d' \neq 0$, where $d, d' \in B$, contains each nonzero element of Γ precisely λ times.*

PROPOSITION 4.2 (see [1, p. 264]). *Suppose $N \equiv 3 \pmod 4$ is a prime power. Then there exists an (N, ℓ, λ) difference set B in $(\mathbb{F}_N, +)$ such that $N = 4t - 1, \ell = 2t - 1$, and $\lambda = t - 1$, where t is a positive integer.*

Example 4.2 (see [1, p. 262]). *$B = \{1, 3, 4, 5, 9\}$ is an $(11, 5, 2)$ difference set in $(\mathbb{F}_{11}, +)$.*

4.2. Optimum scheme based on planar difference set. Corollary 3.3 proves that we should take $|\mathcal{V}_i|$ much larger than $|\mathcal{S}|$ in order to have a secure scheme. The simplest idea of constructing a secure scheme is to use Shamir's scheme with a random polynomial $f(x)$ over $\mathbb{F}_q$, where $q \geq (|\mathcal{S}|-1)/\delta + 1$. Suppose that the value $f(0)$ reconstructed in the reconstruction phase is accepted as a secret if and only if $f(0) \in \{0, 1, \ldots, |\mathcal{S}|-1\}$ (then $s = f(0)$).

However, for this scheme, the probability of successful cheating can be larger than $\delta = (|\mathcal{S}|-1)/(q-1)$. For example, consider a Shamir $(2,2)$ threshold scheme with a polynomial $f(x)$ over $\mathbb{Z}_7$, and let S be a uniformly distributed secret over $\{0,1,2\}$. Then $(|\mathcal{S}|-1)/(q-1) = 1/3$.

Suppose that P_2 opens $v_2' = v_2 + 1$. Then P_1 is cheated with probability $2/3$, since the secret $s = f(0)$ is reconstructed using the formula $s = 2v_1 - v_2 \bmod 7$. In fact, this cheating strategy is the optimal one for P_1. Therefore, this scheme is only a $(2, 2, 2/3)$ secure secret sharing scheme.

This example suggests that we should not assign valid secret values to continuous values in a larger domain, such as $\{0,1,2\}$ in $\mathbb{Z}_7$.

In this subsection, we show that if there exists a planar difference set modulo $N = \ell(\ell-1)+1$ such that N is a prime, then there exists a (k, n, δ) secure secret sharing scheme with $|\mathcal{S}| = \ell$, $\delta = 1/\ell$, and $n < N$ which meets the bound proven in Corollary 3.3.

Let $B = \{d_0, \ldots, d_{\ell-1}\}$ be a planar difference set modulo $N = \ell(\ell-1)+1$ such that N is a prime. We construct a (k, n, δ) secure secret sharing scheme for S, assuming a uniformly distributed secret over $\mathcal{S} = B$. In what follows, we assume that all operations are done over $\mathbb{Z}_N$.

Distribution phase. For a secret $d_s \in \mathcal{S}$ $(= B)$, the dealer D chooses a random polynomial $f(x)$ of degree at most $k-1$ over $\mathbb{Z}_N$ such that $f(0) = d_s$. The share for P_i is given as $v_i = f(i)$. Note that

$$|\mathcal{V}_i| = N = \ell(\ell-1)+1 \tag{12}$$

for all i.

Reconstruction phase. Suppose that $P_{i_1}, \ldots, P_{i_k}$ open (correct or faulty) shares $v_{i_1}', \ldots, v_{i_k}'$. Each participant can compute

$$d_s' = \mathsf{Rec}(v_{i_1}', \ldots, v_{i_k}') = \sum_{j=1}^{k} c_j v_{i_j}',$$

where

$$c_j = \prod_{l \neq j} \frac{-i_l}{i_j - i_l},$$

$1 \leq j \leq k$. If $d_s' \in B$, then $\mathsf{Sec}(v_{i_1}', \ldots, v_{i_k}') = d_s'$; otherwise, $\mathsf{Sec}(v_{i_1}', \ldots, v_{i_k}') = \perp$.

Note that, for any k honest shares $v_{i_1} = f(i_1), \ldots, v_{i_k} = f(i_k)$, we have that

$$\mathsf{Rec}(v_{i_1}, \ldots, v_{i_k}) = \mathsf{Sec}(v_{i_1}, \ldots, v_{i_k}) = \sum_{j=1}^{k} c_j v_{i_j}, \tag{13}$$

which follows from the Lagrange interpolation formula (see [21, p. 331]).

LEMMA 4.3. *The proposed scheme is a (k, n, δ) secure secret sharing scheme for a uniform distribution over $\mathcal{S}$ with $|\mathcal{S}| = \ell$, $\delta = 1/\ell$, and $n < N$.*

Proof. First of all, it is obvious that the proposed scheme is a (k, n) threshold secret sharing scheme, since it is basically Shamir's scheme where the domain of the secrets is not the entire field. It is also clear that $|\mathcal{S}| = \ell$ and $n < N$.

Next, we prove it is secure; that is, the success probability of $k-1$ cheaters is at most $1/\ell = 1/|\mathcal{S}|$. Suppose that cheaters $P_{i_1}, \ldots, P_{i_{k-1}}$ have shares $\mathbf{b} = (v_{i_1}, \ldots, v_{i_{k-1}})$. Let the share of P_{i_k} be denoted by x. Then, from (13), we have

$$\mathsf{Rec}(\mathbf{b}, x) = \sum_{j=1}^{k-1} c_j v_{i_j} + c_k x. \tag{14}$$

Note that $\mathsf{Rec}(\mathbf{b}, x) \in B$. Now define

$$T_{\mathbf{b}} = \{x' : \mathsf{Rec}(\mathbf{b}, x') \in B\}.$$

Then $T_{\mathbf{b}}$ is the set of all possible shares held by P_{i_k}, given that the $k-1$ cheaters hold the shares in $\mathbf{b}$. Since the secret is chosen uniformly at random, it follows that

$$\Pr[x' \mid \mathbf{b}] = \frac{1}{\ell}$$

for all $x' \in T_{\mathbf{b}}$.

For any $k-1$ tuple $\mathbf{b}' = (v'_{i_1}, \ldots, v'_{i_{k-1}})$, define

$$C(\mathbf{b}') = \sum_{j=1}^{k-1} c_j v'_{i_j}.$$

Now, consider the effect of changing $\mathbf{b}$ to $\mathbf{b}'$. It is not hard to verify the following two facts:

1. If $C(\mathbf{b}) = C(\mathbf{b}')$, then $\mathsf{Rec}(\mathbf{b}, x) = \mathsf{Rec}(\mathbf{b}', x)$. In this case, P_{i_k} is not cheated if $\mathbf{b}'$ is opened.
2. If $C(\mathbf{b}) \neq C(\mathbf{b}')$, then P_{i_k} is cheated by opening $\mathbf{b}'$ if and only if $x \in T_{\mathbf{b}} \cap T_{\mathbf{b}'}$. Moreover, $|T_{\mathbf{b}} \cap T_{\mathbf{b}'}| = 1$ and

$$\Pr[x \in (T_{\mathbf{b}} \cap T_{\mathbf{b}'}) \mid \mathbf{b}] = \frac{1}{\ell}.$$

In the case of fact 2, we have

$$\sum_{x} (\Pr[x \mid \mathbf{b}]\, \gamma(\mathbf{b}, \mathbf{b}', x)) = \frac{1}{\ell}.$$

For every $\mathbf{b}$ with $\Pr[\mathbf{b}] > 0$, an optimal strategy is to choose a $\mathbf{b}'$ such that $C(\mathbf{b}) \neq C(\mathbf{b}')$. The success of this strategy can be computed, using (8), to be $1/\ell$. Thus this scheme is a (k, n, δ) secure scheme such that $\delta = 1/\ell$.

It is clear that $|\mathcal{S}| = |B| = \ell$. □

Now the following theorem is obtained from Lemma 4.3.

THEOREM 4.4. *If there exists a planar difference set modulo $N = \ell(\ell-1)+1$ such that N is a prime, then there exists a (k, n, δ) secure secret sharing scheme for a uniformly distributed secret over $\mathcal{S}$ which meets the bound* (11)*, such that $|\mathcal{S}| = \ell$, $\delta = 1/\ell$, and $n < N$.*

Proof. Finally, from (12), $|\mathcal{V}_j| = N = (\ell-1)\ell+1 = (|\mathcal{S}|-1)/\delta+1$ for all j. Hence, this scheme meets the bound (11). □

From Proposition 4.1, we obtain the following corollary.

COROLLARY 4.5. *Let q be a prime power such that q^2+q+1 is a prime. Then, there exists a (k,n,δ) secure secret sharing scheme for a uniform distribution over $\mathcal{S}$ which meets the bound* (11) *such that $|\mathcal{S}| = q+1$, $\delta = 1/(q+1)$, and $n < q^2+q+1$.*

Remark 4.1. The above theorem holds only if the secret is uniformly distributed. If S is not a uniform distribution, it is easier for cheaters to guess the share of an honest participant and to succeed in cheating him. For example, consider the following situation: $\mathcal{S} = \{0,1\}$, $\Pr[S=0] = 2/3, \Pr[S=1] = 1/3$, and $k=2$. Since $|\mathcal{S}| = 2$, we can use a planar difference set $B = \{0,1\}$ modulo $N = 3$. Assume that P_1 tries to cheat P_2. The best strategy of P_1, given his share v_1, is to find v_2' such that $\mathsf{Rec}(v_1, v_2') = 0$ and to find v_1' such that $\mathsf{Rec}(v_1', v_2') = 1$. If P_2 has v_2' as her share, she is cheated by v_1'. Further, P_2 has v_2' with probability $2/3$. Therefore, this strategy succeeds with probability $2/3$.

Remark 4.2. Instead of publicizing a $(q^2+q+1, q+1, 1)$ difference set B, it is enough to publicize two points α^0 and α^1 of $PG(2, |\mathcal{S}|-1)$. According to Proposition 4.1, B can be obtained from (α^0, α^1).

4.3. Optimum scheme based on an (N, ℓ, λ) difference set. Theorem 4.4 is generalized as follows.

THEOREM 4.6. *If there exists an (N,ℓ,λ) difference set B in $(\mathbb{F}_N, +)$, then there exists a (k,n,δ) secure secret sharing scheme which meets the equality of our bound* (11) *such that $|\mathcal{S}| = \ell$, $\delta = \lambda/\ell$, and $n < N$.*

Proof. We use the same argument as in the proof of Lemma 4.3. In this case, however, $|T_{\mathbf{b}} \cap T_{\mathbf{b}'}| = \lambda$. Therefore we use $\delta = \lambda/\ell$ instead of $1/\ell$. It is clear that $|\mathcal{V}_i| = N = (\ell-1)\ell/\lambda + 1 = (|\mathcal{S}|-1)/\delta + 1$. □

The following corollary is obtained from Proposition 4.2.

COROLLARY 4.7. *For a positive integer t such that $4t-1$ is a prime power, there exists a (k,n,δ) secure secret sharing scheme which meets the equality of our bound* (11)*, such that $|\mathcal{S}| = 2t-1$, $\delta = (t-1)/(2t-1)$, and $n < 4t-1$.*

5. Symmetric BIBD and secret sharing secure against cheaters. Theorem 4.4 shows that, if there exists a certain planar difference set, then there exists an optimal (k,n,δ) secure scheme. In this section, we prove a weak converse which shows that, if there exists an optimal (k,n,δ) secure scheme, then there exists a certain symmetric BIBD. (Note that a difference set is equivalent to a symmetric BIBD having a certain automorphism—see Proposition 5.1.)

DEFINITION 5.1 (see [8, p. 3]). *A* balanced incomplete block design *(BIBD, for short) is a pair $(V, \mathcal{B})$ where V is an N-set and $\mathcal{B}$ is a collection of b ℓ-subsets of V (blocks) such that every 2-subset of V is contained in exactly λ blocks. If $N = b$, the BIBD is called a* symmetric *BIBD.*

PROPOSITION 5.1 (see [8, p. 298]). *The existence of an (N,ℓ,λ) difference set over an abelian group G is equivalent to the existence of a symmetric BIBD(N,ℓ,λ) admitting G as a point regular automorphism group; i.e., for any two points p and q, there is a unique group element g which maps p to q.*

We now proceed to develop the tools needed for our proof.

LEMMA 5.2. *Suppose that equality in* (11) *holds for any i. Then for any $\mathbf{b} = (v_{i_1}, \ldots, v_{i_{k-1}})$ where $v_{i_1} \in \mathcal{V}_{i_1}, \ldots, v_{i_{k-1}} \in \mathcal{V}_{i_{k-1}}$, such that $\Pr[\mathbf{b}] > 0$, it holds that*

$$|\{v_{i_k} \in \mathcal{V}_{i_k} : \mathsf{Sec}(\mathbf{b}, v_{i_k}) \in \mathcal{S}\}| = |\mathcal{S}|.$$

Proof. This easily follows from the proof of Lemma 3.1. □

THEOREM 5.3. *Suppose that there exists a (k, n, δ) secure secret sharing scheme such that equality holds in* (11) *for uniformly distributed secrets. Define*

$$N = |\mathcal{V}_{i_1}| = \frac{|\mathcal{S}| - 1}{\delta} + 1$$

and

$$\lambda = \frac{|\mathcal{S}|(|\mathcal{S}| - 1)}{N - 1}.$$

Then there exists a symmetric BIBD$(N, |\mathcal{S}|, \lambda)$.

Proof. Let $\mathcal{V}_{i_1} = \mathcal{V} = \{1, 2, \ldots, N\}$ and let $\mathcal{S} = \{1, \ldots, \ell\}$. Fix k participants $P_{i_1}, \ldots, P_{i_k}$ and choose any list of $k-2$ shares $v'_{i_2}, \ldots, v'_{i_{k-1}}$ such that $\Pr[(v'_{i_2}, \ldots, v'_{i_{k-1}})] > 0$. Now define an $N \times N$ matrix $D = (d_{i,j})$ such that

$$d_{i,j} = \begin{cases} 1 & \text{if } \mathsf{Sec}(i, v'_{i_2}, \ldots, v'_{i_{k-1}}, j) \in \mathcal{S}, \\ 0 & \text{otherwise.} \end{cases}$$

We will show that D is an incidence matrix of a symmetric BIBD(N, ℓ, λ) (where $d_{i,j} = 1$ if and only if the ith point is in the jth block). Note that the Hamming weight of a column in D corresponds to the cardinality of a block, a row in D corresponds to a point, and the inner product of two rows is the number of blocks that contain the corresponding 2-subset. Thus we have to show that (a) any column of D has Hamming weight equal to ℓ and (b) any two rows of D have inner product equal to λ.

From Lemma 5.2, it is easy to see that any row or column of D has Hamming weight equal to 0 or ℓ. If we delete the rows and columns having Hamming weight 0, then we obtain a submatrix of D, denoted D', in which every row and column has Hamming weight ℓ.

Suppose that D' has dimensions N' by N''. The total number of 1's in D' is $\ell N' = \ell N''$, which implies that $N' = N''$. We will eventually show that $N' = N$ and hence $D' = D$. Assume, by relabelling if necessary, that D' consists of the first N' rows and columns of D and let $d'_{i,j}$ denote the entry of D' in row i and column j for all i and j.

We will now show that any two rows of D' have a constant inner product equal to λ. Pick a row r of D'. Let $c_1, \ldots, c_\ell$ be the columns such that

$$d'_{r,c_1} = \cdots = d'_{r,c_\ell} = 1.$$

Let E_r be the $(N' - 1) \times \ell$ submatrix of D' formed by deleting row r and deleting all the columns except the columns $c_1, \ldots, c_\ell$. Every column of E_r has Hamming weight $\ell - 1$, so the total number of 1's in E_r is $\ell(\ell - 1)$. Let s_r be the row of E_r having the largest Hamming weight, and denote the Hamming weight of this row by w_r. Then

$$w_r \geq \frac{\ell(\ell - 1)}{N' - 1}.$$

For each value of r $(1 \leq r \leq N')$, we can define E_r, s_r, and w_r as described above. Now we define a certain cheating strategy C as follows:

1. Given a list of shares $\mathbf{b} = (r, v'_{remunerate}, \ldots, v'_{i_{k-1}})$ for $P_{i_1}, \ldots, P_{i_{k-1}}$, respectively, open the list of shares $(s_r, v'_{i_2}, \ldots, v'_{i_{k-1}})$. That is, for each of these N' values of r (where r is the share for P_{i_1}), we change r to the false value s_r.

2. Given a list of shares $\mathbf{b} = (v_{i_1}, v_{i_2}, \ldots, v_{i_{k-1}})$ other than the ones considered in case 1 (i.e., a list of shares such that $(v_{i_2}, \ldots, v_{i_{k-1}}) \neq (v'_{i_2}, \ldots, v'_{i_{k-1}})$), use the strategy C_0 described in section 3.2.

We refer to (8) and (9) to compute a bound on $\mathsf{Succ}(\mathsf{C})$. First, suppose we are in case 1. Then C defines a unique $\mathbf{b}'$ such that $\Pr[\mathbf{b}'|\mathbf{b}] = 1$. For this $\mathbf{b}'$ we have

$$|\{x : \gamma(\mathbf{b}, \mathbf{b}', x) = 1\}| = w_r.$$

Since the secrets are equiprobable, it follows that

$$\sum_x (\Pr[x \mid \mathbf{b}]\, \gamma(\mathbf{b}, \mathbf{b}', x)) = \frac{w_r}{\ell} \geq \frac{\ell - 1}{N' - 1}.$$

Hence,

$$\sum_{\mathbf{b}'} \left(\Pr[\mathbf{b}' \mid \mathbf{b}] \sum_x (\Pr[x \mid \mathbf{b}]\, \gamma(\mathbf{b}, \mathbf{b}', x)) \right) \geq \frac{\ell - 1}{N' - 1} \geq \frac{\ell - 1}{N - 1}.$$

In case 2, we apply Lemma 3.1, obtaining the following bound:

$$\sum_x \left(\Pr[x \mid \mathbf{b}] \sum_{\mathbf{b}'} (\Pr[\mathbf{b}' \mid \mathbf{b}]\, \gamma(\mathbf{b}, \mathbf{b}', x)) \right) \geq \frac{\ell - 1}{N - 1}.$$

Now, combining (8) and (9), we have

$$\mathsf{Succ}(\mathsf{C}) \geq \sum_{\mathbf{b}} \left(\Pr[\mathbf{b}] \times \frac{\ell - 1}{N - 1} \right) = \frac{\ell - 1}{N - 1}.$$

However, from Corollary 3.3, the optimal cheating strategy succeeds with probability at most $(\ell - 1)/(N - 1)$. Therefore, we can conclude that $N' = N$ and $w_r = \ell(\ell - 1)/(N - 1)$. Hence, every row of E_r has Hamming weight equal to

$$\frac{\ell(\ell - 1)}{N - 1} = \lambda.$$

This means that the inner product of row r and any other row of D' is equal to λ. Since r is an arbitrary row of D', any two distinct rows of D' have inner product equal to λ. This completes the proof. □

6. Tighter bounds under the CDV assumption. In this section, we consider a model where the CDV assumption holds (that is, $k - 1$ cheaters $P_{i_1}, \ldots, P_{i_{k-1}}$ somehow know the value of the secret s). For this model, we show a lower bound on $|\mathcal{V}_i|$ that is stronger than (3).

Before proving our new bound, we reformulate the cheating model under the CDV assumption. We introduce the "CDV cheating game," which is slightly different from the cheating game presented in section 3.1. The boxed text indicates where the model differs from the previous model.

1. $k-1$ cheaters and the target participant are fixed. That is, we fix $i_1, \ldots, i_{k-1}$ and i_k.

2. The dealer picks $s \in \mathcal{S}$ according to distribution S, and uses D to compute shares $v_1, \ldots, v_n$ for the n participants. v_i is given to P_i for $i \in \{1, \ldots, n\}$. Also, s is given to $P_{i_1}, \ldots, P_{i_{k-1}}$.

3. Let $\mathbf{b} = (v_{i_1}, \ldots, v_{i_{k-1}})$. The cheaters use a probabilistic algorithm A to compute forged shares $\mathbf{b}' = (v'_{i_1}, \ldots, v'_{i_{k-1}})$ from $\mathbf{b}$ and s.

4. The cheaters open the forged shares $\mathbf{b}'$. If P_{i_k} is cheated by $\mathbf{b}'$, then we say that the cheaters win the CDV cheating game.

We now modify our previously defined notation which we use to analyze cheating probabilities. First, define

$$\gamma_{\mathrm{cdv}}(\mathbf{b}, s, \mathbf{b}', x) = \begin{cases} 1 & \text{if } \mathsf{Sec}(\mathbf{b}, x) = s \text{ and } \mathsf{Sec}(\mathbf{b}', x) \notin \{s, \perp\}, \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

and

$$\gamma_{\mathrm{cdv}}(\mathbf{b}, s, \mathbf{b}') = \sum_x \left(\gamma_{\mathrm{cdv}}(\mathbf{b}, s, \mathbf{b}', x)\ \Pr[V_{i_k} = x \mid \mathbf{b}, s]\right). \tag{16}$$

The value $\gamma_{\mathrm{cdv}}(\mathbf{b}, s, \mathbf{b}')$ is the probability that the cheaters win if they change $\mathbf{b}$ to $\mathbf{b}'$ when the secret is s.

A *cheating strategy* C defines conditional probabilities $\Pr[\mathbf{b}' \mid \mathbf{b}, s]$ for every $(\mathbf{b}, s)$ such that $\Pr[\mathbf{b}, s] > 0$. The success of the cheating strategy C is computed to be

$$\mathsf{Succ}_{\mathrm{cdv}}(\mathsf{C}) = \sum_{\mathbf{b},s} \left(\Pr[\mathbf{b}, s] \sum_{\mathbf{b}'} \left(\Pr[\mathbf{b}' \mid \mathbf{b}, s]\, \gamma_{\mathrm{cdv}}(\mathbf{b}, s, \mathbf{b}')\right)\right). \tag{17}$$

The probabilities $\Pr[\mathbf{b}, s]$ are determined by the dealer's secret sharing algorithm, while the probabilities $\Pr[\mathbf{b}' \mid \mathbf{b}, s]$ are chosen by the cheaters.

Here are some equivalent formulations of $\mathsf{Succ}_{\mathrm{cdv}}(\mathsf{C})$:

$$\begin{aligned} \mathsf{Succ}_{\mathrm{cdv}}(\mathsf{C}) &= \sum_{\mathbf{b},s} \left(\Pr[\mathbf{b}, s] \sum_{\mathbf{b}'} \left(\Pr[\mathbf{b}' \mid \mathbf{b}, s] \sum_x \left(\Pr[x \mid \mathbf{b}, s]\, \gamma_{\mathrm{cdv}}(\mathbf{b}, s, \mathbf{b}', x)\right)\right)\right) \\ &= \sum_{\mathbf{b},s} \left(\Pr[\mathbf{b}, s] \sum_x \left(\Pr[x \mid \mathbf{b}, s] \sum_{\mathbf{b}'} \left(\Pr[\mathbf{b}' \mid \mathbf{b}, s]\, \gamma_{\mathrm{cdv}}(\mathbf{b}, s, \mathbf{b}', x)\right)\right)\right). \end{aligned} \tag{18}$$

Suppose we fix $i_1, \ldots, i_{k-1}$, and $\mathbf{b}$ represents the vector of shares given to $P_{i_1}, \ldots, P_{i_{k-1}}$, as usual. For any pair $(\mathbf{b}, s)$, define

$$\hat{\mathcal{V}}_{i_k}(\mathbf{b}, s) = \{v_{i_k} \in \mathcal{V}_{i_k} : \mathsf{Sec}(\mathbf{b}, v_{i_k}) = s\}. \tag{19}$$

Observe that $\hat{\mathcal{V}}_{i_k}(\mathbf{b}, s)$ denotes the set of possible shares held by P_{i_k}, given values for s and the vector $\mathbf{b}$ as discussed above.

We first consider a certain cheaters' strategy which we denote by $\mathsf{C}_{\mathsf{guess}}$. For any $(\mathbf{b}, s)$, choose $\hat{x} \in \hat{\mathcal{V}}_{i_k}(\mathbf{b}, s)$ such that $\Pr[V_{i_k} = \hat{x} \mid \mathbf{b}, s]$ is maximized. Then choose any vector $\mathbf{b}'$ such that $\mathsf{Sec}(\mathbf{b}', \hat{x}) \notin \{s, \perp\}$. Then replace $\mathbf{b}$ by $\mathbf{b}'$. Thus, for any $(\mathbf{b}, s)$, the result of $\mathsf{C}_{\mathsf{guess}}$ is to choose a certain $\mathbf{b}' = \mathbf{b}'(\mathbf{b}, s)$ with probability equal to 1 (i.e., $\mathbf{b}'$ is a function of $(\mathbf{b}, s)$).

It is obvious from (16) that the following equation holds for any $(\mathbf{b}, s)$:

$$\gamma_{\mathrm{cdv}}(\mathbf{b}, s, \mathbf{b}'(\mathbf{b}, s)) \geq \frac{1}{|\hat{\mathcal{V}}_{i_k}(\mathbf{b}, s)|}. \tag{20}$$

Hence, we can compute $\mathsf{Succ}_{\mathrm{cdv}}(\mathsf{C}_{\mathsf{guess}})$ using (17):

$$\mathsf{Succ}_{\mathrm{cdv}}(\mathsf{C}_{\mathsf{guess}}) \geq \sum_{\mathbf{b},s} \left(\Pr[\mathbf{b}, s] \times \frac{1}{|\hat{\mathcal{V}}_{i_k}(\mathbf{b}, s)|} \right). \tag{21}$$

Using the fact that the function $x \mapsto 1/x$ is convex, and applying Jensen's inequality[1], we have the following:

$$\sum_{\mathbf{b},s} \left(\Pr[\mathbf{b}, s] \times \frac{1}{|\hat{\mathcal{V}}_{i_k}(\mathbf{b}, s)|} \right) \geq \frac{1}{\sum_{\mathbf{b},s} \left(\Pr[\mathbf{b}, s] \times |\hat{\mathcal{V}}_{i_k}(\mathbf{b}, s)| \right)}. \tag{22}$$

Now, $\mathsf{Succ}_{\mathrm{cdv}}(\mathsf{C}_{\mathsf{guess}}) \leq \epsilon$. Combining this fact with (21) and (22), the following result is immediate.

LEMMA 6.1. *In a (k, n, ϵ) robust secret sharing scheme under the CDV assumption,*

$$\sum_{\mathbf{b},s} \left(\Pr[\mathbf{b}, s] \times |\hat{\mathcal{V}}_{i_k}(\mathbf{b}, s)| \right) \geq \frac{1}{\epsilon}.$$

We will actually apply Lemma 6.1 in an equivalent form, where we interchange the roles of P_{i_1} and P_{i_k}. Suppose we define $\mathbf{d}$ to be the vector of shares given to $P_{i_2}, \ldots, P_{i_k}$. Then we have the following.

LEMMA 6.2. *In a (k, n, ϵ) robust secret sharing scheme under the CDV assumption,*

$$\sum_{\mathbf{d},s} \left(\Pr[\mathbf{d}, s] \times |\hat{\mathcal{V}}_{i_1}(\mathbf{d}, s)| \right) \geq \frac{1}{\epsilon}.$$

Now we recall the cheating strategy C_0 which we considered in section 3.2:

$$\Pr[\mathbf{b}' = (v'_{i_1}, v_{i_2}, \ldots, v_{i_{k-1}}) \mid \mathbf{b}] = \begin{cases} \frac{1}{|\mathcal{V}_{i_1}|-1} & \text{if } v'_{i_1} \neq v_{i_1}, \\ 0 & \text{if } v'_{i_1} = v_{i_1}. \end{cases}$$

Let x be the share held by P_{i_k} and define $\mathbf{d}$ to be the list of shares held by $P_{i_2}, \ldots, P_{i_k}$. That is,

$$\mathbf{d} = (v_{i_2}, \ldots, v_{i_{k-1}}, x).$$

Then it is clear from Figure 1 that the following equation holds:

$$\sum_{\mathbf{b}'} \left(\Pr[\mathbf{b}' \mid \mathbf{b}, s] \, \gamma_{\mathrm{cdv}}(\mathbf{b}, s, \mathbf{b}', x) \right) = \sum_{s' \neq s} \frac{|\hat{\mathcal{V}}_{i_1}(\mathbf{d}, s')|}{|\mathcal{V}_{i_1}| - 1}. \tag{23}$$

We now consider $\mathsf{Succ}_{\mathrm{cdv}}(\mathsf{C}_0)$. From (18) and (23), this is computed to be

$$\mathsf{Succ}_{\mathrm{cdv}}(\mathsf{C}_0) = \sum_{\mathbf{b},s,x} \left(\Pr[\mathbf{b}, s, x] \sum_{s' \neq s} \frac{|\hat{\mathcal{V}}_{i_1}(\mathbf{d}, s')|}{|\mathcal{V}_{i_1}| - 1} \right). \tag{24}$$

[1] Jensen's inequality is as follows. Suppose f is a continuous strictly convex function on the interval I. Suppose further that $\sum_{i=1}^{n} a_i = 1$ and $a_i > 0$, $1 \leq i \leq n$. Then $\sum_{i=1}^{n} a_i f(x_i) \geq f\left(\sum_{i=1}^{n} a_i x_i\right)$, where $x_i \in I$, $1 \leq i \leq n$.

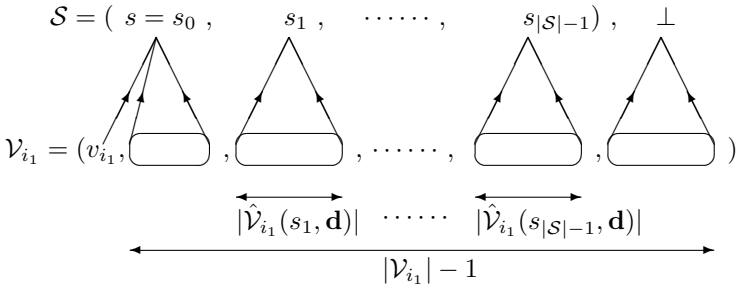


FIG. 1. *Success of strategy* C_0 *(CDV setting).*

If we denote $y = v_{i_1}$, then $(\mathbf{b}, x) = (y, \mathbf{d})$, and (24) becomes

$$\mathsf{Succ}_{\mathrm{cdv}}(\mathsf{C}_0) = \sum_{\mathbf{d},s,y} \left(\Pr[\mathbf{d}, s, y] \sum_{s' \neq s} \frac{|\hat{\mathcal{V}}_{i_1}(\mathbf{d}, s')|}{|\mathcal{V}_{i_1}| - 1} \right). \tag{25}$$

The innermost sum of (25) is independent of y, so (25) can be rewritten as

$$\mathsf{Succ}_{\mathrm{cdv}}(\mathsf{C}_0) = \sum_{\mathbf{d},s} \left(\Pr[\mathbf{d}, s] \sum_{s' \neq s} \frac{|\hat{\mathcal{V}}_{i_1}(\mathbf{d}, s')|}{|\mathcal{V}_{i_1}| - 1} \right). \tag{26}$$

By the defining property of a (k, n) threshold scheme, $\mathbf{d}$ and s are independent, so

$$\Pr[\mathbf{d}, s] = \Pr[\mathbf{d}] \Pr[s].$$

Under the assumption that the secret is uniformly distributed, we have

$$\Pr[s] = \frac{1}{|\mathcal{S}|}$$

for all s. Using these properties, we can recompute (26) as follows:

$$\begin{aligned}
\mathsf{Succ}_{\mathrm{cdv}}(\mathsf{C}_0) &= \sum_{\mathbf{d},s} \left(\Pr[\mathbf{d}, s] \sum_{s' \neq s} \frac{|\hat{\mathcal{V}}_{i_1}(\mathbf{d}, s')|}{|\mathcal{V}_{i_1}| - 1} \right) \\
&= \frac{1}{|\mathcal{S}|} \times \sum_{\mathbf{d}} \left(\Pr[\mathbf{d}] \times \sum_{s} \sum_{s' \neq s} \frac{|\hat{\mathcal{V}}_{i_1}(\mathbf{d}, s')|}{|\mathcal{V}_{i_1}| - 1} \right) \\
&= \frac{|\mathcal{S}| - 1}{|\mathcal{S}|} \times \sum_{\mathbf{d}} \left(\Pr[\mathbf{d}] \times \sum_{s} \frac{|\hat{\mathcal{V}}_{i_1}(\mathbf{d}, s)|}{|\mathcal{V}_{i_1}| - 1} \right) \\
&= \frac{|\mathcal{S}| - 1}{|\mathcal{V}_{i_1}| - 1} \times \sum_{\mathbf{d},s} \left(\Pr[\mathbf{d}, s] \times |\hat{\mathcal{V}}_{i_1}(\mathbf{d}, s)| \right) \\
&\geq \frac{|\mathcal{S}| - 1}{(|\mathcal{V}_{i_1}| - 1)\,\epsilon},
\end{aligned}$$

where the last inequality comes from Lemma 6.2.

TABLE 1
Reconstruction rule.

S		V_1			
		0	1	2	3
V_2	0	0	1	1	1
	1	1	1	1	0
	2	1	1	0	1
	3	1	0	1	1

Finally, using the fact that

$$\mathsf{Succ}_{\mathrm{cdv}}(\mathsf{C}_0) \le \epsilon,$$

we obtain our main theorem.

THEOREM 6.3. *In a (k, n, ϵ) robust secret sharing scheme under the CDV assumption, if S is uniformly distributed, then*

$$|\mathcal{V}_i| \ \ge \ \frac{|\mathcal{S}| - 1}{\epsilon^2} + 1 \tag{27}$$

for any i.

In Theorem 6.3, we cannot remove the condition that the secret is uniformly distributed. We show examples below such that (27) does not hold for a nonuniformly distributed secret.

Example 6.1. Consider a $(2, 2)$ threshold secret sharing scheme constructed as follows: $|\mathcal{S}| = 2$ and $|\mathcal{V}_i| = 4$. The dealer D chooses $v_1 \in \{0, 1, 2, 3\}$ randomly. If $S = 0$, let $v_2 = -v_1 \bmod 4$. If $S = 1$, D chooses v_2 such that $v_2 \neq -v_1 \bmod 4$ randomly. Then D distributes v_1, v_2 to P_1 and P_2. In this scheme, reconstruction is done by using Table 1.

Let the probability distribution of S be $\Pr[S = 0] = 1/4$ and $\Pr[S = 1] = 3/4$. Suppose that P_1 is a cheater. Let's compute the cheating probability of this scheme. We can assume that the cheating algorithm A satisfies that $A(s, v_1) \neq v_1$. First, suppose that $s = 0$. Then P_2 is cheated with probability one:

$$\Pr[P_2 \text{ is cheated by } A(0, v_1) \mid P_1 \text{ has } v_1, S = 0] = 1.$$

Next suppose that $s = 1$. From the table, we obtain

$$\begin{aligned}\Pr[P_2 \text{ is cheated by } 1 \mid P_1 \text{ has } 0, S = 1] &= \Pr[V_2 = 3 \mid P_1 \text{ has } 0, S = 1] \\ &= 1/3.\end{aligned}$$

Similarly, we obtain

$$\Pr[P_2 \text{ is cheated by } v_1' \mid P_1 \text{ has } v_1, S = 1] = 1/3$$

for any $v_1 \in \mathcal{V}_1$ and $v_1' \neq v_1$. Therefore, for any v_1,

$$\begin{aligned}&\Pr[P_2 \text{ is cheated by } A(1, v_1) \mid P_1 \text{ has } v_1, S = 1] \\ &= \sum_{v_1' \neq v_1} \left(\Pr[A(1, v_1) = v_1'] \times \Pr[P_2 \text{ is cheated by } v_1' \mid P_1 \text{ has } v_1, S = 1]\right) \\ &= \sum_{v_1' \neq v_1} \left(\Pr_A[A(1, v_1) = v_1'] \times (1/3)\right) \\ &= 1/3.\end{aligned}$$

Therefore, the cheating probability is

$$1/4 \times 1 + 3/4 \times 1/3 = 1/2.$$

This means that $\epsilon = 1/2$ and therefore (27) does not hold.

Example 6.2. Consider a $(2, 2)$ threshold secret sharing scheme constructed as follows. $|\mathcal{S}| = 2$ and $|\mathcal{V}_i| = 13$. The dealer D chooses $v_1 \in \mathbb{Z}_{13}$ randomly. If $S = 0$, D chooses $r \in \{0, 1\}$ randomly and lets $v_2 = r - v_1 \bmod 13$. If $S = 1$, D chooses $r \in \{2, 4, 6, 8, 10, 12\}$ randomly and lets $v_2 = r - v_1 \bmod 13$. Then D distributes v_1, v_2 to P_1 and P_2. We can prove that this scheme is $\epsilon = 1/4$ robust when $\Pr[S = 0] = 1/4$ and $\Pr[S = 1] = 3/4$. However, (27) does not hold.

REFERENCES

[1] T. Beth, D. Jungnickel, and H. Lenz, *Design Theory*, Cambridge University Press, Cambridge, UK, 1993.

[2] G. R. Blakely, *Safeguarding cryptographic keys*, in Proceedings of the AFIPS 1979 National Computer Conference, 1979, pp. 313–317.

[3] E. F. Brickell and D. R. Stinson, *The detection of cheaters in threshold schemes*, SIAM J. Discrete Math., 4 (1991), pp. 502–510.

[4] M. Carpentieri, *A perfect threshold secret sharing scheme to identify cheaters*, Des., Codes Cryptogr., 5 (1995), pp. 183–187.

[5] M. Carpentieri, A. De Santis, and U. Vaccaro, *Size of shares and probability of cheating in threshold schemes*, in Advances in Cryptography—Eurocrypt '93, Lecture Notes in Comput. Sci. 765, Springer-Verlag, New York, 1994, pp. 118–125.

[6] D. Chaum, C. Crepeau, and I. Damgard, *Multiparty unconditionally secure protocols*, in Proceedings of the 20th Annual ACM Symposium on the Theory of Computing, 1988, pp. 11–19.

[7] B. Chor, S. Goldwasser, S. Micali, and B. Awerbuch, *Verifiable secret sharing and achieving simultaneity in the presence of faults*, in Proceedings of the 26th IEEE Annual Symposium on the Foundations of Computer Science, 1985, pp. 383–395.

[8] C. J. Colbourn and J. H. Dinitz, eds., *The CRC Handbook of Combinatorial Designs*, CRC Press, Boca Raton, FL, 1996.

[9] R. Cramer, I. Damgård, and U. M. Maurer, *General secure multi-party computation from any linear secret-sharing scheme*, in Advances in Cryptography—Eurocrypt '00, Lecture Notes in Comput. Sci. 1807, Springer-Verlag, New York, 2000, pp. 316–334.

[10] Y. Desmedt, K. Kurosawa, and T. V. Le, *Error correcting and complexity aspects of linear secret sharing schemes*, in Information Security: 6th International Conference, ISC 2003, Lecture Notes in Comput. Sci. 2851, Springer-Verlag, New York, 2003, pp. 396–407.

[11] P. Feldman, *A practical scheme for non-interactive verifiable secret sharing*, in Proceedings of the 28th Annual IEEE Symposium on the Foundations of Computer Science, 1987, pp. 427–437.

[12] M. Ito, A. Saito, and T. Nishizeki, *Multiple assignment scheme for sharing secret*, J. Cryptology, 6 (1993), pp. 15–20.

[13] E. D. Karnin, J. W. Greene, and M. E. Hellman, *On secret sharing systems*, IEEE Trans. Inform. Theory, 29 (1982), pp. 35–41.

[14] K. Kurosawa, S. Obana, and W. Ogata, *t-cheater identifiable* (k, n) *threshold secret sharing schemes*, in Advances in Cryptography—Crypto '95, Lecture Notes in Comput. Sci. 963, Springer-Verlag, New York, 1995, pp. 410–423.

[15] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North–Holland, Amsterdam, 1977.

[16] R. J. McEliece and D. V. Sarwate, *On sharing secrets and Reed-Solomon codes*, Comm. ACM, 24 (1981), pp. 583–584.

[17] T. Pedersen, *Non-interactive and information-theoretic secure verifiable secret sharing*, in Advances in Cryptography—Crypto '91, Lecture Notes in Comput. Sci. 576, Springer-Verlag, New York, 1991, pp. 129–149.

[18] T. Rabin and M. Ben-Or, *Verifiable secret sharing and multiparty protocols with honest majority*, in Proceedings of the 21st Annual ACM Symposium on the Theory of Computing, 1989, pp. 73–85.

[19] R. Rees, D. R. Stinson, R. Wei, and G. H. J. van Rees, *An application of covering designs: Determining the maximum consistent set of shares in a threshold scheme*, Ars Combin., 53 (1999), pp. 225–237.
[20] A. Shamir, *How to share a secret*, Comm. ACM, 22 (1979), pp. 612–613.
[21] D. R. Stinson, *Cryptography: Theory and Practice*, CRC Press, Boca Raton, FL, 1995.
[22] D. R. Stinson and R. Wei, *Unconditionally secure proactive secret sharing scheme with combinatorial structures*, in Selected Areas in Cryptography: 6th Annual International Workshop, SAC'99, Lecture Notes in Comput. Sci. 1758, Springer-Verlag, New York, 2000, pp. 200–214
[23] M. Tompa and H. Woll, *How to share a secret with cheaters*, J. Cryptology, 1 (1988), pp. 133–138.
[24] R. Tso, Y. Miao, and E. Okamoto, *A new algorithm for searching a consistent set of shares in a threshold scheme with cheaters*, in Information Security and Cryptology—ICISC 2003, Lecture Notes in Comput. Sci. 2971, Springer-Verlag, New York, 2004, pp. 377–385.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 96–104

DIFFERENTIAL METHODS FOR FINDING INDEPENDENT SETS IN HYPERGRAPHS*

YUSHENG LI† AND WENAN ZANG‡

Abstract. It is shown by using differential methods that if $\mathcal{H}$ is a double linear, r-uniform hypergraph with degree sequence $\{d_v\}$ such that any subhypergraph induced by a neighborhood has maximum degree less than m, then its independence number is at least $\sum_v f_{r,m}(d_v)$, where $f_{r,m}(x)$ is a convex function satisfying $f_{r,m}(x) \sim (\log x)/x$ if $r = 2$ and $c/x^{1/(r-1)}$ if $r \geq 3$, as $x \to \infty$, and $c = c(r,m) > 0$ is a constant. The proof yields a polynomial-time algorithm for finding such an independent set in $\mathcal{H}$.

Key words. hypergraph, independent set, differential method, convex function, algorithm

AMS subject classifications. 05C65, 05C69, 05C85

DOI. 10.1137/S0895480104442571

1. Introduction.

Hypergraphs are systems of sets which are conceived as natural extensions of graphs: elements correspond to vertices and sets correspond to edges which are allowed to connect more than two vertices. Hypergraph theory is a part of the general study of combinatorial properties of families of sets; for in-depth accounts of the subject, see Berge [5] and Duchet [8]. The present paper concerns itself with the independent set problem on hypergraphs.

A *hypergraph* $\mathcal{H} = (V, \mathcal{E})$ consists of a vertex set V and an edge set $\mathcal{E}$ such that each edge is a nonempty subset of V. Throughout this paper we assume that each edge contains at least two vertices. For each vertex v, the *degree* of v, denoted by d_v, is the number of edges containing v, and the *neighborhood* of v, denoted by $N(v)$, is the set of all neighbors of v, where a vertex u is a *neighbor* of (or is *adjacent* to) v if $u \neq v$ and there is an edge that contains both u and v. Let U be a subset of V. Set $\mathcal{E}_U = \{E \in \mathcal{E} : E \subseteq U\}$. The hypergraph $(U, \mathcal{E}_U)$ is called the *subhypergraph of* $\mathcal{H}$ *induced by* U. We say that U is an *independent set* of $\mathcal{H}$ if it contains no edge. The *independence number* of $\mathcal{H}$, denoted by $\alpha(\mathcal{H})$, is the maximum number of vertices in an independent set of $\mathcal{H}$. The *independent set problem* is to find an independent set with the largest size. As is well known, this *NP*-hard problem arises in a rich variety of applications, so it has attracted tremendous research efforts.

Let $G = (V, E)$ be a graph on N vertices with average degree d. A classical theorem of Turán asserts that $\alpha(G) \geq \frac{N}{d+1}$, which was strengthened independently by Caro [6] and Wei [17] as $\alpha(G) \geq \sum_{v \in V} \frac{1}{d_v+1}$ (this bound is better than the former since function $1/(1+x)$ is strictly convex); a nice probabilistic proof of this theorem can be found in Alon and Spencer [4]. In case G is triangle-free, Turán's lower bound can be improved substantially. As shown by Ajtai et al. [1, 2] and Ajtai, Komlós, and Szemerédi [3], $\alpha(G) \geq \frac{cN \log d}{d}$, where (and throughout this paper) $\log x$ stands for the

*Received by the editors March 27, 2004; accepted for publication (in revised form) November 11, 2005; published electronically February 21, 2006.
http://www.siam.org/journals/sidma/20-1/44257.html

†Department of Mathematics, Tongji University, Shanghai 200092, China (li_yusheng@mail.tongji.edu.cn). The work of this author was supported in part by the National Natural Science Foundation of China.

‡Department of Mathematics, University of Hong Kong, Hong Kong, China (wzang@maths.hku.hk). The work of this author was supported in part by the Research Grants Council of Hong Kong.

natural logarithmic function, and constant c can be set equal to $1/2.4$ (cf. Griggs [9]). Shearer [15] confirmed a conjecture of Ajtai, Komlós, and Szemerédi [3] and managed to improve c to $1 - o(1)$ by establishing that $\alpha(G) \geq Ng(d)$, where

$$g(x) = \frac{x \log x - x + 1}{(x-1)^2}; \tag{1}$$

he [16] further improved the bound as $\alpha(G) \geq \sum_v \bar{g}(d_v)$, where the function $\bar{g}(x)$ is asymptotically equal to $g(x)$ as $x \to \infty$. In his proofs, Shearer first introduced the appealing differential methods, which are proved to be very powerful in applications. Shearer's results can be extended [11, 12, 13] as follows: if in a graph G with N vertices and average degree d, any subgraph induced by a neighborhood has no vertex of degree at least m, then $\alpha(G) \geq \sum_v g_m(d_v) \geq Ng_m(d)$, where

$$g_m(x) = \int_0^1 \frac{(1-t)^{1/m}}{m + (x-m)t} dt. \tag{2}$$

(Notice that $g_1(x)$ is exactly Shearer's function $g(x)$ as specified in (1).) This result has interesting applications in Ramsey theory [12, 14]; for instance, it yields $R(m,n) \leq (1+o(1))n^{m-1}/(\log n)^{m-2}$, where Ramsey number $R(m,n)$ is the smallest integer N such that for any graph G of order N, either $\alpha(G) \geq m$ or $\alpha(\bar{G}) \geq n$ holds. It is worthwhile pointing out that since the order of magnitude of $R(3,n)$ is $n^2/\log n$ (see Kim [10]), the above-mentioned lower bound due to Ajtai, Komlós, and Szemerédi [3] cannot be improved more than a constant factor; we believe Shearer's bound is asymptotically sharp on extremal graphs for $R(3,n)$.

The independent set problem on hypergraphs is much more difficult and intractable than that on graphs. So it is natural to restrict our attentions to some special classes of hypergraphs. A hypergraph $\mathcal{H}$ is called *r-uniform* if each edge of $\mathcal{H}$ contains exactly r vertices (so a 2-uniform hypergraph is a graph), and called *triangle-free* if $\mathcal{H}$ contains no three distinct vertices v_1, v_2, v_3 and three distinct edges E_1, E_2, E_3 such that $\{v_1, v_2, v_3\} - \{v_i\}$ is a subset of E_i for $i = 1, 2, 3$. We say that a hypergraph $\mathcal{H}$ is *linear* if any two edges of $\mathcal{H}$ have at most one vertex in common. A linear hypergraph $\mathcal{H}$ is said to be *double linear* if for any two nonadjacent vertices u and v, each edge containing u contains at most one neighbor of v. Caro and Tuza [7] proposed a problem on extending the lower bound of Ajtai, Komlós, and Szemerédi [3] to triangle-free hypergraphs; as a solution to this problem, Zhou and Li [18] proved that every r-uniform linear triangle-free hypergraph $\mathcal{H}$ satisfies $\alpha(\mathcal{H}) \geq Nf_{r,1}(d)$, where function $f_{r,1}(x)$ is much bigger than $(\log x)/x$ when $r \geq 3$. Observe that if a linear hypergraph $\mathcal{H}$ is triangle-free, then its subhypergraph induced by any neighborhood has maximum degree zero. However, the converse need not hold in general. In this paper we consider hypergraphs whose subhypergraphs induced by neighborhoods may have edges.

Let us define some functions before presenting our main result. As usual, let $B(p,q) = \int_0^1 (1-t)^{p-1} t^{q-1} dt$ denote the beta function with $p, q > 0$. For integers $r \geq 2$ and $m \geq 1$, set constants

$$a = \frac{1}{(r-1)^2},\ b = \frac{r-2}{r-1},$$

and

$$B = B(a/m, 1-b) = \int_0^1 (1-t)^{a/m-1} t^{-b} dt.$$

Clearly, $0 < a \le 1$, $0 \le b < 1$, and $B > 0$. For the above r, m and $x \ge 0$, define

$$f_{r,m}(x) = \frac{m}{B} \int_0^1 \frac{(1-t)^{a/m}}{t^b[m + (x-m)t]} dt.$$

Since

$$\frac{(1-t)^{a/m}}{t^b[m + (x-m)t]} \le \frac{(1-t)^{a/m}}{t^b[m(1-t)]} = \frac{1}{m}(1-t)^{a/m-1}t^{-b},$$

we see that $f_{r,m}(x)$ is bounded above by 1 and thus is well defined.

THEOREM. *Let $\mathcal{H} = (V, \mathcal{E})$ be an r-uniform, double linear hypergraph with degree sequence $\{d_v\}$. If the maximum degree of any subhypergraph induced by a neighborhood is less than m, then*

$$\alpha(\mathcal{H}) \ge \sum_{v \in V} f_{r,m}(d_v).$$

Note that if $r = 2$, then $a = 1$, $b = 0$, and $B = m$, so $f_{2,m}$ is the function $g_m(x)$ defined in (2). And $f_{r,1}(x)$ is precisely the function involved in the above Zhou–Li bound. Since any graph and any linear triangle-free hypergraph are double linear, our theorem generalizes all the results cited above, including Turán's theorem and the Caro–Wei theorem as long as graphs in consideration satisfy the conditions.

For any fixed integers $r \ge 3$ and $M \ge 1$, it was shown in [18] that $f_{r,1}(x) \ge (\log^M x)/x$ provided x is large enough. We shall verify that $f_{r,m}(x)$ is a convex function for $x \ge 0$ and that $f_{r,m}(x) \sim (\log x)/x$ if $r = 2$ and $c/x^{1/(r-1)}$ if $r \ge 3$, as $x \to \infty$, where $c = c(r, m) > 0$ is a constant and $\sim$ means an asymptotic equality. By convexity of $f_{r,m}(x)$, we have $f_{r,m}(d) \le \frac{1}{|V|} \sum_{v \in V} f_{r,m}(d_v)$, where $d = \frac{1}{|V|} \sum_{v \in V} d_v$. Thus the following is an immediate consequence of the above theorem.

COROLLARY. *For fixed integers $r \ge 3$ and $m \ge 1$, let $c = c(r, m) > 0$ be the constant as described above. Then for any $\epsilon > 0$, there exists a constant $D = D(r, m, \epsilon)$ such that if a hypergraph $\mathcal{H} = (V, \mathcal{E})$ is double linear, r-uniform, and the subhypergraph induced by any neighborhood has maximum degree less than m, then*

$$\alpha(\mathcal{H}) \ge (1 - \epsilon) \frac{cN}{d^{1/(r-1)}},$$

where $N = |V|$ and d is the average degree of $\mathcal{H}$ with $d \ge D$.

2. Properties of the function $f_{r,m}$. The purpose of this section is to exhibit some properties satisfied by the function $f_{r,m}$ defined in the preceding section.

LEMMA 1. *For fixed integers $r \ge 2$ and $m \ge 1$ and for $x \ge 0$, the function $f(x) = f_{r,m}(x)$ satisfies the differential equation*

$$(r-1)^2 x(x-m) f'(x) + [(r-1)x + 1] f(x) = 1. \tag{3}$$

Moreover, $f(x)$ is strictly and completely monotonic, that is, $(-1)^k f^{(k)}(x) > 0$ for all $x \ge 0$. In particular, $f(x)$ is positive, strictly decreasing, and strictly convex.

Proof. By differentiating x under the integral and then integrating by parts,

we have

$$\begin{aligned}
&x(x-m)f'(x)\\
&= \frac{-mx(x-m)}{B}\int_0^1 \frac{(1-t)^{a/m}t^{1-b}}{[m+(x-m)t]^2}dt\\
&= \frac{mx}{B}\int_0^1 (1-t)^{a/m}t^{1-b}\frac{d}{dt}\left(\frac{1}{m+(x-m)t}\right)dt\\
&= \frac{-mx}{B}\int_0^1 \frac{1}{m+(x-m)t}\left[(1-b)(1-t)^{a/m}t^{-b}-\frac{a}{m}(1-t)^{a/m-1}t^{1-b}\right]dt\\
&= -(1-b)xf(x)+\frac{ax}{B}\int_0^1 \frac{(1-t)^{a/m-1}t^{1-b}}{m+(x-m)t}dt\\
&= -(1-b)xf(x)+\frac{am}{B}\int_0^1 \left(\frac{1}{m(1-t)}-\frac{1}{m+(x-m)t}\right)(1-t)^{a/m}t^{-b}dt\\
&= -(1-b)xf(x)+\frac{a}{B}\int_0^1 (1-t)^{a/m-1}t^{-b}dt-af(x)\\
&= -(1-b)xf(x)+a-af(x)\\
&= -\left(\frac{x}{r-1}+\frac{1}{(r-1)^2}\right)f(x)+\frac{1}{(r-1)^2},
\end{aligned}$$

so the desired differential equation follows. The strict and complete monotonicity of $f(x)$ can be seen by repeatedly differentiating x under the integral. □

Let us now proceed to the asymptotic behavior of the function $f_{2,m}(x)$.

LEMMA 2. *For any fixed integer $m \geq 1$ and for $x > 1$, we have*

$$\frac{\log(x/m)-1}{x} \leq f_{2,m}(x) \leq \frac{x\log x - x + 1}{(x-1)^2}.$$

Therefore $f_{2,m}(x) \sim (\log x)/x$ as $x \to \infty$.

Proof. We first claim that for fixed $x \geq 1$, function

$$f_{2,m}(x) = \int_0^1 \frac{(1-t)^{1/m}dt}{m+(x-m)t} = \int_0^1 \frac{t^{1/m}dt}{mt+x(1-t)}dt$$

decreases as $m \geq 1$ increases. To justify the claim, setting $t = u^m$ gives

$$f_{2,m}(x) = \int_0^1 \frac{mu^m du}{mu^m + x(1-u^m)}.$$

So it suffices to show that if $\delta > 0$ and $0 < u < 1$, then

$$\frac{mu^m}{mu^m + x(1-u^m)} > \frac{(m+\delta)u^{m+\delta}}{(m+\delta)u^{m+\delta} + x(1-u^{m+\delta})}.$$

Equivalently,

$$\delta u^{m+\delta} + m - (m+\delta)u^\delta > 0.$$

For this purpose, set $h(u) = \delta u^{m+\delta} + m - (m+\delta)u^\delta$. Then $h(1) = 0$ and $h'(u) = \delta(m+\delta)u^{\delta-1}(u^m - 1) < 0$ for $0 < u < 1$, and thus the claim follows.

Since for $x > 1$, we have

$$f_{2,1}(x) = \int_0^1 \frac{(1-t)dt}{1+(x-1)t} = \frac{x\log x - x + 1}{(x-1)^2},$$

by the above claim $f_{2,m}(x) \le f_{2,1}(x)$, and so the upper bound is established.

To derive the lower bound, note that

$$\begin{aligned} f_{2,m}(x) &= \int_0^1 \frac{(1-t)^{1/m}dt}{m+(x-m)t} > \int_0^1 \frac{(1-t)dt}{m+(x-m)t} \\ &= \frac{x\log(x/m) - x + m}{(x-m)^2} \ge \frac{\log(x/m) - 1}{x}, \end{aligned}$$

where the last inequality amounts to $(2x-m)\log(x/m) \ge x - m$, or equivalently $(2t-1)\log t \ge t - 1$. Set $\phi(t) = (2t-1)\log t - t + 1$. Then $\phi(1) = 0$ and $\phi'(t) = 2\log t + (1 - 1/t)$, which is less than 0 if $0 < t < 1$, equal to 0 if $t = 1$, and greater than 0 if $t > 1$. Hence $\phi(t) \ge 0$ for $t > 0$, implying the lower bound. □

Our next lemma concerns the case when $r \ge 3$; it shows that the asymptotic behavior of $f_{r,m}$ is dramatically different from that of $f_{2,m}$.

LEMMA 3. *For fixed integers $r \ge 3$ and $m \ge 1$, function $f_{r,m}(x) \sim \frac{c}{x^{1/(r-1)}}$ as $x \to \infty$, where $c = c(r,m) > 0$ is defined to be*

$$\frac{m}{B(m+1)^{a/m}} \int_0^1 \frac{(1-t)^{a/m}}{t^b(m+t)}dt + a\int_{m+1}^{\infty} \frac{dt}{t^{1+a/m}(t-m)^{b-a/m}}.$$

Proof. Our proof relies heavily on the theorem that a linear first-order differential equation

$$\frac{dy}{dx} = p(x)y + q(x)$$

has a unique solution

$$y = e^{\phi(x)}\left(y_0 + \int_{x_0}^{x} q(t)e^{-\phi(t)}dt\right)$$

satisfying $y_0 = y(x_0)$, where $\phi(x) = \int_{x_0}^{x} p(t)dt$. Now let us transform the differential equation (3) in Lemma 1 into the above standard form. Then we get

$$p(x) = -\frac{a((r-1)x+1)}{x(x-m)} \quad \text{and} \quad q(x) = \frac{a}{x(x-m)}.$$

Set $x_0 = m + 1$ and

$$y_0 = f_{r,m}(m+1) = \frac{m}{B}\int_0^1 \frac{(1-t)^{a/m}}{t^b(m+t)}dt.$$

It follows from the uniqueness of the solution that

$$f_{r,m}(x) = e^{\phi(x)}\left(y_0 + \int_{m+1}^{x} q(t)e^{-\phi(t)}dt\right) \quad \text{for } x \ge m+1.$$

Since

$$\begin{aligned}\phi(x) &= -\int_{m+1}^{x} \frac{a((r-1)t+1)}{t(t-m)}dt \\ &= -a\log\left(\left(\frac{m+1}{x}\right)^{1/m}(x-m)^{r-1+1/m}\right),\end{aligned}$$

we obtain

$$e^{\phi(x)} = \frac{x^{a/m}}{(m+1)^{a/m}(x-m)^{1/(r-1)+a/m}} \tag{4}$$

$$\sim \frac{1}{(m+1)^{a/m}x^{1/(r-1)}}, \tag{5}$$

and hence

$$e^{-\phi(x)} \sim (m+1)^{a/m}x^{1/(r-1)}.$$

Thus there exists a constant $M > 0$ such that for all $t \geq m+1$,

$$0 \leq q(t)e^{-\phi(t)} \leq \frac{Mt^{1/(r-1)}}{t^2} = \frac{M}{t^{1+b}}.$$

Recall that $b > 0$ as $r \geq 3$, so $\int_{m+1}^{\infty} q(t)e^{-\phi(t)}dt < \infty$ and

$$\int_{m+1}^{x} q(t)e^{-\phi(t)}dt = \int_{m+1}^{\infty} q(t)e^{-\phi(t)}dt - o(1)$$

as $x \to \infty$. It follows from (5) that

$$\begin{aligned}f_{r,m}(x) &= e^{\phi(x)}\left(y_0 + \int_{m+1}^{\infty} q(t)e^{-\phi(t)}dt - o(1)\right) \\ &\sim \frac{c}{x^{1/(r-1)}},\end{aligned}$$

where $c = \frac{1}{(m+1)^{a/m}}(y_0 + \int_{m+1}^{\infty} q(t)e^{-\phi(t)}dt)$. Using (4) and plugging y_0, we see that c is as defined in the lemma. □

3. Proof of the theorem. Let us introduce some notions before presenting the proof. For each $v \in V$, let $\mathcal{H}_v$ be the subhypergraph of $\mathcal{H}$ induced by $V-(N(v)\cup\{v\})$, and let $\{d'_u\}$ denote the degree sequence of $\mathcal{H}_v$. For simplicity, write $f_{r,m}(x)$ as $f(x)$. Set $S(\mathcal{H}) = \sum_{u\in V(\mathcal{H})} f(d_u)$ and $S(\mathcal{H}_v) = \sum_{u\in V(\mathcal{H}_v)} f(d'_u)$. The default value of $S(\mathcal{H}_v)$ is zero if $V - (N(v)\cup\{v\}) = \emptyset$.

The key step of our proof is to establish the following statement.

LEMMA 4. *There exists a vertex v in $\mathcal{H}$ such that $1 + S(\mathcal{H}_v) \geq S(\mathcal{H})$.*

To show that $\mathcal{H}$ contains an independent set I with size at least $\sum_{v\in V} f(d_v)$, we may apply the following algorithm: Initially set $I = \emptyset$. Let v be the vertex exhibited in Lemma 4. Set $I = I \cup \{v\}$ and $\mathcal{H} = \mathcal{H}_v$. Repeat the process until $\mathcal{H}$ contains no vertex.

So Lemma 4 serves as a criterion for selecting vertices in I. Let us now prove that such an independent set I is indeed as desired.

Proof of the Theorem (assuming Lemma 4). We apply induction on $|V|$, the number of vertices in $\mathcal{H}$. Since $f(0) = 1$ by (3), the assertion holds trivially for $|V| = 1$. So we proceed to the induction step.

Note that $\alpha(\mathcal{H}) \geq 1+\alpha(\mathcal{H}_u)$ for any vertex u of $\mathcal{H}$. Let v be a vertex as described in Lemma 4. Then, by induction hypothesis, we have $\alpha(\mathcal{H}) \geq 1+\alpha(\mathcal{H}_v) \geq 1+S(\mathcal{H}_v) \geq S(\mathcal{H})$, completing the proof.

It therefore remains to prove the above lemma.

Proof of Lemma 4. For each $v \in V$, set

$$N_2(v) = \{x \in V - (N(v) \cup \{v\}) : N(x) \cap N(v) \neq \emptyset\}$$

and

$$\begin{aligned} Y(v) &= 1 + S(\mathcal{H}_v) - S(\mathcal{H}) \\ &= 1 + \sum_{x \in V(\mathcal{H}_v)} [f(d'_x) - f(d_x)] - f(d_v) - \sum_{u \in N(v)} f(d_u). \end{aligned}$$

Besides, for each $x \in N_2(v)$, set $n_{v,x} = |N(v) \cap N(x)|$. Let us consider the terms in $Y(v)$. Since any vertex $x \in V(\mathcal{H}_v) - N_2(v)$ satisfies $d'_x = d_x$ and any vertex $x \in N_2(v)$ satisfies $d'_x = d_x - n_{v,x}$ (for $\mathcal{H}$ is double linear),

$$Y(v) = 1 - f(d_v) - \sum_{u \in N(v)} f(d_u) + \sum_{x \in N_2(v)} [f(d_x - n_{v,x}) - f(d_v)].$$

Clearly, (6) is equivalent to saying that $Y(v) \geq 0$ for some vertex v of $\mathcal{H}$. So to prove the lemma it suffices to show that

$$\sum_{v \in V(\mathcal{H})} Y(v) \geq 0. \tag{6}$$

Since $\mathcal{H}$ is linear and r-uniform,

$$\sum_{v \in V(\mathcal{H})} \sum_{u \in N(v)} f(d_u) = (r-1) \sum_{v \in V(\mathcal{H})} d_v f(d_v).$$

So

$$\begin{aligned} &\sum_{v \in V(\mathcal{H})} Y(v) \\ &= \sum_{v \in V(\mathcal{H})} \{1 - [(r-1)d_v + 1] f(d_v)\} + \sum_{v \in V(\mathcal{H})} \sum_{x \in N_2(v)} [f(d_x - n_{v,x}) - f(d_x)]. \end{aligned}$$

Observe that $x \in N_2(v)$ if and only if $v \in N_2(x)$ and that $n_{v,x} = n_{x,v}$; exchanging the variables in the sum gives

$$\sum_{v \in V(\mathcal{H})} \sum_{x \in N_2(v)} [f(d_x - n_{v,x}) - f(d_x)] = \sum_{v \in V(\mathcal{H})} \sum_{x \in N_2(v)} [f(d_v - n_{v,x}) - f(d_v)].$$

Let

$$Z(v) = \sum_{x \in N_2(v)} [f(d_v - n_{v,x}) - f(d_v)].$$

Then

$$\sum_{v\in V(\mathcal{H})} Y(v) = \sum_{v\in V(\mathcal{H})} \{1-[(r-1)d_v+1]f(d_v)\} + \sum_{v\in V(\mathcal{H})} Z(v). \tag{7}$$

Now comes the technical part of our proof, the analysis of the term $\sum_{v\in V(\mathcal{H})} Z(v)$.

Since $f(x)$ is convex, we have

$$f(x-1)-f(x) \geq f(y-1)-f(y) \text{ whenever } 1\leq x\leq y. \tag{8}$$

(To see this, write $x=\alpha(x-1)+(1-\alpha)y$ and $y-1=\beta(x-1)+(1-\beta)y$, where $0\leq\alpha,\beta\leq 1$. By convexity, $f(x)\leq\alpha f(x-1)+(1-\alpha)f(y)$ and $f(y-1)\leq\beta f(x-1)+(1-\beta)f(y)$. Summing up these two inequalities yields $f(x)+f(y-1)\leq f(x-1)+f(y)$ as $\alpha+\beta=1$.) From (8) we deduce that

$$f(d_v-n_{v,x})-f(d_v)=\sum_{i=1}^{n_{v,x}}[f(d_v-i)-f(d_v-(i-1))]\geq[f(d_v-1)-f(d_v)]n_{v,x},$$

and so

$$Z(v)\geq[f(d_v-1)-f(d_v)]\sum_{x\in N_2(v)} n_{v,x}.$$

Note that $\mathcal{H}$ is double linear, r-uniform, and each vertex $u\in N(v)$ is incident to at most $m-1$ edges in $N(v)$. Moreover, there is precisely one edge in $\mathcal{H}$ containing both u and v. So

$$\sum_{x\in N_2(v)} n_{v,x}\geq(r-1)\sum_{u\in N(v)}(d_u-m).$$

Write $A_v=f(d_v-1)-f(d_v)$. Then

$$\begin{aligned}\sum_{v\in V(\mathcal{H})} Z(v) &\geq (r-1)\sum_{v\in V(\mathcal{H})}\sum_{u\in N(v)}(d_u-m)A_v\\ &=(r-1)\sum_{E\in\mathcal{E}}\sum_{u,v\in E}\{(d_u-m)A_v+(d_v-m)A_u\}\\ &=(r-1)\sum_{E\in\mathcal{E}}\sum_{u,v\in E}\{(d_v-m)A_v+(d_u-m)A_u+(d_u-d_v)(A_v-A_u)\}.\end{aligned}$$

By (8), we get $(d_u-d_v)(A_v-A_u)\geq 0$. Thus

$$\begin{aligned}\sum_{v\in V(\mathcal{H})} Z(v) &\geq (r-1)\sum_{E\in\mathcal{E}}\sum_{u,v\in E}\{(d_v-m)A_v+(d_u-m)A_u\}\\ &=(r-1)\sum_{v\in V(\mathcal{H})}\sum_{u\in N(v)}(d_v-m)A_v\\ &=(r-1)^2\sum_{v\in V(\mathcal{H})} d_v(d_v-m)A_v.\end{aligned}$$

From the convexity of $f(x)$, it follows that $f(y)\geq f(x)+f'(x)(y-x)$ for any $x,y\geq 0$. So $A_v=f(d_v-1)-f(d_v)\geq -f'(d_v)$ and hence

$$\sum_{v\in V(\mathcal{H})} Z(v)\geq -(r-1)^2\sum_{v\in V(\mathcal{H})} d_v(d_v-m)f'(d_v). \tag{9}$$

Finally, combining (7) with (9) and using differential equation (3) in Lemma 1, we obtain

$$\begin{aligned}&\sum_{v\in V(\mathcal{H})} Y(v)\\ &\geq \sum_{v\in V(\mathcal{H})} \{1-[(r-1)d_v+1]f(d_v)-(r-1)^2 d_v(d_v-m)f'(d_v)\}\\ &= 0.\end{aligned}$$

This completes the proof of (6) and hence the lemma.

It is easy to see that our proof yields a polynomial-time algorithm for finding an independent set in $\mathcal{H}$ with at least $\sum_{v\in V} f_{r,m}(d_v)$ vertices.

REFERENCES

[1] M. Ajtai, P. Erdős, J. Komlós, and E. Szemerédi, *On Turán's theorem for sparse graphs*, Combinatorica, 1 (1981), pp. 313–317.

[2] M. Ajtai, P. Erdős, J. Komlós, and E. Szemerédi, *A dense infinite Sidon sequence*, European J. Combin., 2 (1981), pp. 1–11.

[3] M. Ajtai, J. Komlós, and E. Szemerédi, *A note on Ramsey numbers*, J. Combin. Theory Ser. A, 29 (1980), pp. 354–360.

[4] N. Alon and J. Spencer, *The Probabilistic Method*, Wiley-Interscience, New York, 1992.

[5] C. Berge, *Hypergraphs*, North-Holland, Amsterdam, 1989.

[6] Y. Caro, *New Results on the Independence Number*, Technical report, Tel-Aviv University, Tel-Aviv, Israel, 1979.

[7] Y. Caro and Z. Tuza, *Improved lower bounds on k-independence*, J. Graph Theory, 15 (1991), pp. 99–107.

[8] P. Duchet, *Hypergraphs Handbook of Combinatorics*, R. L. Graham, M. Grötschel, and L. Lovász, eds., Elsevier, Amsterdam, 1995, pp. 381–432.

[9] J. R. Griggs, *An upper bound on the Ramsey number $R(3,k)$*, J. Combin. Theory Ser. A, 35 (1983), pp. 145–153.

[10] J. Kim, *The Ramsey number $R(3,t)$ has order of magnitude $t^2/\log t$*, Random Structures Algorithms, 7 (1995), pp. 174–207.

[11] Y. Li and C. Rousseau, *On book-complete graph Ramsey numbers*, J. Combin. Theory Ser. B, 68 (1996), pp. 36–44.

[12] Y. Li, C. Rousseau, and W. Zang, *Asymptotic upper bounds for Ramsey functions*, Graphs Combin., 17 (2001), pp. 123–128.

[13] Y. Li, C. Rousseau, and W. Zang, *The lower bound on independence number*, Sci. China Ser. A, 45 (2002), pp. 64–69.

[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] J. Shearer, *A note on the independence number of triangle-free graphs*, Discrete Math., 46 (1983), pp. 83–87.

[16] J. Shearer, *A note on the independence number of triangle-free graphs*, II, J. Combin. Theory Ser. B, 53 (1991), pp. 300–307.

[17] A. K. Wei, *A Lower Bound on the Stability Number of a Simple Graph*, Bell Laboratories Technical Memorandum, No. 81-11217-9, Murray Hill, NJ, 1981.

[18] G. Zhou and Y. Li, *Independence numbers of hypergraphs with sparse neighborhoods*, European J. Combin., 25 (2004), pp. 355–362.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 105–118

LINEAR ORDERINGS OF SUBFAMILIES OF AT-FREE GRAPHS*

DEREK G. CORNEIL†, EKKEHARD KÖHLER‡, STEPHAN OLARIU§, AND LORNA STEWART¶

Abstract. Asteroidal triple free (AT-free) graphs have been introduced as a generalization of interval graphs, since interval graphs are exactly the chordal AT-free graphs. While for interval graphs it is obvious that there is always a linear ordering of the vertices, such that for each triple of independent vertices the middle one intercepts any path between the remaining vertices of the triple, it is not clear that such an ordering exists for AT-free graphs in general.

In this paper we study graphs that are defined by enforcing such an ordering. In particular, we introduce two subfamilies of AT-free graphs, namely, path orderable graphs and strong asteroid free graphs. Path orderable graphs are defined by a linear ordering of the vertices that is a natural generalization of the ordering that characterizes cocomparability graphs. On the other hand, motivation for the definition of strong asteroid free graphs comes from the fundamental work of Gallai on comparability graphs.

We show that cocomparability graphs $\subset$ path orderable graphs $\subset$ strong asteroid free graphs $\subset$ AT-free graphs. In addition, we settle the recognition question for the two new classes by proving that recognizing path orderable graphs is NP-complete, whereas the recognition problem for strong asteroid free graphs can be solved in polynomial time.

Key words. graph algorithms, complexity, asteroidal triple free graphs, recognition algorithm, linear ordering

AMS subject classifications. 05C75, 05C85, 68R10

DOI. 10.1137/S0895480104445307

1. Introduction.

We say that a vertex in a graph $G = (V, E)$ *intercepts* a path in G if it is adjacent to at least one vertex of the path, and it *misses* the path otherwise. An *asteroidal triple (AT)* is an independent set of three vertices such that, between every pair, there is a path that is missed by the third. A graph is *AT-free* if it does not contain an AT.

One of the most compelling motivations for the study of AT-free graphs is the idea that these graphs exhibit a type of linear structure. Indeed, the linear structure exhibited by AT-free graphs is explained, in part, in [1], where it is shown that every connected AT-free graph contains a *dominating pair* (two vertices such that every path connecting them is a dominating set) and a type of linear "shelling sequence" called a *spine*.

The original motivation for the results of the present paper was the idea that AT-free graphs might be characterized by the existence of a vertex ordering satisfying

*Received by the editors July 26, 2004; accepted for publication (in revised form) August 12, 2005; published electronically March 3, 2006. An expanded abstract of this paper appeared as *On subfamilies of AT free graphs*, in Graph-Theoretic Concepts in Computer Science (Boltenhagen, 2001), A. Brandstädt and V. B Le, eds., Lecture Notes in Comput. Sci. 2204, Springer-Verlag, New York, 2001, pp. 241–253. With kind permission of Springer Science and Business Media.

http://www.siam.org/journals/sidma/20-1/44530.html

†Computer Science Department, University of Toronto, Toronto, ON M55 3G4, Canada (dgc@cs.toronto.edu).

‡Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany (ekoehler@math.TU-Berlin.DE).

§Department of Computer Science, Old Dominion University, Norfolk, VA 23529-0162 (olariu@cs.odu.edu).

¶Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8, Canada (stewart@cs.ualberta.ca).

certain conditions. Looking back at the introduction of AT-free graphs as generalizations of interval graphs, there is an immediate candidate for such an ordering by requiring that for any independent triple in this ordering the central vertex should intercept every path between the remaining vertices of the triple. It is easy to see that interval graphs and even cocomparability graphs have such an ordering of the vertices (see below). However, it is not clear whether every AT-free graph possesses such an ordering.

Vertex orderings have proven to be useful algorithmic tools for several families of graphs. For example, chordal graphs (respectively, cocomparability graphs) are characterized by the existence of vertex orderings that do not contain the forbidden ordered configuration shown in Figure 1 (a) [2] (respectively, (b) [8]). A graph is an *interval graph* if and only if it has a vertex ordering that contains neither of the configurations of Figure 1 (see, for example, [11]). Such vertex orderings are referred to as chordal orderings, cocomparability orderings, and interval orderings, respectively.

FIG. 1. *Forbidden ordered configurations.*

In other words, in an interval ordering, for every path on two vertices (that is, for every edge), the left endpoint of the path is adjacent to all vertices between the two endpoints of the path. In a cocomparability ordering, each vertex between the two endpoints of a P_2 is adjacent to one or both endpoints of the P_2. It is well known that interval graphs are exactly those graphs that are both chordal and cocomparability [5] or, equivalently, both chordal and AT-free [9]. Furthermore, cocomparability graphs are a proper subclass of AT-free graphs [6].

An alternate characterization of the cocomparability ordering is given in Observation 1.1.

OBSERVATION 1.1. *A vertex ordering $v_1, \ldots, v_n$ of graph G is a cocomparability ordering if and only if for all v_i, v_j, v_k with $i < j < k$, vertex v_j intercepts each v_i, v_k-path of G.*

From this, one can easily see that a cocomparability graph must be AT-free since any independent triple occurs in some order, say, $x \prec y \prec z$, in a cocomparability ordering "$\prec$," and thus, there cannot exist an x, z-path missed by y. In an attempt to generalize the cocomparability ordering while retaining the AT-free property, we introduce the following definition.

DEFINITION 1.2. *A graph $G = (V, E)$ is* path orderable *if there is an ordering $v_1, \ldots, v_n$ of the vertices such that for each triple of vertices v_i, v_j, v_k with $i < j < k$ and $v_i v_k \notin E$, vertex v_j intercepts each v_i, v_k-path of G; such an ordering is called a* path ordering.

Observation 1.1 and Definition 1.2 imply that cocomparability graphs are path orderable. C_5, the chordless cycle on five vertices, is a path orderable graph which is not a cocomparability graph. It is clear that path orderable graphs must be AT-free. However, can Definition 1.2 be used for characterizing AT-free graphs? Figure 2 shows an AT-free graph together with an ordering that is not a path ordering. Hence, the question here is whether it can be turned into a path ordering. Unfortunately, we shall see later that path orderable graphs form a strict subset of AT-free graphs; in particular, the graph in Figure 2 will be shown to be not path orderable.

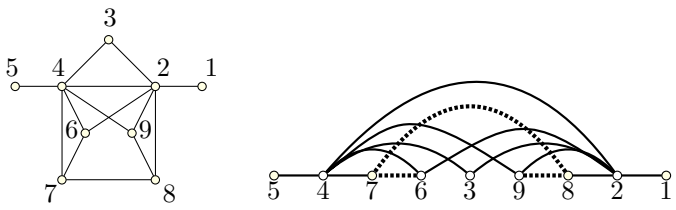


FIG. 2. *AT-free graph G with ordering that is not a path ordering; in particular, path* 6-7-8-9 *is not intercepted by* 3 *(the edges of the path are dashed).*

Nevertheless, since path orderable graphs are interesting in their own right, we attempted to provide a structural characterization of this graph class by identifying a type of forcing relation on nonadjacent pairs of vertices and the type of structure that makes the vertex ordering of Definition 1.2 impossible.

These investigations follow in Gallai's footsteps [3, 10] in that they involve ideas similar to his forcing relation on the edges of a comparability graph (equivalently, the nonedges of a cocomparability graph) and his definitions of wreaths and asteroids. Specifically, we define strong asteroids and show that path orderable graphs are strong asteroid free. However, it turns out that the strong asteroid concept does not provide a characterization of path orderable graphs; we shall see that path orderable graphs form a proper subclass of strong asteroid free graphs which, in turn, form a proper subclass of AT-free graphs.

Thus, we will identify two distinct subclasses of AT-free graphs, both of which contain cocomparability graphs:

cocomparability $\subset$ path orderable $\subset$ strong asteroid free $\subset$ AT-free.

The interest lies, in part, in the natural vertex ordering, in one case, and the relationship with Gallai's work, in the other case. Furthermore, the identification of these graph classes should allow us to narrow the gap between known polynomial and known NP-complete behavior of problems in the domain of AT-free graphs. For example, the complexity status for coloring, Hamiltonian path, and Hamiltonian cycle is still unresolved for AT-free graphs but is in P for cocomparability graphs.

We conclude the paper with a proof that the recognition of path orderable graphs is NP-complete, and with a polynomial time recognition algorithm for strong asteroid free graphs. We note that the NP-completeness result settles an open problem stated in [13].

Background. In his paper on comparability graphs [3, 10], Gallai studies the forcing between the edges imposed by a transitive orientation (to avoid misunderstandings, from now on we will refer to the transitive-forcing as *t-forcing*). Let G be an arbitrary graph. Two edges which share a common endpoint and whose other endpoints are nonadjacent t-force each other directly. That is, in any transitive orientation, either both edges are directed away from the common endpoint or both are directed toward it. The transitive closure of the direct t-forcing relation partitions the edges of G into *t-forcing classes*. Either there are exactly two different transitive orientations of the edges of a t-forcing class, or there is none. The latter case occurs when some edge is t-forced in both directions, in which case G is not a comparability graph. Edges xy and xz are said to be *knotted* if y and z are connected in $\overline{G}[\mathrm{N}(x)]$, the complement of the subgraph of G induced by $\mathrm{N}(x)$, where $\mathrm{N}(x)$, the neighborhood

of x, is defined as $\mathrm{N}(x) = \{u \mid ux \in E\}$.

To capture the t-forcing in a given graph G, Gallai uses the concept of a *knotting graph*: For a graph $G = (V, E)$ the corresponding *knotting graph* is given by $\mathrm{K}[G] = (V_\mathrm{K}, E_\mathrm{K})$, where V_K and E_K are defined as follows. For each vertex v of G there are copies $v_1, v_2, \ldots, v_{i_v}$ in V_K, where i_v is the number of connected components of $\overline{G}[\mathrm{N}(v)]$. For each edge vw of E there is an edge $v_i w_j$ in E_K, where w is contained in the ith connected component of $\overline{G}[\mathrm{N}(v)]$ and v is contained in the jth connected component of $\overline{G}[\mathrm{N}(w)]$. Please refer to Figure 5 for an example of a graph together with its knotting graph.

In this graph two edges are incident if and only if they are knotted. The edges of the t-forcing classes of G are given by the connected components of $\mathrm{K}[G]$. Using this structure, Gallai shows that a graph G is a comparability graph if and only if $\mathrm{K}[G]$ is bipartite.

The following definitions from [3] describe structures which lead to t-forcing classes which cannot be transitively oriented and knotting graphs which are not bipartite.

DEFINITION 1.3. *An* odd wreath *of size k in a graph is a cycle of knotted edges, specifically, a sequence of vertices $v_0, v_1, v_2, \ldots, v_k$, where k is odd, $v_1, \ldots, v_k$ are distinct, $v_0 = v_k$, and for all i, $0 \le i < k$, edges $v_i v_{i+1}$ and $v_{i+1} v_{i+2}$ exist in the graph and are knotted (addition modulo k).*

DEFINITION 1.4. *An* odd asteroid *of size k in a graph is a sequence of vertices $v_0, v_1, v_2, \ldots, v_k$ where k is odd, $v_1, \ldots, v_k$ are distinct, $v_0 = v_k$, and for all i, $0 \le i < k$, there exists a $v_i v_{i+1}$-path in G which is missed by $v_{(i+\frac{k+1}{2})}$ (addition modulo k).*

Gallai points out that an odd asteroid is the complement of an odd wreath and proves that a graph is a comparability graph if and only if it contains no odd wreath or, equivalently, a graph is a cocomparability graph if and only if it contains no odd asteroid. Note also that an AT corresponds to an odd asteroid of size three.

As an example of an odd asteroid, consider the graph G in Figure 2. Here, the sequence of vertices $1, 3, 5, 7, 8, 1$ forms an odd asteroid of size 5 in G. The sequence $1, 5, 8, 3, 7, 1$ of vertices forms an odd wreath of size 5 in $\overline{G}$.

2. Path orderable graphs and strong asteroid free graphs. As we have seen, t-forcing is a fundamental concept for comparability graphs, and thus for cocomparability graphs as well. Given the similarities of the linear ordering characterizations of path orderable graphs and cocomparability graphs, one might expect a similar forcing concept for path orderable graphs. In fact such is the case.

For a graph G and a vertex v of G let $C_1, \ldots, C_k$ be the connected components of $G \setminus \mathrm{N}[v]$ and let $B_i^1, \ldots, B_i^\ell$ be the connected components of the graph induced by the vertices of C_i in $\overline{G}$ $(1 \le i \le k)$; the B_i^j are called the *blobs* of v in G. (Here $\mathrm{N}[v] := \mathrm{N}(v) \cup \{v\}$ denotes the closed neighborhood of vertex v in G.) As an example, consider the graph in Figure 3.

LEMMA 2.1. *Let G be a path orderable graph and let $v_1, \ldots, v_n$ be a path ordering of G. For every vertex v of G and for every blob B of v, the vertices of B occur either all before v in the path ordering or all after v in the path ordering.*

Proof. Suppose there are a vertex v and a blob B of v with $u, w \in B$ and $u \prec v \prec w$ in the path ordering "$\prec$" of G (see Figure 4 for a sketch of this setting). By the definition of blobs, u and w are in the same connected component C of $G \setminus \mathrm{N}[v]$. Since u and w are also in the same connected component B of C in $\overline{G}$, there has to be a path of nonedges in B between u and w. Thus, there is a pair of vertices u', w'

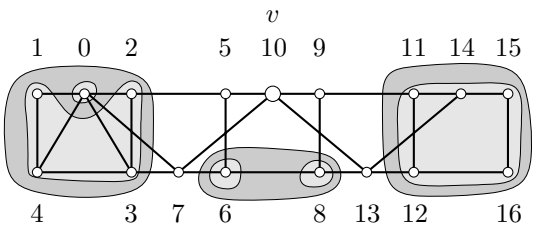


FIG. 3. *The blobs of vertex* $v = 10$ *are given by the sets* $\{0\}$, $\{1,2,3,4\}$, $\{6\}$, $\{8\}$, $\{11,12,14,15,16\}$.

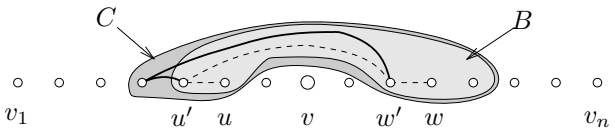


FIG. 4. *Proof idea of Lemma* 2.1.

in B with $u'w' \notin E$ and $u' \prec v \prec w'$. But $u', w' \in C$; therefore there is a u', w'-path in $G \setminus \mathrm{N}[v]$, contradicting the path ordering. □

By Lemma 2.1, any path ordering has to fulfill the property that if one of the vertices u of a blob B of v precedes v in the ordering, then all of the vertices of B occur before v.

Consider now the graph G in Figure 2. Following the above definition of blobs, vertex 3 has the three blobs $\{6,7,8,9\}$, $\{5\}$, $\{1\}$; vertex 7 has the blobs $\{3,1,9\}$, $\{2\}$, $\{5\}$; vertex 8 has the blobs $\{3,5,6\}$, $\{4\}$, $\{1\}$; vertex 5 has only the blob $\{1,2,3,6,7,8,9\}$; and vertex 1 has only the blob $\{3,4,5,6,7,8,9\}$. Suppose there is a path ordering of G. By Lemma 2.1 we can, without loss of generality, assume that 1 precedes all vertices of its blob and thus 5 appears after all vertices of its blob in the path ordering; in particular, vertices 3, 6, 7, 8, 9 are between 1 and 5. Since 7 and 8 are in the same blob of 3, they appear either both before or both after 3 in the path ordering. However, if they both appear before 3, then, again by Lemma 2.1, we have a contradiction because 3 and 1 are in the same blob of 7, but on different sides in the path ordering. On the other hand, if both 7 and 8 appear after 3 in the path ordering we again have a contradiction, since 3 and 5 are in the same blob of 8 but on different sides in the path ordering. Hence there cannot be a path ordering for the graph in Figure 2.

COROLLARY 2.2. *The class of path orderable graphs is strictly contained in the class of AT-free graphs.*

LEMMA 2.3. *If a graph G is path orderable then every induced subgraph of G is path orderable.*

Proof. This follows by the definition of path orderable and since any path in an induced subgraph of graph G is also a path in G. □

When interpreting the constraints of Lemma 2.1 as orientations of the edges of $\overline{G}$, in the sense that edges from the same blob of a vertex v to v in $\overline{G}$ have to have the same orientation (i.e., representing before or after v in the path ordering), one can define the following forcing on the edge set of $\overline{G}$.

Let G be an arbitrary graph and let $e_1 = uv$, $e_2 = vw$ be edges of $\overline{G}$ with a common end-vertex v. Then one can define a relation $\approx$ by $e_1 \approx e_2$ (e_1 and e_2 *force each other* or *are knotted* at v) if and only if u and w are in the same blob of v (possibly $u = w$) in G. The transitive closure of this relation defines a class partition

of the edges of $\overline{G}$, where two edges e_a, e_b are in the same class (*forcing class*) of $\overline{G}$ if there is a sequence $e_1, e_2, \ldots, e_k$ of edges such that $e_a = e_1 \approx e_2 \approx \cdots \approx e_k = e_b$. Observe that the forcing classes are refinements of the t-forcing classes.

An orientation of the edges of $\overline{G}$ is said to *agree with the forcing* if for any vertex v and any blob B of v all edges between B and v are oriented in the same direction (either toward v or away from v). For a graph G a linear ordering $v_1, \ldots, v_n$ of the vertices of G is said to *agree with the forcing* if the corresponding implied orientation of the edges of $\overline{G}$ (uv is oriented from u to v if $u \prec v$ in the linear ordering "$\prec$") agrees with the forcing.

Note that when the orientation of one of the edges of a forcing class is fixed, then the orientation of all the edges of its forcing class is determined; hence, either there are exactly two different orientations of the edges of a forcing class that agree with the forcing, or there is none. In the latter case, some edge is forced to be oriented in both directions, meaning that there is no ordering consistent with the forcing.

LEMMA 2.4. *A graph G is path orderable if and only if there is a linear ordering of the vertices of G agreeing with the forcing.*

Proof. If G is path orderable, then, by Lemma 2.1, the path ordering has to agree with the forcing relation.

Suppose there is a linear ordering "$\prec$" of G that agrees with the forcing relation and suppose there is a triple $u \prec v \prec w$ of vertices that violates the path ordering property, i.e., $uw \notin E$, and there is a u,w-path in $G \setminus \mathrm{N}[v]$. Hence, u and w are in the same connected component C of $G \setminus \mathrm{N}[v]$ and, since $uw \notin E$, u and w are also in the same blob B of v. But then this ordering does not agree with the forcing relation, which is a contradiction. ∎

COROLLARY 2.5. *A graph G is path orderable if and only if there is an acyclic orientation of $\overline{G}$, agreeing with the forcing relation.*

Proof. Determine a topological ordering, using the acyclic orientation of $\overline{G}$; then the corollary follows from Lemma 2.4. ∎

One can define a graph, similar to Gallai's knotting graph, representing the forcing classes of $\overline{G}$. For a graph $G = (V, E)$ the *altered knotting graph* is given by $\mathrm{K}^*[G] = (V_K, E_K)$, where V_K and E_K are defined as follows. For each vertex v of G there are copies $v_1, \ldots, v_{i_v}$ in V_K, where i_v is the number of blobs of v in $\overline{G}$. For each edge vw of E there is an edge $v_i w_j$ in E_K, where w is contained in the ith blob of v in $\overline{G}$ and v is contained in the jth blob of w in $\overline{G}$.

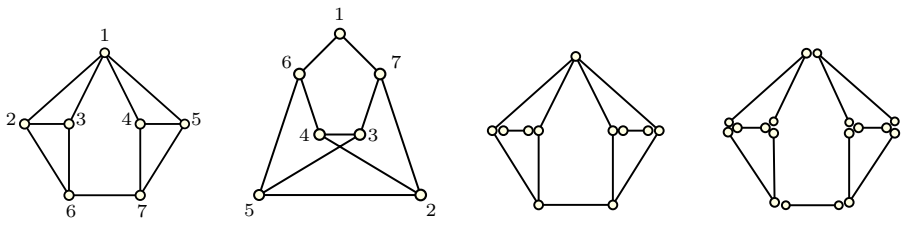


FIG. 5. *A graph G together with its complement $\overline{G}$, $\mathrm{K}[G]$, and $\mathrm{K}^*[G]$.*

As Gallai did for the knotting graph, we draw the altered knotting graph $\mathrm{K}^*[G]$ of a given graph G by putting different copies of the same vertex close together. See Figure 5 for an example of a graph G, together with its complementary graph $\overline{G}$, its knotting graph $\mathrm{K}[G]$, and its altered knotting graph $\mathrm{K}^*[G]$. The blobs of the vertices

of $\overline{G}$ are as follows: vertex 1: $\{2,3\}$, $\{4,5\}$; vertex 2: $\{1\}$, $\{3\}$, $\{6\}$; vertex 3: $\{1\}$, $\{2\}$, $\{6\}$; vertex 4: $\{1\}$, $\{5\}$, $\{7\}$; vertex 5: $\{1\}$, $\{4\}$, $\{7\}$; vertex 6: $\{2,3\}$, $\{7\}$; vertex 7: $\{4,5\}$, $\{6\}$.

Our next task is to examine configurations which cannot occur in path orderable graphs. As a step toward this goal, we define restricted types of odd wreaths and asteroids.

DEFINITION 2.6. *An* odd strong wreath *of size k in a graph G is a sequence of vertices $v_0, v_1, \ldots, v_k$ where k is odd, $v_1, \ldots, v_k$ are distinct, $v_0 = v_k$, and for all i, $0 \leq i < k$, edges $v_i v_{i+1}$ and $v_{i+1} v_{i+2}$ exist in the graph and are knotted in the altered sense; that is, v_i and v_{i+2} are in the same blob of v_{i+1} in $\overline{G}$ (addition modulo k).*

DEFINITION 2.7. *An* odd strong asteroid *of size k in a graph G is a sequence of vertices $v_0, v_1, \ldots, v_k$ where k is odd, $v_1, \ldots, v_k$ are distinct, $v_0 = v_k$, and for all i, $0 \leq i < k$, v_i and v_{i+1} are in the same blob of $v_{(i+\frac{k+1}{2})}$ in G (addition modulo k).*

The two notions are complementary; that is, a graph G has an odd strong wreath if and only if $\overline{G}$ contains an odd strong asteroid. Furthermore, strong asteroids and strong wreaths are restricted types of asteroids and wreaths. We also note that the ATs correspond to the odd strong asteroids of size three. Figure 6 features a graph containing an odd strong asteroid as well as its complement that contains an odd strong wreath.

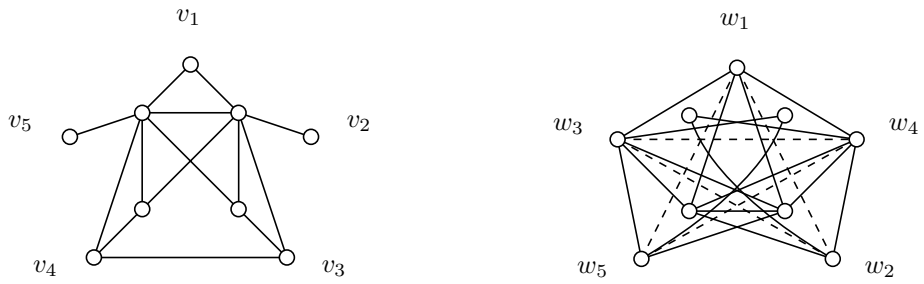


FIG. 6. *Graph of Figure* 2*, containing an odd strong asteroid and its complement, containing an odd strong wreath (vertices of the asteroid and the wreath are marked by $v_1, \ldots, v_5$ and $w_1, \ldots, w_5$, respectively; the edges of the wreath are dashed).*

DEFINITION 2.8. *A graph G is* strong asteroid free *if it does not contain an odd strong asteroid.*

Similar to the t-forcing results, the following holds.

LEMMA 2.9. *The forcing classes of a graph G are precisely the connected components of* $\mathrm{K}^*[G]$.

The next two observations follow from the fact that an odd strong asteroid of size k in G corresponds to an odd cycle of size k in $\mathrm{K}^*[\overline{G}]$.

OBSERVATION 2.10. *A graph G is strong asteroid free if and only if* $\mathrm{K}^*[\overline{G}]$ *is bipartite.*

OBSERVATION 2.11. *A graph G is AT-free if and only if* $\mathrm{K}^*[\overline{G}]$ *is triangle-free.*

LEMMA 2.12. *If a graph G is path orderable then* $\mathrm{K}^*[\overline{G}]$ *is bipartite.*

Proof. Let $v_1, \ldots, v_n$ be a path ordering of G. Now orient the edges of $\mathrm{K}^*[\overline{G}]$ as follows: $v_i v_j$ is oriented from v_i to v_j if $i < j$. Now, by Lemma 2.1, each vertex of $\mathrm{K}^*[\overline{G}]$ has either only incoming or only outgoing edges. Hence, it is bipartite. ∎

Not only does the graph in Figure 2 show that path orderable graphs are strictly contained in AT-free graphs, but it also establishes that strong asteroid free graphs

are strictly contained in AT-free graphs, as shown in the next lemma.

LEMMA 2.13. *The class of strong asteroid free graphs is strictly contained in the class of AT-free graphs.*

Proof. Consider the graphs of Figures 2 and 6. It is easy to check that the vertices named $v_1, \ldots, v_5$ in Figure 6 form an odd strong asteroid in G, and that G is AT-free. □

Similar to Lemma 2.3 one can prove the following lemma.

LEMMA 2.14. *If a graph G is strong asteroid free then every induced subgraph of G is strong asteroid free.*

In the case of comparability graphs, Gallai not only showed that the knotting graph $\mathrm{K}[G]$ of a comparability graph is bipartite but also proved that a bipartite knotting graph $\mathrm{K}[G]$ is a sufficient condition for G being a comparability graph. The major tool that he used for proving this result is a lemma which shows the following. Given a bipartite knotting graph $\mathrm{K}[G]$ consider a triangle of G with the property that at least two of the edges of the triangle are in the same t-forcing class; then in any orientation of G that agrees with the t-forcing, the triangle is not oriented cyclically.

It turns out that a similar lemma holds for strong asteroid free graphs, too. Specifically, for a graph G with a bipartite altered knotting graph $\mathrm{K}^*[\overline{G}]$, any orientation of $\overline{G}$ that agrees with the forcing relation does not contain a cyclically oriented triangle. However, contrary to the t-forcing relation, this lemma is not enough to imply that the orientation is acyclic and, indeed, we shall show that this is not necessarily the case.

OBSERVATION 2.15. *Given a vertex v in a graph H and vertices $u, w \in \mathrm{N}(v)$, which are the endpoints of an induced P_4 in $\mathrm{N}(v)$, then the edges uv and wv force each other (see Figure 7).*

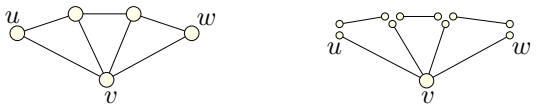


FIG. 7. *Vertex v with P_4 in $\mathrm{N}(v)$ together with the corresponding altered knotting graph.*

Remark 2.16. Using this observation one can create a *forcing path*, i.e., a path P, where each consecutive pair of edges of P is knotted at the common end-vertex by the help of an added P_4 as described in Observation 2.15; see Figure 8 (in the following, edges and vertices of the path P itself are called *original edges/vertices*, and the added edges and vertices are denoted as *auxiliary edges/vertices*). By the forcing, the orientation of any original edge of P forces the orientation of all other original edges of P. Note that the knotting graph of a forcing path does not contain a triangle or any odd cycle. Furthermore, if P has even length, then the end-edges of P are either both oriented toward the inner vertices of P or both oriented outward from the inner vertices of P. Similarly, if P has odd length, the end-edges of P have opposite orientations with respect to the inner vertices of P.

THEOREM 2.17. *The class of path orderable graphs is strictly contained in the class of strong asteroid free graphs.*

Proof. Consider the left graph in Figure 9. This graph is the complement of a strong asteroid free graph G. This is proved by constructing the altered knotting graph $\mathrm{K}^*[\overline{G}]$ (see the right graph in Figure 9). By Observation 2.15, the thick edges force each other, as shown in the altered knotting graph; and, without having a

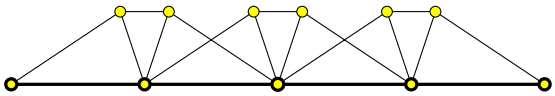


FIG. 8. *A forcing path of length* 4 *(original edges and vertices are bold).*

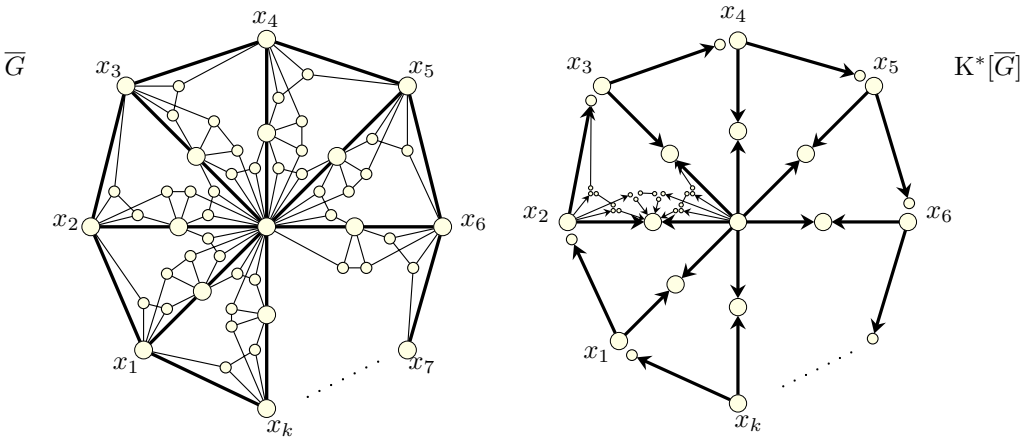


FIG. 9. *Complement of a strong asteroid free graph, which is not path orderable (left), together with its altered knotting graph (right). To ease understanding of its structure, in the knotting graph the corresponding auxiliary* P_4 *vertices are drawn in the figure for only one of the arms of the example. One of the two possible forced orientations of the main forcing class is given in the right picture.*

strong asteroid in G, there is a forced oriented cycle on the vertices $x_1, \ldots, x_k$ in $\overline{G}$. Consequently, by Corollary 2.5, G is not path orderable. This construction holds for any $k \geq 4$. □

3. Recognition of path orderable and strong asteroid free graphs. In this section, we show that the recognition of path orderable graphs is NP-complete. This result answers a question posed by Spinrad in [13]. In contrast, we describe how to recognize strong asteroid free graphs in polynomial time.

First, observe that the recognition problem of path orderable graphs is obviously in NP, since by Lemma 2.1 for a given ordering one can easily check in polynomial time whether it is a path ordering. If there is only one forcing class for the edge set of $\overline{G}$ one can also check in polynomial time whether G is path orderable: Compute $\mathrm{K}^*[\overline{G}]$, check whether it is bipartite, assign an orientation to $\mathrm{K}^*[\overline{G}]$ by orienting all edges from one of the bipartition classes to the other, and check whether this orientation is acyclic on $\overline{G}$.

Similarly one can check whether G is path orderable if the number of forcing classes of $\overline{G}$ is bounded by a constant.

For comparability graphs, Gallai's results for the general case, i.e., where no assumption on the number of edge classes is made, lead to a polynomial time recognition algorithm. For this he introduced the (by now well-known) concept of modular decomposition and proved that, using this decomposition scheme, the problem of

recognizing comparability graphs reduces to the problem of recognizing prime comparability graphs. But what about the recognition of path orderable graphs? Can one extend the decomposition scheme to this problem?

NOT-ALL-EQUAL 3SAT. [4]
INSTANCE: Set U of variables, collection $\mathcal{C}$ of clauses over U such that each clause $c \in \mathcal{C}$ has $|c| = 3$.
QUESTION: Is there a truth assignment A for U such that each clause in C has at least one *true* literal and at least one *false* literal?

Remark 3.1. Without loss of generality one can assume that none of the clauses contains more than one literal of a variable.

To prove the NP-hardness of the recognition problem of path orderable graphs, we use a transformation from NOT-ALL-EQUAL 3SAT (NAE 3SAT). Given an instance I of NAE 3SAT, a graph G is constructed, which is the complement of a path orderable graph if and only if I is NAE 3SAT-satisfiable. In particular, it will be shown that I is NAE 3SAT-satisfiable if and only if there is an acyclic orientation of G that agrees with the forcing. By Corollary 2.5 this is equivalent with $\overline{G}$ being path orderable.

The basic construction of G is as follows. For every variable x of U an edge e_x is created (called a *variable edge* in the following) and the two possible orientations of e_x are associated with the two possible values *true* and *false* of x.

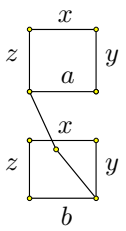


FIG. 10. *Gadget for clause* $x \vee y \vee z$.

For each clause $C = x \vee y \vee z$ with literals x, y, z a gadget is constructed, mainly consisting of two C_4's as shown in Figure 10. In each of the C_4's three of the edges (the *base-edges*) correspond to the three literals x, y, z of C. As will be explained below, a *true* literal of C will correspond to a clockwise orientation of the corresponding base-edges in both of the C_4's, whereas a *false* literal will correspond to a counterclockwise orientation of the corresponding base-edges in both C_4's. Furthermore, in each orientation that agrees with the forcing, the fourth edges of the two C_4's, which will be called the *bridge edges* (edges a and b in Figure 10), will be guaranteed to have opposite orientations in the two C_4's. This is realized by making these bridge edges the end-edges of a forcing path of length 4. Consequently, with this construction, a truth assignment of the variables of U that sets all three literals of C to *true* (*false*) results in a clockwise (counterclockwise) orientation of all three base-edges in both C_4's and, since the bridge edges have opposite orientations in the two C_4's, at least one of the C_4's has a cyclic orientation. On the other hand, by the above correspondence between the orientations of the base-edges and the truth-values of the corresponding literals, each acyclic orientation of G that agrees with the forcing leaves at least one literal of C *true* and one *false*.

Next, it has to be ensured that the value of a variable and the value of the literals of this variable coincide; i.e., the orientation of the variable edge of x for value *true* has to result in a counterclockwise orientation of the base-edges for $\overline{x}$ in all the gadgets

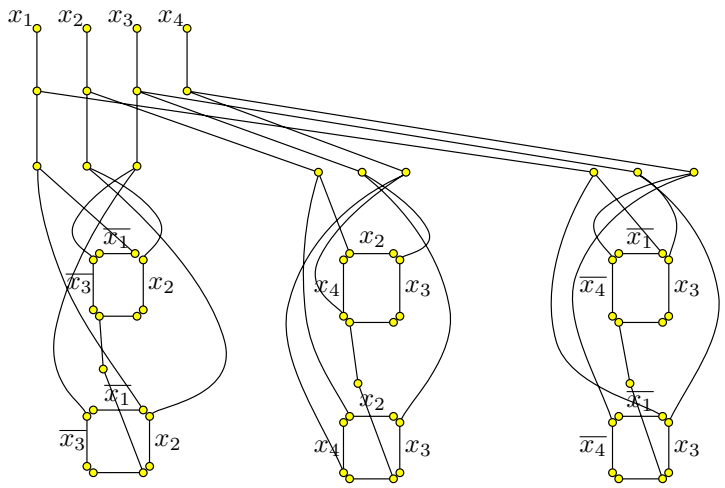


FIG. 11. *General structure of* $\mathrm{K}^*[G]$ *for the instance* $I = (\overline{x_1} \vee x_2 \vee \overline{x_3}) \wedge (x_2 \vee x_3 \vee x_4) \wedge (\overline{x_1} \vee x_3 \vee \overline{x_4})$ *(auxiliary vertices and edges are omitted).*

for clauses containing literal $\overline{x}$ and in a clockwise orientation of the base-edges for x in all the gadgets for clauses containing literal x. This is realized by connecting each variable edge to all corresponding base-edges by the help of forcing paths that are joined appropriately. In other words, for each variable a separate edge class is created, containing the variable edge and all base-edges corresponding to literals of this variable. The general structure of the connection between variable edges and base-edges by forcing paths is shown in Figure 11; for easier understanding the auxiliary edges and vertices of the forcing paths are omitted in this picture. For a variable edge e_x (see top of Figure 11) a downward orientation corresponds to assigning *false* to variable x, whereas an upward orientation corresponds to assigning *true* to x. For each literal x or $\overline{x}$, there is a forcing path of length 4, having e_x and the corresponding base-edge as its end-edges; depending on whether the literal is $\overline{x}$ or x, either the start- or the end-vertex of the base-edge (with respect to a clockwise ordering in the C_4) is made the end-vertex of the forcing path.

Now, by Remark 2.16, assigning an upward orientation to the variable edge e_x results in the desired clockwise orientation of the base-edges of the literals x and a counterclockwise orientation of the base-edges of the literals $\overline{x}$ for any orientation agreeing with the forcing.

In Figure 12 (left) the complete construction of G for a single clause C together with the variable edges and the forcing paths is given, including all auxiliary edges and vertices. In the right part of the figure the corresponding altered knotting graph $\mathrm{K}^*[G]$ is shown.

We now study properties of orientations of G that agree with the forcing. For this it is sufficient to consider $\mathrm{K}^*[G]$. Observe first, that, by the construction, $\mathrm{K}^*[G]$ is bipartite; indeed, $\mathrm{K}^*[G]$ is even a forest and for each of the variables there is exactly one connected component in $\mathrm{K}^*[G]$ that contains both the variable edge and all base-edges corresponding to this variable. Note furthermore that an oriented cycle in an orientation of G can contain neither a source nor a sink vertex of that orientation. Consequently, all the vertices of G, having only one copy in $\mathrm{K}^*[G]$, cannot be contained

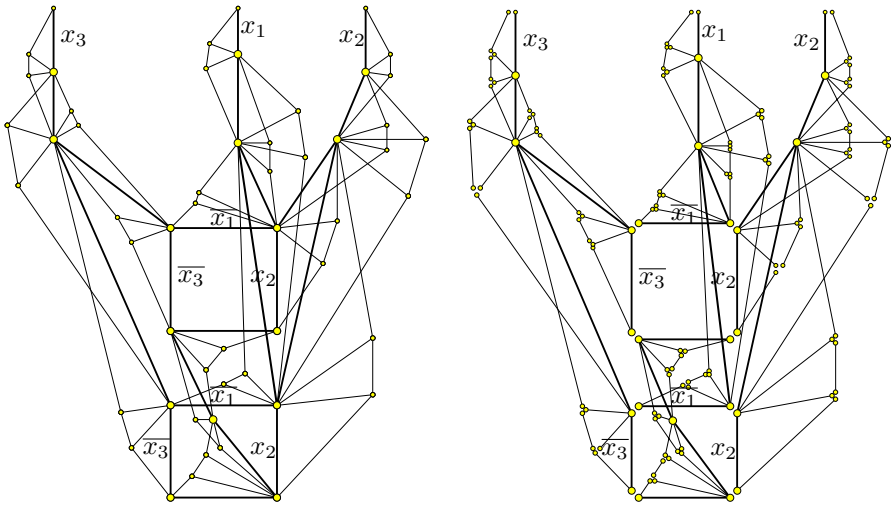


FIG. 12. *Left: Complete construction for a gadget of the clause* $(\overline{x_1} \vee x_2 \vee \overline{x_3})$ *together with the variable edges and the forcing paths. Right: The corresponding altered knotting graph.*

in any such cycle, since they have to be sources or sinks in any orientation of G, which agrees with the forcing. After deleting all those vertices from G, the only cycles of the remaining graph are the two four-cycles per gadget and some triangles, each consisting of auxiliary edges and at most one of the C_4-edges (see Figure 13). Consider any of

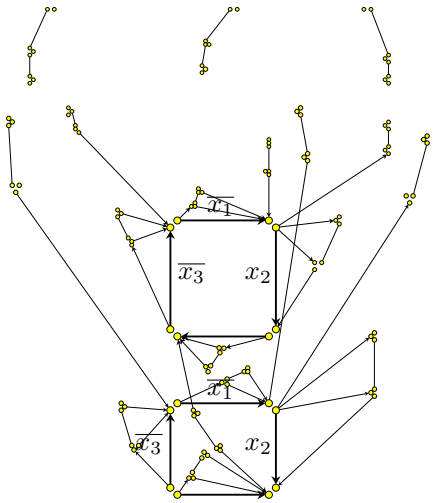


FIG. 13. *A clause-gadget after removing all source and sink vertices.*

those remaining triangles. By the construction, at least two of the three triangle-edges are incident to the same vertex of $\mathrm{K}^*[G]$. Consequently, in any orientation that agrees with the forcing relation, these two edges prevent the triangle from being cyclically oriented. Hence, when checking an orientation (that agrees with the forcing) of the constructed graph G to be acyclic, it is sufficient to show that each of the two C_4's

per gadget is acyclically oriented.

OBSERVATION 3.2. *Given an orientation of G that agrees with the forcing, this orientation is acyclic if and only if it is acyclic on both C_4's of each of the clause gadgets.*

Now we are ready to show the following lemma.

LEMMA 3.3. *There is an acyclic orientation of G agreeing with the forcing relation if and only if $\mathcal{C}$ has an NAE 3SAT satisfying assignment.*

Proof. Suppose that there is an NAE 3SAT satisfying assignment A. An acyclic orientation of G that agrees with the forcing can be constructed as follows. We assign orientations to the variable edges (the edges on top of Figure 11) by orienting an edge downward if the corresponding variable is set *false* in A and upward otherwise. Consequently, all edges of the connected components of those edges in $\mathrm{K}^*[G]$ have a forced orientation as well.

The only edges that have not been assigned an orientation in this way are the forcing classes of the bridge edges of every C_4 and the single edges of the auxiliary P_4's (see connected components of the knotting graph in Figure 12, right). The single edges can be assigned an arbitrary orientation and for each of the bridge edge classes just one edge is oriented arbitrarily, forcing the orientation of all other edges of this class. Obviously, this orientation agrees with the forcing.

By the forcing of the edges and the appropriate knotting of the forcing path from the variable representing edges to the edges representing the literals, each *true* literal in a clause C leads to a clockwise oriented edge, and analogously, each *false* literal implies a counterclockwise oriented edge in the corresponding C_4's. Since every clause has at least one *true* and one *false* literal, each of the C_4's has both an edge that is oriented clockwise and one that is oriented counterclockwise. Hence, none of the C_4's is cyclically oriented and, by Observation 3.2, the orientation is acyclic.

Suppose now that there is an acyclic orientation of G that agrees with the forcing relation. We assign to a variable x of U the value *true* if the edge representing variable x (edges on top of Figure 11) is oriented upward and *false* otherwise. Since the orientation agrees with the forcing relation, all we have to show is that all of the clauses have at least one *true* and one *false* literal. Suppose there is a clause C, which has only *true* (*false*) literals. By the definition of G and the forcing relation, three edges in each of the C_4's in C's gadget are oriented counterclockwise (clockwise). Since the bridge edges have opposite orientations in the two C_4's of C, exactly one of the C_4's is oriented cyclically, contradicting that the orientation of G is acyclic. □

Since it is easy to see that the construction of graph G is polynomial in the size of the input U and $\mathcal{C}$, Lemma 3.3 directly implies the following theorem.

THEOREM 3.4. *The problem of deciding whether a graph is path orderable is NP-complete.*

In contrast to Theorem 3.4, a polynomial time recognition algorithm for strong asteroid free graphs follows from Observation 2.10. Given graph G, the altered knotting graph of $\overline{G}$, $\mathrm{K}^*[\overline{G}]$, can be computed in polynomial time: for each vertex v of G, the blobs of v in G can be computed in $O(n^2)$ time; each vertex has fewer than n blobs. Thus, $\mathrm{K}^*[\overline{G}]$ has $O(n^2)$ vertices and $O(n^2)$ edges (since each edge of $\overline{G}$ corresponds to exactly one edge of $\mathrm{K}^*[\overline{G}]$) and can be constructed in $O(n^3)$ time. To test whether $\mathrm{K}^*[\overline{G}]$ is bipartite can be done in $O(n^2)$ time. Overall, the recognition algorithm requires $O(n^3)$ time.

THEOREM 3.5. *Strong asteroid free graphs can be recognized in time $O(n^3)$.*

4. Concluding remarks. We have defined two graph classes and shown that cocomparability graphs $\subset$ path orderable graphs $\subset$ strong asteroid free graphs $\subset$ AT-free graphs. Furthermore, we have shown that the recognition problem for path orderable graphs is NP-complete, and the recognition of strong asteroid free graphs can be solved in polynomial time. We note that AT-free graph recognition is also in P [1, 7].

Although it is somewhat disappointing that no two of these families are equivalent, these classes may give insight into some open problem complexities on AT-free graphs. By adding graph classes in the hierarchy between cocomparability graphs and AT-free graphs, we may be able to identify more precisely the boundary between polynomial and NP-complete behavior of some of the problems which are known to be polynomially solvable on cocomparability graphs but either NP-complete or unresolved on AT-free graphs. Examples of such problems include graph coloring, clique cover, clique, and the Hamiltonian path and cycle problems. One step in this direction is the observation that the clique problem is NP-complete for path orderable graphs. This follows from the facts that the complements of triangle-free graphs are contained in path orderable graphs, and the independent set problem is known to be NP-complete on triangle-free graphs [12].

Acknowledgment. The authors wish to thank the Natural Science and Engineering Research Council of Canada for financial support.

REFERENCES

[1] D. G. Corneil, S. Olariu, and L. Stewart, *Asteroidal triple-free graphs*, SIAM J. Discrete Math., 10 (1997), pp. 399–430.
[2] R. R. Fulkerson and O. A. Gross, *Incidence matrices and interval graphs*, Pacific J. Math., 15 (1965), pp. 835–855.
[3] T. Gallai, *Transitiv orientierbare Graphen*, Acta Math. Acad. Sci. Hungar., 18 (1967), pp. 25–66.
[4] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman, New York, 1979, ch. A9, p. 259.
[5] P. C. Gilmore and A. J. Hoffman, *A characterization of comparability graphs and of interval graphs*, Canad. J. Math., 16 (1964), pp. 539–548.
[6] M. C. Golumbic, C. L. Monma, and W. T. Trotter, *Tolerance graphs*, Discrete Appl. Math., 9 (1984), pp. 157–170.
[7] E. Köhler, *Graphs without asteroidal triples*, Ph.D. thesis, Technische Universität Berlin, Berlin, Germany, 1999.
[8] D. Kratsch and L. Stewart, *Domination on cocomparability graphs*, SIAM J. Discrete Math., 6 (1993), pp. 400–417.
[9] 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), pp. 45–64.
[10] F. Maffray and M. Preissmann, *A translation of Tibor Gallai's paper: Transitiv orientierbare Graphen*, in Perfect Graphs, J. Ramirez-Alfonsin and B. Reed, eds., John Wiley, New York, 2001, pp. 25–66.
[11] S. Olariu, *An optimal greedy heuristic to color interval graphs*, Inform. Process. Lett., 37 (1991), pp. 65–80.
[12] S. Poljak, *A note on stable sets and colorings of graphs*, Comment. Math. Univ. Carolin., 15 (1974), pp. 307–309.
[13] J. P. Spinrad, *Efficient Graph Representations*, Fields Inst. Monogr. 19, AMS, Providence, RI, 2003.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 119–136

EMBEDDING k-OUTERPLANAR GRAPHS INTO ℓ_1*

CHANDRA CHEKURI†, ANUPAM GUPTA‡, ILAN NEWMAN§, YURI RABINOVICH§, AND ALISTAIR SINCLAIR¶

Abstract. We show that the shortest-path metric of any k-outerplanar graph, for any fixed k, can be approximated by a probability distribution over tree metrics with constant distortion and hence also embedded into ℓ_1 with constant distortion. These graphs play a central role in polynomial time approximation schemes for many NP-hard optimization problems on general planar graphs and include the family of weighted $k \times n$ planar grids.

This result implies a constant upper bound on the ratio between the sparsest cut and the maximum concurrent flow in multicommodity networks for k-outerplanar graphs, thus extending a theorem of Okamura and Seymour [*J. Combin. Theory Ser. B*, 31 (1981), pp. 75–81] for outerplanar graphs, and a result of Gupta et al. [*Combinatorica*, 24 (2004), pp. 233–269] for treewidth-2 graphs. In addition, we obtain improved approximation ratios for k-outerplanar graphs on various problems for which approximation algorithms are based on probabilistic tree embeddings. We conjecture that these embeddings for k-outerplanar graphs may serve as building blocks for ℓ_1 embeddings of more general metrics.

Key words. metric embeddings, k-outerplanar graphs, planar graphs, low-distortion embeddings, probabilistic approximation, metric spaces

AMS subject classifications. 05C12, 05C78, 51F99, 54C25, 54E70, 68R10

DOI. 10.1137/S0895480102417379

1. Introduction.

Many optimization problems on graphs and related combinatorial objects involve some finite metric associated with the object. In particular, the shortest-path metric on the vertices of an undirected graph with nonnegative weights on the edges frequently plays an important role. While for general metric spaces such an optimization problem can be intractable, it is often possible to identify a subset of "nice" metrics for which the problem is easy. Thus, a natural approach to such problems—and one which has proved highly successful in many cases—is to *embed* the original metric into a nice metric, solve the problem for the nice metric, and finally translate the solution back to the original metric.

When the optimization problem is monotone and scalable in the associated metric (as is usually the case), it is natural to restrict one's attention to nice metrics which dominate the original metric, i.e., in which no distances are decreased. The maximum factor by which distances are stretched in the approximating metric is called the *distortion* of the embedding. Typically, the distortion translates more or less directly into the approximation factor that one has to pay in transforming the problem from

*Received by the editors November 7, 2002; accepted for publication (in revised form) August 14, 2005; published electronically March 3, 2006. A preliminary version of this paper appeared in *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2003.

http://www.siam.org/journals/sidma/20-1/41737.html

†Lucent Bell Labs, 600-700 Mountain Avenue, Murray Hill, NJ 07974 (chekuri@research.bell-labs.com).

‡Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 (anupamg@cs.cmu.edu). This work was done while the author was with Lucent Bell Labs, and was visiting the University of California, Berkeley, supported by US-Israeli BSF grant 19999325.

§Computer Science Department, University of Haifa, Haifa 31905, Israel (ilan@cs.haifa.ac.il, yuri@cs.haifa.ac.il). The work of these authors was supported in part by US-Israeli BSF grant 19999325.

¶Computer Science Division, University of California, Berkeley, CA 94720-1776 (sinclair@cs.berkeley.edu). This author's work was supported in part by NSF grants CCR-9820951 and CCR-0121555 and by US-Israeli BSF grant 19999325.

one metric to the other, so obviously we seek an embedding with low distortion. The number of applications of this paradigm has exploded in the past few years, and it has become a versatile and standard part of the algorithm designer's toolkit; see the surveys [21, 22], or the book [25, Chapter 10] for more details. These applications have also given impetus to the study of the underlying theory of finite metric spaces.

In this paper we will be concerned with embedding finite metric spaces into ℓ_1, i.e., real space endowed with the ℓ_1 (or Manhattan) metric. Low-distortion embeddings into ℓ_1 have been recognized, along with embeddings into Euclidean space ℓ_2 and into low-dimensional ℓ_∞, to be of fundamental importance in applications of the above paradigm, as well as for the underlying theory. One of several compelling reasons for studying ℓ_1 embeddings comes from their intimate connection with the maxflow-mincut ratio in a multicommodity flow network. Namely, if every shortest-path metric on a given graph with arbitrary edge lengths can be embedded into ℓ_1 with distortion at most α, then the ratio between the sparsest cut and the maximum concurrent flow for any set of capacities and demands on the graph is bounded by α [23, 4]. In fact, the connection is even stronger: if there is a metric on a graph G that incurs distortion α when optimally embedded into ℓ_1, then there is a setting of capacities and demands on the graph G that achieves a cut-flow ratio of α [19]. For more details on the sparsest cut problem, its relation to embeddings, and its application to the design of a host of divide-and-conquer algorithms, see the survey by Shmoys [32].

A related and equally important tool in algorithmic applications is the notion of approximating a finite metric by a probability distribution over dominating tree metrics [7]. A metric M' *dominates* another metric M if, for every pair $u, v \in M$, the distance between u and v in M' is no smaller than their distance in M. If a metric M is approximated by a distribution over dominating metrics, then the distortion for pair u, v is the ratio of the expected distance between them in the metric chosen according to the distribution and their distance in M. The overall (expected) distortion is defined to be the maximum distortion over all pairs of points in M. We can view these probabilistic approximations as embeddings. We use the term *embedding into random trees* to mean that we approximate a metric by a distribution over dominating tree metrics. Since every tree metric can be embedded isometrically (i.e., exactly, or with distortion 1) into ℓ_1, embedding into random trees with expected distortion α immediately implies an embedding into ℓ_1 with distortion α. As has been recognized in the work of Bartal and others [1, 7], embeddings into random trees have many applications to online and approximation algorithms. Some of these applications are not enjoyed by arbitrary ℓ_1 embeddings.

For *general* metrics the question of embeddability into ℓ_1 is essentially resolved: Bourgain [11] showed that any n-point metric can be embedded into ℓ_1 with $O(\log n)$ distortion, and this result was made algorithmic by Linial, London, and Rabinovich [23] and Aumann and Rabani [4]. A matching lower bound of $\Omega(\log n/p)$ distortion into ℓ_p-spaces was established in [23, 24] for the shortest-path metric of unit-weighted expander graphs. For the case of approximating distances by distributions over dominating trees, a line of work [1, 7, 8, 12, 15] culminated in showing that any n-point metric can be embedded into a distribution over dominating trees with distortion $O(\log n)$ [15]; the lower bound for embeddings into ℓ_1 shows that this is tight.

However, tight bounds on the distortion incurred when embedding into ℓ_1 is still not known for many important classes of graphs, including planar graphs and graphs with bounded treewidth; many such restricted classes are conjectured to be embeddable with constant distortion. Indeed, the general question of how the topology of a graph affects its embeddability into ℓ_1, and into random trees, is one of the most

important open issues in the area of metric embeddings [21, 22]. In addition to its inherent mathematical interest, this question impacts the design of approximation algorithms for many problems on restricted families of graphs and networks.

Some limited but interesting progress has been made on embedding restricted[1] metrics into ℓ_1. Rao [29] showed that the shortest-path metric of any graph that excludes $K_{r,r}$ is embeddable into ℓ_1 with distortion $O(r^3\sqrt{\log n})$. This beats the $\Omega(\log n)$ lower bound for general graphs for any constant r, and also gives $O(\sqrt{\log n})$ distortion embeddings for the classes of planar and bounded-treewidth graphs. However, Rao's approach (of first embedding these graphs into ℓ_2 and then using isometric embeddings of ℓ_2 into ℓ_1) was shown to be tight by Newman and Rabinovich [26], where a lower bound of $\Omega(\sqrt{\log n})$ distortion was shown for embedding even treewidth-2 (and hence also planar) graphs into ℓ_2.

Approaching the question from the other direction, a celebrated theorem of Okamura and Seymour [28] implies that any *outerplanar* metric can be embedded isometrically into ℓ_1.[2] However, it has been shown that outerplanar graphs are essentially the only graphs (with the exception of K_4) that are isometrically embeddable into ℓ_1 [27]. More recently, Gupta et al. [19] showed a constant distortion embedding into ℓ_1 for treewidth-2 graphs (which are essentially series-parallel graphs, and hence also planar). This was the first natural class of graphs shown to be embeddable with constant distortion strictly larger than 1. (For example, the graph $K_{2,n}$ has treewidth 2 but is not isometrically embeddable into ℓ_1; see [2] for a simple proof of this fact.)

Some but not all of the above results carry over to the more restrictive setting of embedding into random trees. In [19] it is shown how to embed outerplanar graphs into random trees with small constant distortion; note that the isometric embedding of Okamura and Seymour is *not* an embedding into random trees. On the other hand, also in [19], it is shown that even series-parallel graphs incur a distortion $\Omega(\log n)$ when embedded into random trees. Despite this limitation, it is worth pointing out that the random tree embeddings of outerplanar graphs played a key role in the development of constant distortion ℓ_1 embeddings of series-parallel graphs in [19]; the trick was to combine the special structure of the tree embeddings with judicious use of random cuts.

1.1. Results. In this paper, we extend the above line of research to a wider class of planar graphs, namely, k-outerplanar graphs for arbitrary constant k. Informally, a planar graph is *k-outerplanar* if it has an embedding with disjoint cycles properly nested at most k deep. A formal definition is given in section 2, while Figure 4.1 shows a simple example; a canonical example of a k-outerplanar family is the sequence of $k \times n$ rectangular grids. k-outerplanar graphs play a central role in polynomial time approximation schemes for many NP-hard optimization problems on general planar graphs (see, e.g., the work of Baker [6]). Our main result is the following.

THEOREM 1.1. *There exists an absolute constant $c > 1$ such that any shortest-path metric of a k-outerplanar graph can be embedded into random trees, and hence into ℓ_1, with distortion c^k. Moreover, such an embedding can be found in randomized polynomial time.*

[1] We emphasize here that our focus is on constraints imposed on metrics by the *topological* properties of the graphs on which they are defined. Thus we exclude from our discussion the extensive recent progress on embedding other types of restricted metrics, such as "negative type metrics," into ℓ_1, as in [3] and related papers.

[2] Their result deals more generally with the cut/flow ratio in planar networks where all terminals lie on a single face; this and other results where restrictions are placed on *both* the supply graph *and* the demand graph can be found in surveys by Frank [16] and Schrijver [31].

Thus, not only do such graphs embed well into ℓ_1, but they even embed well into random trees. This is in contrast to the lower bound of $\Omega(\log n)$ for treewidth-2 graphs [19] mentioned earlier.

Our result immediately implies a constant maxflow-mincut ratio for arbitrary multicommodity flow problems on k-outerplanar graphs. Additionally, because our ℓ_1 embeddings are in fact random tree embeddings, we also obtain as a byproduct improved approximation ratios for a number of algorithms for problems on k-outerplanar graphs, including the buy-at-bulk problem [5] and the group Steiner problem [17]. For any fixed k, the improvement in each case is by an $\Omega(\log n)$ factor.

We should also note that since the maximum treewidth among k-outerplanar graphs is $\Theta(k)$, our result is the first demonstration of ℓ_1 embeddings with constant distortion for a natural family of graphs with arbitrarily large (but bounded) treewidth. Indeed, k-outerplanar graphs are a natural parameterized family of planar graphs having bounded treewidth. (Note that although all treewidth-2 graphs are planar, treewidth-3 graphs include nonplanar examples such as $K_{3,3}$.)

Finally, recall that constant distortion random tree embeddings of 1-outerplanar graphs were a key ingredient in the construction of good ℓ_1 embeddings of series-parallel graphs in [19]. We are therefore optimistic that, with the addition of suitably chosen cuts, our new tree embeddings of k-outerplanar graphs may become a building block for constant distortion ℓ_1 embeddings of wider classes of graphs, such as bounded treewidth graphs or planar graphs.

1.2. Techniques. We start with the approach of trying to extend the random tree embeddings of outerplanar graphs [19] to 2-outerplanar graphs. We do not know a way to solve this problem directly. The first main idea in the paper is to identify a *subclass* of 2-outerplanar graphs that are easier to embed, namely, *Halin graphs* [20]. Informally, a Halin graph is obtained by embedding a tree in the plane and attaching a cycle around the leaves. (The formal definition can be found in section 2.) Halin graphs are useful for the following reason. Given a 2-outerplanar graph, if we remove the outer face we are left with a collection of outerplanar graphs. We can use the embedding of [19] to embed each of these outerplanar graphs into random trees with constant distortion. If we now add the outer face to this collection of trees, we obtain (essentially) a collection of Halin graphs. Hence, if we can embed Halin graphs, we can embed 2-outerplanar graphs. We are then able to extend this approach to embed any k-outerplanar graph by peeling off the outer layer and recursively embedding the inner layers.

The second main idea is a technique for embedding Halin graphs. We note that even for this deceptively simple subclass of 2-outerplanar graphs, it is apparently nontrivial to obtain constant distortion embeddings. To obtain an embedding, we resort to a subtle modification of the algorithm of Gupta [18] which showed how to remove *Steiner vertices*[3] from a tree metric with only a constant factor distortion in distances between the remaining vertices. Though seemingly unrelated to our problem (since we have a priori no Steiner vertices), this algorithm can nonetheless be applied (with suitable modifications) to the tree in the Halin graph, with the effect of reducing the Halin graph to an outerplanar graph on its leaves. This we can once again embed into random trees using [19].

[3]Given an induced metric defined on a subset of vertices of a graph, we call the vertices not belonging to this subset the *Steiner vertices*. Although we are interested only in the metric space induced on the non-Steiner vertices, the Steiner vertices might be necessary in order to define the distances between the non-Steiner vertices.

The rest of the paper is organized as follows. We first fix notation and give essential definitions in section 2. In section 3 we show how to embed Halin graphs into random trees with constant distortion. This is extended to obtain constant distortion embeddings for all k-outerplanar graphs in section 4. In the interest of clarity of exposition, we make no attempt to optimize the constants that arise in the various steps of our procedure.

2. Notation and preliminaries.

Metrics. For general background on finite metrics and embeddings, see [13] or [25, Chapter 15]. Given two metric spaces, (V, ν) and (W, μ), and a map $f : V \to W$, we define the quantities

$$\|f\| = \max_{x,y \in V} \frac{\mu(f(x), f(y))}{\nu(x, y)};$$
$$\|f^{\text{inv}}\| = \max_{x,y \in V} \frac{\nu(x, y)}{\mu(f(x), f(y))}.$$

We say that f has *contraction* $\|f^{\text{inv}}\|$, *expansion* $\|f\|$, and *distortion* $D(f) = \|f\| \cdot \|f^{\text{inv}}\|$. The *distortion between* μ *and* ν is at most r if there exists $f : V \to W$ with $D(f) \leq r$. We often consider two metrics μ and ν over the same vertex set V; in such cases, we assume that f is the identity map. Metric μ is said to *dominate* ν if for all $x, y \in V$, $\mu(x, y) \geq \nu(x, y)$.

Let $G = (V, E)$ be an undirected graph. A metric (V, μ) is *supported on* (or *generated by*) G if it is the shortest-path metric of G w.r.t. some nonnegative weighting of the edges E. Given a graph G with edge weights $w(\cdot)$, d_G denotes the shortest-path metric of G, and we assume that the edge weights satisfy $w(e) = d_G(x, y)$ for $e = \{x, y\} \in E$ unless otherwise stated.

For $S \subseteq V$, the *cut metric* $\delta_S(x, y)$ is defined to be 1 if $|S \cap \{x, y\}| = 1$, and 0 otherwise. It can be shown that a metric is isometrically embeddable into ℓ_1 iff it can be written as a nonnegative linear combination of cut metrics [13].

A metric d_G supported on a graph G is *α-probabilistically approximated* by a distribution $\mathcal{D}$ over trees if the following conditions hold:

1. Each tree T in the distribution $\mathcal{D}$ has $V(G) \subseteq V(T)$.
2. For each tree T in the distribution, the metric d_T *dominates* the metric d_G; i.e., for all nodes $x, y \in V(G)$, $d_G(x, y) \leq d_T(x, y)$.
3. For all $x, y \in V(G)$, the expected distance $\mathrm{E}_{\mathcal{D}}[d_T(x, y)] \leq \alpha \cdot d_G(x, y)$.

We shall also refer to this as an *embedding of G with distortion α into random trees.* (The fact that the distortion is only in expectation will often not be mentioned.) It is known that general graphs can be embedded into random trees with distortion $O(\log n)$ [7, 15].

We state two simple propositions (whose proofs we omit) which we will use extensively in what follows. The first allows us to embed each block (maximal 2-vertex connected subgraph) of a graph separately; the second says that we may always replace a subgraph by its tree embedding without further loss.

PROPOSITION 2.1. *Suppose G has a cut-edge whose removal results in a tree T and a graph H. If H can be embedded into random trees with distortion α, then so can G.*

PROPOSITION 2.2. *Let $H = (V_H, E_H)$ be a subgraph of $G = (V, E)$. Let $H_1, H_2, \ldots, H_s$ be graphs on V_H such that $d_H(u, v) \leq d_{H_i}(u, v) \leq \alpha_i \cdot d_H(u, v)$ for all $u, v \in V_H$, $1 \leq i \leq s$. Then in the graph $G_i = (V, (E \setminus E_H) \cup E_{H_i})$, we*

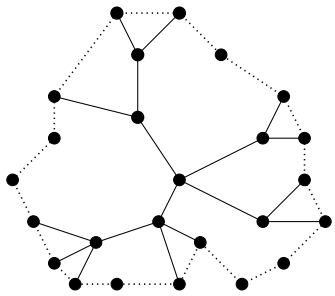


FIG. 2.1. *A Halin graph, with the tree $T = (V, E)$ in solid lines and the cycle $C = (U, E_C)$ in dashed lines.*

have $d_G(u,v) \le d_{G_i}(u,v) \le \alpha_i \cdot d_G(u,v)$ for all $u, v \in V$. Moreover, consider a random variable X taking values in $\{1, 2, \ldots, s\}$, where $\mathbf{Pr}[X = i] = \mu_i$, and let $\alpha = \mathbf{E}[\alpha_X] = \sum_{i=1}^{s} \mu_i \alpha_i$. Then, for any pair $u, v \in V$, the expected distance between u and v in the random graph G_X is at most $\alpha d_G(u,v)$.

Graph-theoretic terms. A graph G' is a *minor* of G if G' is obtained from G by a sequence of edge deletions and contractions. A class of graphs is *closed* under taking minors if for every graph G in the class all its minors are also in the class. For example, planar graphs are minor-closed.

For a formal definition of *treewidth*, the reader is referred to standard graph theory texts such as [14, 34]. Informally, a graph has treewidth k if it can be decomposed recursively by vertex separators where the size of the vertex separator at each stage is at most k.

An embedding in the plane of a graph G is *outerplanar* (or 1-*outerplanar*) if it is planar and all vertices lie on the unbounded face. An embedding of a graph G is k-*outerplanar* if it is planar, and deleting all the vertices on the unbounded face leaves a $(k-1)$-outerplanar embedding of the remaining graph. A graph is k-outerplanar if it has a k-outerplanar embedding. It is known that a k-outerplanar graph has treewidth at most $3k - 1$ [10, 30]; other properties of these graphs and related concepts can be found in [6, 10]. Given a planar graph, a k-outerplanar embedding for which k is minimal can be found in polynomial time [9].

A *Halin graph* [20] is obtained by taking a planar embedding of a tree $T = (V, E)$ and attaching a cycle $C = (U, E_c)$ around the leaves of the tree (in order). If the set of leaves of T is denoted by L, then $V \cap U = L$; note that $U \setminus L$ may not be empty and hence there may be vertices on the cycle C that are not leaves of T. (See Figure 2.1 for an example.) It is known that any Halin graph $G = (V \cup U, E \uplus E_c)$ is 2-outerplanar and has treewidth 3. Many algorithmic problems can be solved efficiently on these graphs (see, e.g., [33] and the references therein). We note that while Halin graphs (as defined here) are not minor-closed, we will not need this property in our algorithms.

3. Embedding a Halin graph. The goal of this section is to prove the following theorem.

THEOREM 3.1. *The shortest-path metric of a Halin graph can be embedded into random trees with distortion at most* 200.

Before embarking on the proof, we give a high-level sketch of our strategy. Given a Halin graph consisting of a tree T and a cycle C, we first process the tree T to obtain a random dominating tree $T^{(1)}$, which approximates distances in T to within a constant factor (in expectation). Furthermore, the tree $T^{(1)}$ has a specific structure:

it consists of a tree $T'' = (L, E'')$ on just the *leaves* L of the original tree T, and the rest of the vertices in $V \setminus L$ lie in subtrees that are attached to vertices in T''. Since we can ensure moreover that the tree T'' is a minor of T, attaching the cycle C back to the vertices in T'' gives us an outerplanar graph. Finally, this outerplanar graph is embedded into random trees with constant distortion using known techniques [19].

We will describe the tree processing procedure (which is the main content of the section) in section 3.1, and in section 3.2 we will explain how to use this to reduce to the outerplanar case.

3.1. Processing the tree. We assume that the tree T is rooted at a root vertex $r \in (V \setminus L)$. This imposes, in the usual manner, an ancestor-descendant relation between the vertices in V. Each vertex v naturally defines a tree $T(v)$, namely, the subtree induced by the vertices that are descendents of v. We will use the following parameters extensively in what follows.

DEFINITION 3.2. *For a vertex $v \in V$, let $l(v)$ be a leaf in $T(v)$* closest *to v, and let $h(v)$ be the distance of v from $l(v)$ in T.*

Note that these functions $h(v)$ and $l(v)$ are fixed given the rooted tree T. Let us first give a brief overview of the processing algorithm, which has two conceptual parts.

- The first step of the algorithm, given in section 3.1.1, returns a tree $T^{(1/2)}$. This tree consists of a tree T' defined on the vertices of L plus some extra (or *Steiner*) vertices, and the vertices of $V \setminus L$ hang off the vertices of T' in the form of (possibly several) subtrees. This is done while incurring a constant expected distortion.
- The second part of the processing, given in section 3.1.2, eliminates the Steiner vertices of T' by contracting some of its edges to yield a tree T'' defined only on the leaves L. As a result, $T^{(1/2)}$ is converted into a tree $T^{(1)}$ with the properties claimed above. This part is shown to incur a further constant factor distortion.

3.1.1. Processing I: Constructing the tree $T^{(1/2)}$. In this section, we will show how to convert the tree T into the tree $T^{(1/2)}$ while incurring only a constant distortion. The procedure Process-Tree to perform this processing cuts off a subtree $\widehat{T}_0$ of T which contains the root but none of the leaves, recursively acts on the subtrees thus created, makes a new root vertex and adds edges from it to the roots of each of the processed subtrees, and finally hangs $\widehat{T}_0$ off this new root. (See Figures 3.3 and 3.4.)

Before we make Process-Tree concrete, we define the auxiliary procedure Cut-Midway in Figure 3.1. This procedure takes as input a tree T which has root r and a set L of leaf nodes. It then cuts a *random* set of edges to separate r from all the leaves in L; in particular, it returns a special tree $\widehat{T}_0$ containing the root r and none of the nodes in L, and a set of subtrees T_i, $1 \le i \le t$, each rooted at some vertex r_i, which between them contain the leaves L. We say that an *edge $e = \{u, v\}$ is at distance d from a vertex r* if e is in the cut defined by the set of vertices whose distance from r is at most d, i.e., if $\mathbf{B}(r, d) \cap \{u, v\}$ has exactly one vertex. (Here $\mathbf{B}(r, d) = \{x \mid d_T(r, x) \le d\}$ is the ball of radius d around the node r.) It should be noted that, in each interation of Cut-Midway, the set $\bar{L}$ decreases in size and the parameter d increases by at least a factor of 2.

The procedure Process-Tree, which outputs a tree $T^{(1/2)}$, is given in Figure 3.2. In this tree $T^{(1/2)}$, we denote by T' the portion formed by the new edges added between r' and r'_i (for $1 \le i \le t$) during the various recursive calls to Process-Tree. (Note that

while a path remains in T from the root r to a vertex in L
  let $\bar{L} \leftarrow$ vertices in L still reachable in T from r
  let $d \leftarrow$ distance in T to the closest vertex in $\bar{L}$
  let $S(d) \leftarrow \{x \in \bar{L} \mid d_T(r,x) \in [d, 2d)\}$
  let $T(d) \leftarrow$ union of the paths from r to vertices in $S(d)$
  choose $D \in_R [d/2, 3d/4)$ uniformly at random
  $E(d) \leftarrow$ edges in $T(d)$ at distance D from r
  delete the edges in $E(d)$ from T
end while
let $\widehat{T}_0 \leftarrow$ component of T containing root r but no leaves of T
let $T_1, T_2, \ldots, T_t \leftarrow$ other components of T
let $d_i \leftarrow$ value of d when edge connecting r to T_i was cut
return $(\widehat{T}_0; \langle T_1, d_1\rangle, \langle T_2, d_2\rangle, \ldots, \langle T_t, d_t\rangle)$

FIG. 3.1. *Procedure* Cut-Midway(T).

apply Cut-Midway(T) to get
    $(\widehat{T}_0, \langle T_1, d_1\rangle, \langle T_2, d_2\rangle, \ldots, \langle T_t, d_t\rangle)$
let r' be a new vertex, called the "Steiner twin" of T's root r
attach r' to r with an edge of length $d_0 = h(r)$
for $1 \le i \le t$ *// We do not have to work on* $\widehat{T}_0$
  if T_i is just a single vertex x (hence $x \in L$) **then**
    let $T_i^{(1/2)} \leftarrow T_i$
  else
    let $T_i^{(1/2)} \leftarrow$ Process-Tree(T_i)
  let r_i' be the root of $T_i^{(1/2)}$
    // r_i' is the Steiner twin of r_i, the root of T_i
  add edge $\{r', r_i'\}$ with length $3d_i$
end for
return tree $T^{(1/2)}$ with r' as its root

FIG. 3.2. *Procedure* Process-Tree(T).

this does not include the edges added between r' and r, i.e., between the original roots and their Steiner twins.) For an example see Figure 3.3, where Cut-Midway performed three cuts, and Process-Tree resulted in the tree in Figure 3.4. The solid edges in the latter tree belong to T, the dashed ones belong to T', and the edge $\{r, r'\}$ is shown as a faint line. We remark that T' includes all the leaves of T, plus all the Steiner twins created during Process-Tree.

Let us call an edge a *candidate to be cut* at some step if it has a nonzero probability of being cut at that step. We show the following bound on the expected distortion incurred by Process-Tree in passing from T to $T^{(1/2)}$.

THEOREM 3.3. *The (expected) distortion introduced by procedure* Process-Tree *is at most* 25.

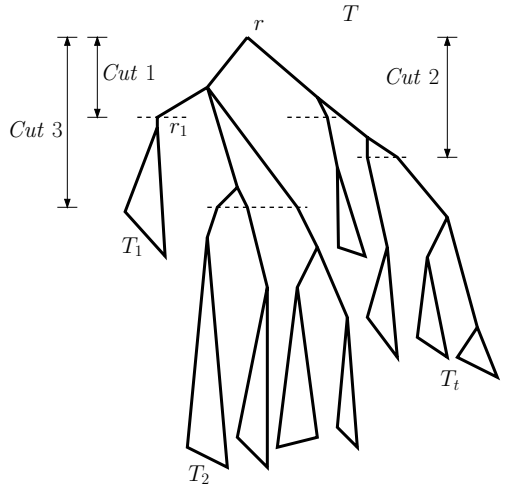


FIG. 3.3. *Cuts obtained by an invocation of* Cut-Midway *on a tree* T.

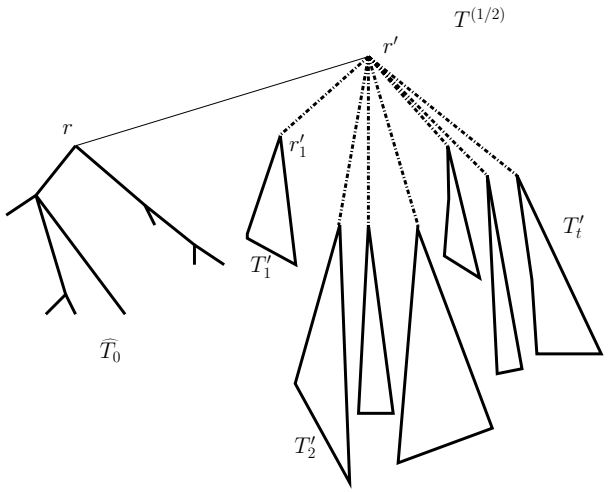


FIG. 3.4. *The tree* $T^{(1/2)}$ *output by* Process-Tree *on the tree* T *from Figure* 3.3. *The dotted lines indicate edges in* T'.

Proof. We first give a high-level sketch. The construction of the tree $T^{(1/2)}$ ensures that distances are not contracted by Process-Tree; the algorithm explicitly ensures this in Process-Tree by the distances it chooses to connect the root r' to each r'_i. Hence it suffices to bound the expected expansion of distances. We do this via two lemmas: first, Lemma 3.4 shows that an edge is a candidate to be cut on at most two (consecutive) occasions. Lemma 3.5 then shows that, when an edge is a candidate to be cut, it suffers only a constant expected expansion. Combining these two results then gives us Theorem 3.3.

LEMMA 3.4. *No edge is a candidate to be cut more than twice during the entire run of the procedure* Process-Tree.

Proof. Let $e = \{u, v\}$ be an edge with u being the parent of v. Consider the first instant in time when the edge e is a candidate to be cut in a call to Cut-Midway.

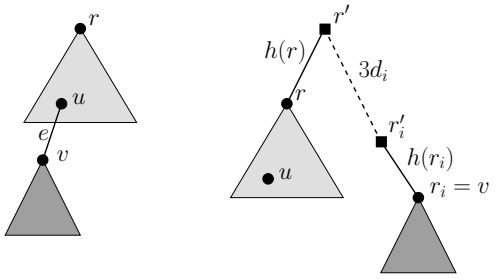


FIG. 3.5. *Illustration for proof of Lemma* 3.5.

Let r be the root at this time, and d^* be the value of the parameter d in the **while** loop of this call to Cut-Midway. In this call of Cut-Midway, it is clear that e cannot be a candidate again. Indeed, after the cut, e will not lie on any path from r to a leaf. A fact that will be useful later is that the portion of e that lies in the distance interval $[d^*/2, 3d^*/4)$ from r is $(\min(d_T(r,v), 3d^*/4) - \max(d_T(r,u), d^*/2))$, and this value multiplied by $4/d^*$ is the probability that e is cut at this time.

The edge e will never be a candidate again if the cut fell "below" v, or if it passed through e, so let us assume that the cut was above u and thus e lies in one of the trees T_i with root r_i. In this case the tree T_i will be passed on to Cut-Midway by Process-Tree. Now e clearly lies on some path from r_i to a leaf, and hence it may be a candidate to be cut again. Let d^{**} be the value of the parameter d in Cut-Midway when this happens for the first time after T_i is formed.

We claim that the cut made at this point must fall below u; i.e., $d^{**}/2 \geq d_T(r_i, u)$. Indeed, such a cut is made at a distance at least $d^{**}/2 = h(r_i)/2$ from the root r_i, where $h(r_i) \geq d^* - d_T(r, r_i)$. Hence, taking distances from r, this cut is at distance at least $d_T(r, r_i) + h(r_i)/2 \geq \frac{1}{2}(d_T(r, r_i) + d^*) \geq 3d^*/4$. But this distance is greater than $d_T(r, u)$, and hence u always lies above this next cut. Thus, when this next cut is made, either e will be deleted (if v lies below this cut), or the cut will fall below v and the edge e will never again be a candidate to be cut, proving the lemma.

Before we end, let us note that the portion of e that lies in distance interval $[d^{**}/2, 3d^{**}/4)$ is disjoint from the portion considered earlier and has a length of at most $\max(d_T(r,v) - 3d^*/4, 0)$. As before, multiplying this by $4/d^{**}$ gives the probability that e is cut if it is a candidate a second time. □

LEMMA 3.5. *Let $e = \{u, v\}$ be an edge in G of length ℓ_e. If e is cut by* Cut-Midway *with parameter d_i, the expected distance between u and v in $T^{(1/2)}$ is at most $6d_i - \ell_e$.*

Proof. Consider an edge $e = \{u, v\}$ of length ℓ_e, with u the parent of v, which is cut in some iteration of Cut-Midway, and let d_i be the value of the parameter d at this time. Consider the distance $d_{T^{(1/2)}}(u, v)$ between u and v in the resulting tree $T^{(1/2)}$.

The vertex u will be in $\widehat{T}_0$ and the vertex v is the root r_i of T_i for some i and hence will be in $T_i^{(1/2)}$ when T_i is processed. From the description of Process-Tree we see that $d_{T^{(1/2)}}(u, v) = d_{T^{(1/2)}}(u, r_i)$ can be expressed as $d_T(u, r) + d_{T^{(1/2)}}(r, r') + d_{T^{(1/2)}}(r', r'_i) + d_{T^{(1/2)}}(r'_i, r_i)$ (see Figure 3.5). From our construction, $d_{T^{(1/2)}}(r, r') = h(r)$, $d_{T^{(1/2)}}(r', r'_i) = 3d_i$, and $d_{T^{(1/2)}}(r'_i, r_i) = h(r_i)$. We observe that $h(r) \leq d_i$ for all i, and that $h(r_i) \leq 2d_i - d_T(r, r_i)$; the latter inequality holds because for e to be cut, r_i must lie on the path from r to a leaf in T of length at most $2d_i$. Note that this calculation also holds in the special case that v is a leaf (when $0 = h(r_i) \leq$

$2d_i - d_T(r, r_i)$).

Putting all these observations together we obtain

$$\begin{aligned} d_{T^{(1/2)}}(u, r_i) &= d_T(u, r) + d_{T^{(1/2)}}(r, r') + d_{T^{(1/2)}}(r', r'_i) + d_{T^{(1/2)}}(r'_i, r_i) \\ &\le d_T(u, r) + h(r) + 3d_i + (2d_i - d_T(r, r_i)) \\ &\le d_T(u, r) + d_i + 3d_i + (2d_i - d_T(r, r_i)) \\ &\le d_T(u, r) - d_T(r, r_i) + 6d_i \\ &= 6d_i - \ell_e. \qquad \square \end{aligned}$$

Now we complete the proof of Theorem 3.3. By Lemma 3.4, the edge $e = \{u, v\}$ is a candidate to be cut at most twice. From the proof of Lemma 3.4, the first time it is a candidate it is cut with probability

$$p_1 = (\min(d_T(r, v), 3d^*/4) - \max(d_T(r, u), d^*/2)) \times 4/d^*;$$

and, by Lemma 3.5, if it is cut, the expected distance between u and v becomes at most $6d^* - \ell_e$. Similarly, the second time the chance of e being cut is

$$p_2 = (\max(d_T(r, v) - 3d^*/4, 0)) \times 4/d^{**},$$

and the expected distance is $6d^{**} - \ell_e$. Finally, the distance remains unchanged at ℓ_e with the remaining probability $(1 - p_1 - p_2)$. Putting these together, we get that the expected distance between u and v after procedure Process-Tree is at most

$$\begin{aligned} 6d^*\, p_1 + 6d^{**}\, p_2 + (1 - 2p_1 - 2p_2)\ell_e & \\ &\le 6(d^*\, p_1 + d^{**}\, p_2) + \ell_e \\ &\le 24\, \big(\min(d_T(r, v), 3d^*/4) - \max(d_T(r, u), d^*/2) \\ &\qquad + \max(d_T(r, v) - 3d^*/4, 0)\big) + \ell_e \qquad (3.1) \\ &\le 24\, \big(d_T(r, v) - \max(d_T(r, u), d^*/2)\big) + \ell_e \qquad (3.2) \\ &\le 24\, \big(d_T(r, v) - d_T(r, u)\big) + \ell_e \\ &\le 24\, \ell_e + \ell_e = 25\, \ell_e, \end{aligned}$$

where we used the simplification $\min(x, y) + \max(x - y, 0) = x$ to obtain (3.2) from (3.1). Thus the expected distortion is at most 25, which proves the theorem. $\square$

Recall that the tree $T^{(1/2)}$ constructed by the procedure Process-Tree includes a tree T' containing the leaves L of the original tree T; we close this subsection with a further observation about T'.

CLAIM 3.6. *The tree T' can be obtained from tree T by edge contractions.*

Proof. In each call to Process-Tree, we progressively construct T' by removing the tree $\widehat{T}_0$ and replacing it with a star connecting r' to the various r_i (for $1 \le i \le t$). But this star could equivalently be obtained by contracting all the edges of the tree $\widehat{T}_0$. (Of course, we are placing new lengths on the remaining edges, but this does not affect the topology.) $\square$

Since L is also the set of leaves of T', and the edge contractions can be performed without changing the planar layout of the trees, adding the cycle C around the leaves of T' also gives us a Halin graph.

3.1.2. Processing II: Removing the Steiner vertices. In this section, we remove the Steiner vertices in the tree T' that were created during runs of Process-Tree, giving us a tree T''. Since $T^{(1/2)}$ consists of T' with several subtrees attached to it via cut-edges, attaching those subtrees to T'' will give us a new tree $T^{(1)}$. The argument in this section is similar in spirit to that in [18]. The Steiner twin vertices from $T^{(1/2)}$ are removed in the same order in which they were created. Consider r', the root of T'; it was created as the Steiner twin of vertex $r \in T$. We now identify all vertices on the path between r' and $l(r)$ with $l(r)$. This process is performed on each of the Steiner twin vertices in turn (in order of their creation), causing each of them to be identified with some vertex in $L \subseteq C$. Call the resulting tree $T^{(1)}$. This has vertex set V, since we removed all the Steiner vertices we created in the previous section. The following lemma proves the main result of this section.

LEMMA 3.7. *The edge-contraction procedure described above ensures that the distance between each pair of vertices of V in $T^{(1)}$ is no shorter than its distance in T.*

Proof. To show that there is no contraction, it suffices to check that no *edge* in $T^{(1)}$ is shorter than the distance between its endpoints in T. There are just three kinds of edges remaining in $T^{(1)}$: those which belong to the trees $\widehat{T}_0$ in the various invocations of Process-Tree, those between some r and $l(r)$,[4] and those between $l(r)$ and $l(r_i)$. Note that the edges of this last type are the only edges that exist between $l(r_a)$ and $l(r_b)$, since such edges (without loss of generality) must be caused by r_a being the root at some invocation of Process-Tree and r_b being one of the r_i's created at this step, and r_a later being identified with $l(r_a)$.

Clearly, the edges in the trees $\widehat{T}_0$ are not changed at all. Now consider an edge between a vertex $l(r)$ and r. The length of this edge in T'' is just $h(r)$, which is also the distance between $l(r)$ and r in T. Finally, for an edge between $l(r)$ and $l(r_i)$ in $T^{(1)}$, the length is just $6d_T(r, r_i)$. However, the distance between these points in T is at most $d_T(r, l(r)) + d_T(r, l_{T_i}(r_i))$, which we upper bound next. Let d^* be the value of d when r_i was separated from r in the procedure Cut-Midway. Then it follows that $d_T(r, l(r)) = h(r) = d^*$; furthermore, the distance $d_T(r, l_{T_i}(r_i)) \leq 2d^*$, since r_i must lie on a root-leaf path of length at most $2d^*$. Hence the distance between $l(r)$ and $l(r_i)$ in T is at most $3d^*$. However, $d_T(r, r_i) \geq d^*/2$, so the distance is at most $6d_T(r, r_i)$ as required. ∎

3.2. Wrapping it all up. We now complete the proof of Theorem 3.1. Let G be the given Halin graph, consisting of a tree $T = (V, E)$ and a cycle $C = (U, E_c)$ around the leaves $L = V \cap U$ of T. We have seen how to transform T into a tree $T^{(1)}$ that consists of a tree $T'' = (L, E'')$ and a collection of trees $T_1, T_2, \ldots, T_j$ each of which is connected by an edge to a vertex in L. Every vertex in $V - L$ is contained in exactly one of $T_1, T_2, \ldots, T_j$. Moreover, the tree T'' is a minor of T. We have also seen that $T^{(1)}$ dominates T and that the expected expansion for any pair in T is at most 25. Now consider the graph $G^{(1)}$ obtained by adding the cycle C to the tree $T^{(1)}$. Let G' be the graph obtained by adding C to T''. (See Figure 3.6.) We claim that G' is an outerplanar graph. Assuming for the moment that this claim is true, we show how we can embed G into trees with the claimed distortion.

First, from Proposition 2.2, it follows that $G^{(1)}$ dominates G and for every pair $u, v \in V_G$, the expected distance in $G^{(1)}$ is at most $25d_G(u, v)$.

Next, note that $G^{(1)}$ consists of G' with the trees $T_1, T_2, \ldots, T_j$ connected to G'

[4] These edges were added between r and r', and the latter has been identified with $l(r)$.

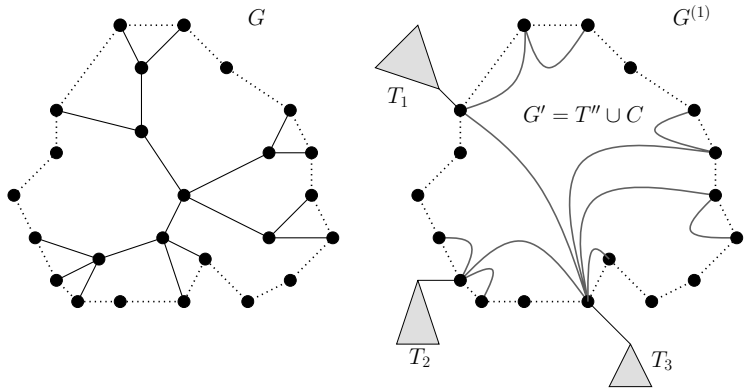


FIG. 3.6. *G is a Halin graph; G' is an outerplanar graph obtained from $T'' \cup C$, and $G^{(1)}$ is obtained by adding the trees T_i to G'.*

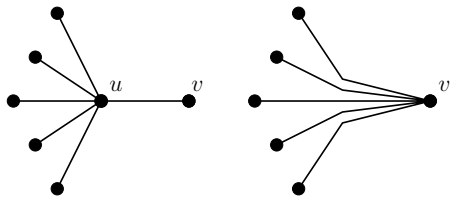


FIG. 3.7. *Contracting edge $\{u, v\}$ and removing u. Obtaining contours for new edges.*

by cut-edges. From Proposition 2.1 it follows that embedding G' into random trees with distortion α produces an embedding of $G^{(1)}$ into random trees with distortion α. Since G' is an outerplanar graph, we can invoke the procedure of [19, Theorem 5.2] to get a random subtree of G' which approximates distances in G' (in expectation) to within a factor of 8. Thus $G^{(1)}$ can be embedded into random trees with distortion 8.

Finally, from Proposition 2.2 we see that embedding $G^{(1)}$ into random trees with distortion 8 implies that G can be embedded into random trees with distortion $8 \cdot 25 = 200$. This completes the proof of Theorem 3.1.

It remains to sketch the proof that G' is outerplanar, as was claimed above. From Claim 3.6, T' is a minor of T, and hence T'', which is obtained by contracting some edges in T', is also a minor of T. Moreover, since no two vertices of L are merged in obtaining T'', G' is a minor of G. Thus we can obtain G' from G by a sequence of edge deletions and contractions. This allows us to obtain an outerplanar embedding of G' from the given planar embedding of G as follows. First, remove any edges of G that are removed in obtaining G'. Then consider the first edge $\{u, v\}$ that is contracted in G. Vertices u and v cannot both be in L, so let u be the vertex outside of L. Let $u_1, u_2, \ldots, u_h$ be the neighbors of u that are not v. The edge $\{u_i, u\}$ is a contour in the planar embedding of G. When $\{u, v\}$ is contracted we remove u and extend the edge $\{u_i, u\}$ to $\{u_i, v\}$. By duplicating the contour of $\{u, v\}$ h times and shifting the resulting contours infinitesimally we can obtain new contours for the edges $\{u_1, v\}, \ldots, \{u_h, v\}$. (See Figure 3.7.) Thus we obtain a planar embedding of the graph with the edge $\{u, v\}$ contracted without changing the position of v. Thus all the vertices remain in their original positions and any edge $\{x, y\}$ that is not contracted or deleted has its contour intact. We can continue this process and obtain a planar embedding of G' such that the vertices $U \supseteq L$ and the contours of edges

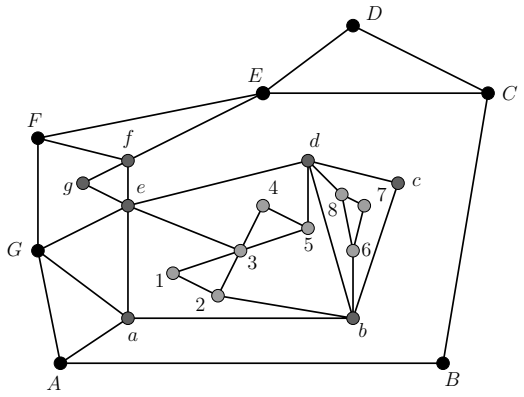


FIG. 4.1. *A 3-outerplanar graph from* [6]. *The three layers A-G, a-g, and 1-8 are shown using different shades of gray.*

in E_c are unchanged from the planar embedding of G that we started with. Since all the vertices of G' are on the outer face C, it follows that we have an outerplanar embedding of G'.

4. On to *k*-outerplanar graphs. In this section, we extend the construction of the previous section to k-outerplanar graphs. Recall that these are graphs embeddable in the plane which are dismantled by k repetitions of the process of removing the vertices on the outermost face. (See Figure 4.1 for an example.)

The main result of this section, and of the paper, is the following.

THEOREM 4.1. *There is a universal constant c such that the shortest-path metric of a k-outerplanar graph can be embedded into random trees with distortion c^k.*

Proof. We begin with a high-level sketch of the proof, which proceeds by induction on k. Since G is k-outerplanar, removing the outer face of G decomposes it into a set of $(k-1)$-outerplanar subgraphs $G_1, \ldots, G_\ell$. Each G_i resides inside a face F_i of the graph induced by the vertices of the outer face of G. (See Figure 4.2.) By the induction hypothesis, each G_i can be embedded into random trees with distortion c^{k-1}; moreover, this can be done leaving the vertices on the outer face of G_i in their original positions. Replacing G_i by its corresponding tree T_i yields a *Halin* graph whose outer cycle is the face F_i (plus possibly some trees attached to internal nodes of T_i); see Figure 4.3. Now the procedure of section 3 can be used to embed this Halin graph into an *outerplanar* graph on F_i (plus some attached trees) with constant distortion c_1. Finally, the union (over i) of all these outerplanar graphs is again outerplanar and so by [19, Theorem 5.2] can be embedded into random trees with constant distortion c_2. The overall distortion incurred in this process is $c^{k-1} \cdot c_1 \cdot c_2 \le c^k$ if we choose $c = c_1 c_2$.

Remark. The reader may recall from section 3 that we can take $c_1 = 25$ and $c_2 = 8$ in the above. Hence Theorem 4.1 holds with the constant $c = 200$.

We now proceed to spell out the details of the above argument. We begin with the induction hypothesis, which needs to be slightly stronger than the statement of the theorem. We assume $G = (V, E)$ is given along with its k-outerplanar embedding, and $F_0(G)$ is the set of vertices on the outer face of G. (In what follows, we will often abuse notation and blur the distinction between a face and the vertices that lie on it.)

> **Induction hypothesis.** Let $G = (V, E)$ be a connected k-outerplanar graph with $F_0(G)$ as the outer face in some k-outerplanar em-

bedding. Then the shortest-path metric of G can be probabilistically approximated by a collection of trees on V with expected distortion at most c^k. Moreover, for each subtree T in this distribution, the vertices of the outer face $F_0(G)$ induce a (connected) subtree that is a minor of G.

The importance of the extra condition placed on the trees T is the following. Let T' be the subtree induced by the vertices of $F_0(G)$; note that the vertices of $V \setminus T'$ reside in subtrees hanging off T' by single edges. Since T' is a minor of G, we can construct it by edge deletions and contractions while leaving the vertices of $F_0(G)$ in their original positions, as explained in section 3.2. This allows us in the induction to replace G by T' without disturbing the outer face $F_0(G)$.

The base case for the induction is $k = 1$, when G is an outerplanar graph. For outerplanar graphs, [19, Theorem 5.2] shows an embedding of G into trees that are *subgraphs* of G with constant distortion (at most 8). Being subgraphs these trees are certainly minors, so the extra condition in the induction is satisfied.

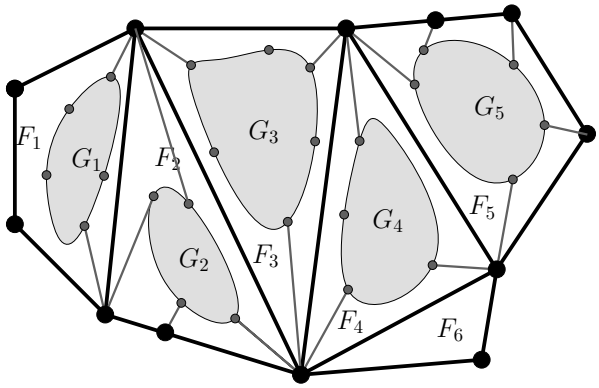


FIG. 4.2. *Partitioning of a k-outerplanar graph G into $(k-1)$-outerplanar graphs $G_1, \ldots, G_5$. The bold lines indicate G_F, the graph induced by the outer face.*

For the induction step, we may assume that G is 2-vertex connected; otherwise we can work with each block of G separately. Let G_F be the subgraph of G induced by $F_0(G)$, the vertices on its outer face; clearly G_F is an outerplanar graph. (See Figure 4.2.) Let $F_1, F_2, \ldots, F_\ell$ be the internal faces of G_F, V_i the subset of $V \setminus F_0(G)$ lying inside the face F_i, and G_i the induced graph on V_i. We assume without loss of generality that G_i is connected, since otherwise we can work with its connected components separately. We make the following assumption for technical reasons: for any vertex $v \in F_i$ there is at most one vertex $u \in V_i$ such that $\{u, v\} \in E$. This is without loss of generality, since if it does not hold for a vertex $v \in F_i$, we can split v into a path of vertices (with the edges between them of length 0) and connect each one to a unique vertex of V_i without violating planarity. Note the following fact, which allows us to use the induction hypothesis.

FACT 4.2. *For $1 \le i \le \ell$, G_i is a $(k-1)$-outerplanar graph.*

Thus, by the induction hypothesis, each G_i can be c^{k-1}-probabilistically approximated by trees satisfying the extra condition. We now give a procedure to extend the embeddings of the various G_i to an embedding of G. For $1 \le i \le \ell$, we independently pick a tree T_i from the distribution over tree metrics for G_i. Let G' be the graph obtained by adding the vertices of $F_0(G)$ and the edges incident to them (in G) to

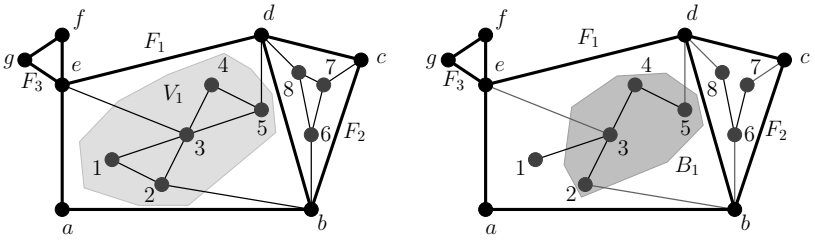


FIG. 4.3. *Returning from the induction: The bold lines denote G_F. The lightly shaded region on the left denotes the vertices V_1 corresponding to the face $F_1 = (abde)$, and the darker shaded region on the right denotes the set B_1. Note that $A_1 = \{2, 3, 5\}$, $B_1 \setminus A_1 = \{4\}$, and $V_1 \setminus B_1 = \{1\}$.*

the trees $T_1, \ldots, T_\ell$. Proposition 2.2 implies that the metric induced by G' is within expected distortion c^{k-1} of d_G, and hence approximating G' by tree metrics with an expected distortion of c will prove the induction hypothesis for G.

Let T'_i be the subtree of T_i that is induced by $F_0(G_i)$; the fact that it is a tree is guaranteed by the extra condition in the induction hypothesis. Let A_i be the vertices in V_i that have an edge to some vertex in F_i; since G is planar, $A_i \subseteq F_0(G_i)$. Let T''_i be the *minimal* connected subtree of T'_i that contains A_i. Let B_i the vertices in T''_i. (Note that B_i may contain vertices not in A_i but by minimality of T''_i, any vertex in $B_i \setminus A_i$ is an internal vertex of T''_i.) The remaining vertices, in $V_i \setminus B_i$, induce a forest in T_i that is connected via cut-edges to T''_i. (An example is given in Figure 4.3.) Using Proposition 2.1, we can eliminate the vertices in $V_i \setminus B_i$ (for $1 \le i \le \ell$) from G'. It now suffices to embed the resulting graph into trees with expected distortion at most c.

The key claim that reduces this problem to the embedding of Halin graphs given in the previous section is the following (see Figure 4.3).

CLAIM 4.3. *Let G'_i be obtained by adding to the tree T''_i the vertices F_i and the edges in G connecting F_i to A_i. Then G'_i is a Halin graph with cycle F_i.*

Proof. By the induction hypothesis, the tree T'_i is a minor of G_i. Since T''_i is a subtree of T'_i it is also a minor of G_i and hence, as in section 3.2, the planar embedding of G_i induces a natural planar embedding of T''_i. Furthermore, by our earlier assumption, each vertex of F_i has at most one edge to T''_i; let E_i be the set of these edges. It follows that T''_i along with these edges E_i still forms a tree. Since T''_i was chosen to be minimal, the leaves in T''_i are a subset of A_i. Therefore the leaves in the tree formed by adding E_i to T''_i are precisely the vertices of F_i incident to an edge in E_i. The edges along the face F_i form a cycle around these leaves, yielding a Halin graph. ∎

Now we can apply the procedure of section 3 to G'_i (omitting the final step of embedding the outerplanar graph into trees). The resulting graph, which we call G''_i, will be an outerplanar graph on F_i, with the vertices of T''_i attached as subtrees; the expected distortion will be at most 25. Using Proposition 2.1 again, we can remove these hanging subtrees to obtain the graph $\mathsf{core}(G''_i)$.

Note that the procedure in section 3 guarantees that $\mathsf{core}(G''_i)$ is a minor of G'_i. Furthermore, each $\mathsf{core}(G''_i)$ is an outerplanar graph on the face F_i of the outerplanar graph G_F. These two facts together imply that $H = \bigcup_i \mathsf{core}(G''_i)$ is also an outerplanar graph. Thus we can use [19, Theorem 5.2] to embed H into random subtrees of H with expected distortion at most 8. Choosing $c = 25 \cdot 8 = 200$, we conclude that G' can be embedded into random trees with expected distortion at most c, and hence

that G can be embeddded with expected distortion at most c^k, as required.

To complete the inductive proof, it remains to verify that the random trees produced by the above procedure satisfy the extra property stated in the induction hypothesis, namely, that the vertices of $F_0(G)$ form a subtree that is a minor of G. The final step of the procedure constructs a subtree T_H of the graph H whose vertex set is exactly $F_0(G)$. Now observe that the procedure discards vertices only when they induce a subtree attached to the rest of the graph (invoking Proposition 2.1 on each occasion to ensure that this introduces no additional distortion). Thus the final tree consists of T_H with other subtrees hanging off it. To see that T_H is a minor of G, it suffices to show that H is a minor of G since T_H is a subtree (and hence a minor) of H. But $H = \bigcup_i \mathsf{core}(G_i'')$, and we already observed above that each $\mathsf{core}(G_i'')$ is a minor of G_i'. Furthermore, G_i' is formed by replacing G_i by the tree T_i'' inside the face F_i, and T_i'' is a subtree of T_i' and hence a minor of T_i'. And we know from the induction hypothesis that T_i' is a minor of G_i; hence G_i' is a minor of G_i. This implies that $\mathsf{core}(G_i'')$ is a minor of G_i, and hence that H is a minor of G, as required. This completes the inductive proof of Theorem 4.1. □

Acknowledgments. We thank Amit Chakrabarti and Amit Kumar for useful discussions and the referees for their suggestions which improved the presentation of the paper.

REFERENCES

[1] N. Alon, R. M. Karp, D. Peleg, and D. West, *A graph-theoretic game and its application to the k-server problem*, SIAM J. Comput., 24 (1995), pp. 78–100.

[2] A. Andoni, M. M. Deza, A. Gupta, P. Indyk, and S. Raskhodnikova, *Lower bounds for embedding edit distance into normed spaces*, in Proceedings of the 14th ACM-SIAM Symposium on Discrete Algorithms (SODA), ACM, New York, SIAM, Philadelphia, 2003, pp. 523–526.

[3] S. Arora, S. Rao, and U. Vazirani, *Expander flows, geometric embeddings, and graph partitionings*, in Proceedings of the 36th ACM Symposium on Theory of Computing (STOC), ACM, New York, 2004, pp. 222–231.

[4] 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.

[5] B. Awerbuch and Y. Azar, *Buy-at-bulk network design*, in Proceedings of the 38th Symposium on Foundations of Computer Science (FOCS), IEEE Computer Society, Los Alamitos, CA, 1997, pp. 542–547.

[6] B. S. Baker, *Approximation algorithms for NP-complete problems on planar graphs*, J. ACM, 41 (1994), pp. 153–180.

[7] Y. Bartal, *Probabilistic approximations of metric spaces and its algorithmic applications*, in Proceedings of the 37th Symposium on Foundations of Computer Science (FOCS), IEEE Computer Society, Los Alamitos, CA, 1996, pp. 184–193.

[8] Y. Bartal, *On approximating arbitrary metrics by tree metrics*, in Proceedings of the 30th ACM Symposium on Theory of Computing (STOC), ACM, New York, 1998, pp. 161–168.

[9] D. Bienstock and C. L. Monma, *On the complexity of embedding planar graphs to minimize certain distance measures*, Algorithmica, 5 (1990), pp. 93–109.

[10] H. L. Bodlaender, *A partial k-arboretum of graphs with bounded treewidth*, Theoret. Comput. Sci., 209 (1998), pp. 1–45.

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

[12] M. Charikar, C. Chekuri, A. Goel, S. Guha, and S. A. Plotkin, *Approximating a finite metric by a small number of tree metrics*, in Proceedings of the 39th Symposium on Foundations of Computer Science (FOCS), IEEE Computer Society, Los Alamitos, CA, 1998, pp. 379–388.

[13] M. M. Deza and M. Laurent, *Geometry of Cuts and Metrics*, Algorithms Combin. 15, Springer-Verlag, Berlin, 1997.

[14] R. Diestel, *Graph Theory*, 2nd ed., Grad. Texts in Math. 173, Springer-Verlag, New York, 2000.

[15] J. Fakcharoenphol, S. Rao, and K. Talwar, *A tight bound on approximating arbitrary metrics by tree metrics*, J. Comput. System Sci., 69 (2004), pp. 485–497.

[16] A. Frank, *Packing paths, circuits, and cuts—a survey*, in Paths, Flows and VLSI-Layout, B. Korte, L. Lovász, H. J. Prömel, and A. Schrijver, eds., Springer-Verlag, New York, 1990, pp. 47–100.

[17] N. Garg, G. Konjevod, and R. Ravi, *A polylogarithmic approximation algorithm for the group Steiner tree problem*, J. Algorithms, 37 (2000), pp. 66–84.

[18] A. Gupta, *Steiner points in tree metrics don't (really) help*, in Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms (Washington, DC, 2001), ACM, New York, SIAM, Philadelphia, 2001, pp. 220–227.

[19] A. Gupta, I. Newman, Y. Rabinovich, and A. Sinclair, *Cuts, trees and ℓ_1-embeddings of graphs*, Combinatorica, 24 (2004), pp. 233–269.

[20] R. Halin, *Studies on minimally n-connected graphs*, in Combinatorial Mathematics and Its Applications, D. J. A. Welsh, ed., Academic Press, London, 1971, pp. 129–136.

[21] P. Indyk, *Algorithmic aspects of geometric embeddings*, in Proceedings of the 42nd Symposium on Foundations of Computer Science (FOCS), IEEE Computer Society, Los Alamitos, CA, 2001, pp. 10–33.

[22] N. Linial, *Finite metric-spaces—combinatorics, geometry and algorithms*, in Proceedings of the International Congress of Mathematicians (Beijing, 2002), Vol. III, Higher Ed. Press, Beijing, 2002, pp. 573–586.

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

[24] J. Matoušek, *On embedding expanders into l_p spaces*, Israel J. Math., 102 (1997), pp. 189–197.

[25] J. Matoušek, *Lectures on Discrete Geometry*, Grad. Texts in Math. 212, Springer-Verlag, New York, 2002.

[26] I. Newman and Y. Rabinovich, *A lower bound on the distortion of embedding planar metrics into Euclidean space*, in Proceedings of the 18th Annual Symposium on Computational Geometry, ACM, New York, 2002, pp. 94–96.

[27] I. Newman, Y. Rabinovich, and M. Saks, *Excluded Minors for Embeddings*, unpublished notes.

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

[29] S. B. Rao, *Small distortion and volume preserving embeddings for planar and Euclidean metrics*, in Proceedings of the 15th Annual ACM Symposium on Computational Geometry, ACM, New York, 1999, pp. 300–306.

[30] N. Robertson and P. D. Seymour, *Graph minors.* III. *Planar tree-width*, J. Combin. Theory Ser. B, 36 (1984), pp. 49–64.

[31] A. Schrijver, *Homotopic routing methods*, in Paths, Flows and VLSI-Layout, B. Korte, L. Lovász, H. J. Prömel, and A. Schrijver, eds., Springer-Verlag, New York, 1990, pp. 329–371.

[32] D. B. Shmoys, *Cut problems and their application to divide-and-conquer*, in Approximation Algorithms for NP-Hard Problems, D. S. Hochbaum, ed., PWS Publishing, Boston, 1997, pp. 192–235.

[33] M. M. Sysło and A. Proskurowski, *On Halin graphs*, in Graph Theory (Łagów, 1981), Springer-Verlag, Berlin, 1983, pp. 248–256.

[34] D. B. West, *Introduction to Graph Theory*, Prentice–Hall, Upper Saddle River, NJ, 1996.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 137–142

RANKING TOURNAMENTS*

NOGA ALON†

Abstract. A tournament is an oriented complete graph. The feedback arc set problem for tournaments is the optimization problem of determining the minimum possible number of edges of a given input tournament T whose reversal makes T acyclic. Ailon, Charikar, and Newman showed that this problem is NP-hard under randomized reductions. Here we show that it is in fact NP-hard. This settles a conjecture of Bang-Jensen and Thomassen.

Key words. tournament, feedback arc set problem

AMS subject classifications. 05C20, 68R10

DOI. 10.1137/050623905

1. Introduction. A *tournament* is an oriented complete graph. A *feedback arc set* in a digraph is a collection of edges whose reversal (or removal) makes the digraph acyclic. The *feedback arc set problem* for tournaments is the optimization problem of determining the minimum possible cardinality of a feedback arc set in a given tournament. The problem for general digraphs is defined analogously. Bang-Jensen and Thomassen conjectured in [7] that this problem is NP-hard, and Ailon, Charikar, and Newman proved in [1] that it is NP-hard under randomized reductions. Here we show how to derandomize a variant of the construction of [1] and prove that the problem is indeed NP-hard. This is based on the known fact that the minimum feedback arc set problem for general digraphs is NP-hard (see [8, p. 192]) and on certain pseudorandom properties of the quadratic residue tournaments described in [5, pp. 134–137]. Similar constructions can be given using any other family of antisymmetric matrices with $\{-1, 1\}$ entries whose rows are nearly orthogonal. We note that unlike the authors of [1], we do not apply the known fact that the minimum feedback arc set problem is APX-hard and need only the simpler fact that it is NP-hard, proved more than thirty years ago. In fact, the proof in [1] can also be modified slightly so as to rely only on this fact (to get hardness of approximation under randomized reductions).

2. Notation. For a digraph $G = (V, E)$ and a permutation π of its vertices, an oriented edge $(u, v) \in E$ is *consistent* with π if u precedes v in π. Let $FIT(G, \pi)$ denote the number of edges whose orientation is consistent with π minus the number of edges whose orientation is not consistent with π. Similarly, if the edges of G are weighted, we let $FIT(G, \pi)$ denote the total weight of the edges whose orientation is consistent with π minus the total weight of the edges whose orientation is not consistent with π. It is convenient to consider unweighted digraphs as weighted digraphs in which the weight of each edge is 1, and the weight of each nonedge is 0. Most of the weighted digraphs we use here have weights in $\{0, 1, -1\}$, but it is helpful to use weights that can be added and subtracted in order to simplify notation.

*Received by the editors May 14, 2004; accepted for publication September 6, 2005; published electronically March 3, 2006. This research was supported in part by the Israel Science Foundation, by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University, and by the Von Neumann Fund.

http://www.siam.org/journals/sidma/20-1/62390.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).

Returning to unweighted digraphs, let $FA(G)$ denote the minimum size of a feedback arc set of $G = (V, E)$. It is easy to see that $FA(G) = (|E| - \max_\pi FIT(G, \pi))/2$, where the maximum is taken over all permutations π of V. This is because omitting a feedback arc set leaves the remaining graph acyclic, ensuring that there is a permutation π consistent with the orientation of all edges left, and similarly, for any π one can omit all edges not consistent with π and get an acyclic digraph.

If $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ are two (weighted) digraphs on the same set of vertices, the sum $G_1 + G_2$ is the digraph on V in which the weight of each edge is the sum of its weights in G_1 and in G_2. The difference $G_1 - G_2$ is defined in a similar manner. Note that for every permutation π on V, $FIT(G_1 + G_2, \pi) = FIT(G_1, \pi) + FIT(G_2, \pi)$ and $FIT(G_1 - G_2, \pi) = FIT(G_1, \pi) - FIT(G_2, \pi)$.

If G is a digraph, and $U \subset V$, then $G[U]$ denotes the induced subgraph of G on U. We consider this subgraph, however, as a digraph whose vertex set is V, where all vertices in $V - U$ are isolated. If U and W are two disjoint subsets of V, then $G[U, W]$ denotes the subgraph of G consisting of all edges of G with an end in U and an end in W. Here, too, the vertex set is the original set V. Let $e(U, W)$ denote the total number of edges of G that start at U and end at W. Thus, the total number of edges of $G[U, W]$ is $e(U, W) + e(W, U)$.

3. The quadratic residue tournaments. Let $p \equiv 3 \bmod 4$ be a prime, and let $T = T_p$ be the tournament whose vertices are all elements of the finite field $GF(p)$, in which (i, j) is a directed edge iff $i - j$ is a quadratic residue. In [5, pp. 134–137] it is shown that for every permutation π of the vertices of T_p, $FIT(T, \pi) \leq O(p^{3/2} \log p)$. Here we need a stronger result, providing a similar bound for certain subgraphs of T.

We need the following known fact, proved, for example, in [2] (see also [5, Lemma 9.1.2]).

LEMMA 3.1. *Let $T = T_p = (V, E)$ be the quadratic residue tournament defined above. Then, for every two disjoint sets U_1, U_2 of T,*

$$e(U_1, U_2) - e(U_2, U_1) \leq |U_1|^{1/2} |U_2|^{1/2} p^{1/2}.$$

Therefore, if $|U_1|$ and $|U_2|$ are large, then the number of edges of G oriented from U_1 to U_2 is roughly the number of edges oriented from U_2 to U_1, as the difference between these two numbers is at most $|U_1|^{1/2} |U_2|^{1/2} p^{1/2}$, whereas their sum is $|U_1||U_2|$. We next observe that this property implies that for every large set of vertices U of T, and for every permutation π, $FIT(T[U], \pi)$ is small.

COROLLARY 3.2. *Let $T = T_p = (V, E)$ be as above, let $U \subset V$ be a set of vertices of T, and let $T[U]$ denote the induced subgraph of T on U. Then, for every permutation π of V,*

$$|FIT(T[U], \pi)| \leq |U| \lceil \log_2 |U| \rceil p^{1/2} \leq |U| \log_2 (2|U|) p^{1/2}.$$

Proof. We prove that for every set U of a most 2^r vertices, and for every permutation π

$$FIT(T[U], \pi) \leq r 2^{r-1} p^{1/2}. \tag{3.1}$$

Note that if $\pi = \pi_1, \pi_2, \ldots, \pi_p$ and $\overline{\pi} = \pi_p, \pi_{p-1}, \ldots, \pi_1$, then $FIT(T[U], \overline{\pi}) = -FIT(T[U], \pi)$, and hence the validity of (3.1) implies the assertion of the corollary (including the absolute value). We prove (3.1) by induction on r. The result is trivial for $r = 1$. Assuming it holds for $r - 1$ we prove it for r. Suppose $|U| \leq 2^r$.

Given π, split U into two disjoint sets U_1, U_2, each of size at most 2^{r-1}, so that all the elements of U_1 precede all those of U_2 in the permutation π. Clearly

$$FIT(T[U], \pi) = e(U_1, U_2) - e(U_2, U_1) + FIT(T[U_1], \pi) + FIT(T[U_2], \pi).$$

By Lemma 3.1 and the induction hypothesis, the right-hand side is at most

$$2^{r-1}p^{1/2} + 2(r-1)2^{r-2}p^{1/2} = r2^{r-1}p^{1/2}.$$

This completes the proof. □

COROLLARY 3.3. *Let $T = T_p = (V, E)$ be as above, let U, W be two disjoint subsets of vertices of T, and let $T[U, W]$ denote the bipartite subgraph of T consisting of all edges of T with an end in U and an end in W. Then, for every permutation π of V,*

$$|FIT(T[U,W], \pi)| \leq [\ (|U| + |W|)\lceil \log_2(|U| + |W|)\rceil + |U|\lceil \log_2 |U|\rceil + |W|\lceil \log_2 |W|\rceil\ ]p^{1/2}.$$

In particular, if $|U| \leq a$ and $|W| \leq a$, then $|FIT(T[U,W], \pi)| \leq 4a \log_2(4a)p^{1/2}$.

Proof. In the notation of section 2, $T[U,W] = T[U \cup W] - T[U] - T[W]$. Therefore, for every π,

$$|FIT(T[U,W], \pi)| = |FIT(T[U \cup W], \pi) - FIT(T[U], \pi) - FIT(T[W], \pi)|,$$

and the desired result follows from the triangle inequality and three applications of the previous corollary. □

The *a-blow-up* of a digraph H, which we denote by $H(a)$, is the digraph obtained by replacing each vertex v of H by an independent set $I(v)$ of size a, and each directed edge (u, v) of H by a complete bipartite digraph containing all a^2 edges from the members of $I(u)$ to those of $I(v)$. It is easy to check that the minimum size of a feedback arc set of $H(a)$ satisfies $FA(H(a)) = a^2 FA(H)$. Indeed, this follows from the fact that if π is a permutation of the vertices of the blow-up $H(a)$ that maximizes $FIT(H(a), \pi)$, and if x, y are two vertices of $H(a)$ that lie in the same $I(v)$, then one may always shift either x to lie right next to y in π or vice versa without decreasing the number of consistent edges.

Our main technical lemma is the following.

LEMMA 3.4. *Let $H = (U, F)$ be a digraph, let $p \equiv 3 \bmod 4$ be a prime, and let $T = T_p = (V, E)$ be the quadratic residue tournament described above. Let a be an integer and suppose that $a|U| \leq p$. For each $u \in U$, let $I(u)$ be an arbitrary subset of size a of V, where all $|U|$ sets $I(u)$ are pairwise disjoint, and let T' be the tournament obtained from T as follows: for each edge $(u, v) \in F$ of H, omit all edges of T that connect members of $I(u)$ with those of $I(v)$, and replace them with all the a^2 directed edges that start at a member of $I(u)$ and end at one of $I(v)$. Then, for every permutation π of V,*

$$|FIT(T', \pi) - FIT(H(a), \pi)| \leq p^{3/2} \log_2(2p) + 4|F|a \log_2(4a)p^{1/2}.$$

Proof. Consider $H(a)$ as a digraph on the sets of vertices $I(u)$, $u \in U$. By construction,

$$T' = T - \sum_{(u,v)\in F} T[I(u), I(v)] + H(a).$$

Therefore, for every π,

$$FIT(T', \pi) = FIT(T, \pi) - \sum_{(u,v)\in F} FIT(T[I(u), I(v)], \pi) + FIT(H(a), \pi).$$

It follows that

$$|FIT(T', \pi) - FIT(H(a), \pi)| \leq |FIT(T, \pi)| + \sum_{(u,v)\in F} |FIT(T[I(u), I(v)], \pi)|,$$

and the desired result follows from Corollary 3.2, which implies that $|FIT(T, \pi)| \leq p^{3/2}\log_2(2p)$, and from Corollary 3.3, which implies that for each fixed $(u, v) \in F$, $|FIT(T[I(u), I(v)], \pi)| \leq 4a\log_2(4a)p^{1/2}$. □

4. The main result.

THEOREM 4.1. *The minimum feedback arc set problem for tournaments is NP-hard.*

Proof. It is known (cf., e.g., [8, p. 192]) that the minimum feedback arc set problem is NP-hard, even for digraphs H in which all outdegrees and indegrees are at most 3 (this last point is not essential here, but we use it to make the computation explicit). Given a digraph $H = (U, F)$ as above, let $a = |U|^c$, where $c > 3$ is a fixed integer, and let $p \equiv 3 \bmod 4$ be a prime between $|U|a$ and, say, $2|U|a$. Such a prime always exists, by the known results on primes in arithmetic progressions, and it is easy to find such a prime in time polynomial in $|U|$, by exhaustive search. Let T' be the tournament constructed from T_p and the blow-up $H(a)$ of H as described in Lemma 3.4. Computing $FA(T')$ is equivalent to computing $\max_\pi FIT(T', \pi)$, where the maximum is taken over all permutations π of V. However, by Lemma 3.4 it follows that the value of $\max_\pi FIT(T', \pi)$ provides an approximation up to an additive error of $p^{3/2}\log_2(2p) + |F|4a\log_2(4a)p^{1/2} \leq 13p^{3/2}\log_2(4p)$ for $\max_\pi FIT(H(a), \pi)$, where here we used the fact that $|F| \leq 3|U|$ and the fact that $|U|a \leq p$. Since, as explained after the proof of Corollary 3.3, $\max_\pi FIT(H(a), \pi) = a^2 \max_\sigma FIT(H, \pi)$, where the last maximum is taken over all permutations σ of the vertices of H, we conclude that if $a^2 > 13p^{3/2}\log_2(4p)$, this approximation will enable us to determine $\max_\sigma FIT(H, \pi)$ (and hence also $FA(H)$) precisely. Since $a = |U|^c$ and $p \leq 2|U|a \leq 2|U|^{c+1}$, this is the case provided $c \geq 4$, completing the proof. □

5. Remarks and problems.

- By choosing c appropriately in the above proof it follows that for every fixed $\epsilon > 0$, it is NP-hard to approximate $FA(T)$ for a tournament on n vertices up to an additive error of $n^{2-\epsilon}$. Note that approximating it up to an additive error of ϵn^2 can be done in polynomial time using the algorithmic version of the regularity lemma (for digraphs), or the methods of [6].
- It will be interesting to decide if the minimum feedback arc set problem for tournaments is APX-hard. The authors of [1] describe a randomized algorithm that provides a constant approximation of this quantity.
- The assertion of Lemma 3.1 here follows from the fact that the absolute value of the sum of entries in any submatrix of the p by p matrix B in which $B_{ij} = \chi(i - j)$, where χ is the quadratic character, can be bounded as described in the lemma. If $G = (V, E)$ is a general directed graph, with weights on

its edges, let $A = A_G$ be a matrix whose rows and columns are indexed by the vertices of G, in which for each $u, v \in V$, $A(u,v) = w(u,v) - w(v,u)$ is the difference between the weight of the directed edge from u to v and that from v to u (0 if both these edges are missing). Thus, the matrix B above is the matrix A_{T_p}, where T_p is the quadratic residue tournament described in section 3.

The *cutnorm* $||A||_C$ of a real matrix A is the maximum absolute value of the sum of entries in a submatrix of A. Note that if $A = A_G$, where G is a weighted directed graph, then for two subsets $U_1, U_2 \subset V$, the sum $\sum_{u_1 \in U_1, u_2 \in U_2} A(u_1, u_2)$ can be expressed as follows. Put $U_3 = U_1 \cap U_2$. For two disjoint subsets X, Y of V let $D(X,Y)$ denote the total weight of all edges oriented from X to Y minus the total weight of all edges oriented from Y to X. Then

$$\sum_{u_1 \in U_1, u_2 \in U_2} A(u_1, u_2) = D(U_1 - U_2, U_2) + D(U_3, U_2 - U_1). \tag{5.1}$$

The authors of [3, 4] describe a polynomial time algorithm that finds, given a matrix A, two subsets U_1, U_2 such that $|\sum_{u_1 \in U_1, u_2 \in U_2} A(u_1, u_2)|$ is at least $\alpha ||A||_C$ for some absolute constant $\alpha > 0$ (for randomized algorithms $\alpha > 0.56$). As in our case the matrix A is antisymmetric, the algorithm provides U_1, U_2 so that the above sum (with no absolute value) approximates the maximum cutnorm. In view of the expression (5.1) this supplies an $\alpha/2$ approximation for the maximum possible value of $D(X,Y)$, as X and Y range over all pairs of disjoint subsets of V.

- The bound in Corollary 3.2 can be slightly improved, using the expression in (5.1) and the fact that for the matrix $A = A_{T_p}$ of the quadratic residue tournament, the absolute value of the sum of entries of any submatrix with s rows and t columns is at most $\sqrt{stp}$. Indeed, plugging this fact into a simple modified version of the proof of Corollary 3.2 one can prove the following: If U is a set of vertices of T_p, and $|U| \leq 3^r$ for some integer r, then for every permutation π of the vertices of T_p, $FIT(T[U], \pi) \leq 2r3^{r-1}p^{1/2}$.
- The basic approach of proving hardness results for dense instances of computational problems by reducing the task of solving precisely sparse instances to dense ones, adding a pseudorandom collection of edges to a blow-up of a sparse instance, can be applied to various additional similar problems. Several far reaching applications of this approach, combined with some additional ideas, will appear in subsequent joint work with Ailon and in another work with Shapira and Sudakov.

REFERENCES

[1] N. Ailon, M. Charikar, and A. Newman, *Aggregating inconsistent information: Ranking and clustering*, in Proceedings of the 37th ACM STOC, Baltimore, ACM, New York, 2005, pp. 684–693.

[2] N. Alon, *Eigenvalues, geometric expanders, sorting in rounds and Ramsey theory*, Combinatorica, 6 (1986), pp. 207–219.

[3] N. Alon and A. Naor, *Approximating the cut-norm via Grothendieck's inequality*, in Proceedings of the 36th ACM STOC, Chicago, ACM, New York, 2004, pp. 72–80.

[4] N. Alon and A. Naor, *Approximating the cut-norm via Grothendieck's inequality*, SIAM J. Comput., 35 (2006), pp. 787–803.

[5] N. Alon and J. H. Spencer, *The Probabilistic Method*, 2nd ed., Wiley, New York, 2000.

[6] S. Arora, D. Karger, and M. Karpinski, *Polynomial time approximation schemes for dense instances of NP-hard problems*, in Proceedings of the 27th ACM STOC, ACM, New York, 1995, pp. 284–293.

[7] J. Bang-Jensen and C. Thomassen, *A polynomial algorithm for the* 2*-path problem for semicomplete digraphs*, SIAM J. Discrete Math., 5 (1992), pp. 366–376.

[8] M. R. Garey and D. S. Johnson, *Computers and Intractability, A Guide to the Theory of NP-Completeness*, W. H. Freeman and Company, New York, 1979.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 143–159

THE LINKING PROBABILITY OF DEEP SPIDER-WEB NETWORKS*

NICHOLAS PIPPENGER†

Abstract. We consider crossbar switching networks with base b (that is, constructed from $b \times b$ crossbar switches), scale k (that is, with b^k inputs, b^k outputs, and b^k links between each consecutive pair of stages), and depth l (that is, with l stages). We assume that the crossbars are interconnected according to the spider-web pattern, whereby two diverging paths reconverge only after at least k stages. We assume that each vertex is independently idle with probability q, the vacancy probability. We assume that $b \geq 2$ and the vacancy probability q are fixed, and that k and $l = ck$ tend to infinity with ratio a fixed constant $c > 1$. We consider the linking probability Q (the probability that there exists at least one idle path between a given idle input and a given idle output). In a previous paper [*Discrete Appl. Math.,* 37/38 (1992), pp. 437–450] it was shown that if $c \leq 2$, then the linking probability Q tends to 0 if $0 < q < q_c$ (where $q_c = 1/b^{(c-1)/c}$ is the critical vacancy probability) and tends to $(1-\xi)^2$ (where ξ is the unique solution of the equation $\bigl(1 - q(1-x)\bigr)^b = x$ in the range $0 < x < 1$) if $q_c < q < 1$. In this paper we extend this result to all rational $c > 1$. This is done by using generating functions and complex-variable techniques to estimate the second moments of various random variables involved in the analysis of the networks.

Key words. communiation networks, crossbar switching networks, blocking probability

AMS subject classifications. 94C15, 60C05

DOI. 10.1137/050624376

1. Introduction.

We deal in this paper with linking in crossbar switching networks, a phenomenon not dissimilar to that of percolation in lattices (as introduced by Broadbent and Hammersley [B] and surveyed by Grimmett [G]). An important difference, however, is that while percolation can be studied in finite subgraphs of a single infinite graph modeling the lattice, there is no single graph that naturally hosts the graph modeling crossbars switching networks in which we are interested. Our first order of business will be to describe these graphs.

A *crossbar graph* is characterized by three parameters: its *base*, $b \geq 2$, its *scale*, $k \geq 0$, and its *depth*, $l \geq 0$. Its vertices are partitioned into $l+1$ *ranks*, each containing b^k vertices, which are labeled with the strings of length k over the alphabet $\{0, \ldots, b-1\}$. The vertices in rank 0 are called *inputs*, those in rank l are called *outputs*, and those in all other ranks are called *links*. The edges of the graph are partitioned into l *stages*, each containing b^{k+1} edges. For $1 \leq m \leq l$, the edges of stage m are directed out of vertices in rank $m-1$ and into vertices in rank m. In a *spider-web* crossbar graph, which is our main concern in this paper, there is an edge of stage m from vertex v of rank $m-1$ to vertex w of rank m if and only if v and w are labeled by strings that differ at most in position j, where $j \equiv m \pmod{k}$. The edges of each stage are thus partitioned into b^{k-1} $b \times b$ complete bipartite graphs (called *crossbars*). The spider-web crossbar graph with base b, scale k, and depth l will be denoted $G_{b,k,l}$. We shall see in section 2 that if $l \geq k$, there are b^{l-k} paths from a given input to a given output; if $l < k$, there is at most one path from a given input to a given output. Our main interest is in spider-web crossbar graphs with $l \geq k$, since

*Received by the editors February 14, 2005; accepted for publication (in revised form) September 15, 2005; published electronically March 3, 2006. This work was supported in part by NSF grant CCF-0430656.

http://www.siam.org/journals/sidma/20-1/62437.html

†Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08540 (njp@princeton.edu).

in these graphs any input can be connected by a path to any output; in our analysis, however, graphs with $l < k$ will occur as subgraphs, so it will be necessary to allow this case in some of our intermediate results.

We shall assume that each vertex in the graph $G_{b,k,l}$ is independently assigned the status *idle*, with probability q (called the *vacancy* probability), or *busy*, with the complementary probability $p = 1 - q$ (called the *occupancy* probability). This random assignment of a status to each vertex in a graph will be called the *state* of the graph. Given an input v and output w, let $Q_{v,w}$ (called the *linking* probability) denote the probability that there exists a path consisting entirely of idle links from v to w. (In this paper, "path" will always mean "directed path." In general, the linking probability is defined as the *conditional* probability that there exists an idle path, *given* that v and w are themselves idle, but for the probabilistic model that we are using, this condition is independent.) We shall see in section 2 that if $l \geq k$, the probability $Q_{v,w}$ does not depend on the choice of the input-output pair (v, w), so we shall let Q denote the common value of these probabilities. The complementary probability $P = 1 - Q$ (called the *blocking* probability) is the probability that all paths between a given input-output pair (v, w) are broken by a set of busy links.

In practice, the parameter p represents the amount of traffic being carried by a crossbar network (which one would like to maximize), and the parameter P represents the fraction of arriving traffic lost due to congestion within the network (which one would like to minimize). In analysis, however, it is almost always more convenient to work with the complementary parameters q and Q, so we shall work exclusively with these parameters in what follows.

In practice, a graph $G_{b,k,l}$ would be fixed, and the linking probability Q would be studied as a function of the vacancy probability q. It is found that Q undergoes a rapid transition from a value near zero to a significantly positive value as q passes through a neighborhood of $1/b^{(l-k)/(l-1)}$. This is easily understood in the following way.

Let the random variable $X_{v,w}$ denote the number of idle paths from v to w. We shall see in section 2 that if $l \geq k$, the distribution of $X_{v,w}$ does not depend on the choice of the input-output pair (v, w), so we shall let X denote a random variable with this common distribution. Each of the b^{l-k} paths from v to w contains $l - 1$ links, which are all idle with probability q^{l-1}. Thus we have

$$\mathrm{Ex}[X] = b^{l-k}\, q^{l-1}. \tag{1.1}$$

Thus as q passes through $1/b^{(l-k)/(l-1)}$, the expected number of idle paths from v to w (called the *specific transparency*) goes from an exponentially decreasing to an exponentially increasing function of k and l. This suggests that if k and l tend to infinity in such a way that their ratio $c = l/k > 1$ remains fixed while b and q are also held fixed, then Q will tend to a limit, and this limit will have a discontinuity as q passes through the critical value

$$q_c = 1/b^{(c-1)/c}.$$

(We note that $1 < c < \infty$ implies $1/b < q_c < 1$.) Our goals in this paper are to confirm this conjecture and to determine the limiting value of Q.

Our first step toward these goals, taken in section 2, will be to derive the following estimate for the second moment $\mathrm{Ex}[X^2]$ of X.

THEOREM 1.1. *Let both $b \geq 2$ and $1/b < q < 1$ be fixed. Then*

$$\mathrm{Ex}[X^2] = \mathrm{Ex}[X] \cdot \left(\left(\frac{b-1}{bq-1} \right)^2 b^{l-k}\, q^{l+1} + 1 + O(l\, b^{l-2k}\, q^l) + O(l\, q^k) \right)$$

as $k, l \to \infty$ with $l \geq k$ and $\big(\log(l+1)\big)/(k+1) \to 0$. (The constants in the O-terms may depend on b and q, but are independent of k and l.)

We observe that this estimate is enough to establish that the limiting value (if it exists) of Q for $k \to \infty$ and $l = ck$ cannot be a continuous function of Q as q passes through q_c. Indeed, from Markov's inequality and (1.1), we have

$$Q = \Pr[X \geq 1] \leq \mathrm{Ex}[X] = b^{l-k}\, q^{l-1} \to 0 \tag{1.2}$$

for $q < q_c$. On the other hand, (1.1) and Theorem 1.1, together with the inequality

$$\Pr[X \geq 1] \geq \frac{\mathrm{Ex}[X]^2}{\mathrm{Ex}[X^2]}, \tag{1.3}$$

imply

$$Q = \Pr[X \geq 1] \geq \frac{\mathrm{Ex}[X]^2}{\mathrm{Ex}[X^2]} = \frac{(bq-1)^2}{(b-1)^2 q^2 + (bq-1)^2 q} \left(1 + O\left(\frac{l}{b^k} \right) + O\left(l\, q^k \right) \right)$$

$$\to \frac{(bq-1)^2}{(b-1)^2 q^2 + (bq-1)^2 q} > 0 \tag{1.4}$$

for $q = q_c$. (To verify (1.3), we consider the distribution of X conditioned on the event $X \geq 1$. Since x^2 is a convex function of x, we have

$$\mathrm{Ex}[X^2 \mid X \geq 1] \geq \mathrm{Ex}[X \mid X \geq 1]^2.$$

Multiplying by $\Pr[X \geq 1]^2$ yields

$$\begin{aligned} \mathrm{Ex}[X^2]\, \Pr[X \geq 1] &= \mathrm{Ex}[X^2 \mid X \geq 1]\, \Pr[X \geq 1]^2 \\ &\geq \mathrm{Ex}[X \mid X \geq 1]^2\, \Pr[X \geq 1]^2 \\ &= \mathrm{Ex}[X]^2, \end{aligned}$$

which is equivalent to (1.3).) The inequalities in (1.2) and (1.4) show that the inferior limit of Q for $q = q_c$ is strictly greater than the limiting value for $q < q_c$, as claimed.

The argument of the preceding paragraph also sheds some light on the condition $\big(\log(l+1)\big)/(k+1) \to 0$ in Theorem 1.1. (This condition involves $k+1$ and $l+1$ rather than k and l simply to avoid dividing by or taking the logarithm of 0.) This condition is not the weakest one sufficient to give an estimate of the form $\mathrm{Ex}[X^2] = O\big(\mathrm{Ex}[X]^2\big)$, but it is clear that some upper bound on the growth of l must be imposed, for with probability $(1-q)^{b^k}$ all the links in a given rank are busy, disconnecting all input-output pairs. Thus if $l \cdot (1-q)^{b^k} \to \infty$, we have $Q \to 0$, contradicting the implication of (1.3) when $\mathrm{Ex}[X^2] = O\big(\mathrm{Ex}[X]^2\big)$.

In section 3, we shall combine Theorem 1.1 with branching-process arguments from Pippenger [P3] to establish the existence and determine the limiting value of Q for $q > q_c$.

THEOREM 1.2. *Let $b \geq 2$ and $0 < q < 1$ be fixed, and let $c > 1$ be rational. Then as $k \to \infty$ with $l = ck$, we have*

$$Q \to \begin{cases} 0 & \text{if } 0 < q < q_c, \\ \\ (1 - \xi)^2 & \text{if } q_c < q < 1, \end{cases}$$

where ξ is the unique solution of the equation $x = \bigl(1 - q(1 - x)\bigr)^b$ in the range $0 < x < 1$.

A comment is in order concerning the behavior of $(1-\xi)^2$ as a function of q. The function $f(x) = \bigl(1 - q(1 - x)\bigr)^b$ is a strictly convex function of x for $0 < q \leq 1$, since $f''(x) = b(b-1)q^2\bigl(1 - q(1 - x)\bigr)^{b-2} > 0$ in this range. Thus the graph of $f(x)$ can intersect the diagonal at most twice in this range. There is one intersection at $x = 1$, and the conditions $f(0) = (1 - q)^b > 0$ and $f'(1) = bq > 1$ imply that there is at least one intersection in the range $0 < x < 1$ when $1/b < q < 1$. Thus there is indeed a unique solution of the equation $x = \bigl(1 - q(1 - x)\bigr)^b$ in the range $0 < x < 1$ when $1/b < q < 1$, and this latter condition is implied by $q_c < q < 1$. The degree of this equation can be reduced by 1 (because of the solution $x = 1$), and it is easy to see that the resulting equation is irreducible over the field of rational functions of q; thus ξ is an algebraic function of q of degree $b - 1$. Since $(1 - \xi)^2$ is a polynomial in ξ, it is also an algebraic function of q of degree $b - 1$. Straightforward analysis shows that $Q \to 1$ as $q \to 1$ with $1 - Q = 1 - (1 - \xi)^2 \sim 1 - 2(1 - q)^b$, which may be interpreted as saying that the main obstacle to linking when $q \to 1$ is complete occupation either of the b links adjacent to the input in rank 1, or of the b links adjacent to the output in rank $l - 1$. As $q \to 1/b$ from above, we have $(1 - \xi)^2 \sim (bq - 1)^2 / \binom{b}{2}^2$.

Theorem 1.2 was proved, under the additional restriction $c \leq 2$, by Pippenger [P3], so the additional contribution of the current paper consists of lifting this restriction. Nevertheless, the techniques used in the current paper go considerably beyond those employed in the previous paper in that the proof of Theorem 1.1 starts with a detailed combinatorial examination of the intersections between paths, then uses complex-variable techniques to determine the asymptotics of the quantities involved.

Spider-web networks were introduced by Ikeno [I] (though the term *spider-web* has sometimes been used to refer to a broader class of networks). They have several optimality properties among networks constructed from the same type and number of crossbars. Takagi [T] showed that they have the largest linking probability in a large class of crossbar networks called "rhyming" networks. Chung and Hwang [C] showed that, surprisingly, these networks are *not* optimal in the larger class of "balanced" networks. But Pippenger [P3] showed that they are *asymptotically* optimal in this class for $1 < c \leq 2$, and the current paper extends this result to all $c > 1$.

The probability distribution on states that we use was introduced by Lee [L1] and Le Gall [L2, L3]. It is by far the easiest to use for analytical purposes, but it suffers from the defect that the set of busy vertices does not form a set of coherent paths from inputs to outputs. Models addressing this defect have been introduced by Koverninskiĭ [K] and Pippenger [P1], and the results in [P3] have been extended to these models in [P2]. It seems likely that the results of the the present paper can be similarly extended.

The current paper is self-contained, except for some estimates concerning branching processes taken from Pippenger [P3]. We have followed the notation of that paper, except that the base, which was denoted d in that paper, is now denoted b (to free

the symbol d for its traditional use in the calculus).

2. The second moment. Our goal in this section is to prove Theorem 1.1. We begin with a combinatorial result concerning spider-web graphs.

LEMMA 2.1. *The automorphism group of $G_{b,k,l}$ acts transitively on the paths from inputs to outputs.*

Proof. Since an automorphism must permute the vertices within each rank, an automorphism ϑ may be regarded as a sequence $\vartheta = (\vartheta_0, \ldots, \vartheta_l)$ of permutations, one for each rank. We shall focus on automorphisms in which each ϑ_m (for $0 \le m \le l$) is characterized by a string $\vartheta_{m,1} \cdots \vartheta_{m,k}$ of k digits from the alphabet $\{0, \ldots, b-1\}$ and acts on the vertices of rank m by carrying the vertex labeled $a_1 \cdots a_k$ to the vertex labeled $a'_1 \cdots a'_k$, where $a'_j \equiv a_j + \vartheta_{m,j} \pmod b$ for $1 \le j \le k$. If, for $1 \le m \le l$, the string ϑ_{m-1} differs from the string ϑ_m in at most position j, where $j \equiv m \pmod k$, then the sequence $\vartheta = (\vartheta_0, \ldots, \vartheta_l)$ will constitute an automorphism.

To show that the automorphisms act transitively on the paths, it will suffice to show, for some fixed path u^*, that for every path u, there is an automorphism that carries u^* to u (since then the inverse of such an automorphism can be used to carry any other path u' to u^*). A path u may be regarded as a sequence $u = (u_0, \ldots, u_l)$ of vertex labels in which, for $1 \le m \le l$, the string u_{m-1} differs from the string u_m in at most position j, where $j \equiv m \pmod k$. We shall choose for u^* the path $u^* = (0^k, \ldots, 0^k)$. Then clearly the automorphism $\vartheta = (\vartheta_0, \ldots, \vartheta_l)$ defined by $\vartheta_m = u_m$ for $0 \le m \le l$ carries u^* to u. □

COROLLARY 2.2. *If $l \ge k$, the graph $G_{b,k,l}$ contains b^{l-k} paths from any given input to any given output; if $l < k$, there is at most one path from any given input to any given output.*

Proof. If $l \ge k$, every input-output pair is joined by at least one path, since every position in the strings labeling vertices has an opportunity to change at least once. Thus, by Lemma 2.1 every input is joined by the same number of paths. Since each of the b^k inputs is the origin of b^l paths to outputs, there are a total of b^{l+k} paths joining inputs to outputs, and thus b^{l-k} paths joining each of the b^{2k} input-output pairs. If $l < k$, there is a path from input v to output w only if the labels of v and w agree in the last $k - l$ positions. Thus $G_{b,k,l}$ breaks into b^{k-l} disjoint components, each containing b^l vertices in each rank; there is a unique path joining input v to output w if they belong to the same component, but no path joining them if they belong to different components. □

COROLLARY 2.3. *If $l \ge k$, the automorphism group of $G_{b,k,l}$ acts transitively on the input-output pairs.*

Proof. If $k \ge k$, each input-output pair is joined by a path, so the corollary follows from Lemma 2.1. □

This corollary, together with the fact that the probability distribution on states of the graph is invariant under automorphisms of the graph, justifies our earlier assertion that the linking probability $Q_{v,w}$ and the distribution of the random variable $X_{v,w}$ are independent of the choice of the input-output pair (v, w) when $l \ge k$. Henceforth we shall focus our attention on the input-output pair $(v^*, w^*) = (0^k, 0^k)$. If $l \ge k$, this entails no loss of generality. When $l < k$, we shall deal only with cases in which the input and output of interest are joined by a path, and in these cases there is again no loss of generality.

Fix $b \ge 2$ and $k \ge 1$. For $l \ge 0$, let $\varphi_l(y)$ denote the generating function for the number of paths from the input $v^* = 0^k$ to the output $w^* = 0^k$ classified according to the number of links that have labels different from 0^k; that is, the coefficient of y^m in

$\varphi_l(y)$ is the number of paths from v^* to w^* that have $l-1-m$ links in common with the path $u^* = (0^k, \ldots, 0^k)$. Clearly $\varphi_l(y) = 1$ for $0 \le l \le k$, and $\varphi_l(y)$ is a polynomial in y of degree $l-1$ if $l \ge k+1$.

We are interested in the polynomials $\varphi_l(y)$ for various values of $l \ge 0$, with b and k fixed. To determine them, it will be convenient to work with a graph $G_{b,k}$ that contains as subgraphs all the graphs $G_{b,k,l}$ for various values of l. For any $m \ge l \ge 0$, $G_{b,k,l}$ may be regarded as the subgraph comprising the vertices in ranks 0 through l and the edges in stages 1 through l of $G_{b,k,m}$. Thus we may define the infinite graph

$$G_{b,k} = \bigcup_{l \ge k} G_{b,k,l}$$

as the union (inductive limit) of all these graphs. The graph $G_{b,k}$ has inputs in rank 0, but all other vertices will be referred to as links.

For $l \ge 0$, the polynomial $\varphi_l(y)$ is the generating function for the number of paths from the input $v^* = 0^k$ to the link labeled 0^k in rank l classified according to the number of links that have labels different from 0^k.

Let

$$\psi(y,z) = \sum_{l \ge 0} \varphi_l(y)\, z^l$$

be the generating function for the polynomials $\varphi_l(y)$. The key to our estimate for the second moment of X is the following proposition.

PROPOSITION 2.4. *We have*

$$\psi(y,z) = \frac{1 - byz + (b-1)(yz)^{k+1}}{(1-z)(1-byz) - (b-1)z(1-y)(yz)^k}.$$

Proof. In this proof, we shall employ a concise alternative representation of a path $u = (u_0, \ldots, u_l)$ of length $l \ge 0$ as a string $t = t_1 \cdots t_{k+l}$ of length $k+l$ over the alphabet $B = \{0, \ldots, b-1\}$. The first k digits $t_1 \cdots t_k$ of t will be the k digits of the label u_0. For $1 \le m \le l$, t_{k+m} will be the digit in position j of u_m, where $j \equiv m \pmod{k}$ (the digit of u_m that might be different from that of u_{m-1}). Then for $0 \le m \le l$, u_m is the string $t_{m+1} \cdots t_{m+k}$. In particular, the last k digits of t are the k digits of the label u_l of the link in rank l, and the paths from the input $v^* = 0^k$ to the link labeled 0^k in rank l are in one-to-one correspondence with the strings of length $k+l$ over the alphabet B, whose first k digits and last k digits are 0's.

Given a path $t = 0^k t_{k+1} \cdots t_{l-k} 0^k$, let us overline each digit t_{k+m} $(1 \le m \le l)$ for which $u_{m-1} \ne 0^k$. The result is a string over the alphabet $B \cup \overline{B}$, where $\overline{B} = \{\overline{0}, \ldots, \overline{b-1}\}$ is the set of overlined digits. For $l \ge 0$, let the language $K_l \subseteq (B \cup \overline{B})^{k+l}$ comprise the strings obtained in this way for all paths from the input $v^* = 0^k$ to the link labeled 0^k in rank l, and define $K \subseteq (B \cup \overline{B})^*$ by

$$K = \bigcup_{l \ge 0} K_l.$$

Then $\psi(y,z)$ is the power series in y and z in which the coefficient of $y^j z^l$ is the number of strings of length $k+l$ in K in which j digits are overlined. Let

$$L = 0^{-k} K = \{t \in (B \cup \overline{B})^* : 0^k t \in K\}$$

be the language obtained from K by deleting the k initial 0's from each string. Since none of this initial 0's are overlined, $\psi(y,z)$ is the power series in y and z in which the coefficient of $y^j z^l$ is the number of strings of length l in L in which j digits are overlined.

Our next step is to write a regular expression for the language L. Define the alphabets $B' = \{1,\dots,b-1\}$ and $\overline{B}' = \{\overline{1},\dots,\overline{b-1}\}$. Then L is described by the regular expression

$$\left(\left(\Lambda + \left(\overline{B}'\left(\Lambda + \overline{0} + \cdots + \overline{0}^{\,k-1}\right)\right)^* \overline{B}'\overline{0}^{\,k-1}\right)0\right)^*, \tag{2.1}$$

where Λ denotes the empty string. To see this, we observe that a string in L can be uniquely parsed into zero or more *stretches*, each of which ends with an unoverlined 0. A stretch consists of an unoverlined 0 optionally preceded by an *excursion*. An excursion consists of a *final segment* preceded by zero or more *preliminary segments*. A final segment consists of a digit from $\overline{B}'$ followed by exactly $k-1$ overlined 0's. A preliminary segment consists of a digit from $\overline{B}'$ followed by at most $k-1$ overlined 0's. Clearly a final segment is described by the regular expression $\overline{B}'\overline{0}^{\,k-1}$, and a preliminary segment is described by the regular expression $\overline{B}'\big(\Lambda + \overline{0} + \cdots + \overline{0}^{\,k-1}\big)$. Thus an excursion is described by the regular expression

$$\left(\overline{B}'\left(\Lambda + \overline{0} + \cdots + \overline{0}^{\,k-1}\right)\right)^* \overline{B}'\overline{0}^{\,k-1},$$

and a stretch is described by the regular expression

$$\left(\Lambda + \left(\overline{B}'\left(\Lambda + \overline{0} + \cdots + \overline{0}^{\,k-1}\right)\right)^* \overline{B}'\overline{0}^{\,k-1}\right)0.$$

Thus the strings in L are described by the regular expression (2.1).

We now observe that the regular expression (2.1) is *unambiguous* in the following sense: A string described by a subexpression $R+S$ is described by R or by S (but not both); a string t described by a subexpression RS has a unique parsing $t = rs$ such that r is described by R and s is described by S; and a string t described by a subexpression S^* has a unique parsing $s = s_1 \cdots s_n$ with $n \geq 0$ such that $s_1,\dots,s_n$ are described by S.

For an unambiguous regular expression, if $\psi_R(y,z)$ and $\psi_S(y,z)$ are the generating functions counting the strings described by subexpressions R and S, respectively, then $\psi_R(y,z)+\psi_S(y,z)$, $\psi_R(y,z)\,\psi_S(y,z)$, and $1/\big(1-\psi_S(y,z)\big)$ are the generating functions counting the strings described by the subexpressions $R+S$, RS, and S^*, respectively.

Thus the final segments are counted by the generating function $(b-1)(yz)^k$ and the preliminary segments are counted by the generating function

$$(b-1)yz\big(1 + yz + \cdots + (yz)^{k-1}\big) = \frac{(b-1)(yz-(yz)^{k+1})}{1-yz}.$$

The excursions are counted by

$$\frac{(b-1)(yz)^k}{1-\frac{(b-1)(yz-(yz)^{k+1})}{1-yz}} = \frac{(b-1)((yz)^k-(yz)^{k+1})}{1-yz-(b-1)(yz-(yz)^{k+1})},$$

and the stretches are counted by

$$\left(1 + \frac{(b-1)((yz)^k-(yz)^{k+1})}{1-yz-(b-1)(yz-(yz)^{k+1})}\right)z = \frac{z-yz^2-(b-1)z(yz-(yz)^k)}{1-yz-(b-1)(yz-(yz)^{k+1})}.$$

Thus the strings in L are counted by

$$\frac{1}{1-\frac{z-yz^2-(b-1)z(yz-(yz)^k)}{1-yz-(b-1)(yz-(yz)^{k+1})}} = \frac{1-byz+(b-1)(yz)^{k+1}}{(1-z)(1-byz)-(b-1)z(1-y)(yz)^k},$$

which completes the proof of the proposition. □

PROPOSITION 2.5. *Let $b \geq 2$ and $0 < q < 1$ be fixed. Then as $k \to \infty$, and as $l \geq 0$ behaves in such a way that $\big(\log(l+1)\big)/(k+1) \to 0$, we have*

$$\varphi_l(q) = \left(\frac{b-1}{bq-1}\right)^2 b^{l-k} q^{l+1} + 1 + O(lb^{l-2k}q^l) + O(lq^k) + O(lq^l).$$

(The constants in the O-terms may depend on b and q, but are independent of k and l.)

Proof. Write $A(z) = 1 - bqz + (b-1)(qz)^{k+1}$ and $B(z) = (1-z)(1-bqz) - (b-1)z(1-q)(qz)^k$ so that $\psi(q,z) = A(z)/B(z)$. Then from Cauchy's formula we have

$$\begin{aligned} \varphi_l(q) &= \frac{1}{2\pi i} \oint_{\Gamma_0} \frac{\psi(q,z)\,dz}{z^{l+1}} \\ &= \frac{1}{2\pi i} \oint_{\Gamma_0} \frac{A(z)}{B(z)} \frac{dz}{z^{l+1}}, \end{aligned} \tag{2.2}$$

where Γ_0 is a contour taken counterclockwise around a circle $|z| = \varepsilon$ centered at 0 and having radius ε sufficiently small to exclude all other singularities of the integrand.

To make further progress, we must estimate the locations of these other singularities, which are poles at the values of z for which the denominator $B(z)$ vanishes. One such singularity is at $z = 1/q$. Let

$$\zeta_1 = \frac{1}{q}\left(1 - \frac{1}{l}\right),$$

and let Γ_1 be a contour taken counterclockwise around the circle $|z| = \zeta_1$ centered at 0 and having radius ζ_1. As z traverses this contour, the magnitude of the first term $(1-z)(1-bqz)$ of $B(z)$ satisfies the lower bound

$$\begin{aligned} |(1-z)(1-bqz)| &= |1-z| \cdot |1-bqz| \\ &\geq \left(\frac{1}{q} - 1 - \frac{1}{ql}\right)\left(b - 1 - \frac{b}{l}\right) \\ &\geq \left(\frac{1}{q} - 1\right)(b-1) - \frac{bq-1}{ql}, \end{aligned}$$

since the minimum occurs when z is real and positive. The magnitude of the second term, $(b-1)z(1-q)(qz)^k$, on the other hand, satisfies the upper bound

$$\begin{aligned} |(b-1)z(1-q)(qz)^k| &\leq (b-1)\left(\frac{1}{q} - 1\right)\left(1 - \frac{1}{l}\right)^{k+1} \\ &\leq (b-1)\left(\frac{1}{q} - 1\right) e^{-k/l} \\ &\leq (b-1)\left(\frac{1}{q} - 1\right)\left(1 - \frac{(e-1)k}{el}\right). \end{aligned}$$

Here we have used the inequality $1-x \le e^{-x}$, which holds for all x because the graph of the convex function e^{-x} lies above that of $1-x$, its tangent at $x=0$, and the inequality $e^{-x} \le 1-(e-1)x/e$, which holds for $0 \le x \le 1$ because the graph of the convex function e^{-x} lies below that of $1-(e-1)x/e$, its chord across the interval $0 \le x \le 1$. Thus for all sufficiently large k (specifically, for $k > (bq-1)e/(b-1)(1-q)(e-1)$), we have the bound

$$|B(z)| = \Omega\left(\frac{1}{l}\right)$$

for z on the contour Γ_1. Since we also have $A(z) = O(1)$ for z on Γ_1, we have the estimate

$$\frac{1}{2\pi i}\oint_{\Gamma_1} \frac{A(z)}{B(z)}\ \frac{dz}{z^{l+1}} = O(l\,q^l). \tag{2.3}$$

Furthermore, as z traverses the contour Γ_1, the value of the first term, $(1-z)(1-bqz)$, in $B(z)$ circles the origin twice, since it is a quadratic polynomial. Since the second term, $(b-1)z(1-q)(qz)^k$, has strictly smaller magnitude, the value of $B(z)$ also circles the origin twice. It follows that the denominator of $B(z)$ has exactly two zeros inside the contour Γ_1. These are perturbations of the zeros of the first term: the zero of the first term at $z=1$ is perturbed to one at

$$z = \zeta_2 = 1 + O(q^k), \tag{2.4}$$

and the zero of the first term at $z = 1/bq$ is perturbed to one at

$$z = \zeta_3 = \frac{1}{bq}\left(1 - \frac{(b-1)(1-q)}{(bq-1)\,b^k} + O\left(\frac{k}{b^{2k}}\right)\right). \tag{2.5}$$

The condition $\big(\log(l+1)\big)/(k+1) \to 0$ ensures that the O-terms in (2.4) and (2.5) have smaller orders of magnitude than the terms preceding them. We observe that $0 < \zeta_3 < \zeta_2 < \zeta_1$, and thus 0, ζ_3, and ζ_2 lie inside Γ_1 and lie in that order along the real axis. Let Γ_2 be a contour taken counterclockwise around a circle $|z-\zeta_2| = \varepsilon$ centered at ζ_2 and having radius ε sufficiently small to exclude all other singularities of the integrand, and let Γ_3 be a contour taken counterclockwise around a circle $|z-\zeta_3| = \varepsilon$ centered at ζ_3 and having radius ε sufficiently small to exclude all other singularities of the integrand. Since the integral of an analytic function around a contour depends only on the homology class of the contour in the domain of analyticity of the function, and since Γ_0 is homologous to $\Gamma_1 - \Gamma_2 - \Gamma_3$ (indeed, Γ_0 is homotopic to a contour that joins a forward traversal of Γ_1 with reverse traversals of Γ_2 and Γ_3 by canceling traversals of segments $[\zeta_3+\varepsilon, \zeta_2-\varepsilon]$ and $[\zeta_2+\varepsilon, \zeta_1]$ of the real axis), from (2.2) we have

$$\begin{aligned}\varphi_l(q) = {} & \frac{1}{2\pi i}\oint_{\Gamma_1} \frac{A(z)}{B(z)}\ \frac{dz}{z^{l+1}} \\ & - \frac{1}{2\pi i}\oint_{\Gamma_2} \frac{A(z)}{B(z)}\ \frac{dz}{z^{l+1}} \\ & - \frac{1}{2\pi i}\oint_{\Gamma_3} \frac{A(z)}{B(z)}\ \frac{dz}{z^{l+1}}.\end{aligned} \tag{2.6}$$

The first integral in (2.6) has already been estimated in (2.3). The remaining integrals circle just one singularity of the integrand, and thus they can be evaluated

by Cauchy's formula. If ζ is a simple pole of the integrand, and if Γ is a contour taken clockwise around just this singularity of the integrand, then we have

$$\begin{aligned}\frac{1}{2\pi i}\oint_\Gamma \frac{A(z)}{B(z)}\ \frac{dz}{z^{l+1}} &= \operatorname*{Res}_{z=\zeta}\ \frac{A(z)}{B(z)}\ \frac{1}{z^{l+1}} \\ &= \frac{A(\zeta)}{B'(\zeta)}\ \frac{1}{\zeta^{l+1}}.\end{aligned}$$

For the integral around Γ_2, we have $A(\zeta_2) = -(bq-1) + O(q^k)$ and $B'(\zeta_2) = (bq-1) + O(k\,q^k)$ so that

$$-\frac{1}{2\pi i}\oint_{\Gamma_2} \frac{A(z)}{B(z)}\ \frac{dz}{z^{l+1}} = 1 + O(l\,q^k). \tag{2.7}$$

For the integral around Γ_3 we have $A(\zeta_3) = (b-1)^2(bq-1)\,b^{k+1} + O(k/b^{2k})$ and $B'(\zeta_3) = -(bq-1) + O(k/b^k)$ so that

$$-\frac{1}{2\pi i}\oint_{\Gamma_3} \frac{A(z)}{B(z)}\ \frac{dz}{z^{l+1}} = \left(\frac{b-1}{bq-1}\right)^2 b^{l-k}\,q^{l+1} + O(l\,b^{l-2k}\,q^l). \tag{2.8}$$

Substituting the estimates (2.3), (2.7), and (2.8) into (2.6) completes the proof of the proposition. □

We observe that by extending the asymptotic expansions in (2.4) and (2.5), it is possible to extend the expansions in (2.7) and (2.8) and thus reduce their contributions to the error terms in Proposition 2.4. The error term in (2.3), however, cannot be improved without taking account of the zeros of $B(z)$ outside the circle $|z| = 1/q$, which will in general contribute oscillatory terms to the expansion of $\varphi_l(q)$.

Proof of Theorem 1.1. By Corollary 2.3, we may take X to be the number of idle paths from $v^* = 0^k$ to $w^* = 0^k$. We then have

$$\begin{aligned}\operatorname{Ex}[X^2] &= \sum_{u':v^*\to w^*}\ \sum_{u:v^*\to w^*} \Pr[u \text{ idle}, u' \text{ idle}] \\ &= \sum_{u':v^*\to w^*} \Pr[u' \text{ idle}] \sum_{u:v^*\to w^*} \Pr[u \text{ idle} \mid u' \text{ idle}],\end{aligned} \tag{2.9}$$

where the sums are over all paths from v^* to w^*. For each path u', we can find by Lemma 2.1 an automorphism ϑ that carries u' to the path u^* in which all links are labeled 0^*. Applying this automorphism to both u and u' gives $\Pr[u \text{ idle} \mid u' \text{ idle}] = \Pr[\vartheta(u) \text{ idle} \mid u^* \text{ idle}]$, since the probability distribution on states of the graph is invariant under automorphisms. Furthermore,

$$\begin{aligned}\sum_{u:v^*\to w^*} \Pr[u \text{ idle} \mid u' \text{ idle}] &= \sum_{u:v^*\to w^*} \Pr[\vartheta(u) \text{ idle} \mid u^* \text{ idle}] \\ &= \sum_{u:v^*\to w^*} \Pr[u \text{ idle} \mid u^* \text{ idle}],\end{aligned}$$

since both right-hand sides sum the same terms in different orders. Thus the inner sum in (2.9) is independent of u', and we have

$$\operatorname{Ex}[X^2] = \sum_{u':v^*\to w^*} \Pr[u' \text{ idle}] \sum_{u:v^*\to w^*} \Pr[u \text{ idle} \mid u^* \text{ idle}]$$

so that $\mathrm{Ex}[X^2]$ factors as the product of two sums. The first sum is just $\mathrm{Ex}[X]$. To evaluate the second sum, we observe that $\Pr[u \text{ idle} \mid u^* \text{ idle}]$ is just q^j, where j is the number of links on u that are not labeled 0^k. Thus the second sum is $\varphi_l(q)$, and the theorem follows from Proposition 2.5. □

3. The linking probability. Our goal in this section is to prove Theorem 1.2. Thus in this section we shall always assume that $b \geq 2$ and $0 < q < 1$ are fixed and that $k \to \infty$ and $l = ck$ for some fixed rational $c > 1$. Thus the constants in O-terms may depend on c as well as on b and q, but not on k or l. We shall also assume that k is even; the case of odd k requires only that $k/2$ be replaced with $\lfloor k/2 \rfloor$ and $\lceil k/2 \rceil$ in appropriate ways.

LEMMA 3.1. *Let $G^*_{b,k,l}$ be the graph obtained from $G_{b,k,l}$ by reversing the direction of its edges and exchanging the roles of its inputs and outputs. Then $G^*_{b,k,l}$ is isomorphic to $G_{b,k,l}$.*

Proof. The isomorphism takes the vertex with label $a_1 \cdots a_k$ in rank m of $G_{b,k,l}$ to the vertex with label $a_1^* \cdots a_k^*$ in rank $l - m$ of $G^*_{b,k,l}$, where $a_i^* = a_j$ with $j \equiv l + 1 - i \pmod{k}$ (and, conversely, as it is an involution). □

Lemma 3.1 establishes a symmetry between $G_{b,k,l}$ and $G^*_{b,k,l}$, which we shall invoke by use of the term "dually." (When l is even, $G_{b,k,l}$ is in fact isomorphic to a graph with manifest bilateral symmetry, as is shown in the appendix of Pippenger [P3].)

LEMMA 3.2. *Let $\langle G_{b,k,l} \rangle_{m,n}$, with $0 \leq m \leq n \leq l$, be the subgraph of $G_{b,k,l}$ comprising the vertices in ranks m (now considered inputs) through n (now considered outputs) and the edges in stages $m + 1$ through n. Then $\langle G_{b,k,l} \rangle_{m,n}$ is isomorphic to $G_{b,k,n-m}$.*

Proof. The isomorphism takes the vertex with label $a_1 \cdots a_k$ in rank h of $G_{b,k,n-m}$ to the vertex with label $a'_1 \cdots a'_k$ in rank $m + h$ of $\langle G_{b,k,l} \rangle_{m,n}$, where $a'_i = a_j$ with $j \equiv i + m \pmod{k}$. □

COROLLARY 3.3. *Between any given input and any given output of $\langle G_{b,k,l} \rangle_{m,n}$, there are b^{n-m-k} paths if $n - m \geq k$, and there is either one path or none if $n - m < k$.*

Proof. The proof is immediate from Lemma 3.2 and Corollary 2.2. □

We begin with the upper bound to Q. For $0 < q < q_c$, where $q_c = 1/b^{(c-1)/c}$, we have $Q \to 0$ by (1.2). For $q_c < q < 1$, we shall use the following lemma from Pippenger [P3, Cor. 4.2].

LEMMA 3.4. *Let T_r be a complete balanced b-ary tree of depth r, and let each vertex of T_r (except for the root) be considered idle with probability q independently. Let the random variable Z_r denote the number of leaves (vertices at depth r) for which every vertex on the path from the root (exclusive) to the leaf (inclusive) is idle. Then we have*

$$\Pr[Z_r = 0] = \xi + O(\eta^r)$$

as $r \to \infty$ with $b \geq 2$ and $1/b < q < 1$ fixed, where ξ is the unique solution of the equation $\bigl(1 - q(1 - \xi)\bigr)^b = \xi$ in the range $0 < \xi < 1$, and $\eta = b\bigl(1 - q(1 - \xi)\bigr)^{b-1} < 1$.

Now set $r = k/2$ and $s = l - k/2$. The paths from an input v to links in rank r of $G_{b,k,l}$ form a tree isomorphic to T_r (if we ignore the directions of the edges), and the paths from links in rank s to an output w form a disjoint tree isomorphic to T_r. Thus the number of links u in rank r for which all the links on the path from v to u are idle is a random variable U with the same distibution as Z_r. Dually, the number of links u in rank s for which all the links on the path from u to w are idle is an independent random variable U' with the same distribution as Z_r. If v and w are linked, then we

must have $U \geq 1$ and $U' \geq 1$, so by Lemma 3.1 we have

$$Q \leq \Pr[U \geq 1, U' \geq 1] = (1 - \xi)^2 + O(\eta^r).$$

This completes the upper bound for Theorem 1.2.

We now turn to the lower bound for Theorem 1.2. Since this result has been proved for $c \leq 2$ in Pippenger [P3], we shall assume that $c > 2$. (This assumption could of course be avoided, but it would require a more complicated choice of parameters and consideration of cases.) For $0 < q < q_c$, there is nothing to prove, since Q is certainly nonnegative. For $q_c < q < 1$, we shall use the following lemma from Pippenger [P3, Lem. 8.1].

LEMMA 3.5. *With Z_r as in Lemma 3.4 and $1 \leq H \leq (bq)^r$, we have*

$$\Pr[Z_r \leq H] \leq \xi + O\left(\left(H/(bq)^r\right)^\alpha\right)$$

as $r \to \infty$ with $b \geq 2$ and $1/b < q < 1$ fixed, where $\alpha = \log(1/\eta)\big/\log(bq)$ and η is as in Lemma 3.4.

Supposing that $q_c < q < 1$, we shall define

$$q_* = q_{c-1}\, q^{1/(c-1)^2}.$$

We observe that $q < 1$ implies $q_* < q_{c-1}$ and that $q_c < q$ implies $q_* < q$.

LEMMA 3.6. *Let $k \to \infty$ and $l = ck$, with $b \geq 2$, $q_c < q < 1$, and $c > 2$ all fixed. Then for all sufficiently large k, we have*

$$\psi_h(q_*) \leq k$$

for all $0 \leq h \leq l - k$.

Proof. From Proposition 2.4 we have

$$\varphi_h(q_*) = \left(\frac{b-1}{bq_* - 1}\right)^2 b^{h-k} q_*^{h+1} + 1 + O(hb^{h-2k} q_*^h) + O(hq_*^k) + O(hq_*^h).$$

Since $q_* < q_{c-1}$ and $h \leq l - k$, each term is O(1), and thus at most k for all sufficiently large k. □

Let

$$H = \lceil (bq_*)^r \rceil.$$

We observe that v and w will be linked if the following three events occur:

I. The input v is joined by paths containing only idle links to all the links in a set V containing at least H idle links in rank r.

II. All the links in a set W containing at least H idle links in rank s are joined by paths containing only idle links to the output w.

III. There is at least one path containing only idle links from some link in V to some link in W.

By Lemma 3.5, we have

$$\Pr[I] \geq 1 - \xi + O\left(\left(q_*/q\right)^r\right),$$

and since $q_* < q$ we have $\Pr[I] \to 1 - \xi$. Dually, we have by Lemma 3.5

$$\Pr[II] \geq 1 - \xi + O\left(\left(q_*/q\right)^r\right),$$

and thus also $\Pr[II] \to 1 - \xi$. Since events I and II are independent, we have $\Pr[I, II] \to (1-\xi)^2$. Thus to complete the proof of the lower bound for Theorem 1.2, it will suffice to show that

$$\Pr[III \mid I, II] \to 1.$$

Event III depends on events I and II through the sets V and W. We can avoid having to consider this dependence by showing that $\Pr[III] \to 1$ for *any* sets V and W each containing at least H links. Thus it will suffice to prove the following propostion.

PROPOSITION 3.7. *Let V and W be any sets of links in ranks r and s, respectively, each containing at least H links. Then*

$$\Pr[III] \to 1$$

as $k \to \infty$ with $l = ck$, and with $b \geq 2$, $c > 2$, and $q_c < q < 1$ all fixed.

Proof. Since $\Pr[III]$ can only increase if links are added to V or W, we may assume that V and W each contain *exactly* H links. Also, since $\Pr[III]$ can only increase if q is increased, it will suffice to estimate $\Pr[III]$, assuming the vacancy probability to be $q_* < q$ rather than q.

Let the random variable Y be the number of paths containing only idle links joining some link in V (exclusive) to some link in W (exclusive). Then event III is equivalent to $Y \geq 1$ and thus it will suffice to show that $\Pr[Y = 0] \to 0$. To do this, we shall use Chebyshev's inequality:

$$\Pr[Y = 0] \leq \frac{\operatorname{Var}[Y]}{\operatorname{Ex}[Y]^2}.$$

Each path from a link in rank r (exclusive) to a link in rank s (exclusive) contains $s - r - 1 = l - k - 1$ links. Since each of these links is independently idle with probability q_*, the probability that such a path contains only idle links is q_*^{l-k-1}. By Corollary 3.3, the number of such paths joining a given link in rank r with a given link in rank s is $b^{s-r-k} = b^{l-2k}$. Since there are H links in each of V and W, we have

$$\operatorname{Ex}[Y] = H^2\, b^{l-2k}\, q_*^{l-k-1}.$$

Next we must estimate $\operatorname{Var}[Y]$. We have

$$\begin{aligned}\operatorname{Var}[Y] &= \sum_{u':V\to W} \sum_{u:V\to W} \big(\Pr[u, u' \text{ idle}] - \Pr[u \text{ idle}]\ \Pr[u' \text{ idle}]\big) \\ &= \sum_{u':V\to W} \Pr[u' \text{ idle}] \sum_{u:V\to W} \big(\Pr[u \text{ idle} \mid u' \text{ idle}] - \Pr[u \text{ idle}]\big).\end{aligned}$$

Here each sum is over all H^2 paths joining a link in V to a link in W, so there are H^4 terms in all. If u is a path from a link in rank r to a link in rank s, let $\varrho(u)$ denote the link in rank r and $\sigma(u)$ the link in rank s. By Lemma 2.1, we may assume (as in the proof of Theorem 1.1) that $u' = u^*$ is part of a path from $v^* = 0^k$ through $\varrho(u') = 0^k$ and $\sigma(u') = 0^k$ to $w^* = 0^k$, in which all the links have label 0^k. Thus we have

$$\operatorname{Var}[Y] = H^2\, b^{l-2k}\, q_*^{l-k-1} \sum_{u:V\to W} \big(\Pr[u \text{ idle} \mid u^* \text{ idle}] - \Pr[u \text{ idle}]\big).$$

The factor $H^2\, b^{l-2k}\, q_*^{l-k-1}$ multiplying the sum is $\operatorname{Ex}[Y]$, so to show that

$$\operatorname{Var}[Y]/\operatorname{Ex}[Y]^2 \to 0,$$

it will suffice to show that $J/\mathrm{Ex}[Y] \to 0$, where

$$J = \sum_{u:V\to W} \big(\Pr[u \text{ idle} \mid u^* \text{ idle}] - \Pr[u \text{ idle}]\big).$$

We now partition the paths u into four classes as follows:

i. those for which $\varrho(u) = \sigma(u) = 0^k$;
ii. those for which $\varrho(u) \neq 0^k$ but $\sigma(u) = 0^k$;
iii. those for which $\varrho(u) = 0^k$ but $\sigma(u) \neq 0^k$; and
iv. those for which $\varrho(u) \neq 0^k$ and $\sigma(u) \neq 0^k$.

We shall denote the contributions to J over these four classes by J_{i}, J_{ii}, J_{iii}, and J_{iv}, respectively, and estimate them in turn.

For J_{i}, we have

$$\begin{aligned} J_{\mathrm{i}} &\le \sum_{u:0^k\to 0^k} \Pr[u \text{ idle} \mid u^* \text{ idle}] \\ &= \varphi_{s-r}(q_*) \\ &\le k \end{aligned}$$

by Lemma 3.6. Thus we have

$$\begin{aligned} \frac{J_{\mathrm{i}}}{\mathrm{Ex}[Y]} &\le \frac{k}{H^2\, b^{l-2k}\, q_*^{l-k-1}} \\ &\le \frac{k}{b^{l-k}\, q_*^{l}} \\ &\to 0, \end{aligned}$$

since $q_* > q_c$.

For J_{ii}, we have

$$J_{\mathrm{ii}} \le \sum_{V\setminus\{0^k\}\to 0^k} \Pr[u \text{ idle} \mid u^* \text{ idle}].$$

To estimate $\Pr[u \text{ idle} \mid u^* \text{ idle}]$, let i be the first rank for which a link in u has label 0^k. Since there are two distinct paths in $\langle G_{b,k,l}\rangle_{0,i}$ from v^* through $\varrho(u^*) = 0^k$ and $\varrho(u) \neq 0^k$ to this link, we must have $i \ge k+1$ by Corollary 3.3. Thus we have

$$\begin{aligned} J_{\mathrm{ii}} &\le (H-1)\left(\sum_{k+1\le i\le k+r} q_*^{i-r-1}\varphi_{s-i}(q_*) + \sum_{k+r+1\le i\le s} b^{i-r-k} q_*^{i-r-1}\varphi_{s-i}(q_*)\right) \\ &\le (H-1)k\left(\sum_{k+1\le i\le k+r} q_*^{i-r-1} + \sum_{k+r+1\le i\le s} b^{i-r-k} q_*^{i-r-1}\right), \end{aligned}$$

where the factor of $H-1$ accounts for the choice of $\varrho(u) \in V \setminus \{0^k\}$, the factors preceding $\varphi_{s-i}(q_*)$ in the sums account for the probability that all the links on u between ranks r (exclusive) and i (exclusive) are idle, the factors of $\varphi_{s-i}(q_*)$ account for the probability that all the links of u between ranks i and s that are not labeled 0^k are idle, and we have bounded $\varphi_{s-i}(q_*)$ using Lemma 3.6. Bounding the sums by the

number of terms (at most $s - r = l - k$) times the largest term (the first for the first sum, and the last for the second), we have

$$J_{\mathrm{ii}} \le (H-1)k\,(l-k)(q_*^{k/2} + b^{l-2k} q_*^{l-k-1}).$$

Thus we have

$$\begin{aligned}\frac{J_{\mathrm{ii}}}{\mathrm{Ex}[Y]} &\le \frac{k(l-k)(q_*^{k/2} + b^{l-2k} q_*^{l-k-1})}{H\, b^{l-2k}\, q_*^{l-k-1}} \\ &\le k(l-k)\left(\frac{1}{b^{l-3k/2}\, q_*^{l-k}} + \frac{1}{(bq_*)^{k/2}}\right) \\ &\to 0,\end{aligned}$$

since $b^{c-3/2} q_*^{c-1} > b^{c-3/2} q_c^{c-1} = b^{-1/2} q_c^{-1} > b^{-1/2} q_2^{-1} = 1$ (because $q_* > q_c$, $b^{c-1} q_c^c = 1$, $q_c < q_2$, and $bq_2^2 = 1$) and $bq_* > 1$ (because $q_* > q_\infty = 1/b$).

Dually, we have $J_{\mathrm{iii}}/\mathrm{Ex}[Y] \to 0$.

Finally, for J_{iv} we have

$$\begin{aligned}J_{\mathrm{iv}} &= \sum_{u: V\setminus\{0^k\} \to W\setminus\{0^k\}} \big(\Pr[u \text{ idle} \mid u^* \text{ idle}] - \Pr[u \text{ idle}]\big) \\ &= \sum_{\substack{u: V\setminus\{0^k\} \to W\setminus\{0^k\} \\ u \cap u^* \ne \emptyset}} \big(\Pr[u \text{ idle} \mid u^* \text{ idle}] - \Pr[u \text{ idle}]\big) \\ &\le \sum_{\substack{u: V\setminus\{0^k\} \to W\setminus\{0^k\} \\ u \cap u^* \ne \emptyset}} \Pr[u \text{ idle} \mid u^* \text{ idle}],\end{aligned}$$

since if $u \cap u^* = \emptyset$, the events "u idle" and "u^* idle" are independent, and the summand $\Pr[u \text{ idle} \mid u^* \text{ idle}] - \Pr[u \text{ idle}]$ vanishes. Given a path u with $u \cap u^* \ne \emptyset$, let i be the first rank in which u has a link with label 0^k, and let $j \ge i$ be the last such rank. As in case ii, we have $k+1 \le i$, and dually we have $j \le l-k-1$. Thus we have

$$\begin{aligned}J_{\mathrm{iv}} \le (H-1)^2 \Bigg(&\sum_{k+1 \le i \le k+r} \; \sum_{\substack{l-k-r \le j \le l-k-1 \\ i \le j}} q_*^{i-r-1} \varphi_{j-i}(q_*) q_*^{s-j-1} \\ &+ \sum_{k+1 \le i \le k+r} \; \sum_{\substack{r \le j \le l-k-r-1 \\ i \le j}} q_*^{i-r-1} \varphi_{j-i}(q_*) b^{s-j-k} q_*^{s-j-1} \\ &+ \sum_{k+r+1 \le i \le s} \; \sum_{\substack{l-k-r \le j \le l-k-1 \\ i \le j}} b^{i-r-k} q_*^{i-r-1} \varphi_{j-i}(q_*) q_*^{s-j-1} \\ &+ \sum_{k+r+1 \le i \le s} \; \sum_{\substack{r \le j \le l-k-r-1 \\ i \le j}} b^{i-r-k} q_*^{i-r-1} \varphi_{j-i}(q_*) b^{s-j-k} q_*^{s-j-1} \Bigg).\end{aligned}$$

Here we have broken the sum into four parts, according to whether $k+1 \le i \le k+r$ or $k+r+1 \le i \le s$, and also according to whether $l-k-r \le j \le l-k-1$ or $r \le j \le l-k-r-1$. (We note that the second and third double sums will vanish unless

$c > 5/2$, and the fourth double sum will vanish unless $c > 3$.) The factor of $(H-1)^2$ accounts for the choice of $\varrho(u) \in V \setminus \{0^k\}$ and $\sigma(u) \in W \setminus \{0^k\}$, the factors preceding $\varphi_{j-i}(q_*)$ in the summands account for the probability that the links of u in ranks less than i are idle, the factors of $\varphi_{j-i}(q_*)$ account for the probability that the links of u between i and j and not labeled 0^k are idle, and the factors following $\varphi_{j-i}(q_*)$ in the summands account for the probability that the links of u in ranks greater than j are idle. Bounding the factors $\varphi_{j-i}(q_*)$ using Lemma 3.6, and bounding each double summation by the number of terms (at most $(l-k)^2$) times the largest term (which occurs for $i = k+1$ and $j = l-k-r$ in the first sum, and for $i = j$ in the remaining three sums), we obtain

$$J_{\mathrm{iv}} \le (H-1)^2 k (l-k)^2 \left(q_*^k + 2b^{l-5k/2-1} q_*^{l-k-2} + b^{l-3k} q_*^{l-k-2} \right).$$

Thus we have

$$\begin{aligned} \frac{J_{\mathrm{iv}}}{\mathrm{Ex}[Y]} &\le k(l-k)^2 \left(\frac{1}{(bq_*)^{l-2k}} + \frac{2}{q_* b^{k/2+1}} + \frac{1}{q_* b^k} \right) \\ &\to 0, \end{aligned}$$

since $bq_* > 1$, $c > 2$, and $b \ge 2$. This completes the proof of the proposition, and with it the proof of Theorem 1.2. □

4. Conclusion. We have determined the limiting value of the linking probability in spider-web networks with scale k and depth l when $l = ck$ with $c > 1$. The same method could be used when $l/k \to \infty$ but $\big(\log(l+1)\big)/(k+1) \to 0$. In this case, the phase transition would be less abrupt: the limiting value of Q, and even its first derivative with respect to q, would be continuous at the critical value $q_\infty = 1/b$, but the second derivative would be discontinuous. Little would be gained by such networks, however, over those with a large fixed value of c: Their great cost would decrease the critical vacancy probability through only a small interval $[q_\infty, q_c]$, and would provide only a small linking probability in this interval.

Another extension of our results would be to consider, instead of the "independent" probability distribution on states introduced by Lee [L1] and Le Gall [L2, L3], the "coherent" distribution introduced by Pippenger [P1]. (The similar distribution introduced by Koverninskiĭ [K] does not have an obvious generalization for $c > 2$, and in any case it does not seem likely that the additional independence in Koverninskiĭ's model would have much effect on its tractability for $c > 2$.)

Yet another line of inquiry would be to consider the computational complexity of path-search problems for spider-web networks with $c > 2$, using the link-probe model introduced by Lin and Pippenger [L4]. Such results were obtained by Pippenger [P4] for $c = 2$ (and these results are easily extended to the case $1 < c < 2$), but even for $c = 2$ the known results are far from definitive.

Acknowledgments. The results reported in this paper were obtained during the fourth meeting of the Institute for Elementary Studies, the Focused Research Group on Problems in Discrete Probability, held July 12–26, 2003, at the Banff International Research Station in Banff, Alberta, Canada. The author is especially grateful to Yuval Peres, one of the organizers of that meeting, for urging continued faith in the power of the second-moment method.

REFERENCES

[B] S. R. Broadbent and J. M. Hammersley, *Percolation processes.* I. *Crystals and mazes*, Proc. Cambridge Philos. Soc., 53 (1957), pp. 629–641.

[C] F. R. K. Chung and F. K. Hwang, *The connection patterns of two complete binary trees*, SIAM J. Alg. Disc. Meth., 1 (1980), pp. 322–335.

[G] G. Grimmett, *Percolation*, 2nd ed., Springer-Verlag, Berlin, 1999.

[I] N. Ikeno, *A limit of crosspoint number*, IEEE Trans. Circuit Theory, 6 (1959), pp. 187–196.

[K] I. V. Koverninskiĭ, *Estimation of the blocking probability for switching circuits by means of probability graphs*, Problems Inform. Transmission, 11 (1975), pp. 63–71.

[L1] C. Y. Lee, *Analysis of switching networks*, Bell System Tech. J., 34 (1955), pp. 1287–1315.

[L2] P. Le Gall, *Étude du blocage dans les systèmes de commutation téléphoniques automatiques utilisant des commutateurs électroniques du type crossbar*, Ann. Télécommun., 11 (1956), pp. 159–171; 180–194; 197.

[L3] P. Le Gall, *Méthod de calcul de l'encombrement dans les systèmes téléphoniques automatiques a marquage*, Ann. Télécommun., 12 (1957), pp. 374–386.

[L4] G. Lin and N. Pippenger, *Routing algorithms for switching networks with probabilistic traffic*, Networks, 28 (1996), pp. 21–29.

[P1] N. Pippenger, *On crossbar switching networks*, IEEE Trans. Comm., COM-23 (1975), pp. 646–659.

[P2] N. Pippenger, *The blocking probability of spider-web networks*, Random Structures Algorithms, 2 (1991), pp. 121–149.

[P3] N. Pippenger, *The asymptotic optimality of spider-web networks*, Discrete Appl. Math., 37/38 (1992), pp. 437–450.

[P4] N. Pippenger, *Upper and lower bounds for the average-case complexity of path-search*, Networks, 33 (1999), pp. 249–259.

[T] K. Takagi, *Design of multi-stage link systems by means of optimum channel graphs*, Electron. Commun. Japan, 51 (1968), pp. 37–46.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 160–170

FULL RANK TILINGS OF FINITE ABELIAN GROUPS*

MICHAEL DINITZ†

Abstract. A *tiling* of a finite abelian group G is a pair (V, A) of subsets of G such that 0 is in both V and A and every $g \in G$ can be uniquely written as $g = v + a$ with $v \in V$ and $a \in A$. Tilings are a special case of normed factorizations, a type of factorization by subsets that was introduced by Hajós [*Casopsis Pěst Path. Rys.*, 74, (1949), pp. 157–162]. A tiling is said to be *full rank* if both V and A generate G. Cohen, Litsyn, Vardy, and Zémor [*SIAM J. Discrete Math.*, 9 (1996), pp. 393–412] proved that any tiling of $\mathbb{Z}_2^n$ can be decomposed into full rank and trivial tilings. We generalize this decomposition from $\mathbb{Z}_2^n$ to all finite abelian groups. We also show how to generate larger full rank tilings from smaller ones, and give two sufficient conditions for a group to admit a full rank tiling, showing that many groups do admit them. In particular, we prove that if $p \geq 5$ is a prime and $n \geq 4$, then $\mathbb{Z}_p^n$ admits a full rank tiling. This bound on n is tight for $5 \leq p \leq 11$, and is conjectured to be tight for all primes p.

Key words. tiling, full rank, finite abelian group, factorization, direct sum, Hamming codes

AMS subject classifications. 05B45, 20K01

DOI. 10.1137/S0895480104445794

1. Introduction.

Throughout this paper G is a finite abelian group. A *factorization* of G is a collection $(A_1, \ldots, A_k)$ of subsets such that every $g \in G$ can be uniquely represented as $a_1 + \cdots + a_k$, where $a_i \in A_i$. A factorization is *normed* if every subset in the factorization contains 0. A *tiling* is a special case of a normed factorization in which there are only two subsets (usually denoted V and A rather than A_1 and A_2). Any subset V for which there exists a subset A such that (V, A) is a tiling of G is called a *tile* of G. Cohen, Litsyn, Vardy, and Zémor first introduced this definition of a tiling in 1996 for the special case of tilings of $\mathbb{Z}_2^n$ in [2], but it extends perfectly well to arbitrary finite abelian groups. Before then, there was no separate term for a normed factorization into two subsets, despite the fact that they had been studied by Hajós [7], Rédei [14], Sands [16], and others. The term "tiling" was a natural choice since all of [2] is phrased in terms of $\mathbb{F}_2^n$ rather than $\mathbb{Z}_2^n$ and a tiling of a vector space is a natural concept. In particular, tilings of the Euclidean space $\mathbb{R}^n$ have been studied extensively (see [15, 18]). But since tilings do not depend on multiplicative structure, $\mathbb{F}_2^n$ is identical to $\mathbb{Z}_2^n$ with respect to tilings, and hence it suffices to look at finite abelian groups rather than vector spaces over finite fields.

The study of factorizations of finite abelian groups by subsets was introduced by Hajós in 1941 [6] as a tool to prove a conjecture on homogenous linear forms posed by Minkowski. Hajós then began to study a certain type of factorization which he called periodic (see [7]). A subset $A \subseteq G$ is *periodic* if there is some nonidentity element $g \in G$ such that $g + A = A$ and a *periodic factorization* is a factorization in which one of the subsets is periodic. Hajós asked for which groups G any factorization into two subsets $G = A + B$ necessarily has either A or B periodic. This question was eventually solved by Sands [16] after major contributions from de Bruijn [3] and Rédei [14].

*Received by the editors August 26, 2004; accepted for publication (in revised form) September 1, 2005; published electronically March 3, 2006. This research was done at the University of Minnesota-Duluth and was supported by NSF grant DMS-0137611 and NSA grant H-98230-04-1-0050.
http://www.siam.org/journals/sidma/20-1/44579.html
†Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213 (mdinitz@cs.cmu.edu).

A group G possesses the *Rédei property* if in every tiling (V, A) of G either V or A is contained in a proper subgroup of G. The question of which groups possess the Rédei property has been investigated since 1979, when Rédei [14] conjectured that $\mathbb{Z}_p^3$ has the Rédei property for all primes p. If G does not possess the Rédei property then there is some tiling (V, A) of G in which $\langle V \rangle = \langle A \rangle = G$, where $\langle S \rangle$ denotes the subgroup generated by S for any $S \subseteq G$. These tilings are said to be *full rank* [2]. Note that having the Rédei property is equivalent to not admitting a full rank tiling. Sands [17] asked whether every group has the Rédei property, which was shown not to be the case by Fraser and Gordon [5], who used results from coding theory to construct a full rank tiling of $\mathbb{Z}_5^6$ as a counterexample.

Until recently the only motivation for studying full rank tilings was to find out which groups had the Rédei property. Then in 1996 Cohen, Litsyn, Vardy, and Zémor [2] found that any tiling of $\mathbb{Z}_2^n$ can be decomposed into full rank tilings and trivial tilings (a tiling is *trivial* if one of V or A is $\mathbb{Z}_2^n$ and the other is just the zero vector). This provided extra motivation for studying which elementary 2-groups (groups of the form $\mathbb{Z}_2^n$) admit full rank tilings (or equivalently do not have the Rédei property). Cohen, Litsyn, Vardy, and Zémor [2] showed that there do not exist full rank tilings of $\mathbb{Z}_2^n$ when $n \leq 7$ and that there do exist full rank tilings of $\mathbb{Z}_2^n$ when $n \geq 112$. Etzion and Vardy [4] then constructed full rank tilings for $n \geq 14$ using techniques that, together with unpublished work of LeVan and Phelps, were used to construct full rank tilings when $n \geq 10$. Trachtenberg and Vardy then proved that $\mathbb{Z}_2^8$ does not admit a full rank tiling [24], and the question of full rank tilings of $\mathbb{Z}_2^n$ was resolved when Östergard and Vardy [9] showed that $\mathbb{Z}_2^9$ does not admit a full rank tiling. All of the work done on full rank tilings of $\mathbb{Z}_2^n$ was actually done in terms of $\mathbb{F}_2^n$, since the authors were approaching the problem from a coding theory perspective and were apparently not aware of much of the work done on the Rédei property or the connection of full rank tilings to it.

It is interesting to note that work on full rank tilings of $\mathbb{F}_2^n$ and work on the Rédei property have proceeded almost independently. In the paper which started work on tilings of $\mathbb{F}_2^n$, Cohen, Litsyn, Vardy, and Zémor [2] reference the work of Hajós on periodic factorizations but do not reference any of the work done on the Rédei property, and neither do any of the papers mentioned above that extend the work of [2]. The only exception to this is a paper by Szabó and Ward [21] in which they reference work done on the Rédei property to prove the existence of full rank tilings of $\mathbb{F}_2^n$ for $n \geq 14$.

We begin in section 2 by generalizing the decomposition of Cohen, Litsyn, Vardy, and Zémor [2, section 6] from $\mathbb{Z}_2^n$ to arbitrary finite abelian groups. Then in section 3 we generalize a construction of Etzion and Vardy [4, section 5] and Szabó and Ward [21, Lemma 1] to create a full rank tiling of a group from a full rank tiling of one of its direct factors. Using this we devise two sufficient conditions for a group to admit a full rank tiling, showing that many groups admit them. The first condition states a group admits a full rank tiling if it contains as a direct factor a subgroup of the type $\mathbb{Z}_a \times \mathbb{Z}_b \times \mathbb{Z}_c$ with $a, b,$ and c composite. This is based on work done by Szabó in [19]. Then in section 4 we extend the work done for $\mathbb{Z}_2^n$ by showing that any group containing $\mathbb{Z}_p^n$ with $p \geq 5$ prime and $n = 4$ as a direct factor admits a full rank tiling. Thus, there exists a full rank tiling of $\mathbb{Z}_p^n$ if $p \geq 5$ is prime and $n \geq 4$. A conjecture of Rédei [14] implies that this is tight for all primes. This conjecture has been verified for primes less than or equal to 11 by Szabó and Ward [22], which completely solves the question of whether there exist full rank tilings of $\mathbb{Z}_p^n$ when p is 5, 7, or 11. We conclude by discussing some remaining open problems on tilings.

2. Decomposition of tilings. In this section we study how tilings of arbitrary finite abelian groups can be recursively decomposed, generalizing some of the work done in [2] for $\mathbb{Z}_2^n$. We first develop a certain characterization of tilings which will prove particularly useful. The notation $V - V$ denotes $\{v_1 - v_2 : v_1, v_2 \in V\}$.

PROPOSITION 1. *Let $V, A \subseteq G$ with $0 \in V$ and $0 \in A$. Then (V, A) is a tiling of G if and only if $(V - V) \cap (A - A) = \{0\}$ and $|V||A| = |G|$.*

Proof. Suppose that $(V - V) \cap (A - A) = \{0\}$ and $|V||A| = |G|$. Assume that $v_1 + a_1 = v_2 + a_2$. Then $v_1 - v_2 = a_2 - a_1 = 0$, so $v_1 = v_2$ and $a_1 = a_2$, and thus the representation of each element of $V + A$ is unique. Since $|V||A| = |G|$, we have that $V + A = G$ and thus (V, A) is a tiling of G.

Now let (V, A) be a tiling of G, and suppose that $(V - V) \cap (A - A) \neq \{0\}$. Then there exist distinct elements v_1 and v_2 in V and a_1 and a_2 in A such that $v_1 - v_2 = a_1 - a_2$, and so $v_1 + a_2 = v_2 + a_1$. Thus (V, A) is not a tiling. If $(V - V) \cap (A - A) = \{0\}$ and $|V||A| \neq |G|$, then clearly $|V||A| < |G|$ so some element of G is not in $V + A$ and thus (V, A) is not a tiling. □

Note that the $|V||A| = |G|$ condition can be replaced with the condition $V + A = G$ if needed. To motivate our discussion of full rank tilings, we give one reason why the subgroup generated by a tile is of interest.

PROPOSITION 2. *A subset $V \subseteq G$ is a tile of G if and only if it is a tile of $\langle V \rangle$.*

Proof. Suppose that V is a tile of $\langle V \rangle$. Since $\langle V \rangle$ is a subgroup of G it is clearly a tile of G. Let $(\langle V \rangle, A_1)$ be a tiling of G and let (V, A_0) be a tiling of $\langle V \rangle$. Then clearly $(V, A_0 + A_1)$ is a tiling of G.

Suppose that (V, A) is a tiling of G. Let $A_0 = A \cap \langle V \rangle$. Since $A_0 \subseteq A$ and $(V - V) \cap (A - A) = \{0\}$, we have that $(V - V) \cap (A_0 - A_0) = \{0\}$. Clearly $V + A_0 \subseteq \langle V \rangle$. Since $\langle V \rangle \subseteq G = V + A$, any $w \in \langle V \rangle$ can be written as $w = v + a$ with $v \in V$ and $a \in A$. Then $a = w - v \in \langle V \rangle$ since $\langle V \rangle$ is a subgroup, and so $a \in A_0$. Hence $\langle V \rangle \subseteq V + A_0$, so $V + A_0 = \langle V \rangle$ and thus (V, A_0) is a tiling of $\langle V \rangle$. □

Because of this proposition we are naturally interested in tilings (V, A) in which $\langle V \rangle = G$. Tilings with this property are called *proper tilings*, a term devised by Cohen, Litsyn, Vardy, and Zémor [2] that was originally used only for tilings of $\mathbb{Z}_2^n$. The following theorem is a generalization to arbitrary finite abelian groups of Theorem 6.2 in [2], the original decomposition showing that every tiling of $\mathbb{Z}_2^n$ can be decomposed into proper tilings of its subgroups. This generalization shows that the classification of all tilings of G can be reduced to the study of all proper tilings of the subgroups of G.

THEOREM 3. *Let V be a tile of G with $\langle V \rangle \neq G$. Let $z = |G|/|V|$, and let $m = |G|/|\langle V \rangle|$. The pair (V, A) is a tiling of G if and only if A has the following form:*

1. *For $i = 0, 1, \ldots, m - 1$, let $A_i \subset \langle V \rangle$ be such that (V, A_i) is a tiling of $\langle V \rangle$.*
2. *Let $c_0 = 0, c_1, \ldots, c_{m-1}$ be a set of coset representatives for $G/\langle V \rangle$.*

Then

$$A = A_0 \cup (c_1 + A_1) \cup \cdots \cup (c_{m-1} + A_{m-1}). \tag{1}$$

Proof. Suppose that A is as in (1). Then $|A_i| = z/m$ so $|A| = z$ and $|V||A| = |G|$. So we just need to show that $(V - V) \cap (A - A) = \{0\}$. Note that any element of $A - A$ has one of the following forms:

1. $(c_i + a_i) - (c_i + a_i) = 0$,
2. $(c_i + a_{i1}) - (c_i + a_{i2}) = a_{i1} - a_{i2}$, or
3. $(c_i + a_i) - (c_j + a_j)$, for $i \neq j$,

where $a_i, a_{i1}, a_{i2}, a_j \in A$. Let U denote the set of elements of type 2, and let W denote the set of elements of type 3. Clearly any element of U is also an element of some $A_i - A_i$, so since $(V - V) \cap (A_i - A_i) = \{0\}$ for all i we have that $(V - V) \cap U = \{0\}$. Since $c_i - c_j \notin \langle V \rangle$ for all $i \neq j$ and $A_i \subset \langle V \rangle$ for all i, it follows that $\langle V \rangle$ and W are disjoint, so $(V - V) \cap W = \emptyset$ and hence (V, A) is a tiling of G.

Now let (V, A) be a tiling of G. Pick a set of representatives $c_0 = 0, c_1, \ldots, c_{m-1}$ of $G/\langle V \rangle$ and let $A_i = -c_i + (A \cap (c_i + \langle V \rangle))$. We start by showing that we can always pick representatives of $G/\langle V \rangle$ so that $0 \in A_i$ for all i. If $0 \notin A_i$ for some i, then let $a_i \in A_i$ and let $c_i' = a_i + c_i$. Note that c_i' represents the same coset of $\langle V \rangle$ as c_i since $a_i \in A_i \subset \langle V \rangle$. If we let A_i' be the set we get by replacing c_i with c_i' in the definition of A_i, then we get that $A_i' = -a_i - c_i + (A \cap (a_i + c_i + \langle V \rangle)) = -a_i + A_i$. Together with the fact that $a_i \in A_i$, this gives us that $0 \in A_i'$, so we could have simply started with c_i' instead of c_i. Thus we can assume the $0 \in A_i$ for all i.

We have that $c_i + A_i = A \cap (c_i + \langle V \rangle)$, so

$$\bigcup_{i=0}^{m-1} (c_i + A_i) = \bigcup_{i=0}^{m-1} (A \cap (c_i + \langle V \rangle)) = A. \tag{2}$$

Now we need to show that (V, A_i) is a tiling of $\langle V \rangle$ for all i. Any element of A_i is of the form $-c_i + a$, so any element of $A_i - A_i$ is of the form $a_1 - a_2$. So $A_i - A_i \subseteq A - A$ and thus $(A_i - A_i) \cap (V - V) = \{0\}$. Note that $A_i \subset \langle V \rangle$, so $V + A_i \subseteq \langle V \rangle$. Thus to establish that (V, A_i) is a tiling of $\langle V \rangle$, it remains to show that $|A_i| = z/m$. Since $(V - V) \cap (A_i - A_i) = \{0\}$ and $V + A_i \subseteq \langle V \rangle$, we obviously have that $|A_i| \leq z/m$. However, $z = |A| \leq \sum_{i=0}^{m-1} |A_i|$ by (2), so $|A_i| = z/m$ for all i. □

Theorem 3 implies that if all of the proper tilings of the subgroups of G are known, then we can construct all the tilings of G. However, proper tilings can be decomposed further by simply switching the roles of V and A. Suppose that (V, A) is a (proper) tiling of $\langle V \rangle$, and consider the tiling (A, V). Unless $\langle A \rangle = \langle V \rangle$ this tiling is not proper, so by the above theorem

$$V = V_0 \cup (c_1 + V_1) \cup \cdots \cup (c_{m-1} + V_{m-1}), \tag{3}$$

where (A, V_i) is a proper tiling of $\langle A \rangle$ for all i and the elements $0, c_1, \ldots, c_m$ are representatives of $\langle V \rangle / \langle A \rangle$. So by using (3), each of the tilings (V, A_i) of Theorem 3 can be decomposed into tilings of subgroups unless $\langle V \rangle = \langle A_i \rangle$. This process can be iterated until the remaining tilings are either trivial or of full rank. So any tiling can be decomposed into full rank and trivial tilings of its subgroups.

We can, however, decompose full rank tilings even further, into nonperiodic full rank tilings. For any subset $A \subseteq G$, let $A_0 = \{g \in G : g + A = A\}$ denote the set of periodic points of A. By definition $A_0 = \{0\}$ if and only if A is nonperiodic. In the literature A_0 is sometimes referred to as the *kernel* of A (see [1, 10, 12]), particularly in regard to tilings derived from codes. Note that if $0 \in A$, then $A_0 \subseteq A$. The following proposition is rather obvious, first appearing in terms of codes over GF(2) [1], but can easily be generalized to finite abelian groups.

PROPOSITION 4. *If $0 \in A$, then A_0 is a subgroup of G contained in A and A is the union of disjoint cosets of A_0.*

Proof. Let $a_1, a_2 \in A_0$. Then $(a_1 + a_2) + A = a_1 + (a_2 + A) = a_1 + A = A$, so $a_1 + a_2 \in A_0$. Since every $a \in A_0$ has some finite order this implies that $-a \in A_0$ and thus A_0 is a subgroup of G. Now let $a \in A$. Then $a + A_0 \in A$ by the definition of

A_0, so A is the union of cosets of A_0. These cosets are clearly disjoint since A_0 is a subgroup of G, proving the proposition. ∎

If $A' \subseteq A$ is a set of representatives for A/A_0, then it follows from this proposition that $A' + A_0 = A$. Now we show how to reduce tilings by the kernel of one of the subsets.

THEOREM 5. *Let (V, A) be a tiling of G, and let A_0 be the kernel of A. Then $(V/A_0, A/A_0)$ is a tiling of G/A_0.*

Proof. Let $\varphi : G \to G/A_0$ be the natural homomorphism. Suppose that the restriction of φ to V (which takes V to V/A_0) is not one-to-one. Then there exist distinct elements v_1 and v_2 in V such that $\varphi(v_1) = \varphi(v_2) = v' + A_0$. So $\varphi(v_1 - v_2) = \varphi(v_1) - \varphi(v_2) = (v' + A_0) - (v' + A_0) = A_0$, which implies that $v_1 - v_2 \in A_0$. This is a contradiction since $A_0 \subseteq A$ and $(V - V) \cap (A - A) = \{0\}$. Hence $|V/A_0| = |V|$, and thus $|V/A_0| \cdot |A/A_0| = |G/A_0|$.

Suppose that there exist distinct elements v_1' and v_2' in V/A_0 and a_1' and a_2' in A/A_0 such that $v_1' - v_2' = a_1' - a_2'$. Then there exist $v_1, v_2 \in V, v_1 \neq v_2$ and $a_1, a_2 \in A, a_1 \neq a_2$ such that $(v_1 - v_2) + A_0 = (a_1 - a_2) + A_0$. So there is some $a_0 \in A_0$ such that $v_1 - v_2 = a_1 - a_2 + a_0$, and since $a_0 \in A_0$ this implies that there is some $a_3 \in A$ such that $v_1 - v_2 = a_1 - a_3$, which is a contradiction since $(V - V) \cap (A - A) = \{0\}$. Thus $(V/A_0, A/A_0)$ is a tiling of G/A_0. ∎

PROPOSITION 6. *If (V, A) is a full rank tiling of G, then $(V/A_0, A/A_0)$ is a full rank tiling of G/A_0.*

Proof. We know from Theorem 5 that $(V/A_0, A/A_0)$ is a tiling of G/A_0, so we just need to show that it is full rank. Let $w + A_0 \in G/A_0$. Since $\langle V \rangle = G$, there are $v_1, \ldots, v_k \in V$, not necessarily distinct, such that $v_1 + \cdots + v_k = w$. Then $(v_1 + A_0) + \cdots + (v_k + A_0) = w + A_0$. Hence $\langle V/A_0 \rangle = G/A_0$. By the same argument, $\langle A/A_0 \rangle = G/A_0$, so $(V/A_0, A/A_0)$ is a full rank tiling of G/A_0. ∎

The following propositions concern the periodicity of the tiling resulting from this decomposition.

PROPOSITION 7. *A/A_0 is nonperiodic.*

Proof. Let a be a periodic point of A/A_0, and let A' be a set of representatives for A/A_0 including 0. Let $c + A_0$ represent a, where $c \in A'$. Then clearly c is a periodic point of A and so is an element of A_0. However, $A' \cap A_0 = \{0\}$, and hence $c = 0$, so $a = 0$. ∎

PROPOSITION 8. *V/A_0 is periodic if V is periodic.*

Proof. Let v_0 be a nonzero periodic point of V. Then since $v_0 + v \in V$ for any $v \in V$, we have that $\varphi(v_0) + \varphi(v) \in \varphi(V)$, so $\varphi(v_0)$ is a periodic point of V/A_0. From the proof of Theorem 5 we know that $|V| = |V/A_0|$, so $\varphi(v_0) \neq 0$ and thus V/A_0 is periodic. ∎

By Proposition 8, after an application of Theorem 5 we can switch V/A_0 and A/A_0 and apply it again. Since at each iteration one of the subsets loses all of its periodic points, this might seem to imply that this recursion never needs to be carried out more than twice, but it turns out that the other subset can acquire new periodic points. Cohen, Litsyn, Vardy, and Zémor [2, section 8] provide an example of this in $\mathbb{Z}_2^7$. The recursion will stop eventually, though, so we are interested not only in full rank tilings but especially in nonperiodic full rank tilings. Note that Proposition 6 also gives us a way to construct smaller full rank tilings from larger ones, which is helpful when trying to determine which groups admit full rank tilings.

3. Constructing full rank tilings of product groups. Etzion and Vardy [4, Construction C] developed a construction to build a full rank tiling of $\mathbb{Z}_2^{n+1}$ from a

full rank tiling of $\mathbb{Z}_2^n$, and Szabó and Ward [21, Lemma 1] developed a similar but more general construction to allow the direct product with arbitrary cyclic groups rather than just $\mathbb{Z}_2$. We generalize both of these to a construction that gives a full rank tiling of any finite abelian group having some direct factor with a full rank tiling.

THEOREM 9. *If there is a full rank tiling (V, A) of G, then there is a full rank tiling of $G \times H$, where G is any nontrivial finite abelian group and H is any finite abelian group.*

Proof. Szabó and Ward [21, Lemma 1] proved this for the case when $H = \langle k \rangle$ is cyclic and there is an element $a \in A \setminus \{0\}$ such that $\langle A \setminus \{a\} \rangle = G$. They did this by letting $V' = \{(v, h) : v \in V, h \in H\}$ and $A' = \{(a', 0) : a' \in (A \setminus \{a\})\} \cup \{(a, k)\}$ and proving that (V', A') is a full rank tiling of $G \times H$. Note that the element $(0, k)$ is not necessary for $\langle V' \rangle$ to equal $G \times H$ since (v, k) and $(v, k + k)$ are both elements of V' and $(v, k + k) - (v, k) = (0, k)$. So we can switch the roles of V and A and repeat for another cyclic group by letting $(0, k)$ play the role of a. Since any finite abelian group can be decomposed into the direct product of cyclic groups, if there is initially some $a \in A \setminus \{0\}$ that is not necessary for A to generate G, then there is a full rank tiling of $G \times H$ for any finite abelian group H.

The only case when there is not such an a is when both $A \setminus \{0\}$ and $V \setminus \{0\}$ are minimal generating sets of G. Let m equal the sum of the multiplicities of the prime divisors of $|G|$. We first show that any minimal generating set of G has at most m elements. Let $A = \{a_1, \ldots, a_k\}$ be a minimal generating set of G. Let $G_i = \langle a_1, \ldots, a_i \rangle$, where $G_0 = \{0\}$. Note that $\prod_{i=0}^{k-1} |G_{i+1}|/|G_i| = |G|$. We know that G_i is a proper subgroup of G_{i+1} since A is a minimal generating set, which means that $|G_{i+1}|/|G_i| > 1$ for all i. Hence $k \leq m$. So if (V, A) is a full rank tiling and $V \setminus \{0\}$ and $A \setminus \{0\}$ are both minimal generating sets, then $(m + 1)^2 \geq |G|$. Clearly $m \leq \lfloor \log_2 |G| \rfloor$, so $(\lfloor \log_2 |G| \rfloor + 1)^2 \geq |G|$. This is true only if $1 \leq |G| \leq 36$. Since $|G| = |V||A|$, it is only possible for both V and A to have at most $m + 1$ elements when $|G|$ is 2, 4, 6, 8, 9, 12, or 16, so we consider the finite abelian groups of those orders. Clearly any tiling of $\mathbb{Z}_2$ is trivial. Rédei [13] proved that if both V and A have prime order, then one of them is a subgroup of G, which implies that there are no full rank tilings of $\mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{Z}_6$, $\mathbb{Z}_3 \times \mathbb{Z}_3$, $\mathbb{Z}_4$, or $\mathbb{Z}_9$.

For the $|G| = 8$, $|G| = 12$, and $|G| = 16$ cases we need a few results on the Hajós property. We say that a finite abelian group G has the Hajós property if in any tiling (V, A) of G at least one of V and A is periodic. Groups with the Hajós property have been completely classified [16]. In particular, all finite abelian groups of order 8, 12, or 16 have the Hajós property. Szabó [20, Lemma 1] has shown that if a finite abelian group has the Hajós property, then it has no full rank tilings. □

Szabó [19, section 4] has proven that there exists a full rank tiling of the direct product of at least three cyclic groups of composite orders other than 4 or 6. We remove the restriction that the orders not be 4 or 6 and combine it with Theorem 9 to get the following theorem.

THEOREM 10. *If G has $\mathbb{Z}_a \times \mathbb{Z}_b \times \mathbb{Z}_c$ as a direct factor, where a, b, c are composite, then G has a full rank tiling.*

Proof. Let G be the direct product of cyclic groups of orders m_1, m_2, m_3 (all composite) and generators g_1, g_2, g_3 respectively. Let $v_i = m_i/u_i$, where u_i is the smallest prime divisor of m_i. Also let $[g]_m$ denote the set $\{0, g, 2g, \ldots, (m-1)g\}$. If $V = \{(a, b, c) : a \in [g_1]_{u_1}, b \in [g_2]_{u_2}, c \in [g_3]_{u_3}\}$ and $A = \{(a, b, c) : a \in [u_1 g_1]_{v_1}, b \in [u_2 g_2]_{v_2}, c \in [u_3 g_3]_{v_3}\}$, then it is not hard to see that (V, A) is a tiling of G. Let π be some cyclic permutation of $\{1, 2, 3\}$, and define the following two sets:

$$X = \bigcup_{i=1}^{3} \{(a_1, a_2, a_3) : a_i \in [u_i g_i]_{v_i} \text{ and } a_{\pi(i)} = u_{\pi(i)} g_{\pi(i)} \text{ and } a_{\pi^{-1}(i)} = 0\}$$

$$Y = \bigcup_{i=1}^{3} \{(a_1, a_2, a_3) : a_i \in [u_i g_i]_{v_i} + g_i \text{ and } a_{\pi(i)} = u_{\pi(i)} g_{\pi(i)} \text{ and } a_{\pi^{-1}(i)} = 0\}.$$

Note that $X \subset A$. Szabó [19, section 2] proved that if $A' = A \cup Y \setminus X$ then (V, A') is a tiling of G, and the tiling is full rank if v_i is at least 4 for all i. Note that if $j = \pi(i)$ and $v_j = 3$ then $0, u_j g_j + g_i, 2u_j g_j \in A'$, so $2u_j g_j + (u_j g_j + g_i) = g_i \in \langle A' \rangle$. If $v_j > 3$ then $3u_j g_j \in A'$, so $3u_j g_j - 2u_j g_j = u_j g_j \in \langle A' \rangle$ and thus $u_j g_j + g_i - u_j g_j = g_i \in \langle A' \rangle$. So if every v_i is at least 3 then (V, A') is a full rank tiling of G.

If $v_i = 2$ for all i then $u_i = 2$, and G is the group $\mathbb{Z}_4 \times \mathbb{Z}_4 \times \mathbb{Z}_4$. It is easy to check by hand that Szabó's construction results in a full rank tiling. In the final case, there is some $v_j > 2$. Let $i = \pi^{-1}(j)$. Then by the above argument $g_i \in \langle A' \rangle$. Let $k = \pi(j)$. Since π is cyclic, $k = \pi^{-1}(i)$. By definition $u_i g_i + g_k \in A'$, so since $g_i \in \langle A' \rangle$ we have that $g_k \in \langle A' \rangle$. Also, $u_k g_k + g_j \in A'$, so $g_j \in \langle A' \rangle$. Thus $\langle A' \rangle = G$, so (V, A') is a full rank tiling of G. Now by Theorem 9 any group containing G as a direct factor has a full rank tiling, which proves the theorem. □

4. Constructions using codes. In this section we get another sufficient condition for G to admit a full rank tiling by using codes. We will work in vector spaces over finite fields in this section since we will on occasion use properties of the vector space. However, as noted in the introduction a tiling of a vector space is also a tiling of the additive group associated with that space, so at the end of the section we translate our main result back to groups. Throughout this section p is a prime. The *Hamming distance* of two n-tuples is the number of coordinates in which they differ. A *perfect code* is a subset $C \subset \mathbb{F}_q^n$ such that $(C, S_R(0))$ is a tiling of $\mathbb{F}_q^n$, where $S_R(0)$ is the Hamming ball of radius R centered on 0 [8]. Since a Hamming ball clearly generates the entire space, this gives a full rank tiling if the code itself generates the entire space. An important special case of perfect codes are the Hamming codes, which are the linear perfect codes for $R = 1$ (see [8]). A Hamming code forms a proper subspace of $\mathbb{F}_q^n$, and so does not immediately result in a full rank tiling. However, we will see how to slightly modify a Hamming code to get a full rank tiling.

Sands posed the question of whether every group has the Rédei property in [17]. Answering this question in the negative, Fraser and Gordon [5] constructed a full rank tiling of $\mathbb{F}_5^6$ by applying permutations of GF(5) to a Hamming code. They state that their construction generalizes to provide an infinite number of counterexamples, but they omit the details. We begin by generalizing their argument to show that there exist full rank tilings of $\mathbb{F}_p^{p+1}$, where $p \geq 5$ is prime. We do this by starting out with the same code they do, a Hamming code on $\mathbb{F}_p^{p+1}$, and then permute the values in the first two coordinates of the vectors in the code. Permuting only the first two coordinates is a property that will prove important when computing the kernel. Let $p \geq 7$ be prime and let H be the following $2 \times (p+1)$ matrix:

$$H = \begin{pmatrix} 0 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & 2 & \cdots & p-1 \end{pmatrix}.$$

Let $C = \{u \in \mathbb{F}_p^{p+1} : Hu^T = 0\}$. It is easy to show that C is a Hamming code, which implies that $(C, S_1(0))$ is a tiling of $\mathbb{F}_p^{p+1}$. It is not a full rank tiling since C is a proper subspace of $\mathbb{F}_p^{p+1}$ of dimension $p-1$. Let $u_i = (p-i, p-1, 0, \ldots, 0, 1, 0 \ldots, 0)$, where the 1 is in the $(i+2)$nd coordinate. Note that $\{u_1, u_2, \ldots, u_{p-1}\}$ is a basis for C.

Now let π_i, for $i = 1, \ldots, p+1$, be permutations of the elements of GF(p). Then the map

$$\pi : (x_1, \ldots, x_{p+1}) \mapsto (\pi_1(x_1), \ldots, \pi_{p+1}(x_{p+1}))$$

from $\mathbb{F}_p^{p+1}$ to itself clearly preserves the Hamming distance. Hence $\pi(C)$ is still a perfect code with $R = 1$ for any choice of the π_i's. We will use this fact to construct full rank tilings from C.

PROPOSITION 11. *There exists a full rank tiling of $\mathbb{F}_p^{p+1}$ if $p \geq 5$.*

Proof. Let $\pi_1 = ((p-3)(p-4))$ be the transposition interchanging $p-3$ and $p-4$, let $\pi_2 = ((p-2)(p-3))$ be the transposition interchanging $p-2$ and $p-3$, and let every other π_i be the identity permutation. We claim that $(\pi(C), S_1(0))$ is a full rank tiling of $\mathbb{F}_p^{p+1}$ for $p \geq 7$. The basis we constructed of $\mathbb{F}_p^{p+1}$ gets mapped to $p-1$ linearly independent vectors since only the first two coordinates get permuted. Also, $\pi(u_1 + u_2) = (p-4, p-3, 1, 1, 0, \ldots, 0)$ is another linearly independent vector, since otherwise the 1's in the third and fourth coordinates would force it to equal $\pi(u_1) + \pi(u_2)$, which it does not since $\pi(u_1) + \pi(u_2) = (p-3, p-2, 1, 1, 0, \ldots, 0)$.

Now consider the vector $\pi(u_5 + u_{p-1}) = (p-3, p-3, 0, \ldots, 0, 1, 0, \ldots, 0, 1)$, where the 1's are in the seventh and the $p+1$st coordinates. Assume that this is a linear combination of the previous p vectors. Because of the placement of the 1's it is clear that $\pi(u_5)$ and $\pi(u_{p-1})$ each have a coefficient of 1 in this linear combination, so the remaining parts of the linear combination must sum to $\pi(u_5 + u_{p-1}) - \pi(u_5) - \pi(u_{p-1}) = (1, p-1, 0, \ldots, 0)$. Clearly the remaining $\pi(u_i)$'s other than $\pi(u_1)$ and $\pi(u_2)$ do not appear in the linear combination. The only way $\pi(u_1)$ and $\pi(u_2)$ can contribute is if each has the negative coefficient of $\pi(u_1 + u_2)$. If x is the coefficient of $\pi(u_1 + u_2)$, then we get the following two equations from the first and second coordinate, respectively:

$$(p-1)(-x) + (p-2)(-x) + (p-4)x = 1$$

$$(p-1)(-x) + (p-1)(-x) + (p-3)x = p-1$$

The left-hand side of each equation simplifies to $(p-1)x$, which is a contradiction since $(p-1)x$ cannot equal both 1 and $p-1$. Thus the coefficients of $\pi(u_1)$, $\pi(u_2)$, and $\pi(u_1+u_2)$ are zero, so $\pi(u_5+u_{p-1}) = \pi(u_5)+\pi(u_{p-1})$. However, this is a contradiction since $\pi(u_5) + \pi(u_{p-1}) = (p-4, p-2, \ldots)$. Hence $\{\pi(u_i) : 1 \leq i \leq p-1\} \cup \{\pi(u_1 + u_2)\} \cup \{\pi(u_5 + u_{p-1})\}$ is a linearly independent set of size $p+1$, and therefore forms a basis of $\mathbb{F}_p^{p+1}$. Thus $\langle \pi(C) \rangle = \mathbb{F}_p^{p+1}$, so $(\pi(C), S_1(0))$ is a full rank tiling of $\mathbb{F}_p^{p+1}$. Since we used u_5 this only works when $p \geq 7$, but the full rank tiling of $\mathbb{F}_5^6$ given by Fraser and Gordon starts with the same basis as our construction ($\{u_1, u_2, \ldots, u_{p-1}\}$) and just uses different permutations (still only changing the elements in the first two coordinates). ∎

To get even smaller full rank tilings we find the kernel of $\pi(C)$ and use Proposition 6.

PROPOSITION 12. *There exist full rank tilings of $\mathbb{F}_p^4$ when $p \geq 5$.*

Proof. Since C is a Hamming code, it is a subgroup of $\mathbb{F}_p^{p+1}$, and so every element of C is a periodic point. The map π used in Proposition 11 only changes the first two coordinates of a vector, so any element of C that has 0's in the first and second coordinates is still a periodic point of $\pi(C)$. We claim that these vectors form a subspace of dimension at least $p-3$. To see this, let $u_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0, i-1, p-i)$, where the 1 is in the ith coordinate, for $3 \leq i \leq p-1$. Note that $1 + (i-$

$1) + (p-i) = 0$ and $i - 2 + (i-1)(p-2) + (p-i)(p-1) = 0$, so $Hu_i^T = 0$ for all i, and thus each of these $p-3$ vectors is a periodic point of $\pi(C)$. They are linearly independent, which shows that the periodic points form a subspace of dimension at least $p-3$. Thus by Proposition 6 there is a full rank tiling of $\mathbb{F}_p^4$. □

Now we use Proposition 12 to obtain another sufficient condition for a finite abelian group to admit a full rank tiling.

THEOREM 13. *If G has $\mathbb{Z}_p^4$ with $p \geq 5$ as a direct factor, then G admits a full rank tiling.*

Proof. Proposition 12 proves that there exists a full rank tiling of $\mathbb{F}_p^4$. Since tilings depend only on the additive group structure, this is the same thing as saying that there is a full rank tiling of $\mathbb{Z}_p^4$. Combining this with Theorem 9 we get that any group containing $\mathbb{Z}_p^4$ as a direct factor has a full rank tiling. □

Rédei [14] conjectured that there do not exist full rank tilings of $\mathbb{Z}_p^3$ for any p. This conjecture is still open, but it has been verified for $p \leq 11$ (see [22]), so when p is 5, 7, or 11 we know exactly for which values of n there is a full rank tiling of $\mathbb{Z}_p^n$.

Unfortunately we could not get as strong a bound for the case when $p = 3$. The construction that we have been using does not work when $p = 3$, so we need to use something else. Phelps, Rifa, and Villanueva [11] have recently found full rank perfect codes of $\mathbb{F}_p^n$ when $n = (p^m - 1)/(p-1)$, where $m \geq 4$, with a kernel of dimension $(p-1)^{m-1}$. So when $p = 3$ this gives the existence of full rank tilings for $\mathbb{Z}_3^n$ for all $n \geq ((p^4-1)/(p-1)) - (p-1)^3 = 4p^2 - 2p + 2 = 32$. Thus there exists a full rank tiling of $\mathbb{Z}_3^n$ if $n \geq 32$. This is not nearly as good a bound as we have for either $p = 2$ or $p \geq 5$, so it can almost definitely be improved. The only lower bound in the literature says that there do not exist full rank tilings of $\mathbb{Z}_3^n$ when $n \leq 4$ [23], so it is not known whether $\mathbb{Z}_3^n$ admits a full rank tiling for $5 \leq n \leq 31$.

5. Open problems. Probably the most tractable open problem remaining is the one mentioned at the end of the last section, the existence of full rank tilings of $\mathbb{Z}_3^n$ for $5 \leq n \leq 31$. Since $p = 3$ allows more freedom in the construction than $p = 2$ but less than $p = 5$, we conjecture that there is some k with $4 < k \leq 10$ for which $\mathbb{Z}_3^n$ has a full rank tiling if and only if $n \geq k$. As with other cases of $\mathbb{Z}_p^n$, we suspect that coding theory approaches will prove valuable, in particular finding full rank perfect ternary codes.

A more difficult open question is what conditions on G are necessary for G to admit a full rank tiling. We know that neither of our two sufficient conditions is necessary on its own, and we suspect that it is not necessary for either of them to be satisfied for G to have a full rank tiling. We have shown that many groups admit full rank tilings, so our conditions are close to necessary, but there is no reason to think that we have characterized all groups admitting full rank tilings. An easier subproblem of this is Rédei's conjecture, mentioned previously, that $\mathbb{Z}_p^3$ does not admit a full rank tiling for any prime p. This conjecture is still wide open, with the only progress being a computer check for $p \leq 11$ by Szabó and Ward [22]. This conjecture immediately implies that our bound of $n \geq 4$ for the existence of full rank tilings of $\mathbb{Z}_p^n$ with $p \geq 5$ is tight and so if proved would give a complete characterization of which elementary p-groups ($p \geq 5$) admit full rank tilings.

There are many generalizations of this problem that could also prove to be interesting. Tilings can easily be defined for groups that are not finite or abelian, so removing those constraints gives many questions. We could also extend the work done for $\mathbb{F}_2^n$ in a different direction by considering not more general groups but more general transformations. We have pointed out that vector spaces are equivalent to groups

with respect to tilings, but that is not true if we allow linear or affine transformations other than translation. Define an affine factorization of $\mathbb{F}_q^n$ to be a pair (V, Φ) with V a subset of $\mathbb{F}_q^n$ and $\Phi = \{\phi_i\}$ a set of affine transformations satisfying $\mathbb{F}_q^n = \bigcup_i \phi_i(V)$ and $\phi_i(V) \cap \phi_j(V) = \emptyset$ for all $i \neq j$. Any tiling (V, A) of $\mathbb{F}_q^n$ automatically gives an affine factorization (V', Φ) by letting $V' = V$ and $\phi_i \in \Phi$ be translation by the ith element of A. However, tilings only give a small subset of affine factorizations. Allowing arbitrary affine transformations seems to make the problem very difficult, but perhaps adding some extra restrictions would make it tractable. In particular, requiring that $|\phi_i(V)| = |V|$ for all i might be helpful.

6. Conclusions. We have generalized the notions of tilings and full rank tilings from $\mathbb{F}_2^n$ to general finite abelian groups and have generalized many existing theorems to this new setting. We then combined and extended these results to prove that a group admits a full rank tiling if any of its direct factors do, allowing us to take any sufficient condition for a group to admit a full rank tiling and extend it by simply requiring a group to have a direct factor for which the condition holds. This method results in two such sufficient conditions: a group G admits a full rank tiling if it has a direct factor of the form $\mathbb{Z}_a \times \mathbb{Z}_b \times \mathbb{Z}_c$ with a, b, and c composite, or if it has a direct factor of the form $\mathbb{Z}_p^4$ with $p \geq 5$ prime. Since any finite abelian group can be decomposed into the direct product of finite abelian groups of prime power order, these are obviously quite strong conditions when the size of the group is large, showing that many groups admit a full rank tiling. We have also suggested some open problems in the area that we feel are tractable and could lead to some interesting results.

Acknowledgments. We would like to thank Joe Gallian for his encouragement and support, and Philip Matchett and Melanie Wood for many helpful discussions. We would also like to thank Reid Barton and Geir Helleloid for their insightful comments.

REFERENCES

[1] H. Bauer, B. Ganter, and F. Hergert, *Algebraic techniques for nonlinear codes*, Combinatorica, 3 (1983), pp. 21–33.
[2] G. Cohen, S. Litsyn, A. Vardy, and G. Zemor, *Tilings of binary spaces*, SIAM J. Discrete Math., 9 (1996), pp. 393–412.
[3] N. de Bruijn, *On the factorization of finite abelian groups*, Indag. Math. (N.S.), 15 (1953), pp. 258–264.
[4] T. Etzion and A. Vardy, *On perfect codes and tilings: Problems and solutions*, SIAM J. Discrete Math., 11 (1998), pp. 205–223.
[5] O. Fraser and B. Gordon, *Solution to a problem of A.D. Sands*, Glasg. Math. J., 20 (1977), pp. 115–117.
[6] G. Hajós, *Uber einfache und mehrfache bedeckung des n-dimensionalen raumes mit einem wurfelgitter*, Mathematische Zeitschrift, 47 (1941), pp. 427–467.
[7] G. Hajós, *Sur la factorisation des groupes abéliens*, Časopis Pěst Path. Rys., 74 (1949), pp. 157–162.
[8] F. MacWilliams and N. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam, 1977.
[9] P. R. J. Ostergard and A. Vardy, *Resolving the existence of full-rank tilings of binary hamming spaces*. SIAM J. Discrete Math., 18 (2004), pp. 382–387.
[10] K. Phelps, *Kernels of nonlinear Hamming codes*, Des. Codes Cryptogr., 6 (1995), pp. 247–257.
[11] K. Phelps, J. Rifa, and M. Villanueva, *Kernels and p-kernels of p^r-ary 1-perfect codes*. Des. Codes Cryptogr., 37 (2005), pp. 243–261.
[12] K. Phelps and M. Villanueva, *Ranks of q-ary 1-perfect codes*, Des. Codes Cryptogr., 27 (2002), pp. 139–144.
[13] L. Rédei, *Die neue theorie der endlichen abelschen gruppen und verallgemeinerung des hauptsatzes von Hajós*, Acta. Math. Acad. Sci. Hungar., 16 (1965), pp. 329–373.
[14] L. Rédei, *Lacunary Polynomials Over Finite Fields*, American Elsevier, New York, 1973.

[15] C. Rogers, *Packing and Covering*, Cambridge University Press, London, 1964.
[16] A. Sands, *On the factorization of finite abelian groups* II, Acta. Math. Acad. Sci. Hungar., 13 (1962), pp. 153–169.
[17] A. Sands, *On a conjecture of G. Hajós*, Glasg. Math. J., 15 (1974), pp. 88–89.
[18] S. Stein, *Tiling space by congruent polyhedra*, Bull. Amer. Math. Soc., 80 (1974), pp. 819–820.
[19] S. Szabó, *A type of factorization of finite abelian groups*, Discrete Math., 54 (1985), pp. 121–124.
[20] S. Szabó, *Groups with the Rédei property*, Matematiche, 52 (1997), pp. 357–364.
[21] S. Szabó and C. Ward, *Factoring abelian groups and tiling binary spaces*, Pure Math. Appl., 8 (1997), pp. 111–115.
[22] S. Szabó and C. Ward, *Factoring elementary groups of prime cube order into subsets*, Math. Comp., 67 (1998), pp. 1199–1206.
[23] S. Szabó and C. Ward, *Factoring groups having periodic maximal subgroups*, Bol. Soc. Mat. Mexicana (3), 5 (1999), pp. 327–333.
[24] A. Trachtenberg and A. Vardy, *Full-rank tilings of* $\mathbb{F}_2^8$ *do not exist*, SIAM J. Discrete Math., 16 (2003), pp. 390–392.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 171–188

DISCRETE POINT X-RAYS*

PAOLO DULIO†, RICHARD J. GARDNER‡, AND CARLA PERI§

Abstract. A discrete point X-ray of a finite subset F of $\mathbb{R}^n$ at a point p gives the number of points in F lying on each line passing through p. A systematic study of discrete point X-rays is initiated, with an emphasis on uniqueness results and subsets of the integer lattice.

Key words. convex lattice set, discrete tomography, geometric tomography, lattice, X-ray

AMS subject classifications. Primary: 05B50, 52C05, 52C07; Secondary: 52B20

DOI. 10.1137/040621375

1. Introduction. The (continuous) *parallel X-ray* of a convex body K in $\mathbb{R}^n$ in a direction $u \in S^{n-1}$ gives the lengths of all the intersections of K with lines parallel to u, and the (continuous) *point X-ray* of K at a point $p \in \mathbb{R}^n$ gives the lengths of all the intersections of K with lines passing through p. (See section 2 for all terminology.) In 1963, P. C. Hammer asked: How many parallel (or point) X-rays are needed to determine any convex body among all convex bodies? Answers to these questions are now known and are surveyed in [11, Chapters 1 and 5]. The topic of determining convex bodies and more general sets by their X-rays forms part of a larger area of inverse problems called geometric tomography, which concerns the retrieval of information about a geometric object via measurements of its sections by lines or planes or its projections on lines or planes. It is also clearly related to computerized tomography, where sets are replaced by density functions, and lengths of intersections with lines are replaced by line integrals.

Around 1994, Larry Shepp introduced the term "discrete tomography." Here the focus is on determining finite subsets of the integer lattice $\mathbb{Z}^n$ by means of their discrete parallel X-rays. A *discrete parallel X-ray* of a finite subset F of $\mathbb{Z}^n$ in the direction of a vector $v \in \mathbb{Z}^n$ gives the number of points in F lying on each line parallel to v. The points in F can model the atoms in a crystal, and indeed there is a genuine application of discrete tomography in high resolution transmission electron microscopy (HRTEM); see, for example, [15]. New techniques in HRTEM effectively allow the discrete parallel X-rays of a crystal to be measured, and the main goal of discrete tomography is to use these X-rays to determine the position of the atoms, with a view to applications in the material sciences.

By now there are many results available on continuous parallel or point X-rays of sets and on discrete parallel X-rays of finite subsets of the integer lattice. It is the purpose of this paper to initiate a study of the obvious remaining category of X-rays, namely, discrete point X-rays. The definition is the natural one: A *discrete point X-ray* of a finite subset F of $\mathbb{R}^n$ at a point $p \in \mathbb{R}^n$ gives the number of points

*Received by the editors December 23, 2004; accepted for publication (in revised form) September 21, 2005; published electronically March 3, 2006.
http://www.siam.org/journals/sidma/20-1/62137.html
†Dipartimento di Matematica "F. Brioschi," Politecnico di Milano, Piazza Leonardo da Vinci 32, I-20133 Milano, Italy (paolo.dulio@polimi.it).
‡Department of Mathematics, Western Washington University, Bellingham, WA 98225-9063 (Richard.Gardner@wwu.edu). The work of this author was supported in part by U.S. National Science Foundation grant DMS-0203527.
§Università Cattolica S.C., Largo Gemelli 1, I-20123 Milano, Italy (carla.peri@unicatt.it).

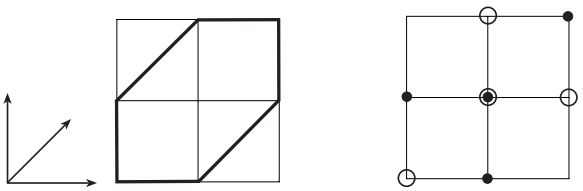


FIG. 1. *A U-hexagon (left) and two convex lattice sets with equal discrete parallel X-rays in the directions in U (right).*

in F lying on each line passing through p. Note that the above definition of discrete parallel X-ray also extends readily to finite subsets of $\mathbb{R}^n$, but as in that case, the main interest here is with discrete point X-rays of finite subsets of $\mathbb{Z}^n$ at points in $\mathbb{Z}^n$.

In order to describe our results, it is useful to briefly recall the corresponding results for discrete parallel X-rays. Firstly, given any finite set U of lattice directions in $\mathbb{Z}^2$, there are two different finite subsets of $\mathbb{Z}^2$ with equal discrete parallel X-rays in the directions in U (see [11, Lemma 2.3.2] or [13, Theorem 4.3.1]). In view of this, Gardner and Gritzmann [12] focused on convex lattice sets, employing the notion of a U-polygon in $\mathbb{R}^2$ for a given set U of directions. See section 2 for the formal definition; an example for a set of three lattice directions is shown at the left of Figure 1.

When a lattice U-polygon exists, it is easy to construct two different convex lattice sets with equal discrete parallel X-rays in the directions in U, as on the right of Figure 1 (one set indicated by black dots, and the other by circles). In [12] it was proved that in fact the nonexistence of a lattice U-polygon is necessary and sufficient for the discrete parallel X-rays in the directions in U to determine convex lattice sets (provided U has at least two nonparallel directions). It is easy to see that when $|U| = 3$, lattice U-polygons always exist. With tools from p-adic number theory, it was shown in [12] that they do not exist for certain sets of four lattice directions and *any* set of at least seven lattice directions, but can exist for certain sets of six lattice directions. Corresponding uniqueness results for discrete parallel X-rays follow immediately.

Our investigation begins in section 3, where we prove that for discrete point X-rays, there is also a general lack of uniqueness: Given any finite set P of points in $\mathbb{Z}^2$, there are two different finite subsets of $\mathbb{Z}^2$ with equal discrete point X-rays at the points in P. This is more involved than the corresponding result for discrete parallel X-rays, requiring the solution of a system of linear congruences. Our proof makes an unexpected use of the existence of arbitrarily long arithmetic progressions of relatively prime numbers. Thereafter we focus on convex lattice sets in $\mathbb{Z}^2$. Section 4 provides a rather complete analysis when discrete point X-rays are taken at two points; in fact, no open problems remain, in contrast to the continuous case (compare [11, Problems 5.1 and 5.2]). The discussion in section 4 also shows that it is hopeless to obtain uniqueness results unless the class of convex lattice sets is restricted to those not meeting any line through two of the points at which the X-rays are taken, a condition that we shall assume for the remainder of this introduction.

As with parallel discrete X-rays, uniqueness results hinge on the nonexistence of special lattice polygons we call lattice P-polygons, for finite subsets P of $\mathbb{Z}^2$. (However, the connection is less clear than in the parallel case.) See section 2 for the formal definition, and Figure 3 for an example. The construction of lattice P-polygons for sets of three collinear lattice points again requires the solution of a system of lin-

ear congruences; see section 5. It follows that for uniqueness when the points in P are collinear, P must contain at least four points. In section 6, the above results on discrete parallel X-rays, combined with the use of a new measure and projective transformations, lead to Corollary 6.6, which states that when the points in P are collinear, uniqueness is obtained for certain sets of four points and any set of at least seven points, while six points are generally not enough.

The final two sections concern noncollinear sets P. By appealing to classic theorems of projective geometry, we show in section 7 that, somewhat surprisingly, for any set P of less than five noncollinear points in $\mathbb{Z}^2$ there is a *rational* P-polygon. It follows that there are sets P of four noncollinear points in $\mathbb{Z}^2$ such that there is a lattice P-polygon, and we also show that there is a set of six noncollinear points in $\mathbb{Z}^2$ such that there is a lattice P-polygon. Corresponding nonuniqueness results are deduced in section 8, including Theorem 8.1, a result quite different from the analogous one concerning continuous point X-rays.

While there is at present no known application of discrete point X-rays in HRTEM, we feel that this study is warranted by their natural role in the general theory of X-rays and by the increasing attention to convex lattice sets (see, for example, [3], [7], [8], and [9]). Moreover, the central role of P-polygons highlights these intriguing structures, which should be of independent interest in incidence geometry. They may also have consequences for number theory, as the analogous U-polygons do. (See, for example, [2], where U-polygons are used to make progress on the Prouhet-Tarry-Escott problem concerning multigrades. The connection between multigrades and discrete tomography was first noticed by Ron Graham.) Much remains to be done. For example, it is open at present whether convex lattice sets in $\mathbb{Z}^2$ are determined by their discrete point X-rays at some set of three noncollinear lattice points or at any set of seven lattice points (provided they do not meet any line joining two of the points).

An extended abstract of this paper appeared in [10].

2. Definitions and preliminaries. As usual, S^{n-1} denotes the unit sphere and o the origin in Euclidean n-space $\mathbb{R}^n$. If $u \in \mathbb{R}^n$, we denote by $u^\perp$ the $(n-1)$-dimensional subspace orthogonal to u. The standard orthonormal basis for $\mathbb{R}^n$ will be $\{e_1, \ldots, e_n\}$. The line segment with endpoints x and y is denoted by $[x, y]$, and we write $L[x, y]$ for the line through x and y.

If A is a set, we denote by $|A|$, ∂A, and $\operatorname{conv} A$ the *cardinality*, *boundary*, and *convex hull* of A, respectively. The notation for the usual orthogonal *projection* of A on a subspace S is $A|S$. The *symmetric difference* of two sets A_1 and A_2 is $A_1 \triangle A_2 = (A_1 \setminus A_2) \cup (A_2 \setminus A_1)$.

We denote n-dimensional projective space by $\mathbb{P}^n$, and regard it as $\mathbb{R}^n \cup H_\infty$, where H_∞ is the hyperplane at infinity. Points in H_∞ can be associated with a pair $\{u, -u\}$ of directions in S^{n-1}. If ϕ is a projective transformation from $\mathbb{P}^n$ onto $\mathbb{P}^m$ mapping a point p in H_∞ to a finite point ϕp in $\mathbb{P}^m$ (i.e., a point in $\mathbb{R}^m$), then lines in $\mathbb{R}^n$ parallel to a direction u associated with p map to lines passing through ϕp. If $E \subset \mathbb{P}^n$ is such that $\phi E \subset \mathbb{R}^m$ (i.e., ϕE does not meet the hyperplane at infinity in $\mathbb{P}^m$), then ϕ is called *permissible* for E. Note that ϕ preserves the convexity of sets in $\mathbb{R}^n$ for which it is permissible. See [11, pp. 2 and 7] for more details.

The *cross ratio* $\langle p_1, p_2, p_3, p_4 \rangle$ of four points p_i, $i = 1, \ldots, 4$ in a line L is given by

$$\langle p_1, p_2, p_3, p_4 \rangle = \frac{(x_3 - x_1)(x_4 - x_2)}{(x_4 - x_1)(x_3 - x_2)}, \tag{1}$$

where x_i is the coordinate of p_i, $i = 1, \dots, 4$ in some fixed Cartesian coordinate system in L. See, for example, [4, section 6.2].

A *convex polytope* is the convex hull of a finite subset of $\mathbb{R}^n$. We sometimes refer to a finite subset of the n-dimensional integer lattice $\mathbb{Z}^n$ as a *lattice set*. A *convex lattice set* is a finite subset F of $\mathbb{Z}^n$ such that $F = (\operatorname{conv} F) \cap \mathbb{Z}^n$. A *lattice polygon* is a convex polygon with its vertices in $\mathbb{Z}^2$. A polygon is called *rational* if its vertices have rational coordinates. A *lattice line* is a line containing at least two points in $\mathbb{Z}^2$.

Call a vector $u \in \mathbb{Z}^n$ *primitive* if the line segment $[o, u]$ contains no lattice points other than o and u.

Let F be a finite subset of $\mathbb{R}^n$ and let $u \in \mathbb{R}^n \setminus \{o\}$. The *discrete parallel X-ray of F parallel to u* is the function $X_u F$ defined by

$$X_u F(v) = |F \cap (L[o, u] + v)|,$$

for each $v \in u^\perp$. The function $X_u F$ is in effect the projection, counted with multiplicity, of F on $u^\perp$. For an introduction to the many known results on discrete parallel X-rays and their applications, see [5], [12], [13], and [15].

Let F be a finite subset of $\mathbb{R}^n$ and let $p \in \mathbb{R}^n$. The *discrete point X-ray of F at p* is the function $X_p F$ defined by

$$X_p F(u) = |F \cap (L[o, u] + p)|,$$

for each $u \in \mathbb{R}^n \setminus \{o\}$.

Let U be a finite set of vectors in $\mathbb{R}^2$. We call a nondegenerate convex polygon Q a *U-polygon* if it has the following property: If v is a vertex of Q, and $u \in U$, then the line $v + L[o, u]$ meets a different vertex v' of Q.

Let P be a finite set of points in $\mathbb{R}^2$. A nondegenerate convex polygon Q is a *P-polygon* if it satisfies the following property: If v is a vertex of Q, and $p \in P$, then the line $L[p, v]$ meets a different vertex v' of Q.

Note that in view of these definitions, a lattice P-polygon is a convex subset of $\mathbb{R}^2$, while a convex lattice polygon is a finite subset of $\mathbb{Z}^2$.

There is a convenient common generalization of the previous two definitions. Consider $\mathbb{P}^2 = \mathbb{R}^2 \cup H_\infty$ and let P be a finite set of points in $\mathbb{P}^2$. A nondegenerate convex polygon Q in $\mathbb{R}^2$ is a *P-polygon* if it satisfies the following property: If v is a vertex of Q, and $p \in P$, then the line $L[p, v]$ in $\mathbb{P}^2$ meets a different vertex v' of Q. Note that if $P \subset H_\infty$, then the P-polygon Q is also a U-polygon for the set U of unit vectors associated with points in P.

Let $\mathcal{F}$ be a class of finite sets in $\mathbb{R}^n$ and P a finite set of points in $\mathbb{R}^n$. We say that $F \in \mathcal{F}$ is *determined* by the discrete point X-rays at the points in P if whenever $F' \in \mathcal{F}$ and $X_p F = X_p F'$ for all $p \in P$, we have $F = F'$.

The *greatest common divisor* of integers m and n is denoted by $\gcd(m, n)$. We need the following strong form of the Chinese Remainder Theorem (see, for example, [1, pp. 46 and 56]).

PROPOSITION 2.1. *Let $a_i \in \mathbb{Z}$ and $n_i \in \mathbb{N}$, $i = 1, \dots, k$. The system*

$$x \equiv a_i \pmod{n_i}, \ i = 1, \dots, k$$

has a solution $x \in \mathbb{Z}$ if and only if

$$\gcd(n_i, n_j) | (a_i - a_j) \tag{2}$$

for all $1 \le i \ne j \le k$. Moreover, if (2) holds, there are infinitely many solutions, each pair of them congruent modulo $n_1 n_2 \cdots n_k$.

3. The general case. The purpose of this section is to prove the following result.

THEOREM 3.1. *For each finite subset P of $\mathbb{Z}^2$, there are two different finite subsets of $\mathbb{Z}^2$ with the same discrete point X-rays at the points in P.*

Proof. Suppose that $p_j = (p_{1j}, p_{2j})$, $j = 1 \ldots, m$ are distinct points in $\mathbb{Z}^2$. Let C_1 and C_2 be the two disjoint sets of 2^{m-1} alternate vertices of the unit cube $[0,1]^m$ in $\mathbb{R}^m$. Then C_1 and C_2 have the same discrete parallel X-rays in the m coordinate directions in $\mathbb{R}^m$. We aim to define a suitable projective transformation $\phi : \mathbb{P}^m \to \mathbb{P}^2$ that maps the jth coordinate direction in $\mathbb{R}^m$ to p_j in such a way that $\phi(C_1)$ and $\phi(C_2)$ are disjoint subsets of $\mathbb{Z}^2$ with equal discrete point X-rays at each p_j.

To this end, let $a, b \in \mathbb{N}$ and $c_1, c_2 \in \mathbb{Z}$ (all to be chosen later) and define $\phi : \mathbb{P}^m \to \mathbb{P}^2$, using homogeneous coordinates in both $\mathbb{P}^m$ and $\mathbb{P}^2$, by

$$\begin{aligned}&\phi(x_1, \ldots, x_m, x_{m+1})\\ &\quad = \left(\sum_{i=1}^{m} 2^{i-1} b\, p_{1i} x_i + c_1 x_{m+1}, \sum_{i=1}^{m} 2^{i-1} b\, p_{2i} x_i + c_2 x_{m+1}, \sum_{i=1}^{m} 2^{i-1} b\, x_i + a x_{m+1} \right).\end{aligned}$$

Let e_j be the jth vector in the standard orthonormal basis for $\mathbb{R}^{m+1}$. Then

$$\phi(e_j) = \left(2^{j-1} b\, p_{1j}, 2^{j-1} b\, p_{2j}, 2^{j-1} b \right),$$

for $j = 1, \ldots, m$, and this shows that ϕ maps the jth coordinate direction in $\mathbb{R}^m$ to p_j, $j = 1, \ldots, m$.

As a map from $\mathbb{R}^m$ to $\mathbb{R}^2$, ϕ is given by

$$\phi(x_1, \ldots, x_m) = \left(\frac{\sum_{i=1}^{m} 2^{i-1} b\, p_{1i} x_i + c_1}{\sum_{i=1}^{m} 2^{i-1} b\, x_i + a}, \frac{\sum_{i=1}^{m} 2^{i-1} b\, p_{2i} x_i + c_2}{\sum_{i=1}^{m} 2^{i-1} b\, x_i + a} \right). \tag{3}$$

Denote an arbitrary vertex of the unit cube $[0,1]^m$ in $\mathbb{R}^m$ by v_I, $I \subset \{1, \ldots, m\}$, where the ith component of v_I is 1 if $i \in I$ and 0 otherwise. In view of (3), we have $\phi(v_I) \in \mathbb{Z}^2$ if and only if

$$c_k \equiv -\sum_{i \in I} 2^{i-1} b\, p_{ki} \pmod{\sum_{i \in I} 2^{i-1} b + a}, \tag{4}$$

for $k = 1, 2$. It is easy to check that

$$\left\{ \sum_{i \in I} 2^{i-1} : I \subset \{1, \ldots, m\} \right\} = \{0, 1, \ldots, 2^m - 1\}.$$

Therefore the set of possible moduli in the congruences (4) is precisely the arithmetic progression

$$\{a, a + b, \ldots, a + (2^m - 1)\, b\}.$$

Sierpiński [16] noted that if $a = 1$ and $b = (2^m - 1)!$, each pair of this arithmetic progression is relatively prime. It follows from the Chinese Remainder Theorem (see Proposition 2.1) that for each $k = 1, 2$, the system (4), where $I \subset \{1, \ldots, m\}$, has a solution $c_k \in \mathbb{Z}$. Then ϕ maps each vertex of $[0,1]^m$ to a point in $\mathbb{Z}^2$.

The basic properties of projective transformations guarantee that if C_1 and C_2 are the two disjoint sets of 2^{m-1} alternate vertices of $[0,1]^m$, then $\phi(C_1)$ and $\phi(C_2)$

have equal discrete point X-rays at each p_i. We will also have $\phi(C_1) \neq \phi(C_2)$ if ϕ is injective on the set of vertices of $[0,1]^m$. If this is not the case, then there are different subsets I and J of $\{1,\dots,m\}$ such that $\phi(v_I) = \phi(v_J)$. By (3), this implies that

$$X_I(Y_J + c_1) = X_J(Y_I + c_1), \tag{5}$$

where

$$X_I = \sum_{i\in I} 2^{i-1}b + a \quad \text{and} \quad Y_I = \sum_{i\in I} 2^{i-1}bp_{1i},$$

and X_J and Y_J are obtained by replacing I with J. Since $I \neq J$, $X_I \neq X_J$ and from (5) we obtain

$$c_1 \leq |X_IY_J - X_JY_I| \leq 2\,(a + (2^m - 1)\,b)\,(2^m - 1)\,b \max_{1\leq i\leq m} |p_{1i}|.$$

By Proposition 2.1 we can choose a solution c_1 to the system (4) so large that this inequality is false, and the injectivity of ϕ on the set of vertices of $[0,1]^m$ follows. □

Note that since Sierpiński's result is constructive, the previous proof is also. An alternative approach is to apply instead the remarkable recent result of Green and Tao [14], who establish the existence of arbitrarily long arithmetic progressions of primes; however, this proof is not constructive.

4. Discrete point *X*-rays at two points. We begin this section with the following simple observation.

THEOREM 4.1. *Let p_1 and p_2 be distinct points in $\mathbb{Z}^2$. Then there are two different convex lattice sets that meet $L[p_1,p_2]$ and have equal discrete point X-rays at p_1 and p_2.*

Proof. Without loss of generality, let $p_1 = (0,0)$ and $p_2 = (k,0)$ for some $k > 0$. Suppose that $m \in \mathbb{N}$. Then the sets $K_1 = \{(k+i,0) : i = 1,\dots,m\}$ and $K_2 = \{(k+i,0) : i = 2,\dots,m+1\}$ have equal discrete point X-rays at p_1 and p_2. By adjoining the point $(k+m,1)$ to both sets we can obtain two-dimensional examples with the same property. □

Note that the sets K_1 and K_2 in the previous theorem also have the same discrete point X-rays at any lattice point on the x-axis.

THEOREM 4.2. *Let K_i, $i = 1,2$ be convex lattice sets in $\mathbb{Z}^2$ with equal discrete point X-rays at distinct points $p_1, p_2 \in \mathbb{Z}^2$. Suppose that*

(i) *$L[p_1,p_2] \cap K_i = \emptyset$, $i = 1,2$, and*

(ii) *$\operatorname{conv} K_1$ and $\operatorname{conv} K_2$ either both meet $[p_1,p_2]$ or both meet $L[p_1,p_2] \setminus [p_1,p_2]$.*

Then $K_1 = K_2$.

Proof. By (i) and the fact that K_i, $i = 1,2$ are convex lattice sets, we have $p_i \notin \operatorname{conv} K_1 \cup \operatorname{conv} K_2$, $i = 1,2$. Suppose that $\operatorname{conv} K_1$ and $\operatorname{conv} K_2$ both meet $L[p_1,p_2] \setminus [p_1,p_2]$. If p_1 and p_2 lie between $\operatorname{conv} K_1$ and $\operatorname{conv} K_2$, these sets cannot have equal supporting lines from p_1 and p_2, contradicting the equality of the discrete point X-rays of K_1 and K_2 at p_1 and p_2. Then we may assume that p_1, p_2, and $L[p_1,p_2] \cap \operatorname{conv} K_i$, $i = 1,2$ are in that order on $L[p_1,p_2]$. Suppose that $K_1 \neq K_2$. Without loss of generality, we may assume that $L[p_1,p_2]$ is the x-axis. Then by (i), we can assume that $(K_1 \triangle K_2) \cap \{y > 0\} \neq \emptyset$. Let L_1 be the line through p_2 and containing a point of $K_1 \triangle K_2$, with minimal positive angle with the x-axis. Since K_1 and K_2 have equal discrete point X-rays at p_2, there are points $v_1 \in K_1 \setminus K_2$ and $v_2 \in K_2 \setminus K_1$ on L_1, and we can assume that p_2, v_1, and v_2 are in that order on L_1. Since K_1 and K_2 have equal discrete point X-rays at p_1, the line L_2 through p_1 and

v_1 must meet $K_2 \backslash K_1$ in a point v_3. If p_1, v_1, and v_3 are in that order on L_2, then the line through p_2 and v_3 has a smaller positive angle with the x-axis than L_1. Therefore $v_3 \in [p_1, v_1]$. Assumptions (i) and (ii) imply that there is a point $c \in K_2 \cap \{y < 0\}$, but then $v_1 \notin K_2$ lies in the interior of the triangle with vertices v_2, v_3, and c, all of which lie in K_2. This contradicts the fact that K_2 is a convex lattice set, and proves that $K_1 = K_2$.

The case when $\operatorname{conv} K_1$ and $\operatorname{conv} K_2$ both meet $[p_1, p_2]$ is proved in similar fashion. □

The next result shows that the assumption (ii) in Theorem 4.2 is necessary.

THEOREM 4.3. *Let p_1 and p_2 be distinct points in $\mathbb{Z}^2$. Then there are different convex lattice sets K_1 and K_2 such that $L[p_1, p_2] \cap K_i = \emptyset$ and $L[p_1, p_2] \cap \operatorname{conv} K_i \neq \emptyset$, $i = 1, 2$, and with equal discrete point X-rays at p_1 and p_2.*

Proof. Let $p_1 = (0, 0)$, and let $p_2 = ku$, where $u \in \mathbb{Z}^2$ is primitive and $k \in \mathbb{N}$. Then there is a $v \in \mathbb{Z}^2$ such that $\{u, v\}$ is a basis in $\mathbb{R}^2$. The unimodular affine transformation mapping $\{u, v\}$ to $\{e_1, e_2\}$ is a bijection of $\mathbb{Z}^2$ onto itself preserving convexity and incidence. Therefore we may, without loss of generality, take $p_1 = (0, 0)$ and $p_2 = (k, 0)$ for some $k > 0$.

Suppose that $k = 1$. Then the sets $K_1 = \{(2, 3), (-1, -2)\}$ and $K_2 = \{(3, 6), (-2, -3)\}$ fulfill the requirements of the theorem.

Now suppose that $k > 1$. Let $a = (k, k)$, $b = (k, k+1)$, $c = (-k(k-1), 1-k^2)$, and $d = (-k(k^2-1), -k(k^2-1))$. Let $K_1 = \{a, c\}$ and $K_2 = \{b, d\}$. It is easy to check that $L[p_1, p_2] \cap \operatorname{conv} K_i \neq \emptyset$, $i = 1, 2$ and that the sets K_1 and K_2 have equal discrete point X-rays at p_1 and p_2. It remains to show that K_1 and K_2 are convex lattice sets. To this end, note that the line $L[a, c]$ has slope $(k^2+k-1)/k^2$. Moreover, k^2+k-1 and k^2 are relatively prime; otherwise, if $p > 1$ is prime, $p|(k^2+k-1)$, and $p|k^2$, then $p|(k-1)$, so p does not divide k, contradicting $p|k^2$. It follows that $K_1 = (\operatorname{conv} K_1) \cap \mathbb{Z}^2$, as required. The line $L[b, d]$ has slope $(k^3+1)/k^3$, and since k^3 and k^3+1 are consecutive integers, they are relatively prime. Consequently, $K_2 = (\operatorname{conv} K_2) \cap \mathbb{Z}^2$, and the proof is complete. □

The next two lemmas are rather general and will be useful also in subsequent sections of the paper.

LEMMA 4.4. *If Q is a P-polygon such that $|P| \geq 2$ and $P \cap Q = \emptyset$, then Q does not meet any line through two points in P.*

Proof. Let p_1 and p_2 be different points in P, and without loss of generality, suppose that they lie on the x-axis and that Q is a P-polygon whose interior meets the upper open half plane. Suppose that $Q \cap [p_1, p_2] \neq \emptyset$. Let L_1 be the lattice line through p_1 with minimal positive angle with the x-axis such that L_1 contains vertices v_1 and v_2 of Q. Without loss of generality suppose that p_1, v_2, and v_1 lie on L_1 in that order. Since Q meets $[p_1, p_2]$, by convexity the line L_2 through p_2 and v_2 contains a vertex v_3 of Q with p_2, v_3, and v_2 in that order on L_2. But then the line L_3 through p_1 and v_3 has a smaller positive angle with the x-axis than L_1, a contradiction. A similar argument applies to the case when Q meets the x-axis outside the segment $[p_1, p_2]$. □

LEMMA 4.5. *Let P be a set of points in $\mathbb{Z}^2$. If there is a lattice P-polygon Q, then there are different convex lattice sets K_1 and K_2 with equal discrete point X-rays at the points in P. Moreover, if $P \cap Q = \emptyset$, then in addition* $\operatorname{conv} K_1$ *and* $\operatorname{conv} K_2$ *do not meet any line through two points of P.*

Proof. Let Q be a lattice P-polygon. Partition the vertices of Q into two disjoint sets V_1 and V_2, where the members of each set are alternate vertices in a clockwise

ordering around ∂Q. Let

$$C = (\mathbb{Z}^2 \cap Q) \setminus (V_1 \cup V_2),$$

and let $K_i = C \cup V_i$, $i = 1, 2$. Then K_1 and K_2 are different convex lattice sets with equal discrete point X-rays at the points in P.

If $P \cap Q = \emptyset$, then by Lemma 4.4, Q does not meet any line through two points of P and the second statement follows immediately. □

THEOREM 4.6. *Let p_1 and p_2 be distinct points in $\mathbb{Z}^2$ and let $P = \{p_1, p_2\}$. Then there is a lattice P-polygon Q with $P \cap Q = \emptyset$, and hence two different convex lattice sets, with convex hulls disjoint from $L[p_1, p_2]$ and with equal discrete point X-rays at the points in P.*

Proof. Without loss of generality, let $p_1 = (0, 0)$ and $p_2 = (k, 0)$ for some $k > 0$. Then one can check that $(2k, 2k), (3k, 3k), (3k, 4k)$, and $(9k, 12k)$ are the vertices of a lattice P-quadrilateral. The conclusion follows from Lemma 4.5. □

5. Lattice P-hexagons for collinear sets P. This section is devoted to the proof of the following result.

THEOREM 5.1. *If P is a set of three collinear points in $\mathbb{Z}^2$, there exists a lattice P-hexagon.*

Proof. As in the proof of Theorem 4.3, we can assume, without loss of generality, that the points in P lie on the x-axis. More precisely, we may take $P = \{(-a, 0), (0, 0), (b, 0)\}$, $a, b \in \mathbb{N}$, where $\gcd(a, b) = 1$ and b is odd, since the general case then follows by applying a suitable dilatation and/or reflection in the y-axis, if necessary. We suppose henceforth that a and b are fixed positive integers satisfying these conditions.

Let $r, s, t \in \mathbb{Z}$ and consider the projective transformation ϕ of $\mathbb{P}^2$ given in homogeneous coordinates (x, y, z) by

$$\phi(x, y, z) = (abx - aby + rz, sz, ax + by).$$

Then $\phi(1, 1, 0) = (0, 0, a + b)$, $\phi(1, 0, 0) = (ab, 0, a)$, and $\phi(0, 1, 0) = (-ab, 0, b)$. It follows that ϕ is a projective transformation that takes lines parallel to $u_1 = (1, 1)$ (or parallel to $u_2 = (1, 0)$ or parallel to $u_3 = (0, 1)$) to lines through $(0, 0)$ (or through $(b, 0)$ or through $(-a, 0)$, respectively). As a map from $\mathbb{R}^2$ into itself, ϕ can be written as

$$\phi(x, y) = \left(\frac{abx - aby + r}{ax + by}, \frac{s}{ax + by} \right). \tag{6}$$

Let $U = \{u_1, u_2, u_3\}$. If $k, l \in \mathbb{Z}$ are such that $ak + bl > 0$ and if $m, n \in \mathbb{N}$, then the points

$$(k, l), (k, l + m), (k + n, l), (k + n, l + m + n), (k + m + n, l + m), (k + m + n, l + m + n) \tag{7}$$

are the vertices of a U-hexagon Q such that ϕ is permissible for Q (i.e., Q does not meet the line $ax + by = 0$). It follows that ϕQ is a P-hexagon, and remains to show that $k, l, r, s \in \mathbb{Z}$ with $ak + bl > 0$ and $m, n \in \mathbb{N}$ can be chosen so that the vertices of ϕQ have integer coordinates. Since obviously s can be chosen so that the y-coordinates of these vertices are integers, we have only to consider the x-coordinates.

Let

$$c = ab(k-l) + r \quad \text{and} \quad d = ak + bl. \tag{8}$$

Then, by (6), (7), and (8), the x-coordinates of the vertices of ϕQ are

$$\frac{c}{d},\ \frac{c-abm}{d+bm},\ \frac{c+abn}{d+an},\ \frac{c-abm}{d+(a+b)n+bm},\ \frac{c+abn}{d+(a+b)m+an},\ \text{and}\ \frac{c}{d+(a+b)(m+n)}.$$

Therefore we seek $d, m, n \in \mathbb{N}$ such that there is a solution $c \in \mathbb{Z}$ to the following system of congruences (which we have rearranged for our convenience):

$$c \equiv 0 \pmod{d} \tag{9}$$
$$c \equiv 0 \pmod{d+(a+b)(m+n)} \tag{10}$$
$$c \equiv abm \pmod{d+bm} \tag{11}$$
$$c \equiv abm \pmod{d+(a+b)n+bm} \tag{12}$$
$$c \equiv -abn \pmod{d+an} \tag{13}$$
$$c \equiv -abn \pmod{d+(a+b)m+an}. \tag{14}$$

By the Chinese Remainder Theorem (see Proposition 2.1), we have at first sight to consider the division criterion (2) for 15 pairs of the congruences (9)–(14). However, the following pairs can be eliminated: (9) and (10) (obviously), (9) and (11) (since if $j|d$ and $j|d+bm$, then $j|bm$, so $\gcd(d, d+bm)|abm$), (9) and (13) (by a similar argument), (10) and (12) (since if $j|d+(a+b)(m+n)$ and $j|d+(a+b)n+bm$, then $j|am$, so $\gcd(y+(a+b)(m+n), y+(a+b)n+bm)|abm$), (10) and (14) (by a similar argument), (11) and (12) (obviously), and (13) and (14) (obviously). So only the following eight pairs must in general be considered: (9) and (12), (9) and (14), (10) and (11), (10) and (13), (11) and (13), (11) and (14), (12) and (13), and (12) and (14).

We claim that if $d = ab$, $m = a+b$, and $n = a$, there is a solution $c \in \mathbb{Z}$ to the congruences (9)–(14). To see this, note first that since $d = ab$, we can obviously eliminate the pairs of congruences (9) and (12), and (9) and (14). Consider the division criterion (2) for the remaining six pairs of congruences, that is, (10) and (11), (10) and (13), (11) and (13), (11) and (14), (12) and (13), and (12) and (14) in order:

$$\gcd(2a^2+4ab+b^2, 2ab+b^2) \mid ab(a+b) \tag{15}$$
$$\gcd(2a^2+4ab+b^2, a^2+ab) \mid a^2b \tag{16}$$
$$\gcd(2ab+b^2, a^2+ab) \mid ab(2a+b) \tag{17}$$
$$\gcd(2ab+b^2, 2a^2+3ab+b^2) \mid ab(2a+b) \tag{18}$$
$$\gcd(a^2+3ab+b^2, a^2+ab) \mid ab(2a+b) \tag{19}$$
$$\gcd(a^2+3ab+b^2, 2a^2+3ab+b^2) \mid ab(2a+b). \tag{20}$$

Observe that (15) holds since if $j|2a^2+4ab+b^2$ then since b is odd, j is also odd; if also $j|2ab+b^2$, then $j|2a^2+2ab = 2a(a+b)$. Now j odd and $j|2a(a+b)$ imply that $j|a(a+b)$ and hence $j|ab(a+b)$.

For (16), suppose that $j|2a^2+4ab+b^2$ and $j|a^2+ab = a(a+b)$. Then we can write $j = pq$, where $p|a$ and $q|a+b$, and it suffices to show that $q|ab$. Now $q|a+b$ implies $q|a^2+2ab+b^2$, which together with $q|2a^2+4ab+b^2$ gives $q|a^2+2ab$. Since also $q|a^2+ab$, we get $q|ab$ as required.

Conditions (17) and (18) hold since $j|2ab+b^2 = b(2a+b)$ implies that $j|ab(2a+b)$, and (19) holds since if $j|a^2+3ab+b^2$ and $j|a^2+ab$, then $j|2ab+b^2 = b(2a+b)$ and hence $j|ab(2a+b)$.

For (20), suppose that $j|a^2+3ab+b^2$ and $j|2a^2+3ab+b^2 = (a+b)(2a+b)$. Then we can write $j = pq$, where $p|a+b$ and $q|2a+b$, and it suffices to show that $p|ab$. Now $p|a+b$ implies $p|a^2+2ab+b^2$, which together with $p|a^2+3ab+b^2$ gives $p|ab$, as required. This proves the claim.

We still have to prove that for $d = ab$, $m = a+b$, $n = a$, and a corresponding solution $c \in \mathbb{Z}$ to the congruences (9)–(14), there are $k, l, r \in \mathbb{Z}$ with $ak+bl > 0$ so that (8) holds. To see this, use the condition $\gcd(a,b) = 1$ to choose $k', l' \in \mathbb{Z}$ such that $ak'+bl' = 1$ and then let $k = dk'$ and $l = dl'$. Then the second equation in (8) is satisfied and $ak+bl = d > 0$. After this, we can find $r \in \mathbb{Z}$ so that the first equation in (8) is satisfied for this k and l. This completes the proof. □

As an example in which the computations can be done by hand, suppose that $P = \{(-1,0),(0,0),(1,0)\}$. Then $a = b = 1$, so we have $d = 1$, $m = 2$, and $n = 1$. It is easy to see $c = 77$ is a solution of the congruences (9)–(14) and we can take $k = 1$, $l = 0$, and $r = 76$ in order that (8) holds. Moreover, $s = 210$ is a suitable choice. This leads to a P-hexagon with vertices (in counterclockwise order around the hexagon) (11, 30), (13, 35), (39, 105), (77, 210), (25, 70), and (15, 42).

6. Discrete point *X*-rays at collinear points. As we have seen, a convex lattice set is determined by its discrete point X-rays at two different points only in the situation of Theorem 4.2. Thus to have more general uniqueness results we need more than two points. Moreover, the following result is an immediate consequence of Theorem 5.1 and Lemma 4.5.

THEOREM 6.1. *If P is a set of three collinear points in $\mathbb{Z}^2$, then convex lattice sets not meeting the line containing P are not determined by discrete point X-rays at the points in P.*

To make progress, we require the following technical lemmas.

LEMMA 6.2. *Let $p \in \mathbb{Z}^2$ and let F_1 and F_2 be finite subsets of $\mathbb{Z}^2$ such that $p \notin F_1 \cup F_2$ and $X_pF_1 = X_pF_2$. Then $|F_1| = |F_2|$.*

Proof. Since $p \notin F_1 \cup F_2$, we have for $i = 1, 2$,

$$|F_i| = \sum_{u \in S^1} |F_i \cap (L[o,u]+p)| = \sum_{u \in S^1} X_pF_i(u). \qquad \square$$

Let L be a lattice line in $\mathbb{R}^2$, and suppose that L is taken as the x-axis in a Cartesian coordinate system. For each finite set F in $\mathbb{Z}^2$, define

$$\nu(F) = \sum_{(x,y) \in F} \frac{1}{|y|}. \tag{21}$$

Then ν is a measure in $\mathbb{Z}^2$, and we call L the *baseline* of ν.

LEMMA 6.3. *Let ν be a measure defined by* (21) *with respect to the baseline L. Suppose that F_1 and F_2 are finite subsets of $\mathbb{Z}^2$ contained in one of the open half planes bounded by L and with equal discrete point X-rays at $p \in L \cap \mathbb{Z}^2$. Then the centroids of F_1 and F_2 with respect to ν lie on the same line through p.*

Proof. Without loss of generality we may take L to be the x-axis, $p = (0,0)$, and F_1 and F_2 finite subsets of $\mathbb{Z}^2$ contained in the upper open half plane. Let $c_i = (x_i, y_i)$

be the centroid of F_i, for $i = 1, 2$, with respect to the measure ν. Then

$$x_i = \frac{1}{\nu(F_i)} \sum_{(x,y)\in F_i} \frac{x}{y}$$

and

$$y_i = \frac{|F_i|}{\nu(F_i)}$$

for $i = 1, 2$. Therefore

$$\frac{y_i}{x_i} = \frac{|F_i|}{\sum\limits_{(x,y)\in F_i} (x/y)} = \frac{|F_i|}{\sum\limits_{u\in S^1} (X_p F_i(u)) \cot\theta(u)},$$

for $i = 1, 2$, where $\theta(u)$ denotes the angle between the x-axis and a line parallel to u. Since $X_p F_1 = X_p F_2$ and $p \notin F_1 \cup F_2$, we have $|F_1| = |F_2|$ by Lemma 6.2, and hence $y_1/x_1 = y_2/x_2$, as required. □

THEOREM 6.4. *Let P be a set of at least three points in $\mathbb{Z}^2$ lying in a line L. If there are different convex lattice sets not meeting L with equal discrete point X-rays at the points in P, then there is a rational P-polygon disjoint from L.*

Proof. Let K_1 and K_2 be different convex lattice sets not meeting L and with equal discrete point X-rays at the points in P. If $L \cap \operatorname{conv} K_1 \neq \emptyset$, then clearly $L \cap \operatorname{conv} K_2 \neq \emptyset$. Then either for some $1 \leq i \neq j \leq 3$, $\operatorname{conv} K_1$ and $\operatorname{conv} K_2$ both meet $[p_i, p_j]$, or for some $1 \leq i \neq j \leq 3$, $\operatorname{conv} K_1$ and $\operatorname{conv} K_2$ both meet $L[p_i, p_j] \setminus [p_i, p_j]$, contradicting Theorem 4.2.

Consequently $L \cap \operatorname{conv} K_1 = \emptyset$ and therefore $L \cap \operatorname{conv} K_2 = \emptyset$. Then we can follow exactly the proof of [12, Theorem 5.5] for discrete parallel X-rays, on replacing lattice lines parallel to directions in a set with lattice lines through points in P, replacing ordinary centroids with centroids with respect to the measure ν defined by (21) with baseline L, and using Lemma 6.3 instead of [12, Lemma 5.4]. Note that this argument uses only cardinality and collinearity properties and the fact that the centroid of a finite set of lattice points is a point with rational coordinates, a fact that still holds when centroids are taken with respect to ν. Also, note that the observation that $|U| \geq 4$ in the second paragraph of the proof of [12, Theorem 5.5] is not needed. The conclusion is that there is a rational P-polygon disjoint from L. □

THEOREM 6.5.

(i) *Let U be a set of mutually nonparallel vectors in $\mathbb{Z}^2$ such that there exists a lattice U-polygon and let L be a lattice line. Then for some set P of $|U|$ points in L, there exists a lattice P-polygon disjoint from L.*

(ii) *Let P be a set of at least two points in $\mathbb{Z}^2$ in a line L such that there exists a rational P-polygon disjoint from L. Let ϕ be a projective transformation of $\mathbb{P}^2$ taking L to the line at infinity, and let $U = \phi P$. Then there exists a lattice U-polygon.*

Proof.

(i) Let Q be a lattice U-polygon and suppose that L is a lattice line. Let ϕ be a nonsingular projective transformation of $\mathbb{P}^2$ such that $\phi H_\infty = L$, where H_∞ is the line at infinity in $\mathbb{P}^2$, so that $L \cap \phi Q = \emptyset$. If $p \in \mathbb{P}^2$ has rational coordinates (rational slope if $p \in H_\infty$), then ϕp also has rational coordinates. By translating Q, if necessary, we may assume that $Q \cap \phi^{-1} H_\infty = \emptyset$. Then $(\phi Q) \cap H_\infty = \emptyset$, so ϕ is permissible for Q and hence ϕQ is a rational ϕU-polygon, where ϕU is a set of $|U|$

points in L with rational coordinates. Choose an $m \in \mathbb{N}$ so that the $|U|$ points in $m\phi U$ and the vertices of $m\phi Q$ belong to $\mathbb{Z}^2$. Then $m\phi Q$ is a lattice $m\phi U$-polygon, and $m\phi U$ is a subset of the line mL. Let ψ be a translation taking mL onto L and let $P = \psi(m\phi U)$. Then $\psi(m\phi Q)$ is the required lattice P-polygon.

(ii) Let Q be a rational P-polygon disjoint from L. Since the hypotheses ensure that Q is permissible for ϕ and L is a lattice line, ϕQ is a rational U-polygon. Then there is an $m \in \mathbb{N}$ such that $m\phi Q$ is a lattice U-polygon. □

COROLLARY 6.6. *Let P be a set of points in $\mathbb{Z}^2$ in a line L. Then convex lattice sets in $\mathbb{Z}^2$ not meeting L are determined by discrete point X-rays at the points in P if either:*

(i) $|P| \geq 7$, *or*

(ii) $|P| = 4$ *and there is no ordering of points in P such that their cross ratio is* 2, 3, *or* 4.

On the other hand, it is possible that $|P| = 6$ and there exist different convex lattice sets with convex hulls disjoint from L and equal discrete point X-rays at points in P.

Proof. Suppose that P is a set of points in $\mathbb{Z}^2$ in a line L, such that convex lattice sets in $\mathbb{Z}^2$ not meeting L are not determined by discrete point X-rays at the points in P. Then, by Theorem 6.4, there is a rational P-polygon disjoint from L. Theorem 6.5(ii) implies that there is a set U of $|P|$ mutually nonparallel vectors such that there exists a lattice U-polygon. By [12, Theorem 4.5], we have $|U| \leq 6$, so $|P| \leq 6$ and (i) is proved. Moreover, if $|P| = |U| = 4$, [12, Theorem 4.5] implies that there is an ordering of the vectors in U such that their cross ratio is 2, 3, or 4. Since U is obtained from P by a projective transformation, and such transformations preserve cross ratio, the same is true for P. Therefore (ii) is established.

By [12, Example 4.3], there is a set U of six mutually nonparallel vectors such that there exists a lattice U-polygon. It follows from Theorem 6.5(i) that there is a set P of six points in $L \cap \mathbb{Z}^2$ such that there is a lattice P-polygon disjoint from L. The proof is completed by an application of Lemma 4.5. □

In particular it follows from the previous result that convex lattice sets not meeting the x-axis are determined by their discrete point X-rays at points in the set $\{(0,0),(1,0),(2,0),(5,0)\}$.

7. The structure of *P*-polygons. Lemma 4.5 indicates that further progress hinges on a deeper understanding of the structure of lattice P-polygons. In view of the results of section 6, we focus on the case when the points in P are not collinear. This section provides some constructions of P-polygons Q such that $P \cap Q = \emptyset$. Note that Lemma 4.4 then guarantees that Q does not meet any line joining two points of P.

7.1. *P*-hexagons. We begin with the following construction.

THEOREM 7.1. *If P is a set of three points in $\mathbb{R}^2$, there exists a P-hexagon Q such that $P \cap Q = \emptyset$.*

Proof. Let $P = \{p_1, p_2, p_3\}$, and without loss of generality, suppose that the points are labeled so that p_2 and p_3 lie on the x-axis and p_1 is in the closed half plane $\{y \leq 0\}$. We may also assume that there is a line L_1 through p_1 meeting the relative interior of $[p_2, p_3]$; see Figure 2. Let $q_1 \in L_1 \cap \{y > 0\}$. Let q_2 and q_3 be in the relative interior of $[p_3, q_1]$ and $[p_2, q_1]$, respectively, and let $L_i = L[p_i, q_i]$, $i = 2, 3$. Let $L_4 = L[p_1, q_2]$, $L_5 = L[p_1, q_3]$, $\{q_4\} = L_2 \cap L_5$, and $\{q_5\} = L_3 \cap L_4$. Finally, let $L_6 = L[p_2, q_5]$ and $L_7 = L[p_3, q_4]$.

We claim that $L_1 \cap L_6 = L_1 \cap L_7 = \{q_6\}$, say. From this it would follow that the points q_i, $i = 1, \ldots, 6$ form the vertices of the required P-hexagon Q. To prove the

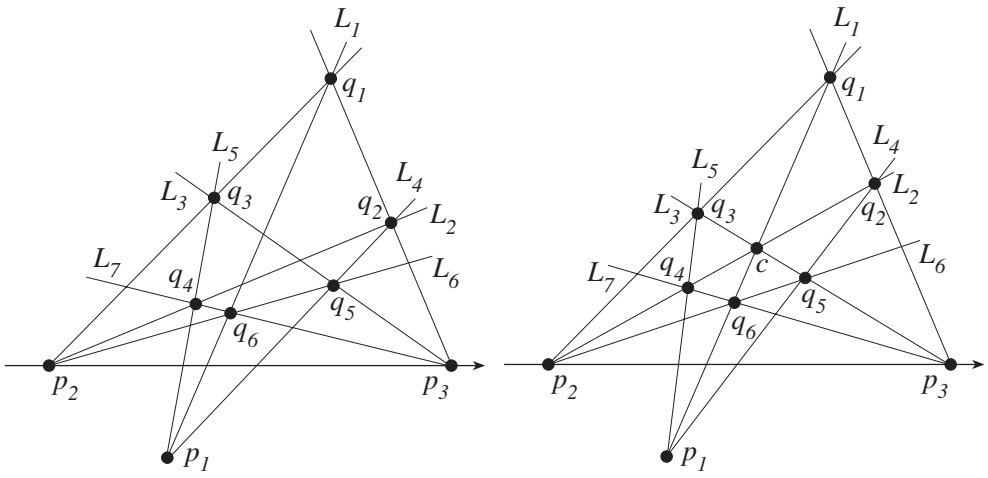


FIG. 2. *A P-hexagon (left) and a special P-hexagon (right) for three points.*

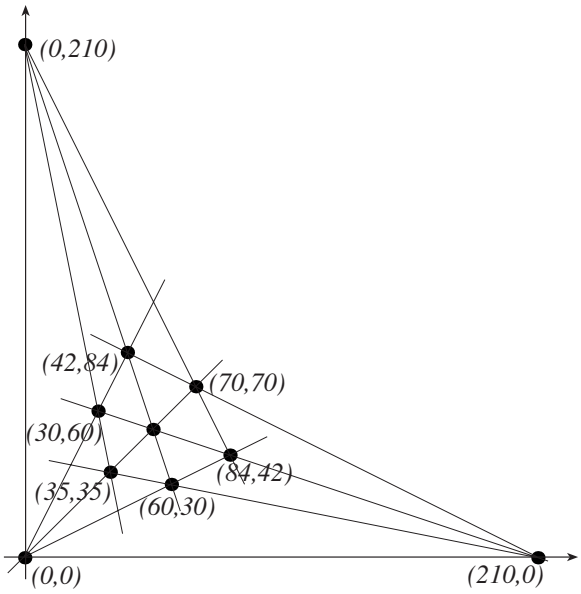


FIG. 3. *A lattice special P-hexagon for three noncollinear points.*

claim, consider the collinear triples (q_1, q_2, p_3) and (q_4, q_3, p_1). Note that $L[q_1, q_3]$ and $L[q_2, q_4]$ intersect at p_2, $L_1 \cap L_7 = [q_1, p_1] \cap [p_3, q_4]$, and $\{q_5\} = [q_2, p_1] \cap [q_3, p_3]$. By Pappus' theorem (see, for example, [6, section 4.3]), it follows that p_2, $L_1 \cap L_7$, and q_5 are collinear. Since p_2 and q_5 lie on L_6, $L_1 \cap L_7$ also lies on L_6 and the claim is proved. □

If in the construction of Theorem 7.1 the lines L_2 and L_3 are chosen so that $L_1 \cap L_2 = L_1 \cap L_3 = \{c\}$, say, we call the P-hexagon a *special P-hexagon* with *center* c. An example is shown in Figure 2.

COROLLARY 7.2. *For every set P of three points in $\mathbb{Z}^2$, there is a rational special*

P-hexagon Q such that $P \cap Q = \emptyset$. Hence there are sets P of three collinear points in $\mathbb{Z}^2$, or three noncollinear points in $\mathbb{Z}^2$, such that there exists a lattice special P-hexagon Q such that $P \cap Q = \emptyset$.

Proof. If the points in $P = \{p_1, p_2, p_3\}$ are lattice points, each line in the construction of Theorem 7.1 may be chosen so that it is represented by a linear equation with integer coefficients. The first statement in the corollary follows immediately. If Q is a rational special P-hexagon, there is an integer k such that if $P' = \{kp_1, kp_2, kp_3\}$, then kQ is a lattice special P'-hexagon. □

Figure 3 depicts a particular lattice special P-hexagon, obtained from a variation of the construction of Theorem 7.1 in which the hexagon is contained in the interior of the triangle with vertices at the points in P. The center of the hexagon has coordinates $(105/2, 105/2)$, so on multiplying each coordinate by 2, we obtain an example where the center of the hexagon is also a lattice point.

The following theorem shows that in the first statement of Corollary 7.2, "rational" cannot be replaced with "lattice." Note that for $P = \{(-1,0),(0,0),(1,0)\}$ a specific example of a lattice P-hexagon was given immediately after Theorem 5.1.

THEOREM 7.3. *If $P = \{(-1,0),(0,0),(1,0)\}$, there does not exist a lattice special P-hexagon.*

Proof. Suppose that Q is a lattice special P-hexagon, and without loss of generality, suppose that it is constructed and labeled as in Theorem 7.1 with $p_2 = (-1,0)$, $p_1 = (0,0)$, and $p_3 = (1,0)$. Let $q_6 = (a,b) \in \mathbb{Z}^2$, where $b \neq 0$, so that $c = (ma, mb)$ for some $m \in \mathbb{N}$, $m > 1$. Then we have

$$q_4 = \left(\frac{2am - m + 1}{m+1}, \frac{2bm}{m+1}\right) \quad \text{and} \quad q_5 = \left(\frac{2am + m - 1}{m+1}, \frac{2bm}{m+1}\right). \tag{22}$$

Subtracting the x-coordinates, we see that $m+1$ divides $2(m-1)$ and hence $m = 3$. Substituting this value of m into (22), we see that $a = 2k+1$ must be odd and $b = 2l$ must be even, and then $c = (6k+3, 6l)$, $q_4 = (3k+1, 3l)$, and $q_5 = (3k+2, 3l)$.

Now repeat the whole argument, replacing c, q_4, q_5, and q_6 by q_1, q_3, q_2, and c, respectively. Since c now plays the role of q_6, the coordinates of q_2 and q_3 are given by the formulas (22) for q_5 and q_4, respectively, with $m = 3$, $a = 6k+3$, and $b = 6l$. So $q_2 = (9k+5, 9l)$ and $q_3 = (9k+4, 9l)$. But then the line through q_3 and q_4 is parallel to the line through q_2 and q_5, impossible since these lines should meet at p_1. □

7.2. *P*-octagons. We start with the following lemma.

LEMMA 7.4. *For every set P of four noncollinear points in $\mathbb{Z}^2$, there exists a convex quadrilateral V with $P \cap V = \emptyset$ such that lines containing opposite edges of V intersect at points in P and the lines containing the diagonals of V each contain one of the remaining points of P.*

Proof. Suppose first that $P = \{p_1, p_2, p_3, p_4\}$ is a set of four points in $\mathbb{Z}^2$ such that the points p_2, p_3, and p_4 lie in a line L, and without loss of generality, suppose $p_3 \in [p_2, p_4]$. See Figure 4 (left). Let $m \in [p_2, p_3]$ be such that $\langle p_2, p_3, p_4, m\rangle = -1$, where $\langle\cdot\rangle$ denotes the cross ratio, as in (1). Let $L_1 = L[m, p_1]$, let $v_1 \in L_1$ be in the relative interior of $[p_1, m]$, and let $L_i = L[p_i, v_1]$, $i = 2, 3$. Let $v_2 \in L_2$ be in the relative interior of $[p_2, v_1]$ and let $L_4 = L[p_4, v_2]$. Let $\{v_3\} = L_3 \cap L_4$ and let $L_5 = L[p_2, v_3]$. Finally, let $L_6 = L[p_3, v_2]$.

We claim that $L_1 \cap L_5 = L_1 \cap L_6 = \{v_4\}$, say. From this it would follow that the points v_i, $i = 1, \ldots, 4$, form the vertices of the required quadrilateral.

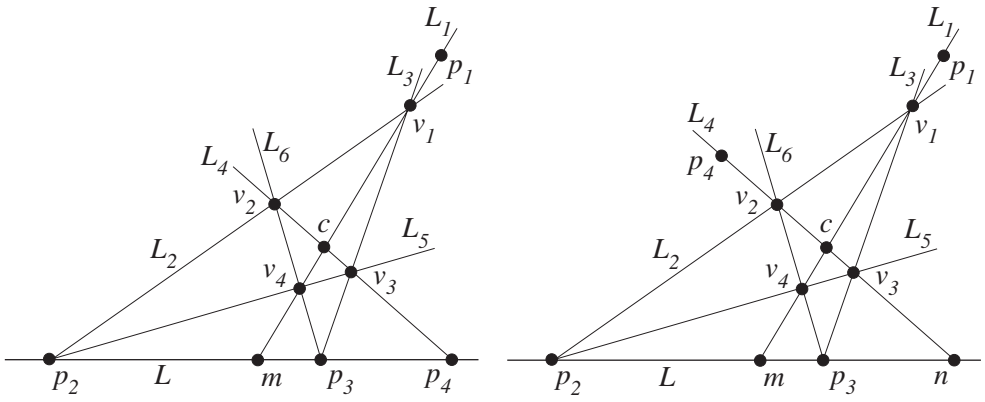


FIG. 4.

To prove the claim, let $\{c\} = L_1 \cap L_4$, let $L_1 \cap L_5 = \{v_4\}$, and let $L_1 \cap L_6 = \{v_4'\}$. The perspectivity with center v_1 takes the points v_2, v_3, p_4, and c on L_4 onto the points p_2, p_3, p_4, and m in L, so we have $\langle v_2, v_3, p_4, c\rangle = \langle p_2, p_3, p_4, m\rangle = -1$. The perspectivity with center p_2 takes the points v_2, v_3, p_4, and c on L_4 onto the points v_1, v_4, m, and c on L_1, so $\langle v_1, v_4, m, c\rangle = \langle v_2, v_3, p_4, c\rangle = -1$. Finally, the perspectivity with center p_3 takes the points v_2, v_3, p_4, and c on L_4 onto the points v_4', v_1, m, and c on L_1, so $\langle v_1, v_4', m, c\rangle = 1/\langle v_4', v_1, m, c\rangle = 1/\langle v_2, v_3, p_4, c\rangle = -1$. Thus $\langle v_1, v_4', m, c\rangle = \langle v_1, v_4, m, c\rangle = -1$, which gives $v_4 = v_4'$, as required.

Now suppose that $P = \{p_1, p_2, p_3, p_4\}$ is a set of four points in $\mathbb{Z}^2$, no three of which are collinear. Suppose that the points are labeled so that $L[p_1, p_4] \cap [p_2, p_3] = \emptyset$. If $L = L[p_2, p_3]$, then either p_1 and p_4 lie on the same side of L, or they lie on opposite sides of L. The former case is illustrated in Figure 4 (right); the latter case corresponds to the situation in which the point p_4 in Figure 4 (right) lies in L_4 below L. Let m be in the relative interior of $[p_2, p_3]$, and suppose $n \in L$ is such that $\langle p_2, p_3, n, m\rangle = -1$. Let $L_1 = L[m, p_1]$ and $L_4 = L[n, p_4]$. We may assume that m and n are chosen so that if $\{c\} = L_1 \cap L_4$, then c lies in the relative interior of $[m, p_1]$. Let $v_1 \in L_1$ be in the relative interior of $[c, p_1]$, and let $L_i = L[p_i, v_1]$, $i = 2, 3$. Let $\{v_i\} = L_i \cap L_4$, $i = 2, 3$. Then v_i lies in the relative interior of $[p_i, v_1]$, $i = 2, 3$, and $p_4 \notin [v_2, v_3]$, since p_4 is not contained in the triangle with vertices p_1, p_2, and p_3. Finally, let $L_5 = L[p_2, v_3]$ and $L_6 = L[p_3, v_2]$.

We claim that $L_1 \cap L_5 = L_1 \cap L_6 = \{v_4\}$, say. From this it would follow that the points v_i, $i = 1, \ldots, 4$, form the vertices of the required quadrilateral. To prove the claim we can follow exactly the argument used in the previous case, on replacing p_4 with n. □

THEOREM 7.5. *For every set P of four noncollinear points in $\mathbb{Z}^2$, there exists a rational P-octagon Q such that $P \cap Q = \emptyset$. Hence there are sets P of four noncollinear points in $\mathbb{Z}^2$ such that there exists a lattice P-octagon Q such that $P \cap Q = \emptyset$.*

Proof. Let $P = \{p_1, p_2, p_3, p_4\}$ be a set of noncollinear points in $\mathbb{Z}^2$. Suppose that the points in P are labeled so that $L[p_1, p_4] \cap [p_2, p_3] = \emptyset$. Let V be a quadrilateral built as in Lemma 7.4. Note that in the proof of Lemma 7.4, we can interchange p_2 and p_3, if necessary, so that the points p_3 and p_4 belong to the same half plane bounded by $L[p_2, v_4]$. Moreover, since $P \subset \mathbb{Z}^2$, each line in the construction of Lemma 7.4 can be chosen so that it is represented by a linear equation with integer coefficients.

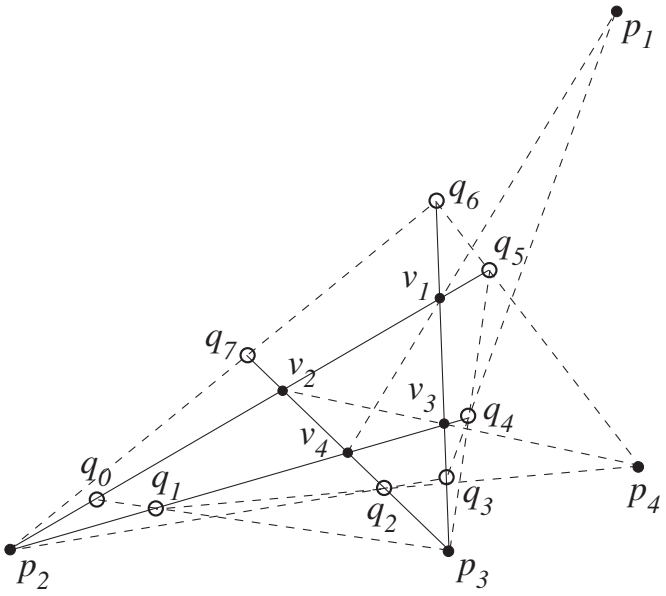


FIG. 5. *Construction of a P-octagon for four noncollinear points.*

Thus, we can assume that the vertices of V have rational coordinates.

We will construct a P-octagon with vertices on the lines containing the edges of V. Let q_0 be a point with rational coordinates in the relative interior of $[p_2, v_2]$, and let q_1 be in the relative interior of $[p_2, v_4]$ such that $\{q_1\} = L[p_2, v_4] \cap L[p_3, q_0]$. See Figure 5. Let $\{q_2\} = L[p_3, v_4] \cap L[p_4, q_1]$, and note that q_0 can be chosen so that q_2 is in the relative interior of $[p_3, v_4]$. Let $\{q_3\} = L[p_3, v_3] \cap L[p_2, q_2]$, and note that q_3 is in the relative interior of $[p_3, v_3]$. Finally, let $\{q_4\} = L[p_2, v_3] \cap L[p_1, q_3]$, $\{q_5\} = L[p_3, q_4] \cap L[p_2, v_1]$, $\{q_6\} = L[p_4, q_5] \cap L[p_3, v_1]$, and $\{q_7\} = L[p_2, q_6] \cap L[p_3, v_2]$.

We claim that the octagon Q with vertices q_i (indicated by white circles in Figure 5), where the subscripts are understood to be integers mod 8, is a P-polygon. The construction ensures that the points p_2, q_i, and q_{5-i} are collinear and that the points p_3, q_i, and q_{1-i} are collinear. It remains to prove that Q is a P'-polygon, where $P' = \{p_1, p_4\}$. Note that for $i = 1$ and 5, the points p_4, q_i, and q_{3-i} are collinear by construction.

Consider the lines $L_1 = L[q_4, q_5]$, $L_2 = L[q_3, q_6]$, and $L_3 = L[q_2, q_7]$ through p_3 and the lines $L_1' = L[q_6, q_7]$, $L_2' = L[q_0, q_5]$, and $L_3' = L[q_1, q_4]$ through p_2. Then $L_1 \cap L_2' = \{q_5\}$, $L_1' \cap L_2 = \{q_6\}$, $L_1 \cap L_3' = \{q_4\}$, $L_1' \cap L_3 = \{q_7\}$, $L_2 \cap L_3' = \{v_3\}$, and $L_2' \cap L_3 = \{v_2\}$, so by the dual of Pappus's theorem ([6, section 4.3]), it follows that the lines $L[q_5, q_6]$, $L[q_4, q_7]$, and $L[v_2, v_3]$ belong to the same pencil. Since $L[q_5, q_6] \cap L[v_2, v_3] = \{p_4\}$, the points p_4, q_4, and q_7 are collinear, as required. In the same way, by applying the dual of Pappus's theorem to the lines $L_1 = L[q_3, q_6]$, $L_2 = L[q_0, q_1]$, and $L_3 = L[q_2, q_7]$ through p_3, and $L_1' = L[q_0, q_5]$, $L_2' = L[q_2, q_3]$, and $L_3' = L[q_1, q_4]$ through p_2, it follows that p_4, q_0, and q_3 are collinear. Therefore p_4, q_i, and q_{3-i} are collinear for every i.

Note that p_1, q_3, and q_4 are collinear by construction. By applying the dual of Pappus's theorem to the lines $L_1 = L[q_4, q_5]$, $L_2 = L[q_3, q_6]$, and $L_3 = L[q_2, q_7]$ through p_3, and the lines $L_1' = L[q_2, q_3]$, $L_2' = L[q_1, q_4]$, and $L_3' = L[q_0, q_5]$ through

p_2, it follows that p_1, q_2, and q_5 are collinear. Consider the triangle T with vertices q_1, q_2, and v_4, and the triangle T' with vertices q_5, q_6, and v_1. Let $L_1 = L[q_1, q_2]$, $L_2 = L[q_2, v_4]$, and $L_3 = L[q_1, v_4]$ be the lines containing the edges of T, and let $L_1' = L[q_5, q_6]$, $L_2' = L[q_5, v_1]$, and $L_3' = L[q_6, v_1]$ be the corresponding lines containing the edges of T'. Since $L_1 \cap L_1' = \{p_4\}$, $L_2 \cap L_2' = \{v_2\}$, and $L_3 \cap L_3' = \{v_3\}$, and the points p_4, v_2, and v_3 are collinear, the lines $L[q_2, q_5]$, $L[q_1, q_6]$, and $L[v_1, v_4]$ belong to the same pencil. Also, $L[q_2, q_5] \cap L[v_1, v_4] = \{p_1\}$, so the points p_1, q_1, and q_6 are collinear. By applying the dual of Pappus's theorem to the lines $L_1 = L[q_3, q_6]$, $L_2 = L[q_0, q_1]$, and $L_3 = L[q_2, q_7]$ through p_3, and the lines $L_1' = L[q_1, q_4]$, $L_2' = L[q_6, q_7]$, and $L_3' = L[q_0, q_5]$ through p_2, it follows that p_1, q_0, and q_7 are collinear. Therefore p_1, q_i, and q_{7-i} are collinear for every i, completing the proof that Q is a P-octagon.

Finally, note that since $P \subset \mathbb{Z}^2$ and the vertices of Q all have rational coordinates, there is an integer k such that if $P' = \{kp_1, kp_2, kp_3, kp_4\}$, then kQ is a lattice P'-octagon. □

The previous result stands in contrast to the situation for collinear sets P. It follows from [12, Theorem 4.5] that there are sets U of four directions in $\mathbb{R}^2$ with rational slopes such that there do not exist *any* U-polygons (lattice, rational, or otherwise). By Theorem 6.5(ii), there are sets P of four collinear points in the x-axis such that there are no P-polygons, and, in particular, no rational P-polygons, disjoint from the x-axis.

7.3. A *P*-dodecagon. Almost nothing seems to be known about P-polygons beyond the material in the previous subsections. The following result is obtained by a construction similar to that of Theorem 7.5, but starting with a certain special P-hexagon instead of a quadrilateral. In view of the isolated nature of the construction, we simply list the relevant points.

THEOREM 7.6. *There is a set P of six points in $\mathbb{Z}^2$, no four of which are collinear, such that there exists a lattice P-dodecagon Q such that $P \cap Q = \emptyset$.*

Proof. Let $P' = \{p_1, p_2, \ldots, p_6\}$, where $p_1 = (0, -12)$, $p_2 = (6, 0)$, $p_3 = (-4, 4)$, $p_4 = (-24, 12)$, $p_5 = (12, 12)$, and $p_6 = (-6, 12)$. Let Q be the dodecagon with vertices $q_0 = (16/5, -14/5)$, $q_1 = (84/29, -96/29)$, $q_2 = (12/107, -672/107)$, $q_3 = (-1/9, -55/9)$, $q_4 = (-7/3, -1/3)$, $q_5 = (-48/19, 6/19)$, $q_6 = (-336/109,\ 330/109)$, $q_7 = (-220/73, 224/73)$, $q_8 = (-4/15, 32/15)$, $q_9 = (3/11, 21/11)$, $q_{10} = (165/41, 3/41)$, and $q_{11} = (112/27, -2/27)$. A computation of the slopes of the segments $[q_i, q_{i+1}]$, where the indices are taken modulo 12, shows that Q is convex. A further computation shows that for $i = 0, \ldots, 5$, the following triples of points are collinear: (p_1, q_i, q_{11-i}), (p_2, q_i, q_{3-i}), (p_3, q_i, q_{7-i}), (p_4, q_i, q_{5-i}), (p_5, q_i, q_{1-i}), and (p_6, q_i, q_{9-i}), where again indices are taken modulo 12. This shows that Q is a P'-polygon. Since the vertices of Q have rational coordinates, there is a $k \in \mathbb{N}$ such that kQ is a lattice P-polygon, where $P = \{kp_1, kp_2, \ldots, kp_6\}$. □

8. Discrete point *X*-rays at noncollinear points. Volčič (see [17] or [11, Chapter 5]) proved that planar convex bodies are determined by their continuous point X-rays at any set of four points, no three of which are collinear. We show in this section that the situation is somewhat different for discrete point X-rays.

By Corollary 7.2, there are sets $P = \{p_1, p_2, p_3\}$ of three noncollinear points in $\mathbb{Z}^2$, such that there exists a lattice special P-hexagon. Moreover, it can be arranged that the center c of the hexagon is also a lattice point, in which case the hexagon is a lattice P'-polygon for the set $P' = \{p_1, p_2, p_3, c\}$ of four noncollinear points. The point c may be in the interior of the triangle formed by the points in P, as in Figure 3, or exterior to this triangle, as in the construction of Theorem 7.1. By Lemma 4.5,

there are different convex lattice sets with equal discrete point X-rays at the points in P'. These examples show that the results of Volčič (see [17] or [11, Theorems 5.3.6 and 5.3.7]) do not hold in the discrete case.

The following direct consequence of Theorem 7.5 and Lemma 4.5 shows that another result of Volčič (see [17] or [11, Theorem 5.3.8]) also does not hold in the discrete case.

THEOREM 8.1. *There is a set P of four points in $\mathbb{Z}^2$, no three of which are collinear, such that convex lattice sets not meeting any line joining two points in P are not determined by discrete point X-rays at the points in P.*

Finally, Theorem 7.6 and Lemma 4.5 immediately yield the following result.

THEOREM 8.2. *There is a set P of six points in $\mathbb{Z}^2$, no four of which are collinear, such that convex lattice sets not meeting any line joining two points in P are not determined by discrete point X-rays at the points in P.*

REFERENCES

[1] A. ADLER AND J. E. COURY, *The Theory of Numbers*, Jones and Bartlett, Boston, 1995.

[2] A. ALPERS AND R. TIJDEMAN, *The two-dimensional Prouhet-Tarry-Escott problem*, J. Number Theory, to appear.

[3] I. BÁRÁNY AND J. MATOUŠEK, *A fractional Helly theorem for convex lattice sets*, Adv. Math., 174 (2003), pp. 227–235.

[4] M. BERGER, *Geometry*, Springer, Berlin, 1987.

[5] S. BRUNETTI AND A. DAURAT, *An algorithm reconstructing lattice convex sets*, Theoret. Comput. Sci., 304 (2003), pp. 35–57.

[6] H. S. M. COXETER, *The Real Projective Plane*, Cambridge University Press, Cambridge, 1961.

[7] V. I. DANILOV AND G. A. KOSHEVOY, *Discrete convexity and unimodularity—I*, Adv. Math., 189 (2004), pp. 301–324.

[8] A. DAURAT, *Connexité et Convexité Directionelle Dans $\mathbb{Z}^2$*, in CNR'IUT2000, Vol. 1, Presses Universitaires d'Orléans, Orléans, France, 2000, pp. 341–350.

[9] I. DEBLED-RENNESSON, J.-L. RÉMY, AND J. ROUYER-DEGLI, *Detection of the discrete convexity of polyominoes*, Discrete Appl. Math., 125 (2003), pp. 115–133.

[10] P. DULIO, R. J. GARDNER, AND C. PERI, *Discrete point X-rays of convex lattice sets*, Electron. Notes Discrete Math., 20 (2005), pp. 1–13.

[11] R. J. GARDNER, *Geometric Tomography*, Cambridge University Press, New York, 1995. Second edition, 2006.

[12] R. J. GARDNER AND P. GRITZMANN, *Discrete tomography: Determination of finite sets by X-rays*, Trans. Amer. Math. Soc., 349 (1997), pp. 2271–2295.

[13] R. J. GARDNER AND P. GRITZMANN, *Uniqueness and complexity in discrete tomography*, in Discrete Tomography: Foundations, Algorithms and Application, G. T. Herman and A. Kuba, eds., Birkhäuser, Boston, 1999, pp. 85–113.

[14] B. GREEN AND T. TAO, *The primes contain arbitrarily long arithmetic progressions*, Ann. of Math., to appear.

[15] G. T. HERMAN AND A. KUBA, *Discrete Tomography: Foundations, Algorithms, and Applications*, Birkhäuser, Boston, 1999.

[16] W. SIERPIŃSKI, *Sur les suites d'entiers deux à deux premiers entre eux*, Enseignement Math., 10 (1964), pp. 229–235.

[17] A. VOLČIČ, *A three-point solution to Hammer's X-ray problem*, J. London Math. Soc., 34 (1986), pp. 349–359.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 189–202

IMPROVED BOUNDS FOR THE CROSSING NUMBERS OF $K_{m,n}$ AND K_n*

E. DE KLERK†, J. MAHARRY‡, D. V. PASECHNIK§, R. B. RICHTER†, AND G. SALAZAR¶

Abstract. It has been long conjectured that the crossing number $\mathrm{cr}(K_{m,n})$ of the complete bipartite graph $K_{m,n}$ equals the Zarankiewicz number $Z(m,n) := \lfloor \frac{m-1}{2} \rfloor \lfloor \frac{m}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor \lfloor \frac{n}{2} \rfloor$. Another longstanding conjecture states that the crossing number $\mathrm{cr}(K_n)$ of the complete graph K_n equals $Z(n) := \frac{1}{4} \lfloor \frac{n}{2} \rfloor \lfloor \frac{n-1}{2} \rfloor \lfloor \frac{n-2}{2} \rfloor \lfloor \frac{n-3}{2} \rfloor$. In this paper we show the following improved bounds on the asymptotic ratios of these crossing numbers and their conjectured values:

(i) for each fixed $m \geq 9$, $\lim_{n\to\infty} \mathrm{cr}(K_{m,n})/Z(m,n) \geq 0.83m/(m-1)$;
(ii) $\lim_{n\to\infty} \mathrm{cr}(K_{n,n})/Z(n,n) \geq 0.83$; and
(iii) $\lim_{n\to\infty} \mathrm{cr}(K_n)/Z(n) \geq 0.83$.

The previous best known lower bounds were $0.8m/(m-1), 0.8$, and 0.8, respectively. These improved bounds are obtained as a consequence of the new bound $\mathrm{cr}(K_{7,n}) \geq 2.1796n^2 - 4.5n$. To obtain this improved lower bound for $\mathrm{cr}(K_{7,n})$, we use some elementary topological facts on drawings of $K_{2,7}$ to set up a quadratic program on 6! variables whose minimum p satisfies $\mathrm{cr}(K_{7,n}) \geq (p/2)n^2 - 4.5n$, and then use state-of-the-art quadratic optimization techniques combined with a bit of invariant theory of permutation groups to show that $p \geq 4.3593$.

Key words. crossing number, semidefinite programming, copositive cone, invariants and centralizer rings of permutation groups

AMS subject classifications. 05C10, 05C62, 90C22, 90C25, 57M15, 68R10

DOI. 10.1137/S0895480104442741

1. Introduction.

In the earliest known instance of a crossing number question, Turán raised the problem of calculating the crossing number of the complete bipartite graphs $K_{m,n}$. Turán's interesting account of the origin of this problem can be found in [27].

We recall that in a *drawing* of a graph in the plane, different vertices are drawn as different points, and each edge is drawn as a simple arc whose endpoints coincide with the drawings of the endvertices of the edge. Furthermore, the interior of the arc for an edge is disjoint from all the vertex points. We often make no distinction between a graph object, such as a vertex, edge, or cycle, and the subset of the plane that represents it in a drawing of the graph.

The *crossing number* $\mathrm{cr}(G)$ of a graph G is the minimum number of pairwise intersections of edges (at a point other than a vertex) in a drawing of G in the plane.

*Received by the editors April 5, 2004; accepted for publication (in revised form) August 29, 2005; published electronically March 3, 2006.

http://www.siam.org/journals/sidma/20-1/44274.html

†Department of Combinatorics and Optimization, Faculty of Mathematics, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e.deklerk@uvt.nl, brichter@math.uwaterloo.ca).

‡Department of Mathematics, The Ohio State University, Columbus, OH 43210 (maharry@math.ohio-state.edu).

§Theoretische Informatik, FB20 Informatik, J.W. Goethe-Universität, Robert-Mayer Str. 11-15, Postfach 11 19 32, 60054 Frankfurt(Main), Germany (Dima@ntu.edu.sg). This author's work was partially supported by DFG grant SCHN-503/2-1. Part of the research was completed while this author was supported by the Mathematical Sciences Research Institute (MSRI) at Berkeley, CA.

¶Instituto de Fisica, Universidad Autonoma de San Luis Potosi, San Luis Potosi, SLP 78000, Mexico (gsalazar@ifisica.uaslp.mx). This author's work was supported by grants CONACYT J32168 and FAI-UASLP. Part of the research was completed during a sabbatical leave at The Ohio State University, Columbus, OH.

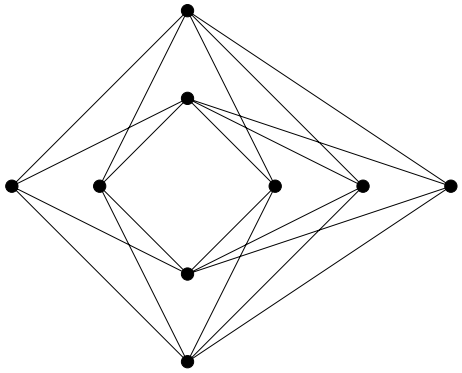


FIG. 1. *A drawing of $K_{4,5}$ with 8 crossings. A similar strategy can be used to construct drawings of $K_{m,n}$ with exactly $Z(m,n)$ crossings.*

Exact crossing numbers of graphs are in general very difficult to compute. Longstanding conjectures involve the crossing numbers of interesting families of graphs, such as $K_{m,n}$ and K_n. On a positive note, it was recently proved by Glebsky and Salazar [9] that the crossing number of the Cartesian product $C_m \times C_n$ of the cycles of sizes m and n equals its long conjectured value, namely $(m-2)n$, at least for $n \geq m(m+1)$. For recent surveys of crossing number results, see [23] or [26].

Zarankiewicz published a paper [29] in which he claimed that $\mathrm{cr}(K_{m,n}) = Z(m,n)$ for all positive integers m, n, where

$$Z(m,n) = \left\lfloor \frac{m-1}{2} \right\rfloor \left\lfloor \frac{m}{2} \right\rfloor \left\lfloor \frac{n-1}{2} \right\rfloor \left\lfloor \frac{n}{2} \right\rfloor. \tag{1}$$

However, several years later Ringel and Kainen independently found a hiatus in Zarankiewicz's argument. A comprehensive account of the history of the problem, including a discussion of the gap in Zarankiewicz's argument, is given by Guy [11].

Figure 1 shows a drawing of $K_{4,5}$ with 8 crossings. As Zarankiewicz observed, such a drawing strategy can be naturally generalized to construct, for any positive integers m, n, drawings of $K_{m,n}$ with exactly $Z(m,n)$ crossings. This observation implies the following well-known upper bound for $\mathrm{cr}(K_{m,n})$:

$$\mathrm{cr}(K_{m,n}) \leq Z(m,n).$$

No one has yet exhibited a drawing of any $K_{m,n}$ with fewer than $Z(m,n)$ crossings. In allusion to Zarankiewicz's failed attempt to prove that this is the crossing number of $K_{m,n}$, the following is commonly known as *Zarankiewicz's crossing–number conjecture*:

$$\mathrm{cr}(K_{m,n}) \stackrel{?}{=} Z(m,n) \qquad \text{for all positive integers } m, n.$$

In 1973, Guy and Erdős [6] wrote, "Almost all questions that one can ask about crossing numbers remain unsolved." More than three decades later, despite some definite progress in our understanding of this elusive parameter, most of the fundamental and more important questions about crossing numbers remain open. Zarankiewicz's conjecture has been verified by Kleitman [13] for $\min\{m,n\} \leq 6$ and by Woodall [28] for the special cases $7 \leq m \leq 8$, $7 \leq n \leq 10$.

Since the crossing number of $K_{m,n}$ is unknown for all other values of m and n, it is natural to ask what are the best general lower bounds known for $\mathrm{cr}(K_{m,n})$. A standard counting argument, together with the fact that $\mathrm{cr}(K_{5,n})$ is as conjectured, yields the best general lower bound (2) known for $\mathrm{cr}(K_{m,n})$. It goes as follows: Suppose we know a lower bound c_r on $\mathrm{cr}(K_{r,n})$ for $2 < r < m \leq n$. Each crossing in the embedding of $K_{m,n}$ lies in $\binom{m-2}{r-2}$ distinct $K_{r,n} \subset K_{m,n}$. As there are in total $\binom{m}{r}$ distinct $K_{r,n}$'s, one obtains

$$\mathrm{cr}(K_{m,n}) \geq \frac{c_r\binom{m}{r}}{\binom{m-2}{r-2}}; \quad \text{for } r = 5 \text{ one derives} \quad \mathrm{cr}(K_{m,n}) \geq 0.8\, Z(m,n). \tag{2}$$

A small improvement on the 0.8 factor (roughly to something around 0.8001) was recently reported by Nahas [18].

Zarankiewicz's conjecture for $K_{7,n}$ states that

$$\mathrm{cr}(K_{7,n}) \stackrel{?}{=} 9\left\lfloor \frac{n-1}{2} \right\rfloor \left\lfloor \frac{n}{2} \right\rfloor = \begin{cases} 2.25n^2 - 4.5n + 2.25, & n \text{ odd}, \quad n \geq 7, \\ 2.25n^2 - 4.5n, & n \text{ even}, \quad n \geq 8. \end{cases}$$

As we observed above, this has been verified only for $n = 7, 8, 9$, and 10. Using $\mathrm{cr}(K_{7,10}) = 180$, a standard counting argument gives the best known lower bounds for $\mathrm{cr}(K_{7,n})$ for $11 \leq n \leq 22$. However, for $n \geq 23$, the best known lower bounds for $\mathrm{cr}(K_{7,n})$ are obtained by the same counting argument, but using the known value of $\mathrm{cr}(K_{5,n})$ instead of $\mathrm{cr}(K_{7,10})$. Summarizing, previous to this paper, the best known lower bounds for $\mathrm{cr}(7,n)$ were

$$\mathrm{cr}(K_{7,n}) \geq \begin{cases} 2n(n-1), & 11 \leq n \leq 22, \\ 2.1n^2 - 4.2n + 2.1, & \text{odd } n \geq 23, \\ 2.1n^2 - 4.2n, & \text{even } n \geq 24. \end{cases} \tag{3}$$

In this paper we prove the following theorem.

THEOREM 1. *For all integers n,*

$$\mathrm{cr}(K_{7,n}) > 2.1796n^2 - 4.5n.$$

An elementary calculation shows that this is an improvement, for all $n \geq 23$, on the bounds for $\mathrm{cr}(K_{7,n})$ given in (3).

The strategy of the proof can be briefly outlined as follows. Let (A, B) be the bipartition of the vertex set of $K_{7,n}$, where $|A| = 7$ and $|B| = n \geq 2$. Let b, b' be vertices in B. In any drawing $\mathcal{D}$ of $K_{7,n}$, the number of crossings that involve an edge incident with b and an edge incident with b' is bounded from below by a function of the cyclic rotation schemes of b and b'. This elementary topological observation on drawings of $K_{2,7}$ naturally yields a standard quadratic (minimization) program whose minimum p satisfies $\mathrm{cr}(K_{7,n}) \geq (p/2)n^2 - 4.5n$ (see Lemma 2). We then use state-of-the-art quadratic programming techniques to show that $p \geq 4.3593$ (see Proposition 3), thus implying Theorem 1.

The rest of this paper is organized as follows. In section 2, we review some elementary topological observations about drawings of $K_{2,n}$ and use these facts to set up the quadratic program mentioned in the previous paragraph. The bound for $\mathrm{cr}(K_{7,n})$ in terms of the minimum of this quadratic program is the content of Lemma 2. In section 3 we prove Proposition 3, which gives a lower bound for the quadratic program. As we observe at the end of section 3, Theorem 1 is an obvious consequence of Lemma 2 and Proposition 3. In section 4 we discuss consequences of Theorem 1: The improved bound for $\mathrm{cr}(K_{7,n})$ implies improved asymptotic bounds for the crossing numbers of $\mathrm{cr}(K_{m,n})$ and $\mathrm{cr}(K_n)$.

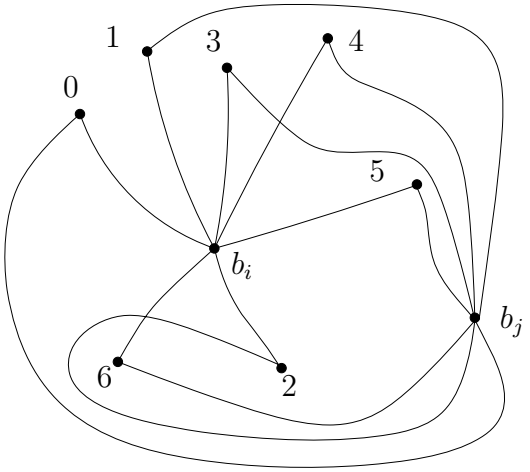


FIG. 2. *Here $m = 7$. Vertices b_i and b_j have cyclic orderings* $(0\,1\,3\,4\,5\,2\,6)$ *and* $(0\,2\,6\,5\,3\,4\,1)$*, respectively (we write i for a_i for the sake of brevity). It is easy to check that the minimum number of interchanges among adjacent elements in* $(0\,1\,3\,4\,5\,2\,6)$ *required to obtain* $(0\,2\,6\,5\,3\,4\,1)^{-1}$ *(namely* $(0\,1\,4\,3\,5\,6\,2)$*) is* 2*. Thus, $Q((0\,1\,3\,4\,5\,2\,6),(0\,1\,3\,4\,5\,2\,6)) = 2$. Therefore, there must be at least* 2 *crossings (as is indeed the case in the drawing above) that involve edges incident with b_i and b_j.*

2. Quadratic optimization problem yielding a lower bound for $\mathbf{cr}(K_{m,n})$. Our goal in this section is to establish Lemma 2, a statement that gives a lower bound for $\mathrm{cr}(K_{m,n})$ for $m \le n$ (and thus for $\mathrm{cr}(K_{7,n})$) in terms of the solution of a quadratic minimization problem on $(m-1)!$ variables.

Let $n \ge m$ be fixed. Let V denote the vertex set of $K_{m,n}$, and let (A, B) denote the bipartition of V such that each vertex of $A = \{a_0, a_1, \ldots, a_{m-1}\}$ is adjacent to each vertex of $B = \{b_0, b_1, \ldots, b_{n-1}\}$.

Consider a fixed drawing $\mathcal{D}$ of $K_{m,n}$. To each vertex b_i we associate a cyclic ordering $\pi_{\mathcal{D}}(b_i)$ of the elements in A, defined by the (clockwise) cyclic order in which the edges incident with b_i leave b_i toward the vertices in A (see Figure 2). Let Π denote the set of all cyclic orderings of $\{a_0, a_1, \ldots, a_{m-1}\}$. Note that $|\Pi| = m!/m = (m-1)!$.

Following Kleitman [13], let $\mathrm{cr}_{\mathcal{D}}(b_i, b_j)$ denote the number of crossings in $\mathcal{D}$ that involve an edge incident with b_i and an edge incident with b_j. Further, let $\rho_1, \rho_2 \in \Pi$ and $Q(\rho_1, \rho_2)$ be the minimum number of interchanges of adjacent elements of ρ_1 required to produce ρ_2^{-1}. Then, for all b_i, b_j with $b_i \ne b_j$,

$$\mathrm{cr}_{\mathcal{D}}(b_i, b_j) \ge Q(\pi_{\mathcal{D}}(b_i), \pi_{\mathcal{D}}(b_j)). \tag{4}$$

This inequality is stated in [13] and proved in [28]. This observation alone yields a lower bound for $\mathrm{cr}(K_{m,n})$, as follows. Fix any drawing $\mathcal{D}$ of $K_{m,n}$. For each $\rho \in \Pi$, let

$$x_\rho := \frac{1}{n}|\{b_i \in B \mid \pi_{\mathcal{D}}(b_i) = \rho\}|.$$

The matrix Q can be viewed as the matrix of quadratic form $Q(\cdot,\cdot)$ on the space $\mathbb{R}^{|\Pi|}$.

It follows from (4) that

$$\begin{aligned}\operatorname{cr}(\mathcal{D}) &\geq \sum_{\substack{\rho,\rho'\in\Pi\\ \rho\neq\rho'}} Q(\rho,\rho')(x_\rho n)(x_{\rho'} n) + \sum_{\rho\in\Pi} Q(\rho,\rho)\binom{x_\rho n}{2}\\ &= \frac{n}{2}\left(n\sum_{\rho,\rho'\in\Pi} Q(\rho,\rho')x_\rho x_{\rho'} - \left\lfloor\frac{m}{2}\right\rfloor\left\lfloor\frac{m-1}{2}\right\rfloor\right),\end{aligned}$$

using the (easily verifiable; see, e.g., [28]) fact that $Q(\rho,\rho) = \lfloor m/2\rfloor\lfloor(m-1)/2\rfloor$ for every $\rho\in\Pi$.

Since the drawing $\mathcal{D}$ was arbitrary, we have proved the following lemma.

LEMMA 2. *Let Q be the $(m-1)!\times(m-1)!$ matrix of the form $Q(\cdot,\cdot)$, and let e denote the all ones vector. Then, for every integer $n\geq m\geq 2$,*

$$\operatorname{cr}(K_{m,n}) \geq \frac{n}{2}\left(n\min\{x^TQx \;\mid\; x\in\mathbb{R}_+^{(m-1)!}, e^Tx=1\} - \left\lfloor\frac{m}{2}\right\rfloor\left\lfloor\frac{m-1}{2}\right\rfloor\right),$$

$$\operatorname{cr}(K_{7,n}) \geq \frac{n}{2}(n\min\{x^TQx \;\mid\; x\in\mathbb{R}_+^{6!}, e^Tx=1\} - 9).$$

Remark. In this paper we focus on the case $m=7$. For obvious reasons (for $m=7$, Q is a 720×720 matrix) we do not include in this paper the matrix Q in table form. As we mentioned above, $Q(\rho,\rho)=9$ for every $\rho\in\Pi$, and therefore all the diagonal entries of Q are 9. It is not difficult to show that $Q(\rho,\rho')\leq 8$ if $\rho\neq\rho'$, so every nondiagonal entry of Q is at most 8. The calculation of the entries of Q, using the definition of $Q(\cdot,\cdot)$ and taking its symmetries into account (see section 3.2), takes only a few seconds of computer time.

3. Finding a lower bound for the optimization problem. Our aim in this section is to find a (reasonably good) lower bound for the quadratic programming problem with $m=7$ given in Lemma 2, in order to obtain a (reasonably good) lower bound for $\operatorname{cr}(K_{7,n})$. The main result in this section is the following.

PROPOSITION 3. *Let Q be the $6!\times 6!$ matrix of the quadratic form $Q(\cdot,\cdot)$. Then*

$$\min\{x^TQx \;\mid\; x\in\mathbb{R}_+^{6!}, e^Tx=1\} \geq 4.3593.$$

We devote this section to the proof of Proposition 3. It involves computer calculations; more details on this are given in section 3.8.

3.1. The standard quadratic programming problem. The problem we have formulated is known as *standard quadratic optimization problem*. The standard quadratic optimization problem (standard QP) is to find the global minimizers of a quadratic form over the standard simplex; i.e., we consider the global optimization problem

$$\underline{p} := \min_{x\in\Delta} x^TQx, \tag{5}$$

where Q is an arbitrary symmetric $d\times d$ matrix, e is the all ones vector, and Δ is the standard simplex in $\mathbb{R}^d$,

$$\Delta = \{x\in\mathbb{R}_+^d : e^Tx=1\}.$$

We will now reformulate the standard QP as a convex optimization problem in conic form. First, we will review the relevant convex cones as well as the duality theory of conic optimization. We define the following convex cones:

- the $d \times d$ symmetric matrices:

$$\mathcal{S}_d = \{X \in \mathbb{R}^d \times \mathbb{R}^d,\ X = X^T\};$$

- the $d \times d$ symmetric positive semidefinite matrices:

$$\mathcal{S}_d^+ = \{X \in \mathcal{S}_d,\ y^T X y \geq 0\ \ \forall y \in \mathbb{R}^d\};$$

- the $d \times d$ symmetric copositive matrices:

$$\mathcal{C}_d = \{X \in \mathcal{S}_d,\ y^T X y \geq 0\ \ \forall y \in \mathbb{R}^d,\ y \geq 0\};$$

- the $d \times d$ symmetric completely positive matrices:

$$\mathcal{C}_d^* = \left\{ X = \sum_{i=1}^{k} y_i y_i^T,\ y_i \in \mathbb{R}^d,\ y_i \geq 0\ (i = 1, \ldots, k) \right\};$$

- the $d \times d$ symmetric nonnegative matrices:

$$\mathcal{N}_d = \{X \in \mathcal{S}_d,\ X_{ij} \geq 0\ \ (i, j = 1, \ldots, d)\}.$$

Recall that the completely positive cone is the dual of the copositive cone [12], and that the nonnegative and semidefinite cones are self-dual for the inner product $\langle X, Y \rangle := \mathrm{Tr}(XY)$, where "Tr" denotes the trace operator.

For a given cone $\mathcal{K}_d$ and its dual cone $\mathcal{K}_d^*$ we define the primal and dual pair of conic linear programs:

$$\text{(P)} \qquad p^* := \inf_{X \in \mathcal{K}_d} \left\{ \mathrm{Tr}(CX) \,|\, \mathrm{Tr}(A_i X) = b_i\ (i = 1, \ldots, M) \right\},$$

$$\text{(D)} \qquad d^* := \sup_{y \in \mathbb{R}^m} \left\{ b^T y \Big| \sum_{i=1}^{M} y_i A_i + S = C,\ S \in \mathcal{K}_d^* \right\}.$$

If $\mathcal{K}_d = \mathcal{S}_d^+$, we refer to semidefinite programming; if $\mathcal{K}_d = \mathcal{N}_d$, to linear programming; and if $\mathcal{K}_d = \mathcal{C}_d$, to copositive programming.

The well-known conic duality theorem (see, e.g., Renegar [20]) gives the duality relations between (P) and (D).

THEOREM 4 (conic duality theorem). *If there exists an interior feasible solution* $X^0 \in \mathrm{int}(\mathcal{K}_d)$ *of* (P) *and a feasible solution of* (D), *then* $p^* = d^*$ *and the supremum in* (D) *is attained. Similarly, if there exist feasible* y^0, S^0 *for* (D), *where* $S^0 \in int(\mathcal{K}_d^*)$, *and a feasible solution of* (P), *then* $p^* = d^*$ *and the infimum in* (P) *is attained.*

Optimization over the cones $\mathcal{S}_d^+$ and $\mathcal{N}_d$ can be done in polynomial time (to compute an ϵ-optimal solution), but some NP-hard problems can be formulated as copositive programs; see, e.g., de Klerk and Pasechnik [14].

3.1.1. Convex reformulation of the standard QP. We rewrite problem (5) in the following way:

$$\underline{p} := \min_{x \in \Delta} \mathrm{Tr}(Q x x^T).$$

Now we define the cone of matrices

$$\mathcal{K} = \left\{ X \in \mathcal{S}_d \ : \ X = xx^T, \ x \geq 0 \right\}.$$

Note that the requirement $x \in \Delta$ corresponds to $X \in \mathcal{K}$ with $\mathrm{Tr}(ee^T X) = 1$.

We arrive at the following reformulation of problem (5):

$$\underline{p} = \min \left\{ \mathrm{Tr}(QX) \ : \ \mathrm{Tr}(ee^T X) = 1, \ X \in \mathcal{K} \right\}. \tag{6}$$

The last step is to replace the cone $\mathcal{K}$ by its convex hull, which is simply the cone of completely positive matrices, i.e.,

$$\mathrm{conv}\,(\mathcal{K}) = \mathcal{C}_d^* = \left\{ X = \sum_{i=1}^{k} y_i y_i^T, \ y_i \in \mathbb{R}^n, \ y_i \geq 0 \ (i = 1, \ldots, k) \right\}.$$

Replacing the feasible set by its convex hull does not change the optimal value of problem (6), since its objective function is linear. Thus we obtain the well-known convex reformulation

$$\underline{p} = \min \left\{ \mathrm{Tr}(QX) \mid \mathrm{Tr}\big(ee^T X\big) = 1, \ X \in \mathcal{C}_d^* \right\}. \tag{7}$$

The dual problem takes the form

$$\underline{p} = \max \left\{ t \mid Q - tee^T \in \mathcal{C}_d \right\}, \tag{8}$$

where $\mathcal{C}_d$ is the cone of copositive matrices, as before. Note that both problems have the same optimal value, in view of the conic duality theorem.

3.2. Exploiting group symmetries. We can reduce considerably the number of variables in the optimization problems in (7), (8) by exploiting the invariance properties of the quadratic function $x^T Q x$. This will also prove to be computationally necessary for the problems we intend to solve.

Consider the situation where the matrix Q is invariant under the action of a group G of order $k = |G|$ of permutation matrices $P \in G$, in the sense that

$$Q = P^T Q P \quad \forall \ P \in G.$$

Then we have

$$\begin{aligned}
\underline{p} &= \min \left\{ \mathrm{Tr}(QX) \mid \mathrm{Tr}\big(ee^T X\big) = 1, \ X \in \mathcal{C}_d^* \right\} \\
&= \min \left\{ \mathrm{Tr}\big(P^T Q P X\big) \mid \mathrm{Tr}\big(Pee^T P X\big) = 1, \ X \in \mathcal{C}_d^* \right\} \ \text{for any } P \in G \\
&= \min \left\{ \mathrm{Tr}\big(QP^T X P\big) \mid \mathrm{Tr}\big(ee^T P^T X P\big) = 1, \ X \in \mathcal{C}_d^* \right\} \ \text{for any } P \in G \\
&= \min \left\{ \mathrm{Tr}\left(Q \frac{1}{k} \left[\sum_{P \in G} P^T X P \right] \right) \Bigg| \mathrm{Tr}\left(ee^T \left[\frac{1}{k} \sum_{P \in G} P^T X P \right] \right) = 1, \ X \in \mathcal{C}_d^* \right\}.
\end{aligned}$$

We can therefore restrict the optimization to the subset of the feasible set obtained by replacing each feasible X by the *group average* $\frac{1}{k} \sum_{P \in G} P^T X P$, i.e., replacing X by its image under what is known in invariant theory as the Reynolds operator. Note that if $X \in \mathcal{C}_d^*$, then so is its image under the group average.

In particular, we wish to compute a basis for the so-called *fixed point subspace*

$$\mathcal{A} := \left\{ Y \in \mathcal{S}_d \mid Y = \frac{1}{k} \sum_{P \in G} P^T X P,\ X \in \mathcal{S}_d \right\}.$$

Note that Q and ee^T are elements of $\mathcal{A}$ (set $X = Q$, respectively, $X = ee^T$). Hence $Q - tee^T \in \mathcal{A}$ for any t, and

$$\underline{p} = \max\left\{ t \mid Q - tee^T \in \mathcal{C}_d \right\} = \max\left\{ t \mid Q - tee^T \in \mathcal{C}_d \cap \mathcal{A} \right\}.$$

The right-hand side here is the dual of the primal problem when it is restricted to $\mathcal{A}$ as above.

The next step is to compute a basis for the subspace $\mathcal{A}$.

3.3. Computing a basis for the fixed point subspace. We assume for simplicity that G acts transitively as a permutation group on the standard basis vectors. (This holds in our setting. A more general, and computationally less efficient, setting can be found in Gatermann and Parrilo [8].) The theory here is well known and goes back to Burnside, Schur, and Wielandt. See, e.g., Cameron [5] for details. Although we need a basis of $\mathcal{A}$, the subspace of *symmetric* matrices fixed by G, it is more natural to compute the basis $\mathcal{X}$ of the subspace $\mathcal{B}$ of *all* matrices fixed by G and then pass on to $\mathcal{A}$.

The dimension of $\mathcal{B}$ equals the number r of orbits of G on the Cartesian square of the standard basis. The set of the latter orbits, also known as 2-orbits, naturally corresponds to certain set $\mathcal{X}$ of $d \times d$ zero-one matrices. Namely, for each $X \in \mathcal{X}$ one has $X_{ij} = 1$ if and only if $X_{P(i),P(j)} = 1$ for all $P \in G$ and all $1 \le i \le j \le |\Pi|$. As G is transitive on the standard basis vectors, the identity matrix I belongs to $\mathcal{X}$. We also have $\sum_{X \in \mathcal{X}} X = ee^T$.

As $\mathcal{X}$ is closed under the matrix transposition, i.e., $X^T \in \mathcal{X}$ for any $X \in \mathcal{X}$,

$$\mathcal{X}_{\mathcal{A}} = \{A_1, \ldots, A_M\} = \{X \mid X = X^T \in \mathcal{X}\} \cup \{X + X^T \mid X \in \mathcal{X},\ X \neq X^T\}$$

is a basis of $\mathcal{A}$. Each $A \in \mathcal{X}_{\mathcal{A}}$ is a symmetric zero-one matrix, and $\sum_{A \in \mathcal{X}_{\mathcal{A}}} A = ee^T$. Moreover,

$$\left\{ Y \in \mathcal{S}_d \mid Y = \sum_{i=1}^{M} y_i A_i \right\} = \mathcal{A} \equiv \left\{ Y \in \mathcal{S}_d \mid Y = \frac{1}{k} \sum_{P \in G} P^T X P,\ X \in \mathcal{S}_d \right\}.$$

Since $Q \in \mathcal{A}$, we will write $Q = \sum_{i=1}^{M} b_i A_i$.

It is worth mentioning that algebraically the vector space $\mathcal{B}$ behaves very nicely: it is closed under multiplication. In other words, $\mathcal{B}$ is a matrix algebra of dimension r, also known as the *centralizer ring* of the permutation group G.

We proceed to describe G and $\mathcal{B}$ in our case. For us G is isomorphic to the direct product $\mathrm{Sym}(m) \times \mathrm{Sym}(2)$ of symmetric groups $\mathrm{Sym}(m)$ and $\mathrm{Sym}(2)$, where $\mathrm{Sym}(m)$ acts (as a permutation group) by conjugation on the $d = (m-1)!$ elements of Π, and $\mathrm{Sym}(2)$ acts (as a permutation group) on Π by switching $\pi \in \Pi$ with $\pi^{-1} \in \Pi$.

Computing $\mathcal{X}$ is an elementary combinatorial procedure, which can be found in one form or another in many computer algebra systems, so one does not have to program this again. First, the permutations that generate $\mathrm{Sym}(m) \times \mathrm{Sym}(2)$ in its action on Π are computed. The action of $\mathrm{Sym}(2)$ is already known, and is described

by the permutation g_0, say. In its usual action on m symbols, $\mathrm{Sym}(m)$ is generated by $h_1 = (0, 1, \ldots, m-1)$ and $h_2 = (0, 1)$. These h_i (for $i = 1, 2$) act on Π by mapping each $\pi \in \Pi$ to $h_i \pi h_i^{-1}$. Denote by g_i (for $i = 1, 2$) the permutations of Π that realize these actions.

Next, one computes the orbits of the permutation group $\mathrm{Sym}(m) \times \mathrm{Sym}(2) = \langle g_0, g_1, g_2 \rangle$ on the Cartesian square $\Pi \times \Pi$ of Π, by "spinning" $(\pi_i, \pi_j) \in \Pi \times \Pi$: Begin with $S_{ij} = \{(\pi_i, \pi_j)\}$ and apply the generators g_i, $0 \le i \le 2$, in a loop until S_{ij} stops growing. Then one sets $\Pi := \Pi - S_{ij}$ and repeats until Π is exhausted.

When $m = 7$, one has $r = 78$ and $M = 56$. Note that here the algebra $\mathcal{B}$ is not commutative.

When $m = 5$, one has $r = M = 6$, and $\mathcal{B}$ is commutative.

3.4. Reformulation of the optimization problem. We can now reformulate the dual problem by using the basis of $\mathcal{A}$ to obtain

$$\underline{p} = \max\left\{t \mid Q - tee^T \in \mathcal{C}_d \cap \mathcal{A}\right\} = \max\left\{t \,\middle|\, \sum_{i=1}^{M} (b_i - t) A_i \in \mathcal{C}_d\right\}.$$

We will now proceed to derive a lower bound on $\underline{p}$ by solving the dual problem approximately.

3.5. Approximations of the copositive cone. The problem of determining whether a matrix is not copositive is NP-complete, as shown by Murty and Kabadi [17]. We therefore wish to replace the copositive cone $\mathcal{C}_d$ by a conic subset, in such a way that the resulting optimization problem becomes tractable. We can represent the copositivity requirement for a $d \times d$ symmetric matrix S as

$$P(x) := (x \circ x)^T S (x \circ x) = \sum_{i,j=1}^{d} S_{ij} x_i^2 x_j^2 \ge 0 \quad \forall x \in \mathbb{R}^d, \tag{9}$$

where "$\circ$" indicates the componentwise (Hadamard) product. We therefore wish to know whether the polynomial $P(x)$ is nonnegative for all $x \in \mathbb{R}^d$. Although one apparently cannot answer this question in polynomial time in general, as it is an NP-hard problem, one can decide using semidefinite programming whether $P(x)$ can be written as a sum of squares.

Parrilo [19] showed that $P(x)$ in (9) allows a sum of squares decomposition if and only if $S \in \mathcal{S}_d^+ + \mathcal{N}_d$, which is a well-known sufficient condition for copositivity. Set $\mathcal{K}_d^0$ to be the convex cone $\mathcal{K}_d^0 = \mathcal{S}_d^+ + \mathcal{N}_d$.

Higher order sufficient conditions can be derived by considering the polynomial

$$P^{(\ell)}(x) = P(x) \left(\sum_{i=1}^{d} x_i^2\right)^{\ell} = \left(\sum_{i,j=1}^{d} S_{ij} x_i^2 x_j^2\right) \left(\sum_{i=1}^{d} x_i^2\right)^{\ell}, \tag{10}$$

and asking whether $P^{(\ell)}(x)$—which is a homogeneous polynomial of degree $2(\ell+2)$—has a sum of squares decomposition, or whether it has only nonnegative coefficients.

For $\ell = 1$, Parrilo [19] showed that a sum of squares decomposition exists if and

only if[1] the following system of linear matrix inequalities has a solution:

$$S - S^{(i)} \in \mathcal{S}_d^+, \quad i = 1, \ldots, d, \tag{11}$$

$$S_{ii}^{(i)} = 0, \quad i = 1, \ldots, d, \tag{12}$$

$$S_{jj}^{(i)} + 2S_{ij}^{(j)} = 0, \quad i \neq j, \tag{13}$$

$$S_{jk}^{(i)} + S_{ik}^{(j)} + S_{ij}^{(k)} \geq 0, \quad i < j < k, \tag{14}$$

where $S^{(i)}$ $(i = 1, \ldots, d)$ are symmetric matrices. Similar to the $\ell = 0$ case, we define $\mathcal{K}_d^1$ as the (convex) cone of matrices S for which the above system has a solution.

We will consider the lower bounds we get by replacing the copositive cone by either $\mathcal{K}_d^0$ or $\mathcal{K}_d^1$:

$$\underline{p} \geq p_\ell := \max\left\{t \mid Q - tee^T \in \mathcal{K}_d^\ell\right\}, \qquad \ell \in \{0, 1\}. \tag{15}$$

3.6. Approximations (relaxations) of the copositive cone. We will now study the relaxation obtained by replacing the copositive cone by its proper subset $\mathcal{K}_d^0$. In other words, we study the relaxation

$$\begin{aligned}\underline{p} &= \max\left\{t \,\middle|\, \sum_{i=1}^M (b_i - t)A_i \in \mathcal{C}_d\right\} \\ &\geq p_0 := \max\left\{t \,\middle|\, \sum_{i=1}^M (b_i - t)A_i \in \mathcal{K}_d^0 = \mathcal{S}_d^+ + \mathcal{N}_d\right\}.\end{aligned}$$

We rewrite $\sum_{i=1}^M (b_i - t)A_i \in \mathcal{K}_d^0$ as

$$\sum_{i=1}^M (b_i - t)A_i = \sum_{i=1}^M y_i A_i + \sum_{i=1}^M z_i A_i,$$

where $\sum_{i=1}^M y_i A_i \in \mathcal{S}_d^+$ and $\sum_{i=1}^M z_i A_i \in \mathcal{N}_d$.

Note that, since the A_i's are zero-one matrices that sum to ee^T, it follows that $z_i \geq 0$. Moreover,

$$b_i - t = y_i + z_i \qquad \text{implies} \qquad b_i - t - y_i \geq 0.$$

We obtain the relaxation

$$p_0 = \max\left\{t \mid b_i - t - y_i \geq 0 \ (i = 1, \ldots, M),\ \sum_{i=1}^M y_i A_i \in \mathcal{S}_d^+\right\}. \tag{16}$$

3.7. Block factorization. The next step in reducing the problem size is to perform a similarity transformation that simultaneously block-diagonalizes the matrices $A_1, \ldots, A_M$. In particular, we want to find an orthogonal matrix V such that the matrices

$$\tilde{A}_i := VA_iV^{-1}, \quad i = 1, \ldots, M,$$

[1] In fact, Parrilo [19] proved only the "if" part; the converse is proved in Bomze and de Klerk [4].

all have the same block-diagonal structure, and the maximum block size is as small as possible. Note that the conjugation preserves spectra, and orthogonality of V preserves symmetry.

This will further reduce the size of the relaxation (16) via

$$\begin{aligned} p_0 &= \max\left\{ t \mid b_i - t - y_i \geq 0 \ (i = 1, \ldots, M),\ \sum_{i=1}^{M} y_i A_i \in \mathcal{S}_d^+ \right\} \\ &= \max\left\{ t \mid b_i - t - y_i \geq 0 \ (i = 1, \ldots, M),\ \sum_{i=1}^{M} y_i V A_i V^{-1} \in \mathcal{S}_d^+ \right\} \\ &= \max\left\{ t \mid b_i - t - y_i \geq 0 \ (i = 1, \ldots, M),\ \sum_{i=1}^{M} y_i \tilde{A}_i \in \mathcal{S}_d^+ \right\}. \end{aligned}$$

The necessity to restrict to orthogonal V's lies in the fact that there is currently no software (or algorithms) available that would be able to deal with nonsymmetric $\tilde{A}_i$'s.

Computing the finest possible block decomposition (this would mean finding explicitly the orthogonal bases for the irreducible submodules of the natural module of G in its action by the matrices P) is computationally not easy, especially due to the orthogonality requirement on V. We restricted ourselves to decomposing into two blocks of equal size $\frac{d}{2} \times \frac{d}{2}$. Namely, each row corresponds to a cyclic permutation $g \in \Pi$, and the natural pairing (g, g^{-1}) can be used to construct $V = \frac{\sqrt{2}}{2} V'$ as follows:

- the first half of the rows of V' are characteristic vectors of the 2-subsets $\{g, g^{-1}\}$, $g \in \Pi$;
- the second half of the rows of V' consists of "twisted" rows from the first half: namely, one of the two 1's is replaced by -1.

It is obvious that $V' V'^T = 2I$ and thus V is orthogonal.

Remark. It is worth mentioning that in [22] Schrijver essentially dealt, in a different context, with a similar setup, except that in his case the elements of the basis $\mathcal{X}$ of $\mathcal{B}$ were symmetric and (hence) the algebra $\mathcal{B}$ commutative. In such a situation the elements of $\mathcal{X}$ can be simultaneously diagonalized, and the corresponding optimization problem becomes a linear programming problem.

3.8. Computational results: Proof of Theorem 3. The combinatorial/group theoretic part of the computations, namely of the A_i's, V, and $Q = \sum_i b_i A_i$, was performed using a computer algebra system GAP [7], version 4.3, and its shared package GRAPE by Soicher [24]. Semidefinite programs (SDPs) were solved by Sturm [25] using SeDuMi, version 1.05 under MATLAB 6.5. The biggest SDP took about 10 minutes of CPU time of a Pentium 4 with 1 GB of RAM.

In addition, the results were verified using MAPLE. Namely, for $t = p_0$ and y, the variables computed upon solving (16), we checked that the corresponding (matrix and scalar) inequalities in (16) hold. As p_0 is a lower bound on $\underline{p}$, we thus validated the computed value of p_0 independently of the SDP solver used.

For the test case of $K_{5,n}$ we solved the relaxed problem (15) with $\ell = 1$ to obtain

$$p_1 \approx 1.9544, \quad \text{that is,} \quad \operatorname{cr}(K_{5,n}) \geq \frac{1}{2}(1.9544) n^2 = 0.9772 n^2,$$

asymptotically. The correct asymptotic value is known to be $\operatorname{cr}(K_{5,n}) = n^2$, which shows the quality of the bound. In fact, we could show that $p_1 \approx 1.9544$ corresponds

to the optimal value of the first optimization problem in Lemma 2 for $m = 5$. This shows that the optimal value of this optimization problem is a strict lower bound of the crossing number of $K_{m,n}$, even for $m = 5$.

The weaker bound for $\ell = 0$ in (15) yields, still quite tight,

$$p_0 \approx 1.94721, \quad \text{that is,} \quad \operatorname{cr}(K_{5,n}) \geq \frac{1}{2}(1.94721)n^2 = 0.973605n^2.$$

For the case $K_{7,n}$ we solved the relaxed problem (15) with $\ell = 0$ to obtain

$$p_0 \approx 4.3593, \quad \text{that is,} \quad \operatorname{cr}(K_{7,n}) \geq \frac{1}{2}(4.3593)n^2 = 2.1796n^2,$$

asymptotically.

Proof of Theorem 1. For the sake of completeness, we close this section with the observation that Theorem 1 has been proved. It follows from Lemma 2 and Proposition 3. □

4. Improved bounds for the crossing numbers of $K_{m,n}$ and K_n. Perhaps the most appealing consequence of our improved bound for $\operatorname{cr}(K_{7,n})$ is that it also allows us to give improved lower bounds for the crossing numbers of $K_{m,n}$ and K_n. The quality of the new bounds is perhaps best appreciated in terms of the following asymptotic parameters:

$$A(m) := \lim_{n\to\infty} \frac{\operatorname{cr}(K_{m,n})}{Z(m,n)}, \qquad B := \lim_{n\to\infty} \frac{\operatorname{cr}(K_{n,n})}{Z(n,n)},$$

(see Richter and Thomassen [21]). These natural parameters give us a good idea of our current standing with respect to Zarankiewicz's conjecture. It is not difficult to show that $A(m)$ (for every integer $m \geq 3$) and B both exist [21].

Previous to the new bound we report in Theorem 1, the best known lower bounds for $A(m)$ and B were $A(m) \geq 0.8\frac{m}{m-1}$ and (consequently) $B \geq 0.8$. Both bounds were obtained by using the known value of $\operatorname{cr}(K_{5,n})$ and applying a standard counting argument.

By applying the same counting argument but instead using the bound given by Theorem 1, we improve these asymptotic quotients to $A(m) > 0.83\frac{m}{m-1}$ and $B > 0.83$.

The improved lower bound for B has an additional, important application. It has been long conjectured that $\operatorname{cr}(K_n) = Z(n)$, where

$$Z(n) = \frac{1}{4}\left\lfloor \frac{n}{2} \right\rfloor \left\lfloor \frac{n-1}{2} \right\rfloor \left\lfloor \frac{n-2}{2} \right\rfloor \left\lfloor \frac{n-3}{2} \right\rfloor,$$

but this has been verified only for $n \leq 10$ (see, for instance, [6]). As we did with $K_{m,n}$, it is natural to inquire about the asymptotic parameter

$$C := \lim_{n\to\infty} \frac{\operatorname{cr}(K_n)}{Z(n)}.$$

In [21] it is proved that C exists, and, moreover, that $C \geq B$. In view of this, our improved lower bound for B yields $C > 0.83$.

We summarize these results in the following statement.

THEOREM 5. *With $Z(m,n)$ and $Z(n)$ as above,*

$$\lim_{n\to\infty}\frac{\operatorname{cr}(K_{m,n})}{Z(m,n)} \geq 0.83\frac{m}{m-1}, \quad \lim_{n\to\infty}\frac{\operatorname{cr}(K_{n,n})}{Z(n,n)} \geq 0.83,$$
$$\text{and} \quad \lim_{n\to\infty}\frac{\operatorname{cr}(K_n)}{Z(n)} \geq 0.83. \qquad \square$$

Recall that these results followed from an improved lower bound on $\operatorname{cr}(K_{7,n})$ obtained by solving the optimization problem (15) for $m = 7$. The results can be further improved by solving (15) for larger values of m. After the first submission of the present work, the optimization problem was successfully solved for $m = 9$ by de Klerk, Pasechnik, and Schrijver [15], by using a more sophisticated way of exploiting the algebraic symmetry. In particular, the constant 0.83 in Theorem 5 could thus be improved to 0.859.

We close this section with a few words on some important recent developments involving the *rectilinear* crossing number of K_n.

The *rectilinear crossing number* $\overline{\operatorname{cr}}(G)$ of a graph G is the minimum number of pairwise intersections of edges in a drawing of G in the plane, with the additional restriction that all edges of G must be drawn as straight segments.

It is known that $\overline{\operatorname{cr}}(K_n)$ and $\operatorname{cr}(K_n)$ may be different (for instance, $\overline{\operatorname{cr}}(K_8) = 19$, whereas $\operatorname{cr}(K_8) = 18$; see [10]). While we have a (nonrectilinear) way of drawing K_n that shows $\operatorname{cr}(K_n) \leq Z(n)$ (equality is conjectured to hold, as we observed above), good upper bounds for $\overline{\operatorname{cr}}(K_n)$ are notoriously difficult to obtain. Currently, the best upper bound known is $\overline{\operatorname{cr}}(K_n) \leq 0.3807\binom{n}{4}$ (see Aichholzer, Aurenhammer, and Krasser [2]).

For many years the best lower bounds known for $\overline{\operatorname{cr}}(K_n)$ were considerably smaller (around $0.32\binom{n}{4}$) than the best upper bounds available (currently around $0.3807\binom{n}{4}$). However, remarkably better lower bounds have been recently proved independently by Ábrego and Fernández–Merchant [1] and Lovász et al. [16], and refined by Balogh and Salazar [3]. In [1], the technique of allowable sequences was used to show that $\overline{\operatorname{cr}}(K_n) \geq 0.375\binom{n}{4}$. Lovász et al. used similar methods to prove $\overline{\operatorname{cr}}(K_n) > 0.37501\binom{n}{4} + O(n^3)$. Recently, Balogh and Salazar improved this to $\overline{\operatorname{cr}}(K_n) > 0.37553\binom{n}{4} + O(n^3)$ [3]. The importance of establishing that $\overline{\operatorname{cr}}(K_n)$ is strictly greater than $0.375\binom{n}{4} + O(n^3)$ is that it effectively shows that the ordinary and the rectilinear crossing numbers of K_n are different in the asymptotically relevant term, namely n^4.

Acknowledgment. Etienne de Klerk would like to thank Pablo Parrilo for his valuable comments.

REFERENCES

[1] B. M. ÁBREGO AND S. FERNÁNDEZ-MERCHANT, *A lower bound for the rectilinear crossing number*, Graphs Combin., 21 (2005), pp. 293–300.

[2] O. AICHHOLZER, F. AURENHAMMER, AND H. KRASSER, *On the crossing number of complete graphs*, Computing, 76 (2006), pp. 165–176.

[3] J. BALOGH AND G. SALAZAR, *On k-sets, convex quadrilaterals, and the rectilinear crossing number of K_n*, Discrete Comput. Geom., to appear.

[4] I. M. BOMZE AND E. DE KLERK, *Solving standard quadratic optimization problems via linear, semidefinite and copositive programming*, J. Global Optim., 24 (2002), pp. 163–185.

[5] P. J. CAMERON, *Permutation Groups*, Cambridge University Press, Cambridge, UK, 1999.

[6] P. ERDŐS AND R. K. GUY, *Crossing number problems*, Amer. Math. Monthly, 80 (1973), pp. 52–58.

[7] THE GAP GROUP, *GAP—Groups, Algorithms, and Programming, Version* 4.3, http://www.gap-system.org (2002).

[8] K. Gatermann and P. A. Parrilo, *Symmetry groups, semidefinite programs, and sums of squares*, J. Pure Appl. Algebra, 192 (2004), pp. 95–128.
[9] L. Glebsky and G. Salazar, *The crossing number of $C_m \times C_n$ is as conjectured for $n \geq m(m+1)$*, J. Graph Theory, 47 (2005), pp. 53–72.
[10] R. K. Guy, *Latest results on crossing numbers*, in Recent Trends in Graph Theory, Springer, New York, 1971, pp. 143–146.
[11] R. K. Guy, *The decline and fall of Zarankiewicz's theorem*, in Proof Techniques in Graph Theory (Ann Arbor, MI, 1968), Academic Press, New York, 1969, pp. 63–69.
[12] M. Hall, Jr., and M. Newman, *Copositive and completely positive quadratic forms*, Proc. Cambridge Philos. Soc., 59 (1963), pp. 329–339.
[13] D. J. Kleitman, *The crossing number of $K_{5,n}$*, J. Combin. Theory, 9 (1970), pp. 315–323.
[14] E. de Klerk and D. V. Pasechnik, *Approximation of the stability number of a graph via copositive programming*, SIAM J. Optim., 12 (2002), pp. 875–892.
[15] E. de Klerk, D. V. Pasechnik, and A. Schrijver, *Reduction of symmetric semidefinite programs using the regular *-representation*, Math. Program., to appear.
[16] L. Lovász, K. Vesztergombi, U. Wagner, and E. Welzl, *Convex quadrilaterals and k–sets*, in Towards a Theory of Geometric Graphs, Contemp. Math. 342, János Pach, ed., American Mathematical Society, Providence, RI, 2004, pp. 139–148.
[17] K. G. Murty and S. N. Kabadi, *Some NP-complete problems in quadratic and linear programming*, Math. Programming, 39 (1987), pp. 117–129.
[18] N. Nahas, *On the crossing number of $K_{m,n}$*, Electron. J. Combin., 10 (2003), Note 8.
[19] P. A. Parrilo, *Structured Semidefinite Programs and Semi-Algebraic Geometry Methods in Robustness and Optimization*, Ph.D thesis, California Institute of Technology, Pasadena, CA, 2000.
[20] J. Renegar, *A Mathematical View of Interior-Point Methods in Convex Optimization*, MPS/SIAM Ser. Optim. 3, SIAM, Philadelphia, 2001.
[21] R. B. Richter and C. Thomassen, *Relations between crossing numbers of complete and complete bipartite graphs*, Amer. Math. Monthly, 104 (1997), pp. 131–137.
[22] A. Schrijver, *A comparison of the Delsarte and Lovász bounds*, IEEE Trans. Inform. Theory, 25 (1979), pp. 425–429.
[23] F. Shahrokhi, O. Sýkora, L. A. Székely, and I. Vrťo, *Crossing numbers: Bounds and applications*, in Intuitive Geometry, Bolyai Soc. Math. Stud. 6, János Bolyai Math. Soc., Budapest, 1997, pp. 179–206.
[24] L. H. Soicher, *GRAPE: A system for computing with graphs and groups*, in Groups and Computation, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 11, L. Finkelstein and W. M. Kantor, eds., 1991, pp. 287–291; also available online from http://www.gap-system.org/Packages/grape.html.
[25] J. F. Sturm, *Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11/12 (1999), pp. 625–653; also available online from http://sedumi.mcmaster.ca.
[26] L. A. Székely, *A successful concept for measuring non-planarity of graphs: The crossing number*, Discrete Math., 276 (2004), pp. 331–352.
[27] P. Turán, *A note of welcome*, J. Graph Theory, 1 (1977), pp. 7–9.
[28] D. R. Woodall, *Cyclic-order graphs and Zarankiewicz's crossing-number conjecture*, J. Graph Theory, 17 (1993), pp. 657–671.
[29] K. Zarankiewicz, *On a problem of P. Turán concerning graphs*, Fund. Math., 41 (1954), pp. 137–145.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 203–212

PAIRWISE COLLIDING PERMUTATIONS AND THE CAPACITY OF INFINITE GRAPHS*

JÁNOS KÖRNER† AND CLAUDIA MALVENUTO†

Abstract. We call two permutations of the first n naturals colliding if they map at least one number to consecutive naturals. We give bounds for the exponential asymptotics of the largest cardinality of any set of pairwise colliding permutations of $[n]$. We relate this problem to the determination of the Shannon capacity of an infinite graph and initiate the study of analogous problems for infinite graphs with finite chromatic number.

Key words. extremal combinatorics, Shannon capacity of graphs, permutations, infinite graphs

AMS subject classifications. 05D05, 05C69, 05A15, 94A24

DOI. 10.1137/050632877

1. Introduction.

Let n be an arbitrary natural number and let $[n]$ be the set of all natural numbers from 1 to n. We will say that two permutations of $[n]$ are *colliding* if they map at least one element of $[n]$ into two consecutive numbers, i.e., into numbers differing by 1. It is then natural to ask for the determination of the maximum cardinality $\rho(n)$ of a set of pairwise colliding permutations of $[n]$. One easily sees that this number grows exponentially with n and its asymptotic exponent lies between $\log_2 \frac{1+\sqrt{5}}{2}$ and 1. We will prove this and some better bounds later on.

Certain graphs having as vertex set the permutations of $[n]$ have been introduced before by Cameron and Ku [1] and Larose and Malvenuto [10], cf. also Ku and Leader [9] for a generalization. These authors considered Kneser-type graphs in which they studied the growth of stable sets describing sets of permutations that are "similar" in some sense, whereas our definition of adjacency corresponds to being "different" and distinguishable in some other, particular sense. In fact, the above Kneser-type problems, unlike ours, have no immediate relation to capacity in the Shannon sense.

In this paper we will generalize our introductory problem in several ways. We will consider arbitrary infinite graphs over the natural numbers and introduce various new concepts of capacity. As always, graph capacity measures the exponential growth rate of the largest cliques induced on the Cartesian powers of the vertex set of a graph. In case of an infinite vertex set such as the naturals this is not always interesting, for the graph in itself might have infinite cliques. Then it is reasonable to restrict our attention to particular subsets of the power sets, e.g., those representing permutations. We will present some simple bounds for the value of the so obtained new capacities.

2. Permutation capacity.

Let G be an arbitrary graph with a countable set of vertices. Without loss of generality we can suppose that the vertex set $V(G)$ of G is the set $\mathbb{N}$ of natural numbers. Further, let us denote by $G[A]$ the subgraph of G induced by an arbitrary subset A of the vertex set of G. As usual, we also consider, for every natural $n \in \mathbb{N}$, the power graph G^n whose vertex set is $\mathbb{N}^n$, the set of n–length sequences of natural numbers. Two such sequences $\mathbf{x} \in \mathbb{N}^n$ and $\mathbf{y} \in \mathbb{N}^n$ are adjacent

*Received by the editors June 1, 2005; accepted for publication (in revised form) September 27, 2005; published electronically March 3, 2006.
http://www.siam.org/journals/sidma/20-1/63287.html
†Dipartimento di Informatica, Università di Roma "La Sapienza", Via Salaria 113 - 00198 Rome, Italy (korner@di.uniroma1.it, claudia@di.uniroma1.it).

in G^n if $\mathbf{x} = x_1x_2\dots x_n$ and $\mathbf{y} = y_1y_2\dots y_n$ have at least one coordinate $i \in [n]$ for which $\{x_i, y_i\} \in E(G)$, i.e., if the vertices x_i and y_i are adjacent in G. (This concept of power graph is rooted in information theory. If we interpret adjacency of vertices of a graph as a relation of distinguishability, it is very intuitive to extend such a notion to strings of vertices in the above way, with the meaning that two strings are distinguishable if we can distinguish them in at least one of their coordinates.)

Throughout this paper we write $\binom{X}{n}$ for the family of all n–element subsets of X. For an arbitrary set $A \in \binom{\mathbb{N}}{n}$ we write $R(A)$ for the set of all the n-length sequences without repetitions on the alphabet A. As usual, we can think of a sequence in $R(A)$ as a permutation of the set A. In particular, when $A = [n]$, the sequence $\mathbf{x} = x_1 \dots x_n \in R([n])$ represents the permutation of $[n]$ which maps i into x_i.

We denote by $G(A)$ the subgraph of the power G^n induced by $R(A)$ and by $\rho(G, A)$ its clique number. We set $\rho(G, n)$ for the largest cardinality of a clique induced by G^n on the sequences corresponding to the permutations of an n-set in $V(G)$, i.e.,

$$\rho(G, n) = \max_{A \in \binom{\mathbb{N}}{n}} \rho(G, A).$$

Finally, we define

$$\rho(G) = \limsup_{n\to\infty} \frac{1}{n} \log_2 \rho(G, n)$$

and call it the *permutation capacity* of the graph G.

In this paper we consider some infinite graphs and try to determine their permutation capacity. Since our graphs have a countable set of vertices, the value of their permutation capacity might well be infinite. The same is true for Shannon capacity. In fact, to our knowledge Shannon capacity of infinite graphs has not been considered so far, even though it makes perfect sense and will be discussed below.

The problem of the asymptotic growth of cliques of particular induced subgraphs of G^n as n goes to infinity is a key ingredient in determining the Shannon capacity of graph families in the sense of Cohen, Körner and Simonyi [2], where the sets inducing the subgraphs are formed by all the sequences "of a given type," in an information theoretic sense (see Csiszár and Körner [3] for a definition and more on this). All the sequences of a given "type" form a minimal set that is invariant under the action of all the permutations of the coordinates of the sequences. Our present concepts are natural extensions to the case of infinite graphs of Shannon capacity in a given type, in the sense of [3].

3. Examples. Let us start with an atypical and even somewhat trivial example, just to rephrase the already cited results of [1] and [10] in our present terms.

Consider the graph G, where $V(G) = \mathbb{N}$ and $E(G) = \{\{x, x\} : x \in \mathbb{N}\}$, consisting of loops on the natural numbers. When $A = [n]$, its set $R(A)$ is the set of permutations of $[n]$ and two permutations $\mathbf{x} = x_1 \dots x_n$ and $\mathbf{y} = y_1 \dots y_n$ are adjacent if and only if there is a coordinate $i \in [n]$ such that $x_i = y_i$. We will denote by $\mathbf{x}^{-1}$ the inverse of the permutation represented by the sequence $\mathbf{x}$. With this notation, $\mathbf{x}$ and $\mathbf{y}$ are adjacent if the product $\mathbf{x}\mathbf{y}^{-1}$ is not a derangement. This is the complement of the graph of permutations studied by Cameron–Ku [1] and Larose–Malvenuto [10], that is, the Cayley graph of permutations with generators the derangements. It is obvious that $\rho(G, n) = \rho(G, [n]) = (n-1)!$ and thus the clique number $(n-1)!$ is super-exponential in n. In fact, the above authors show far more than this; they prove that the trivial construction, consisting of the set of all permutations that map an arbitrary fixed

natural l into an arbitrary fixed natural m, is the unique way to achieve the clique number. This graph is somewhat artificial in the present context. If in a graph the only edges are loops, then adjacency corresponds to "being similar." In graph capacity problems one usually considers only graphs without loops and interprets adjacency as some sort of distinguishability between vertices. From now on we will restrict attention to these cases.

One of the simplest and perhaps most natural examples of our present problem is furnished by the *(semi-)infinite path* L whose vertices x and y from $\mathbb{N}$ are adjacent if they are consecutive in the natural order, that is $|y-x| = 1$. Clearly, $\omega(L) = \chi(L) = 2$ and thus the Shannon capacity $\log \lim_{n\to\infty} \sqrt[n]{\omega(L^n)}$ equals 1 (cf. Shannon [13], Lovász [11] and, in particular, Cohen, Körner and Simonyi [2], where the problem is reformulated, geared towards the subsequent generalizations [6] and[7], in the present terms). We will show that

$$\log_2 \frac{1+\sqrt{5}}{2} \leq \rho(L) \leq 1.$$

For the infinite path L, denote simply by $L(n)$ rather than $L([n])$ the subgraph induced by the set $A = [n]$ on the nth power of L. Its vertex set is the set of all the permutations of the set $[n]$ (the permutations of n elements) and two of such permutations $\mathbf{x} = x_1 \dots x_n$ and $\mathbf{y} = y_1 \dots y_n$ are adjacent in $L(n)$ if and only if the following condition holds:

$$\exists i \in [n] : |y_i - x_i| = 1. \tag{1}$$

Note that, as observed in [5], every finite graph is an induced subgraph of L^n for some value of n.

The two graphs above belong to a more general class of graphs $G(\mathcal{D})$ depending on a finite subset $\mathcal{D}$ of $\mathbb{N}$ of "allowed differences" as follows: its vertices are, as before, the natural numbers $\mathbb{N}$ and $\{x, y\} \in G(\mathcal{D})$ if and only if $|x - y| \in \mathcal{D}$. When $\mathcal{D} = \{0\}$ we have the all-loops graph described above; when $\mathcal{D} = \{1\}$ we have $G(\mathcal{D}) = L$.

4. The infinite path. In this section we will study the behavior of the cliques in the powers of the (semi-)infinite path L. In particular, we will derive some recursive inequalities for the value of $\rho(L, n)$.

Observation. For any n-element subset A of the naturals the graph induced on it by L is isomorphic to a subgraph of the path of n vertices induced by L on the set $[n]$. Hence by an obvious monotonicity

$$\rho(L, n) = \max_{A \in \binom{\mathbb{N}}{n}} \rho(L, A) = \rho(L, [n]).$$

In other words, $\rho(L, n)$ is the maximum number of permutations of $[n]$ such that for any two of them, there is an element of $[n]$ mapped into two consecutive integers from $[n]$. Recall that this is the very same problem we introduced at the beginning of this paper, where we wrote $\rho(n)$ for $\rho(L, n)$.

The following recursive inequality will play a key role in our attempt to determine the permutation capacity of the infinite path.

PROPOSITION 4.1. *The function $\rho(L, n)$ is super-multiplicative:*

$$\rho(L, n + m) \geq \rho(L, n) \cdot \rho(L, m).$$

Proof. Take a clique C in $L(n)$ of maximal size $\rho(L,n)$ and a clique D of maximal size $\rho(L,m)$ in $L(m)$. Denote by $D+n$ the set obtained from D by adding n to each element of the sequences of D:

$$D+n=\{x_1+n\dots x_m+n : x_1\dots x_m\in D\}\subseteq R(\{n+1,\dots,n+m\}).$$

Clearly the size of the clique $D+n$ in $G(\{n+1,\dots,n+m\})$ is the same as that of D. Hence the product construction

$$C\times(D+n)=\{x_1\dots x_{n+m} : x_1\dots x_n\in C; x_{n+1}\dots x_{n+m}\in D+n\}\subseteq R([n+m]),$$

obtained by concatenating sequences from C to sequences from $D+n$, gives a clique in $G(n+m)$ of size $\rho(L,n)\cdot\rho(L,m)$. □

By the well-known elementary inequality called Fekete's lemma (see [14]), the last proposition implies that the limit $\lim_{n\to\infty}\sqrt[n]{\rho(L,n)}$ exists, and its logarithm coincides with the permutation capacity $\rho(L)$.

It is immediately obvious that the capacity $\rho(L)$ is upper bounded by the logarithm of the chromatic number of L, and thus is at most 1. The following non-asymptotic refinement might be interesting.

PROPOSITION 4.2.

$$\rho(L,n)\le\binom{n}{\lfloor\frac{n}{2}\rfloor}.$$

Proof. Call *parity pattern* of a permutation $\mathbf{x}=x_1x_2\dots x_n$ the binary sequence of length n obtained when substituting every entry in $\mathbf{x}$ by its congruence class modulo 2. Now observe that if two permutations $\mathbf{x}$ and $\mathbf{y}$ are colliding, which means that there is a coordinate i such that x_i and y_i are consecutive integers, then in the ith coordinate of the corresponding parity patterns there is a difference in 0 and 1, implying that their parity patterns are different. So in a clique of $L(n)$ there is at most one permutation for any given parity pattern. Finally, the parity pattern of a permutation of $[n]$ has $\lfloor\frac{n}{2}\rfloor$ 0's and $\lceil\frac{n}{2}\rceil$ 1's. □

PROPOSITION 4.3.

$$\rho(L,n)\ge\rho(L,n-1)+\rho(L,n-2).$$

Proof. Take a clique C of maximal size for $L(n-1)$ and a clique D of maximal size for $L(n-2)$. Now set $\hat{C}:=\{x_1\dots x_{n-1}n :\ x_1\dots x_{n-1}\in C\}$ and $\hat{D}:=\{x_1\dots x_{n-2}n(n-1) :\ x_1\dots x_{n-2}\in D\}$. In this way any element from $\hat{C}$ will collide with any element from $\hat{D}$ in the last coordinate because of the edge $\{n,n-1\}$ so that $\hat{C}\cup\hat{D}$ is a clique in $L(n)$ of size $\rho(L,n-1)+\rho(L,n-2)$. □

COROLLARY 4.4.

$$\log_2\left(\frac{1+\sqrt{5}}{2}\right)\le\rho(L)\le 1.$$

Proof. Although the present upper bound to the permutation capacity of the infinite path is obvious as observed before, for the sake of completeness we deduce from Proposition 4.2 that $\frac{1}{n}\log_2\rho(L,n)\le\frac{1}{n}\log_2\binom{n}{\lfloor\frac{n}{2}\rfloor}\le 1$.

For the lower bound, Proposition 4.3 shows, together with $\rho(L,1)=1$ and $\rho(L,2)=2$, that the sequence $\rho(L,n)$ grows at least as fast as the basic Fibonacci sequence $F(n)$. Since $\lim_{n\to\infty}\sqrt[n]{F(n)}=\frac{1+\sqrt{5}}{2}$, we get $\log_2(\frac{1+\sqrt{5}}{2})\le\rho(L)$. □

A nonrecursive way of constructing a clique of size $F(n)$ in $L(n)$ follows. Consider the set S of permutations obtained from the identical permutation by exchanging two consecutive integers, i.e., $S = \{s_i : i = 1, \dots, n-1\}$, where $s_i = (i, i+1)$ is the adjacent transposition, in cyclic notation. For $I = \{i_1 < \dots < i_k\} \subseteq [n-1]$, let $s_I = s_{i_1} \dots s_{i_k}$. Let

$$C(n) = \{J \subseteq [n-1] : \forall i, j \in J \; s_i s_j = s_j s_i\}$$

be the family of subsets of $[n-1]$ whose corresponding adjacent transpositions are pairwise commuting. Since for $i \neq j$ one has $s_i s_j = s_j s_i$ if and only if $|i-j| \geq 2$, we can encode the elements of $C(n)$ as zero-one sequences of length $n-1$ with the property that no consecutive 1's appear in the sequence. Since the number of zero-one sequences of length n without consecutive "1"'s is known to be $F(n)$, and since each of these is in bijection with some element of $C(n)$, we see that in $C(n)$ there are exactly $F(n)$ sequences. Furthermore for $I, J \in C(n)$ with $I \neq J$ one has $\{s_I, s_J\} \in E(L(n))$. Let $h = \min I \Delta J$, where Δ denotes the symmetric difference of sets, and suppose that $h \in I$; then clearly $h \notin J$, $h+1 \notin I$ because of the condition on $C(n)$ and $h-1 \notin J$ by the minimality of h. When we deduce that $s_I(h) = h+1$ and $s_J(h) = h$, s_I and s_J are adjacent.

For $n = 4$, the set of binary sequences $\{000, 100, 010, 001, 101\}$ represents $C(4)$ and the corresponding set of permutations is $\{id; (12); (23); (34); (12)(34)\}$ in cycle notation, i.e.,

$$\{1234; 2134; 1324; 1243; 2143\}.$$

However, we will see very soon that the lower bound in the last corollary can be improved. The asymptotic improvement we obtain will be a direct consequence of the following inequality that follows easily from Proposition 4.1.

PROPOSITION 4.5. *For every $n \in \mathbb{N}$ we have*

$$\rho(L) \geq \log \sqrt[n]{\rho(L, n)}.$$

Proof. By Proposition 4.1 we have $\sqrt[nk]{\rho(L, nk)} \geq \sqrt[n]{\rho(L, n)}$. □

This justifies our interest in calculating $\rho(L, n)$ for the first values of n. The results are shown in the following table.

n	1	2	3	4	5	6	7
$\rho(L, n)$	1	2	3	6	10	20	35

For $n = 7$ we built a clique of size 35 by putting together 7 cliques each of size 5, obtained as cyclic shifts of certain sequences of length 5. Before explaining this construction in more detail, we prove a general result on cyclic shifts for any graph G.

Let $A \subseteq \mathbb{N}$ with $|A| = k$. Let $\pi = a_1 \dots a_k$ be an arrangement of $A = \{a_1, \dots, a_k\}$ on a cycle of length k; we say that π is a *circular arrangement* of A. We define the *circular distance* $\partial_\pi(a_i, a_j)$ of a_i and a_j with respect to π as follows:

$$\partial_\pi(a_i, a_j) = \begin{cases} 0 & \text{if } i = j \\ \min\{j-i, k+i-j\} & \text{if } i < j \\ \partial_\pi(a_j, a_i) & \text{if } i > j. \end{cases}$$

We say that a circular arrangement $\pi = a_1 \dots a_k$ is *complete* if for every $d = 1, \dots, \lfloor \frac{k}{2} \rfloor$ there exists an edge $\{a_i, a_j\} \in E(G)$ with $\partial_\pi(a_i, a_j) = d$.

LEMMA 4.6. *If $\pi = a_1 \dots a_k$ is a complete circular arrangement of A, then the subset $S(\pi)$ of $R(A)$ consisting of all the cyclic shifts of π, i.e.,*

$$S(\pi) = \{\pi^d = a_d a_{d+1} \dots a_{d+(k-1)} : d = 1, \dots, k\},$$

where

$$a_r = a_s \Leftrightarrow r \equiv s(\mod k),$$

is a clique in $G(A)$.

Proof. It is enough to show that for any $t = 2, \dots k$ one has $\{\pi, \pi^t\} \in E(G(A))$. First start with any t such that $t \leq \lfloor \frac{k}{2} \rfloor$. Since π is complete, there exists $\{a_i, a_j\} \in E(G)$ such that $\partial_\pi(a_i, a_j) = t$; we can fix $i < j$. If the circular distance t is achieved as $j - i$, then $\{\pi, \pi^t\} \in E(G(A))$ since in coordinate i one has $\{\pi_i, \pi_i^t\} = \{a_i, a_j\} \in E(G)$ and also $\{\pi, \pi^{k-t+1}\} \in E(G(A))$ since in coordinate j one has $\{\pi_j, \pi_j^{k-t+1}\} = \{a_j, a_i\} \in E(G)$; if the circular distance t is achieved as $k + i - j$, then $\{\pi_j, \pi_j^t\} = \{a_j, a_i\}$ and $\{\pi_i, \pi_i^{k-t+1}\} = \{a_i, a_j\}$. In any case both $\{\pi, \pi^t\}$ and $\{\pi, \pi^{k-t+1}\}$ are edges of $G(A)$; consequently $\{\pi, \pi^t\} \in E(G(A))$ for $t = 2, \dots k$. □

PROPOSITION 4.7.

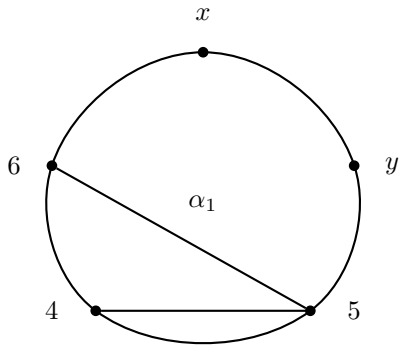
$$\rho(L, 7) = 35.$$

Proof. Let Γ be the set consisting of the following sequences:

$$\begin{array}{rcl} \alpha_1' & = & 23546 \\ \alpha_1'' & = & 32546 \\ \alpha_2' & = & 23547 \\ \alpha_2'' & = & 54237 \\ \beta_1' & = & 34651 \\ \beta_1'' & = & 65341 \\ \beta_2 & = & 14357. \end{array}$$

Each sequence in Γ is a complete circular arrangement of the corresponding set of its entries. By Lemma 4.6 it follows that $S(\gamma)$ is a clique for any $\gamma \in \Gamma$.

First we show that $A_1 := S(\alpha_1') \cup S(\alpha_1'')$ is a clique in $L(\{2, 3, 4, 5, 6\})$.



Observe that any sequence of the form $xy546$ is already a complete circular arrangement of $\{x, y, 4, 5, 6\}$ (we are not using the edges $\{2, 3\}$ and $\{3, 4\}$ to achieve the adjacencies). The same argument as in Lemma 4.6 makes clear that $(\alpha_1')^d$ is adjacent to any element of $S(\alpha_1'')$ of the form $(\alpha_1'')^t$ with $t \neq d$, while when $t = d$, then $(\alpha_1')^d$ and $(\alpha_1'')^d$ will be colliding in coordinate d, because $(\alpha_1'')^d$ is obtained from $(\alpha_1')^d$ by interchanging 2 and 3.

Now we show that $A_2 := S(\alpha_2') \cup S(\alpha_2'')$ is a clique in $L(\{2,3,4,5,7\})$.

Notice that we have $\alpha_2'' = \phi \circ \alpha_2'$, with $\phi = (2,5) \circ (3,4)$, where (i,j) is the transposition of i and j. Clearly $\{\alpha_2', \alpha_2''\} \in E(L(\{2,3,4,5,7\}))$; they collide in the second coordinate through the edge $\{3,4\}$. Observe that we can find two 2-sets $\{a,b\}$ (precisely $\{3,5\}$ and $\{2,4\}$) at circular distance, respectively 1 and 2, in α_2' such that the corresponding 2-sets of the form $\{a,\phi(b)\}$ (precisely $\{3,2\}$ and $\{2,3\}$) are edges of L. Hence the same reasoning as in Lemma 4.6 can be applied to establish that we also have $\{\alpha_2', (\alpha_2'')^d\} \in E(L(\{2,3,4,5,7\}))$ for $1 < d \leq 5$.

As for $B_1 := S(\beta_1') \cup S(\beta_1'')$, we notice that the bijection

$$\phi = \begin{pmatrix} 2 & 3 & 4 & 5 & 7 \\ 3 & 4 & 5 & 6 & 1 \end{pmatrix}$$

is an isomorphism of $L(\{2,3,4,5,7\})$ into $L(\{1,3,4,5,6\})$ such that $\beta_1' = \phi \circ \alpha_2'$ and $\beta_1'' = \phi \circ \alpha_2''$. So the argument used for the set A_2 applies to B_1 as well, showing that the latter is a clique in $L(\{1,3,4,5,6\})$.

Now let $sX := \{s\mathbf{x} : \mathbf{x} \in X\}$ (resp., Xs) be the set of sequences obtained from those in X by prefixing (resp., postfixing) to each of them the symbol s and set

$$A = 1A_1 7 \ \cup \ 1A_2 6,$$

$$B = 2B_1 7 \ \cup \ 2S(\beta_2)6.$$

The sets A and B are both cliques in $L(7)$ since the adjacency between elements of $1A_17$ and $1A_26$, or between those of $2B_17$ and $2S(\beta_2)6$, is guaranteed in the last coordinate, where we use the edge $\{6,7\}$. Finally $C = A \cup B$ is a clique in $L(7)$, since the adjacency between elements of A and B is established in the first coordinate, where we use the edge $\{1,2\}$. Obviously, C has 35 elements. □

Conjecture. Encouraged by the previous clique of size 35 we are tempted to formulate the following conjecture:

$$\rho(L,n) = \binom{n}{\lfloor \frac{n}{2} \rfloor}.$$

Unfortunately, we do not have more serious reasons to believe in it.

Now we are ready to improve the lower bound $\rho(L) \geq \log_2(\frac{1+\sqrt{5}}{2}) = 0.6942\ldots$ of Corollary 4.4.

PROPOSITION 4.8.

$$\rho(L) \geq 0.732\ldots.$$

Proof. Combining Proposition 4.5 with that of Proposition 4.7 we immediately see that $\rho(L,n) \geq 35^{\frac{n}{7}}$, and $\rho(L) \geq \frac{1}{7}\log_2 35 = 0.732\ldots$. □

5. Surprise capacity. Let G be once again an arbitrary graph with countable vertex set $\mathbb{N}$ and let G^n be the same power graph as before. We will say that the set $C \subseteq \mathbb{N}^n$ generates a *surprise clique* in G^n if C generates a clique in G^n with the property that for any $\{\mathbf{x},\mathbf{y}\} \in \binom{\mathbb{N}^n}{2}$ the ordered pairs of coordinates (x_i, y_i) are all different. We shall denote by $S(G,n)$ the maximum cardinality of a surprise clique in G^n and will call the limit

$$\limsup_{n\to\infty} \frac{1}{n} \log_2 S(G,n)$$

the *surprise capacity* of G. Clearly, this quantity is lower bounded by the permutation capacity of the same graph. (At the end of the paper we will comment on the intuitive meaning of this definition.)

PROPOSITION 5.1. *The maximum cardinality of a surprise clique in L^n is 2^n, where L is the infinite path.*

Proof. We shall use the easy and well-known submultiplicativity of the chromatic number, i.e., $\chi(G^n) \le [\chi(G)]^n$. To verify this, let $c : V(G) \to \mathbb{N}$ be an optimal coloring of G. Then the map $c^n : [V(G)]^n \to \mathbb{N}^n$ defined by $c^n(x_1 \dots x_n) = c(x_1) \dots c(x_n)$ is a proper coloring of G^n. The chromatic number of the infinite path L is 2. So for the power graph L^n one has $\chi(L^n) = 2^n$ and consequently $\omega(L^n) \le 2^n$. It follows that if C is a surprise clique in L^n, then

$$|C| \le 2^n,$$

since any surprise clique is in particular a clique.

We now construct a clique of cardinality 2^n, showing that

$$S(L, n) \ge 2^n.$$

Our construction will consist of appropriately chosen sequences of length n, with entries from $[2n]$, where lack of repetition will be ensured by strict monotonicity. For any binary sequence $\mathbf{x}$ of length n, define

$$a(\mathbf{x}) := x_1, x_1 + x_2, \dots, \sum_{j=1}^{n} x_j,$$

that is,

$$a(\mathbf{x})_i := \sum_{j=1}^{i} x_j. \tag{2}$$

Apply this to binary strings on the alphabet $\{1, 2\}$ and set

$$C := \{a(\mathbf{x}) : \mathbf{x} \in \{1, 2\}^n\} \subseteq [2n]^n.$$

By construction one has $a(\mathbf{x}) \ne a(\mathbf{x}')$ if and only if $\mathbf{x} \ne \mathbf{x}'$. Hence C has the same cardinality as the set of all the binary sequences of length n. Now we show that C is a clique in the nth power of L. Take $\mathbf{x}, \mathbf{x}' \in \{1, 2\}^n$ with $\mathbf{x} \ne \mathbf{x}'$ and let s be the first coordinate in which the two binary sequences differ. Then for the corresponding sequences in C we have $\{a(\mathbf{x}), a(\mathbf{x}')\} \in E(L^n)$, since $|a(\mathbf{x})_s - a(\mathbf{x}')_s| = |\sum_{j=1}^{s} x_j - \sum_{j=1}^{s} x'_j| = |x_s - x'_s| = 1$. Finally the condition for C to be a surprise clique holds, since by the definition (2) its elements are strictly increasing sequences; hence there are no repetitions of symbols. □

The above proposition shows that the surprise capacity of the infinite path is 1.

6. Unimodal permutations. Let us return to $L(n)$, the graph induced in the power graph L^n by the set of all permutations of $[n]$. In order to see the wealth of relatively large cliques in $L(n)$ it might be interesting to understand the density of cliques in the relatively small set of unimodal permutations.

We say that a permutation $\mathbf{a}$ of $[n]$ is unimodal if there is an index $h \in [n]$ such that $a_1 < a_2 < \dots < a_h > a_{h+1} > \dots > a_n$. We will introduce a new variant of our

introductory problem of determining the number $\rho(n)$ of the maximum cardinality of a set of pairwise colliding permutations of n (recall that this concept was at the core of our permutation capacity problem).

Let us denote by $U(n)$ the maximum cardinality of a set of pairwise colliding unimodal permutations of n.

THEOREM 6.1.

$$\log_2 \frac{1+\sqrt{5}}{2} \leq \limsup_{n\to\infty} \frac{1}{n} \log_2 U(n) \leq 1.$$

Proof. The upper bound is an obvious consequence of Proposition 4.2, since $U(n) \leq \rho(L, n)$.

Fix $\alpha \in (0,1)$ and $\beta \in (0,1)$ in a way to be specified later. Write $l_n = \lfloor \alpha n \rfloor$ and $k_n = \lfloor \beta \alpha n \rfloor$. Now we shall adapt the construction in the proof of Proposition 5.1 to define our unimodal permutations. To this purpose consider the set B_n of all the sequences in $\{1,2\}^{l_n}$ in which the symbol 2 appears k_n times. Thus by the well-known asymptotics of the binomial coefficients (Lemma 2.3, p. 30 of [4])

$$|B_n| = \binom{l_n}{k_n} \geq \frac{1}{n+1} 2^{l_n h(k_n/l_n)}, \tag{3}$$

where $h(t) = -t \log_2 t - (1-t)\log_2(1-t)$ is the binary entropy function.

To every $\mathbf{x} \in B_n$ we shall associate as before the increasing sequence of natural numbers $a(\mathbf{x})$ whose ith element $a_i(\mathbf{x})$ is as in formula (2). Suffixing to the sequence $a(\mathbf{x})$ the naturals from $[n] \setminus \{a_1(\mathbf{x}), \ldots a_{l_n}(\mathbf{x})\}$ in decreasing order we obtain the unimodal sequence $\widehat{\mathbf{x}}$ of integers from the set $[l_n + k_n]$, where $l_n + k_n = \lfloor \alpha n \rfloor + \lfloor \beta \alpha n \rfloor$. Then in order for $\widehat{\mathbf{x}}$ to be a permutation of $[n]$, we must have $\alpha(1+\beta) \leq 1$, i.e.,

$$\alpha \leq \frac{1}{1+\beta}. \tag{4}$$

Clearly the relation between $\mathbf{x}$ and $\widehat{\mathbf{x}}$ is bijective and therefore we have obtained $|B_n|$ unimodal permutations. As in Proposition 5.1, we have that if $\mathbf{x} \neq \mathbf{x}'$, then $a(\mathbf{x})$ and $a(\mathbf{x}')$ are colliding. So the corresponding sequences $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{x}'}$ will be colliding too.

We just saw that the set $\widehat{B_n} = \{\widehat{\mathbf{x}} : \mathbf{x} \in B_n\}$ has the same cardinality as B_n and it is a clique of unimodal elements of $L(n)$.

Recalling the definition of $U(n)$ and (3), we deduce that

$$U(n) \geq \frac{1}{n+1} 2^{l_n h(k_n/l_n)}$$

when

$$\limsup_{n\to\infty} \frac{1}{n} \log_2 U(n) \geq \limsup_{n\to\infty} \frac{l_n}{n} h\big(k_n/l_n\big) = \alpha h(\beta).$$

Choosing the largest α with respect to the constraint in (4), i.e., $\alpha = \frac{1}{1+\beta}$, and maximizing in β we obtain:

$$\limsup_{n\to\infty} \frac{1}{n} \log_2 U(n) \geq \max_{\beta\in(0,1)} \frac{h(\beta)}{1+\beta} = \log_2 \frac{1+\sqrt{5}}{2}.$$

In order to see that our last entropy expression has as its maximum the logarithm of the golden ratio, as claimed, the reader is referred to [8]. □

It is tempting to believe the lower bound to be tight, even though we have no real reason to do so.

7. Concluding remarks. In this paper we have introduced several closely related concepts of capacity for infinite graphs. It is not clear whether these can have the same interpretation in terms of Shannon's theory of information as do the concepts of Shannon capacity [13] and Sperner capacity [6]. In particular, the Shannon capacity of a finite simple graph is the highest rate at which one can transmit data over a discrete memoryless (stationary) channel with zero probability of error. Recently Nayak and Rose [12] showed that Sperner capacity is the key in determining the analogous transmission rate for compound channels with an uninformed coder-decoder pair.

The common feature of our models is that for no codeword pairs can we transmit the same symbol pair at different instants of time. This restriction might be of relevance if one is to guarantee security of transmission; an intruder can never experience the repetition of a symbol configuration and thereby learn how to adapt to a hitherto unknown communication situation it creates.

Having a disposable symbol set does not necessarily mean that the channel has an infinite input alphabet. In fact, note that in all our code constructions every symbol has at most seven different "successors."

Acknowledgments. We would like to thank Miki Simonovits for his friendly interest.

REFERENCES

[1] P. J. Cameron and C. Y. Ku, *Intersecting families of permutations*, European J. Combin., 2 (2003), pp. 881–890.
[2] G. Cohen, J. Körner, and G. Simonyi, *Zero-error capacities and very different sequences*, in Sequences, Combinatorics, Compression, Security and Transmission, R. Capocelli, ed., Springer-Verlag, New York, 1990, pp. 87–101.
[3] I. Csiszár and J. Körner, *On the capacity of the arbitrarily varying channel for maximum probability error*, Z. Wahrscheinlichkeittheorie verw. Geb., 57 (1981), pp. 87–101.
[4] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1982 and Akadémiai Kiadó, Budapest, Hungary, 1981.
[5] A. Galluccio, L. Gargano, J. Körner, and G. Simonyi, *Different capacities of digraphs*, Graphs Combin., 10 (1994), pp. 105–121.
[6] L. Gargano, J. Körner, and U. Vaccaro, *Sperner capacities*, Graphs Combin., 9 (1993), pp. 31–46.
[7] L. Gargano, J. Körner, and U. Vaccaro, *Capacities: From information theory to extremal set theory*, J. Combin. Theory Ser. A, 68 (1994), pp. 296–316.
[8] J. Körner and G. Simonyi, *A Sperner–type theorem and qualitative independence*, J. Combin. Theory Ser. A, 59 (1992), pp. 90–103.
[9] C. Y. Ku and I. Leader, *An Erdős–Ko–Rado theorem for partial permutations*, Discrete Math., 306 (2006), pp. 74–86.
[10] B. Larose and C. Malvenuto, *Stable sets of maximal size in Kneser-type graphs*, European J. Combin., 25 (2004), pp. 657–673.
[11] L. Lovász, *On the Shannon capacity of a graph*, IEEE Trans. Inform. Theory, 25 (1979), pp. 1–7.
[12] J. Nayak and K. Rose, *Graph capacities and zero-error transmission over compound channels*, IEEE Trans. Inform. Theory, to appear.
[13] C. E. Shannon, *The zero-error capacity of a noisy channel*, IEEE Trans. Inform. Theory, 2 (1956), pp. 8–19.
[14] J. H. Van Lint and R. M. Wilson, *A Course in Combinatorics*, Cambridge University Press, Cambridge, 1992.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 213–226

M-CONVEX FUNCTIONS ON JUMP SYSTEMS: A GENERAL FRAMEWORK FOR MINSQUARE GRAPH FACTOR PROBLEM*

KAZUO MUROTA†

Abstract. The concept of M-convex functions is generalized for functions defined on constant-parity jump systems. M-convex functions arise from minimum weight perfect b-matchings and from a separable convex function (sum of univariate convex functions) on the degree sequences of an undirected graph. As a generalization of a recent result of Apollonio and Sebő for the minsquare factor problem, a local optimality criterion is given for minimization of an M-convex function subject to a component sum constraint.

Key words. jump system, degree sequence, graph factor, discrete convex function, local optimality

AMS subject classifications. 90C10, 90C25, 90C35, 90C27

DOI. 10.1137/040618710

1. Introduction.

A recent paper of Apollonio and Sebő [2] has shown that the minsquare factor problem on a graph can be solved in polynomial time. The problem is, given an undirected graph possibly containing loops and parallel edges, to find a subgraph with a specified number of edges that minimizes the sum of squares of the degrees (= numbers of incident edges) of vertices. The key observation in [2] is that global optimality is guaranteed by local optimality in the neighborhood of ℓ_1-distance at most 4 in the space of degree sequences. It has also been observed in [2] that this local optimality criterion remains valid when the objective function is generalized to a separable convex function (= sum of univariate convex functions) of the degree sequence.

The objective of this paper is to put the above results in a more general context of discrete convex analysis [21] by introducing the concept of M-convex functions on constant-parity jump systems. A separable convex function of the degree sequences of a graph is an M-convex function in this sense.

A jump system [4] is a set of integer points with an exchange property (to be described in section 2); see also [14], [16]. It is a generalization of a matroid [6], [15], a delta-matroid [3], [5], [7], and a base polyhedron of an integral polymatroid (or a submodular system) [11]. Minimization of a separable convex function over a jump system has been studied in [1], where a local criterion for optimality as well as a greedy algorithm is given.

Study of nonseparable nonlinear functions on matroidal structures was started with valuated matroids [8], [9], which have come to be accepted as discrete concave functions; see [18], [20]. This concept has been generalized to M-convex functions on base polyhedra [19], which play a central role in discrete convex analysis [21]. Valuated

*Received by the editors November 10, 2004; accepted for publication (in revised form) November 9, 2005; published electronically March 3, 2006.
http://www.siam.org/journals/sidma/20-1/61871.html
†Graduate School of Information Science and Technology, University of Tokyo, and PRESTO, JST, Tokyo 113-8656, Japan (murota@mist.i.u-tokyo.ac.jp). 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.

delta-matroids [10] afford another generalization of valuated matroids; see also [10], [17], [24], [25]. 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-convex functions on base polyhedra, whereas they fail for valuated delta-matroids. M-convex functions on constant-parity jump systems, to be introduced in this paper, are a common generalization of valuated delta-matroids and M-convex functions on base polyhedra.

In this paper, we investigate the problem of minimizing an M-convex function on a constant-parity jump system. It is shown, in particular, that (i) global optimality for unconstrained minimization is equivalent to local optimality in the neighborhood of ℓ_1-distance 2 (Theorem 3.3), and (ii) global optimality for constrained minimization on a hyperplane of constant component sum is equivalent to local optimality in the neighborhood of ℓ_1-distance at most 4 (Theorem 4.1). The former generalizes the optimality criterion in [1] for separable convex function minimization over a jump system, and the latter generalizes the optimality criterion in [2] for the minsquare factor problem. Theorem 4.3 reveals convexity of the optimal values with respect to the component sum, on the basis of which algorithms are constructed for the constrained minimization in section 5.

2. Exchange axioms. Let V be a finite set. 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 $v \neq u$. For $x = (x(v)), y = (y(v)) \in \mathbf{Z}^V$ define

$$\begin{aligned} x(V) &= \sum_{v \in V} x(v), \\ \|x\|_1 &= \sum_{v \in V} |x(v)|, \\ \mathrm{supp}(x) &= \{v \in V \mid x(v) \neq 0\}, \\ \mathrm{supp}^+(x) &= \{v \in V \mid x(v) > 0\}, \\ \mathrm{supp}^-(x) &= \{v \in V \mid x(v) < 0\}, \end{aligned}$$

$$[x, y] = \{z \in \mathbf{Z}^V \mid \min(x(v), y(v)) \leq z(v) \leq \max(x(v), y(v)), \forall v \in V\}.$$

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 $x, y \in J$.

We introduce 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 the author by J. Geelen (see section 6.1 for a proof).

LEMMA 2.1 (see Geelen [12]). *A nonempty set J is a constant-parity jump system if and only if it satisfies* (J-EXC).

It turns out (see section 6.2 for a proof) that (J-EXC) can be replaced by a weaker axiom:

(J-EXC$_{\rm w}$) For any distinct $x, y \in J$ there exists an (x, y)-increment pair (s, t) such that $x + s + t \in J$ and $y - s - t \in J$.

LEMMA 2.2. *A set J satisfies* (J-EXC) *if and only if it satisfies* (J-EXC$_{\rm w}$).

We call $f : J \to \mathbf{R}$ an *M-convex function* if it satisfies the following exchange axiom:

(M-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

$$f(x) + f(y) \geq f(x + s + t) + f(y - s - t).$$

We adopt the convention that $f(x) = +\infty$ for $x \notin J$.

It turns out that the exchange axiom (M-EXC) is equivalent to a local exchange axiom:

(M-EXC$_{\rm loc}$) 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 2.3. *A function $f : J \to \mathbf{R}$ defined on a constant-parity jump system J satisfies* (M-EXC) *if and only if it satisfies* (M-EXC$_{\rm loc}$).

Proof. The proof is technical and given in section 6.3. □

This implies that (M-EXC) can be replaced by a weaker axiom:

(M-EXC$_{\rm w}$) For any distinct $x, y \in J$ 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 2.4. *A function $f : J \to \mathbf{R}$ satisfies* (M-EXC) *if and only if it satisfies* (M-EXC$_{\rm w}$).

Proof. It suffices to prove the "if" part. (M-EXC$_{\rm w}$) implies (J-EXC$_{\rm w}$) for J, and hence J is a constant-parity jump by Lemma 2.2. Then the claim follows from Theorem 2.3. □

Note that addition of a linear function preserves M-convexity. That is, for an M-convex function f and a vector $p = (p(v)) \in \mathbf{R}^V$, the function $f[-p]$ defined by $f[-p](x) = f(x) - \langle p, x\rangle$ with $\langle p, x\rangle = \sum_{v \in V} p(v)x(v)$ is M-convex.

Remark 2.1. Our 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 [11], and then our M-convex function is the same as the M-convex function investigated in [19], [21]. Case (ii) is equivalent to J being an even delta-matroid [24], [25], and then f is M-convex in our sense if and only if $-f$ is a valuated delta-matroid in the sense of [10]. □

Examples of M-convex functions follow.

Example 2.1. A *separable convex function* on a constant-parity jump system J, i.e., a function $f : J \to \mathbf{R}$ of the form $f(x) = \sum_{v \in V} \varphi_v(x(v))$ with univariate (one-dimensional) convex functions φ_v, is M-convex. In particular, the sum of squares $f(x) = \sum_{v \in V}(x(v))^2$ is M-convex. Such functions have been investigated in [1], [2]. □

Example 2.2. Minimum weight factors in a graph yield an M-convex function. Let $G = (V, E)$ be an undirected graph that may contain loops and parallel edges. For a subgraph $H = (V, F)$, denote its *degree sequence* by $\deg_H = \sum\{\chi_u + \chi_v \mid (u, v) \in F\} \in \mathbf{Z}^V$. It is well known [4], [16] that

$$J = \{\deg_H \mid H \text{ is a subgraph of } G\}$$

forms a constant-parity jump system called the *degree system* of G. Given edge weighting $w : E \to \mathbf{R}$, define a function $f : J \to \mathbf{R}$ by

$$f(x) = \min\{w(F) \mid H = (V, F) \text{ is a subgraph of } G \text{ with } \deg_H = x\}$$

with notation $w(F) = \sum_{e \in F} w(e)$, where $f(x)$ represents the minimum weight of a subgraph with degree sequence x.

This f is an M-convex function. In fact, (M-EXC) can be verified by the alternating path argument as follows. For distinct $x, y \in J$ let F_x and F_y be subsets of edges such that $f(x) = w(F_x)$ and $f(y) = w(F_y)$ with $x = \sum\{\chi_u + \chi_v \mid (u, v) \in F_x\}$ and $y = \sum\{\chi_u + \chi_v \mid (u, v) \in F_y\}$. Let s be an (x, y)-increment, and put $u_* = \mathrm{supp}(s)$. We may assume, without loss of generality, that $s = \chi_{u_*}$. Starting with an edge in $F_y \setminus F_x$ incident to u_* we construct an alternating path P by adding an edge in $F_x \setminus F_y$ and an edge in $F_y \setminus F_x$ alternately. The path P consists of distinct edges but may contain the same vertex more than once. We assume that P is maximal in the sense that it cannot be extended further beyond the end vertex, say, v_*. Then there exists an $(x + \chi_{u_*}, y)$-increment t with $\mathrm{supp}(t) = v_*$; more specifically, $t = \chi_{v_*}$ or $-\chi_{v_*}$ according to whether P consists of an odd or even number of edges. Denote by $F_x \Delta P$ the symmetric difference of F_x and P, and by $F_y \Delta P$ that of F_y and P. Since $x + s + t = \sum\{\chi_u + \chi_v \mid (u, v) \in F_x \Delta P\}$ and $y - s - t = \sum\{\chi_u + \chi_v \mid (u, v) \in F_y \Delta P\}$, we have $f(x + s + t) \le w(F_x \Delta P)$ and $f(y - s - t) \le w(F_y \Delta P)$, whereas $w(F_x \Delta P) + w(F_y \Delta P) = w(F_x) + w(F_y) = f(x) + f(y)$. Hence (M-EXC) holds. Note that the alternating path argument above also serves as a proof of (J-EXC) for J.

Furthermore,

$$f(x) = \min\{w(F) \mid H = (V, F) \text{ is a subgraph of } G \text{ with } \deg_H = x\} + \sum_{v \in V} \varphi_v(x(v))$$

is an M-convex function on the degree system of G, where each φ_v is a univariate convex function. □

Example 2.3. As a variant of the construction from the degree system in Example 2.2, an M-convex function arises from minimum weight perfect b-matchings; see [13], [23] for b-matchings. Let $G = (V, E)$ be an undirected graph that may have loops but no parallel edges, and let $w : E \to \mathbf{R}$ be an edge weighting. Let $J \subseteq \mathbf{Z}^V$ be the set of vectors $x \in \mathbf{Z}^V$ such that a perfect x-matching exists in G, and define a function $f : J \to \mathbf{R}$ by setting $f(x)$ to be the minimum weight of a perfect x-matching:

$$f(x) = \min\left\{ \sum_{e \in E} \lambda(e) w(e) \mid \sum_{e \in \delta(v)} \lambda(e) = x(v)\ (\forall v \in V); \lambda(e) \in \mathbf{Z}_+\ (\forall e \in E) \right\}, \tag{2.1}$$

where $\delta(v)$ denotes the set of edges incident to vertex $v \in V$, and $\mathbf{Z}_+$ the set of nonnegative integers. This function is M-convex as in Example 2.2. □

Example 2.4. Let $A(t)$ be a skew-symmetric polynomial matrix in variable t. The degree in t of the principal minors of $A(t)$ yields a valuated delta-matroid, as is pointed out in [10], [24], and hence the negative of an M-convex function. □

Remark 2.2. Unlike in the previously studied special cases where J is a base polyhedron or an even delta-matroid, an M-convex function on a jump system is not always extensible to a convex function. Nevertheless, our results will provide convincing evidence to indicate its discrete convexity. See also [22]. □

3. Unconstrained minimization. We consider minimization of an M-convex function $f : J \to \mathbf{R}$ defined on a constant-parity jump system $J \subseteq \mathbf{Z}^V$.

First we note a property of an M-convex function that indicates its discrete convexity. Given $f : J \to \mathbf{R}$ and $x, y \in J$, a sequence of points in J, say, $x_0, x_1, \ldots, x_m$, is called a *steepest-descent chain* connecting x to y if $x_0 = x$, $x_m = y$, and for $i = 1, \ldots, m$ we have $x_i = x_{i-1} + s_i + t_i$ for some (x_{i-1}, y)-increment pair (s_i, t_i) such that $f(x_{i-1} + s_i + t_i) \leq f(x_{i-1} + s + t)$ for every (x_{i-1}, y)-increment pair (s, t); we have $m = ||x - y||_1/2$. An M-convex function turns out to be convex along a steepest-descent chain, as follows.

PROPOSITION 3.1. *Let $f : J \to \mathbf{R}$ be an M-convex function, and let $x_0, x_1, \ldots, x_m$ be a steepest-descent chain connecting $x \in J$ to $y \in J$. Then*

$$f(x_{i-1}) + f(x_{i+1}) \geq 2f(x_i) \qquad (i = 1, \ldots, m-1). \tag{3.1}$$

Proof. Put $x_i = x_{i-1} + s + t$ and $x_{i+1} = x_i + s' + t'$. By (M-EXC) we have

$$\begin{aligned} &f(x_{i-1}) + f(x_{i+1}) \\ &\geq \min[f(x_{i-1} + s + t) + f(x_{i-1} + s' + t'), \\ &\qquad f(x_{i-1} + s + t') + f(x_{i-1} + s' + t), \\ &\qquad f(x_{i-1} + s + s') + f(x_{i-1} + t + t')] \ \geq \ 2f(x_i). \end{aligned}$$ □

As an immediate corollary we see that a nonoptimal point can be improved with a suitable increment pair.

PROPOSITION 3.2.

(1) *If $x, y \in J$ and $f(x) > f(y)$, there exists an (x, y)-increment pair (s, t) such that $f(x) > f(x + s + t)$.*

(2) *If $x, y \in J$ and $f(x) \geq f(y)$, there exists an (x, y)-increment pair (s, t) such that $f(x) \geq f(x + s + t)$.*

This implies, in turn, that global optimality (minimality) of an M-convex function is guaranteed by local optimality in the neighborhood of ℓ_1-distance 2.

THEOREM 3.3. *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 $||x - y||_1 \leq 2$.*

Proof. The "only if" part is obvious, and the "if" part follows from Proposition 3.2. □

The minimizers of an M-convex function form a constant-parity jump system, as follows. We denote by $\arg\min f[-p]$ the set of minimizers of function $f[-p]$.

PROPOSITION 3.4. *For any $p \in \mathbf{R}^V$, $\arg\min f[-p]$ is a constant-parity jump system if it is nonempty.*

Proof. Let β denote the minimum value of $f[-p]$, and let $x, y \in \arg\min f[-p]$. Then, in (M-EXC) we have $2\beta = f[-p](x) + f[-p](y) \geq f[-p](x + s + t) + f[-p](y - s - t) \geq 2\beta$, which implies $x + s + t, y - s - t \in \arg\min f[-p]$. □

Remark 3.1. The local optimality criterion for M-convex functions on jump systems in Theorem 3.3 contains a number of previous results as special cases. In the case of constant-sum jump systems, case (i) in Remark 2.1, the present theorem reduces to the optimality criterion for M-convex functions on base polyhedra established in [19] (see Theorem 6.26 of [21]), and, moreover, Proposition 3.2(1) above coincides with Proposition 6.23 of [21]. In the case of constant-parity jump systems contained in $\{0,1\}^V$, case (ii) in Remark 2.1, Theorem 3.3 reduces to the optimality criterion for valuated delta-matroids established in [10]. Both of these are generalizations, in different directions, of the optimality criterion for valuated matroids given in [8], [9]. It is noted that the optimality criterion for valuated matroids given in [8], [9] is the origin of such optimality criteria for nonseparable nonlinear objective functions, and the two special cases above are generalizations in different directions thereof. Separable convex functions on jump systems have been considered in [1]. □

4. Minimization under sum constraint. In this section we investigate the problem of minimizing an M-convex function $f(x)$ when the sum of the components of x is specified. Recalling the notation $x(V)$ for the sum of components of a vector x, we introduce some other notation concerning the feasible regions of our optimization problem:

$$\begin{aligned} k_{\min} &= \min\{x(V) \mid x \in J\}, \\ k_{\max} &= \max\{x(V) \mid x \in J\}, \\ \Lambda &= \{k \mid k_{\min} \le k \le k_{\max},\ k \equiv k_{\min} (\mathrm{mod}\ 2)\}, \\ J_k &= \{x \in J \mid x(V) = k\} \qquad (k \in \Lambda), \end{aligned}$$

where $J_k \neq \emptyset$ for each $k \in \Lambda$ by (J-EXC) and it may be that $k_{\min} = -\infty$ and/or $k_{\max} = +\infty$.

Our problem is to minimize $f(x)$ subject to $x \in J_k$, where $k \in \Lambda$ is a parameter. Denote by f_k and M_k the minimum value and the set of minimizers, respectively, i.e.,

$$\begin{aligned} f_k &= \min\{f(x) \mid x \in J_k\} \qquad (k \in \Lambda), \\ M_k &= \{x \in J_k \mid f(x) = f_k\} \qquad (k \in \Lambda), \end{aligned}$$

where we assume that, for each $k \in \Lambda$, f_k is finite and M_k is nonempty. By convention we put $f_k = +\infty$ for $k \notin \Lambda$.

Global optimality (minimality) on J_k is guaranteed by local optimality in the neighborhood of ℓ_1-distance at most 4. Compare this with the unconstrained optimization treated in Theorem 3.3, which refers to the neighborhood of ℓ_1-distance 2. It is emphasized that J_k is not necessarily a jump system, and accordingly, Theorem 3.3 does not apply to minimization of f over J_k.

THEOREM 4.1. *Let $f : J \to \mathbf{R}$ be an M-convex function on a constant-parity jump system J, and let $x \in J_k$ with $k \in \Lambda$. Then $f(x) \le f(y)$ for all $y \in J_k$ if and only if $f(x) \le f(y)$ for all $y \in J_k$ with $\|x - y\|_1 \le 4$.*

Proof. The "only if" part is obvious. To prove the "if" part by contradiction, assume that $f(x) > f(y)$ for some $y \in J_k$ and take such y with minimum $||y - x||_1$. Since $x(V) = y(V)$ and $x \neq y$, both $\mathrm{supp}^+(y - x)$ and $\mathrm{supp}^-(y - x)$ are nonempty.

Claim 1. If $u \in \mathrm{supp}^+(y - x)$ and $v \in \mathrm{supp}^-(y - x)$, then

$$f(x) + f(y) < f(x + \chi_u - \chi_v) + f(y - \chi_u + \chi_v).$$

Proof of Claim 1. We have $f(x) \le f(x+\chi_u-\chi_v)$ by the assumed local optimality, and $f(x) \le f(y-\chi_u+\chi_v)$ since $y-\chi_u+\chi_v$ is closer to x than y. Adding these two and $f(y) < f(x)$ yields the desired inequality. □

By (M-EXC) for (x, y), together with Claim 1, there exist $u_1 \in \mathrm{supp}^+(y-x)$, $u_2 \in \mathrm{supp}^+(y-x-\chi_{u_1})$, $v_1 \in \mathrm{supp}^-(y-x)$, and $v_2 \in \mathrm{supp}^-(y-x-\chi_{v_1})$ such that

$$f(x)+f(y) \ge f(x+\chi_{u_1}+\chi_{u_2})+f(y-\chi_{u_1}-\chi_{u_2}), \tag{4.1}$$

$$f(x)+f(y) \ge f(x-\chi_{v_1}-\chi_{v_2})+f(y+\chi_{v_1}+\chi_{v_2}). \tag{4.2}$$

By (M-EXC) for $(x+\chi_{u_1}+\chi_{u_2}, x-\chi_{v_1}-\chi_{v_2})$ and the local optimality, we obtain

$$\begin{aligned} &f(x+\chi_{u_1}+\chi_{u_2})+f(x-\chi_{v_1}-\chi_{v_2}) \\ &\ge \min[f(x+\chi_{u_1}-\chi_{v_1})+f(x+\chi_{u_2}-\chi_{v_2}), \\ &\qquad f(x+\chi_{u_1}-\chi_{v_2})+f(x+\chi_{u_2}-\chi_{v_1}), \\ &\qquad f(x)+f(x+\chi_{u_1}+\chi_{u_2}-\chi_{v_1}-\chi_{v_2})] \\ &\ge 2f(x). \end{aligned} \tag{4.3}$$

Similarly, by (M-EXC) for $(y-\chi_{u_1}-\chi_{u_2}, y+\chi_{v_1}+\chi_{v_2})$, we obtain

$$\begin{aligned} &f(y-\chi_{u_1}-\chi_{u_2})+f(y+\chi_{v_1}+\chi_{v_2}) \\ &\ge \min[f(y-\chi_{u_1}+\chi_{v_1})+f(y-\chi_{u_2}+\chi_{v_2}), \\ &\qquad f(y-\chi_{u_1}+\chi_{v_2})+f(y-\chi_{u_2}+\chi_{v_1}), \\ &\qquad f(y)+f(y-\chi_{u_1}-\chi_{u_2}+\chi_{v_1}+\chi_{v_2})] \\ &\ge f(x)+f(y), \end{aligned} \tag{4.4}$$

since $f(y-\chi_{u_i}+\chi_{v_j}) \ge f(x)$ and $f(y-\chi_{u_1}-\chi_{u_2}+\chi_{v_1}+\chi_{v_2}) \ge f(x)$ by the choice of y. Adding (4.1), (4.2), (4.3), and (4.4) yields a contradiction. □

The ℓ_1-distance of 4 in Theorem 4.1 cannot be replaced by the ℓ_1-distance of 2, as we see in the following example.

Example 4.1 (see [2]). Let $J \subseteq \mathbf{Z}^6$ be the degree system (see Example 2.2) of an undirected graph consisting of two vertex-disjoint triangles, and let $f : J \to \mathbf{R}$ be an M-convex function representing the sum of squares of the components (see Example 2.1). Let $k = 8$ and $x = (2, 2, 2, 1, 1, 0)$, for which $f(x) = 14$. For any point $y \in J_8$ with $\|y-x\|_1 = 2$ we have $f(y) = 14$, whereas for $x^* = (2, 1, 1, 2, 1, 1)$ we have $f(x^*) = 12$ and $\|x^*-x\|_1 = 4$. □

Remark 4.1. Theorem 4.1 above is a generalization of Theorem 1 of [2], since the degree system of a graph is a constant-parity jump system (Example 2.2) and a separable convex function on a constant-parity jump system is an M-convex function (Example 2.1). In fact, the result of [2] was the primary motivation behind Theorem 4.1. □

The following theorem reveals a kind of monotonicity of the minimizers of f on J_k.

THEOREM 4.2. *For any $x_k \in M_k$ with $k \in \Lambda$ there exists $(x_l \in M_l \mid l \in \Lambda \setminus \{k\})$ such that $x_{k_{\min}} \le \cdots \le x_{k-2} \le x_k \le x_{k+2} \le \cdots \le x_{k_{\max}}$.*

Proof. We show the existence of such x_{k-2}. Then x_{k+2} can be shown to exist in a similar manner, and the other x_l ($l \le k-4$ or $l \ge k+4$) exist by induction.

Take $y \in M_{k-2}$ with minimum $\|y-x_k\|_1$. If $y \le x_k$, we are done with $x_{k-2} = y$. Otherwise, take $u \in \mathrm{supp}^-(y-x_k)$ and apply (M-EXC) to obtain either

$$\exists v \in \mathrm{supp}^-(y-x_k):\ f_{k-2}+f_k \ge f(y+\chi_u+\chi_v)+f(x_k-\chi_u-\chi_v)$$

or

$$\exists v \in \mathrm{supp}^+(y - x_k) : \ f_{k-2} + f_k \geq f(y + \chi_u - \chi_v) + f(x_k - \chi_u + \chi_v).$$

In the first case the right-hand side is lower bounded by $f_k + f_{k-2}$ and hence $y + \chi_u + \chi_v \in M_k$ and $x_k - \chi_u - \chi_v \in M_{k-2}$; then we can take $x_{k-2} = x_k - \chi_u - \chi_v$. The second case cannot occur, since the right-hand side is lower bounded by $f_{k-2} + f_k$, from which follows $y + \chi_u - \chi_v \in M_{k-2}$, whereas $\|(y + \chi_u - \chi_v) - x_k\|_1 = \|y - x_k\|_1 - 2$, which is a contradiction to the choice of y. □

THEOREM 4.3. *Minimum values f_k form a convex sequence:*

$$f_{k-2} + f_{k+2} \geq 2f_k \qquad (k \in \Lambda \setminus \{k_{\min}, k_{\max}\}). \tag{4.5}$$

Proof. By Theorem 4.2 we can take $x_{k-2} \in M_{k-2}$ and $x_{k+2} \in M_{k+2}$ with $x_{k-2} \leq x_{k+2}$, and also $u \in \mathrm{supp}^+(x_{k+2} - x_{k-2})$. By (M-EXC) there exists $v \in \mathrm{supp}^+(x_{k+2} - x_{k-2})$ such that

$$f_{k-2} + f_{k+2} \geq f(x_{k-2} + \chi_u + \chi_v) + f(x_{k+2} - \chi_u - \chi_v) \geq 2f_k. \qquad \square$$

Convexity of the minimum values motivates us to consider the subgradient. For $\alpha \in \mathbf{R}$ define $f^\alpha : J \to \mathbf{R}$ by

$$f^\alpha(x) = f(x) - \alpha x(V). \tag{4.6}$$

Then we have

$$\min_{x \in J} f^\alpha(x) = \min_{l \in \Lambda} \min_{x \in J_l} f^\alpha(x) = \min_{l \in \Lambda}(f_l - \alpha l). \tag{4.7}$$

By Theorem 4.3, the minimum of $f_l - \alpha l$ over $l \in \Lambda$ is attained by $l = k$ if

$$(f_k - f_{k-2})/2 \leq \alpha \leq (f_{k+2} - f_k)/2. \tag{4.8}$$

Hence

$$f_k = k\alpha + \min\{f^\alpha(x) \mid x \in J\} \tag{4.9}$$

for α in the range of (4.8). This shows that the optimal value f_k can be computed by solving an unconstrained minimization problem for another M-convex function f^α.

Let us note, however, that not every minimizer of f^α belongs to J_k. A point $x \in J$ minimizes $f^\alpha(x)$ if and only if $x \in M_k$ for some k with $k_-(\alpha) \leq k \leq k_+(\alpha)$, where

$$k_-(\alpha) = \min\{k \mid \min_l(f_l - \alpha l) = f_k - \alpha k\}, \tag{4.10}$$

$$k_+(\alpha) = \max\{k \mid \min_l(f_l - \alpha l) = f_k - \alpha k\}. \tag{4.11}$$

THEOREM 4.4. *For each $\alpha \in \mathbf{R}$, $\bigcup\{M_k \mid k_-(\alpha) \leq k \leq k_+(\alpha)\}$ is a constant-parity jump system. In particular, M_k is a base polyhedron if $k = k_{\min}$, or $k = k_{\max}$, or $f_{k-2} + f_{k+2} > 2f_k$ with $k \in \Lambda \setminus \{k_{\min}, k_{\max}\}$.*

Proof. The first statement follows from Proposition 3.4 since, as observed above, $\bigcup\{M_k \mid k_-(\alpha) \leq k \leq k_+(\alpha)\}$ coincides with $\arg\min f^\alpha$. For the second statement it suffices to note that for such k we can choose an α with $k_-(\alpha) = k = k_+(\alpha)$ and that a constant-sum jump system is a base polyhedron. □

Remark 4.2. Theorems 4.2, 4.3, and 4.4 above are natural generalizations of the similar results of [17] for valuated delta-matroids. □

5. Algorithms. The local optimality criteria in Theorems 3.3 and 4.1 for unconstrained and constrained minimization, respectively, naturally suggest descent-type algorithms. At each feasible nonoptimal point, an improved point can be found with $O(|V|^2)$ function evaluations in unconstrained minimization and $O(|V|^4)$ function evaluations in constrained minimization. Although we do not enter into further technical details (see [22]), the number of updates of the solution point may be bounded by the ℓ_1-distance from the initial point to the optimal point, or by the difference of the objective function values at the initial point and at the optimal point if the objective function is integer-valued.

Two other algorithms can be constructed for constrained minimization, to minimize $f(x)$ subject to $x \in J_k$, on the basis of Theorems 4.2 and 4.3. It is assumed that an algorithm is available for unconstrained minimization. For the convenience of descriptions it is also assumed that $k_{\min}$ and $k_{\max}$ are finite.

An increasing sequence of optimal solutions, the existence of which is guaranteed by Theorem 4.2, can be generated by the following algorithm. Once a global minimizer x^* is found, the algorithm computes the whole set of f_k ($k \in \Lambda$) with $\mathrm{O}((k_{\max} - k_{\min})|V|^2)$ evaluations of f. Note that the algorithm works even if $k_{\min}$ and/or $k_{\max}$ are not known in advance.

ALGORITHM I.

Compute $x^* \in J$ that minimizes f;
Set $k^* := x^*(V)$, $x_{k^*} := x^*$, $f_{k^*} := f(x_{k^*})$;
for $k := k^* + 2, k^* + 4, \ldots, k_{\max}$ **do**
  Find $\{u, v\} \subseteq V$ that minimizes $f(x_{k-2} + \chi_u + \chi_v)$
  and put $x_k := x_{k-2} + \chi_u + \chi_v$ and $f_k := f(x_k)$;
for $k := k^* - 2, k^* - 4, \ldots, k_{\min}$ **do**
  Find $\{u, v\} \subseteq V$ that minimizes $f(x_{k+2} - \chi_u - \chi_v)$
  and put $x_k := x_{k+2} - \chi_u - \chi_v$ and $f_k := f(x_k)$.

Convexity of the sequence f_k makes it possible to convert the constrained minimization to an unconstrained minimization of f^α with an appropriate value of α; see (4.9). Here f^α is M-convex and, by our assumption, the minimum of $f^\alpha(x)$ over $x \in J$ can be computed efficiently. We assume that we can find $k_+(\alpha)$ and $k_-(\alpha)$ of (4.11) and (4.10) by maximizing (resp., minimizing) $x(V)$ among the minimizers of $f^\alpha(x)$ by means of some variant of an unconstrained minimization algorithm.

The following algorithm computes $k_{\min}$, $k_{\max}$, and f_k ($k \in \Lambda$) by searching for appropriate values of α. It requires $\mathrm{O}((k_{\max} - k_{\min})|V|^2)$ evaluations of f.

ALGORITHM II.

Let α be sufficiently large;
Minimize f^α to find $k_{\min} = k_+(\alpha) = k_-(\alpha)$ and $f_{k_{\min}}$;
Let α be sufficiently small
  (α is a negative number with a large absolute value);
Minimize f^α to find $k_{\max} = k_+(\alpha) = k_-(\alpha)$ and $f_{k_{\max}}$;
if $k_{\max} - k_{\min} \geq 4$ **then** search$(k_{\min}, k_{\max})$.

Here the procedure "search(k_1, k_2)" is defined when $k_1 + 4 \leq k_2$ as follows:

procedure search(k_1, k_2)
$\alpha := (f_{k_2} - f_{k_1})/(k_2 - k_1)$;
Minimize f^α to find $k_+ = k_+(\alpha)$, $k_- = k_-(\alpha)$, $f_+ = f_{k_+}$ and $f_- = f_{k_-}$;
for $k := k_- + 2, k_- + 4, \ldots, k_+ - 2$ **do**
  $f_k := ((k - k_-)f_+ + (k_+ - k)f_-)/(k_+ - k_-)$;
if $k_1 + 4 \leq k_-$ **then** search(k_1, k_-);
if $k_+ + 4 \leq k_2$ **then** search(k_+, k_2).

The second algorithm, as it stands, computes the values of f_k and not the optimal solutions x_k. If x_k's are wanted, they can be computed easily in procedure "search" by generating a sequence of points $x_k \in J_k \cap \arg\min f^\alpha$ by applying (J-EXC) to the pair of the optimal solutions x_{k_-} and x_{k_+}.

6. Proofs.

6.1. Proof for (J-EXC). A proof of Lemma 2.1, different from that of Geelen [12], is provided here. This proof can be extended to Theorem 2.3.

For a constant-parity system J, the 2-step axiom of a jump system is simplified to:

(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 $x + s + t \in J$.

It suffices to prove (J-EXC$_+$) $\Rightarrow$ (J-EXC), since (J-EXC) $\Rightarrow$ (J-EXC$_+$) is obvious and (J-EXC) implies J being a constant-parity system.

We first note the following fact.

LEMMA 6.1. *Assume* (J-EXC$_+$), *let* $y \in J$, *and let* z *be a point at* ℓ_1*-distance* 4 *from* y, *represented as* $z = y - s_1 - s_2 - s_3 - s_4$ *with* $s_i \in \mathbf{Z}^V$ *and* $||s_i||_1 = 1$ *for* $i = 1, 2, 3, 4$. *If* $z \in J$, *then* $y - s_i - s_j \in J$ *and* $y - s_k - s_l \in J$ *for some* $i, j, k, l \in \{1, 2, 3, 4\}$ *with* $\{i, j, k, l\} = \{1, 2, 3, 4\}$.

Proof. Consider an undirected graph G with vertex-set $\{1, 2, 3, 4\}$ and edge-set $\{(i, j) \mid y - s_i - s_j \in J\}$. It follows from (J-EXC$_+$) for (y, z) with $s = -s_i$ that, for each vertex i, there exists an edge incident to i. Similarly, it follows from (J-EXC$_+$) for (z, y) with $s = s_i$ that, for each vertex i, there exists an edge not incident to i. Such a graph has a perfect matching consisting of two edges, say, (i, j) and (k, l) with $\{i, j, k, l\} = \{1, 2, 3, 4\}$. This means that $y - s_i - s_j \in J$ and $y - s_k - s_l \in J$. □

To prove (J-EXC$_+$) $\Rightarrow$ (J-EXC) by contradiction, we assume that there exists a pair (x, y) for which (J-EXC) fails. That is, we assume that the set of such pairs,

$$\mathcal{D} = \{(x, y) \mid x, y \in J,\ \exists s_* : (x, y)\text{-increment such that} \\ \forall t : (x + s_*, y)\text{-increment} : x + s_* + t \notin J \text{ or } y - s_* - t \notin J\},$$

is nonempty.

Take a pair $(x, y) \in \mathcal{D}$ with minimum $\|x-y\|_1$, where $\|x-y\|_1 \geq 4$, fix s_* satisfying the condition above, and put $u_* = \mathrm{supp}(s_*)$. Denoting the set of $(x+s_*, y)$-increments by I, we have

$$x + s_* + t \notin J \quad \text{or} \quad y - s_* - t \notin J \qquad (t \in I). \tag{6.1}$$

Put $U = \mathrm{supp}(y - x)$ and, for $v \in U$, let t_v denote the (uniquely determined) (x, y)-increment such that $\mathrm{supp}(t_v) = v$; we have $t_v = \sigma(v)\chi_v$ using the notation σ defined by $\sigma(v) = 1$ for $v \in \mathrm{supp}^+(y - x)$ and $\sigma(v) = -1$ for $v \in \mathrm{supp}^-(y - x)$. Define $\alpha \in \mathbf{R}$ by

$$\alpha = \begin{cases} 1/2 & (s_* \in I,\ x + 2s_* \notin J,\ y - 2s_* \in J), \\ 0 & (\text{otherwise}), \end{cases}$$

and $p \in \mathbf{R}^V$ by

$$\sigma(v)p(v) = \begin{cases} \alpha & (v = u_*), \\ -\alpha & (v \in U \setminus \{u_*\}, x + s_* + t_v \in J), \\ -\alpha + 1 & (v \in U \setminus \{u_*\}, x + s_* + t_v \notin J, y - s_* - t_v \in J), \\ 0 & (\text{otherwise}). \end{cases}$$

Claim 1.

$$\langle p, s_* + t\rangle = 0 \qquad \text{if } t \in I,\ x + s_* + t \in J, \tag{6.2}$$

$$\langle p, s_* + t\rangle = 1 \qquad \text{if } t \in I,\ y - s_* - t \in J. \tag{6.3}$$

The equality (6.2) is easy to see, whereas (6.3) can be shown as follows. By (6.1) we have $x + s_* + t \notin J$, and hence

$$\langle p, s_* + t\rangle = \begin{cases} 2\alpha = 1 & \text{if } t = s_*, \\ \alpha + (-\alpha + 1) = 1 & \text{if } t \neq s_*. \end{cases}$$

Next, let P denote the set of $(x + s_*, y)$-increment pairs.

Claim 2. There exists $(s_0, t_0) \in P$ such that $y - s_0 - t_0 \in J$ and

$$\langle p, s_0 + t_0\rangle \leq \langle p, s + t\rangle \quad \text{if } (s, t) \in P,\ y - s - t \in J. \tag{6.4}$$

Since s_* is an (x, y)-increment and J satisfies (J-EXC$_+$), there exists $t_* \in I$ such that $x + s_* + t_* \in J$, where t_* may possibly be identical to s_*. We see that $x + s_* + t_*$ is distinct from y since $\|x - y\|_1 \geq 4$. By (J-EXC$_+$) and the minimal choice of x, y there exists an $(x + s_* + t_*, y)$-increment pair (s, t) such that $y - s - t \in J$. This shows the existence of $(s, t) \in P$ with $y - s - t \in J$. Then (6.4) is satisfied by the pair $(s, t) = (s_0, t_0)$ that minimizes $\langle p, s + t\rangle$ over $(s, t) \in P$ subject to the condition $y - s - t \in J$.

Claim 3. $(x, y') \in \mathcal{D}$ with $y' = y - s_0 - t_0$.

To show this, first note that s_* is an (x, y')-increment, and let t be an $(x + s_*, y')$-increment. We have $t \in I$, $(s_0, t) \in P$, and $(t_0, t) \in P$. Hence, by (6.4), we have

$$\langle p, s_0 + t_0\rangle \leq \langle p, s_0 + t\rangle \qquad \text{if } y - s_0 - t \in J, \tag{6.5}$$

$$\langle p, s_0 + t_0\rangle \leq \langle p, t_0 + t\rangle \qquad \text{if } y - t_0 - t \in J. \tag{6.6}$$

We assume $y' - s_* - t \in J$ and derive $x + s_* + t \notin J$. By Lemma 6.1 with $z = y - s_0 - t_0 - s_* - t$, at least one of the following three cases occurs: (i) $y - s_0 - t_0 \in J$ and $y - s_* - t \in J$, (ii) $y - s_0 - t \in J$ and $y - s_* - t_0 \in J$, and (iii) $y - t_0 - t \in J$ and $y - s_* - s_0 \in J$. In any case we have

$$\langle p, s_* + t\rangle \geq 1, \tag{6.7}$$

since, in case (ii), for example, we have

$$\langle p, s_0 + t_0 + s_* + t\rangle = \langle p, s_0 + t\rangle + \langle p, s_* + t_0\rangle \geq \langle p, s_0 + t_0\rangle + 1$$

by (6.3) and (6.5). By (6.7) and (6.2) we see $x + s_* + t \notin J$. Hence $(x, y') \in \mathcal{D}$.

Finally, since $\|x - y'\|_1 = \|x - y\|_1 - 2$, Claim 3 contradicts our choice of $(x, y) \in \mathcal{D}$. Therefore we conclude $\mathcal{D} = \emptyset$, completing the proof of Lemma 2.1.

6.2. Proof for (J-EXC) $\Leftrightarrow$ (J-EXC$_{\mathrm{w}}$). A proof of Lemma 2.2 is provided here. By the argument in section 6.1, it suffices to show (J-EXC$_{\mathrm{w}}$) $\Rightarrow$ (J-EXC$_+$), which we prove by induction on $\|x - y\|_1$. Take distinct $x, y \in J$ and an (x, y)-increment s. By (J-EXC$_{\mathrm{w}}$) there exists an (x, y)-increment pair (s_1, t_1) such that $x + s_1 + t_1 \in J$ and $y - s_1 - t_1 \in J$. If $s \in \{s_1, t_1\}$, we are done. Otherwise, put $y' = y - s_1 - t_1$. We have $\|x - y'\|_1 = \|x - y\|_1 - 2$ and s is an (x, y')-increment. By the induction hypothesis, (J-EXC$_+$) with (x, y') and s implies $x + s + t \in J$ for some (x, y')-increment t, which is also an (x, y)-increment.

6.3. Proof for (M-EXC) $\Leftrightarrow$ (M-EXC$_{\mathbf{loc}}$). A proof of Theorem 2.3 is provided here. It suffices to prove (M-EXC$_{\mathrm{loc}}$) $\Rightarrow$ (M-EXC). For $x \in J$, $d \in \mathbf{Z}^V$, and $p \in \mathbf{R}^V$, define $f(x,d) = f(x+d) - f(x)$ and $f_p(x,d) = f(x+d) - f(x) - \langle p, d\rangle$. We then have

$$f_p(x,d) + f_p(y,-d) = f(x,d) + f(y,-d). \tag{6.8}$$

We use an abbreviation f_p for $f[-p]$.

LEMMA 6.2. *Assume $x, y \in J$, $\|x-y\|_1 = 4$, and $p \in \mathbf{R}^V$. If* (M-EXC$_{\mathrm{loc}}$) *is satisfied, then*

$$f_p(y) - f_p(x) \geq \min(\pi_{12}+\pi_{34}, \pi_{13}+\pi_{24}, \pi_{14}+\pi_{23}). \tag{6.9}$$

Here $\pi_{ij} = f_p(x, s_i + s_j)$ for $i,j \in \{1,2,3,4\}$, where $y = x + s_1 + s_2 + s_3 + s_4$ with $s_i \in \mathbf{Z}^V$ and $\|s_i\|_1 = 1$ for $i = 1,2,3,4$.

Proof. Note that $x + s_i + s_j = y - s_k - s_l$ if $\{i,j,k,l\} = \{1,2,3,4\}$. (M-EXC$_{\mathrm{loc}}$) for f is equivalent to that for $f[-p]$, which implies

$$\begin{aligned} f_p(y) - f_p(x) \geq \min[&f_p(x, s_1+s_2) + f_p(x, s_3+s_4),\\ &f_p(x, s_1+s_3) + f_p(x, s_2+s_4),\\ &f_p(x, s_1+s_4) + f_p(x, s_2+s_3)]. \end{aligned}$$ ∎

To prove by contradiction, we assume that there exists a pair (x,y) for which (M-EXC) fails. That is, we assume that the set of such pairs,

$$\begin{aligned} \mathcal{D} = \{(x,y) \mid\ & x,y \in J,\ \exists s_* : (x,y)\text{-increment such that}\\ & \forall t : (x+s_*, y)\text{-increment} : f(x, s_*+t) + f(y, -s_*-t) > 0\}, \end{aligned}$$

is nonempty. Take a pair $(x,y) \in \mathcal{D}$ with minimum $\|x-y\|_1$; we have $\|x-y\|_1 > 4$ by (M-EXC$_{\mathrm{loc}}$). Let s_* be an (x,y)-increment satisfying the condition above, and put $u_* = \mathrm{supp}(s_*)$. Denoting the set of $(x+s_*, y)$-increments by I, we have

$$f(x, s_*+t) + f(y, -s_*-t) > 0 \qquad (t \in I). \tag{6.10}$$

Put $U = \mathrm{supp}(y-x)$ and, for $v \in U$, let t_v denote the (uniquely determined) (x,y)-increment such that $\mathrm{supp}(t_v) = v$; we have $t_v = \sigma(v)\chi_v$ using the notation σ defined by $\sigma(v) = 1$ for $v \in \mathrm{supp}^+(y-x)$ and $\sigma(v) = -1$ for $v \in \mathrm{supp}^-(y-x)$. Using this convention, define $\alpha \in \mathbf{R}$ by

$$\alpha = \begin{cases} f(x, 2s_*)/2 & (s_* \in I,\ x + 2s_* \in J),\\ (-f(y, -2s_*) + \varepsilon)/2 & (s_* \in I,\ x + 2s_* \notin J,\ y - 2s_* \in J),\\ 0 & (\text{otherwise}), \end{cases}$$

and $p \in \mathbf{R}^V$ by

$$\sigma(v)p(v) = \begin{cases} \alpha & (v = u_*),\\ f(x, s_*+t_v) - \alpha & (v \in U \setminus \{u_*\}, x + s_* + t_v \in J),\\ -f(y, -s_*-t_v) - \alpha + \varepsilon & (v \in U \setminus \{u_*\}, x + s_* + t_v \notin J,\\ & \quad y - s_* - t_v \in J),\\ 0 & (\text{otherwise}) \end{cases}$$

with some $\varepsilon > 0$.

Claim 1.

$$f_p(x, s_* + t) = 0 \qquad \text{if } t \in I, x + s_* + t \in J, \tag{6.11}$$

$$f_p(y, -s_* - t) > 0 \qquad \text{if } t \in I. \tag{6.12}$$

The equality (6.11) follows from

$$\begin{aligned} &f_p(x, s_* + t) = f(x, s_* + t) - \langle p, s_* \rangle - \langle p, t \rangle \\ &= \begin{cases} f(x, 2s_*) - 2\alpha = 0 & \text{if } t = s_*, \\ f(x, s_* + t) - \alpha - [f(x, s_* + t) - \alpha] = 0 & \text{if } t \neq s_*. \end{cases} \end{aligned}$$

The inequality (6.12) can be shown as follows. We may assume $y - s_* - t \in J$, since otherwise $f_p(y, -s_* - t) = +\infty$. If $x + s_* + t \in J$, we have $f_p(x, s_* + t) = 0$ by (6.11) and

$$f_p(x, s_* + t) + f_p(y, -s_* - t) = f(x, s_* + t) + f(y, -s_* - t) > 0$$

by (6.8) and (6.10). Otherwise ($y - s_* - t \in J$ and $x + s_* + t \notin J$), we have

$$\begin{aligned} &f_p(y, -s_* - t) = f(y, -s_* - t) + \langle p, s_* \rangle + \langle p, t \rangle \\ &= \begin{cases} f(y, -2s_*) + 2\alpha = \varepsilon & \text{if } t = s_*, \\ f(y, -s_* - t) + \alpha + [-f(y, -s_* - t) - \alpha + \varepsilon] = \varepsilon & \text{if } t \neq s_*. \end{cases} \end{aligned}$$

Next, let P denote the set of $(x + s_*, y)$-increment pairs.

Claim 2. There exists $(s_0, t_0) \in P$ such that $y - s_0 - t_0 \in J$ and

$$f_p(y, -s_0 - t_0) \leq f_p(y, -s - t) \qquad (\forall (s, t) \in P). \tag{6.13}$$

Since s_* is an (x, y)-increment and J satisfies (J-EXC), there exists $t_* \in I$ such that $x + s_* + t_* \in J$, where t_* may possibly be identical to s_*. We see that $x + s_* + t_*$ is distinct from y since $\|x - y\|_1 > 4$. By (J-EXC) there exists an $(x + s_* + t_*, y)$-increment pair (s, t) such that $y - s - t \in J$. This shows the existence of $(s, t) \in P$ with $y - s - t \in J$. Then (6.13) is satisfied by the pair $(s, t) = (s_0, t_0)$ that minimizes $f_p(y, -s - t)$ over $(s, t) \in P$.

Claim 3. $(x, y') \in \mathcal{D}$ with $y' = y - s_0 - t_0$.

To show this, first note that s_* is an (x, y')-increment, and let t be an $(x + s_*, y')$-increment. We have $t \in I$, $(s_0, t) \in P$, and $(t_0, t) \in P$. Hence, by (6.13), we have

$$f_p(y, -s_0 - t_0) \leq f_p(y, -s_0 - t), \quad f_p(y, -s_0 - t_0) \leq f_p(y, -t_0 - t). \tag{6.14}$$

Suppose that $x + s_* + t \in J$ and $y' - s_* - t \in J$. From (6.8), (6.11), Lemma 6.2, (6.12), and (6.14) we obtain

$$\begin{aligned} &f(x, s_* + t) + f(y', -s_* - t) \\ &= f_p(x, s_* + t) + f_p(y', -s_* - t) \\ &= f_p(y', -s_* - t) \\ &= f_p(y - s_0 - t_0 - s_* - t) - f_p(y - s_0 - t_0) \\ &\geq \min[f_p(y, -s_0 - t_0) + f_p(y, -s_* - t), \\ &\qquad f_p(y, -s_0 - t) + f_p(y, -s_* - t_0), \\ &\qquad f_p(y, -t_0 - t) + f_p(y, -s_* - s_0)] \\ &\quad - f_p(y, -s_0 - t_0) \\ &> \min[f_p(y, -s_0 - t_0), f_p(y, -s_0 - t), f_p(y, -t_0 - t)] \\ &\quad - f_p(y, -s_0 - t_0) \\ &= 0. \end{aligned}$$

This shows $(x, y') \in \mathcal{D}$.

Finally, since $\|x-y'\|_1 = \|x-y\|_1-2$, Claim 3 contradicts our choice of $(x, y) \in \mathcal{D}$. Therefore we conclude $\mathcal{D} = \emptyset$, completing the proof of Theorem 2.3.

Acknowledgments. The author thanks Jim Geelen and Satoru Iwata for discussions when we were at RIMS, Kyoto University, in April and May 1996. He is also grateful to András Sebő for a stimulating comment that led to Proposition 3.1, and to Akihisa Tamura and Ken'ichiro Tanaka for checking the proofs.

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] A. Bouchet, *Greedy algorithm and symmetric matroids*, Math. Programming, 38 (1987), pp. 147–159.
[4] A. Bouchet and W. H. Cunningham, *Delta-matroids, jump systems, and bisubmodular polyhedra*, SIAM J. Discrete Math., 8 (1995), pp. 17–32.
[5] R. Chandrasekaran and S. N. Kabadi, *Pseudomatroids*, Discrete Math., 71 (1988), pp. 205–217.
[6] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, and A. Schrijver, *Combinatorial Optimization*, John Wiley and Sons, New York, 1998.
[7] A. W. M. Dress and T. Havel, *Some combinatorial properties of discriminants in metric vector spaces*, Adv. Math., 62 (1986), pp. 285–312.
[8] A. W. M. Dress and W. Wenzel, *Valuated matroid: A new look at the greedy algorithm*, Appl. Math. Lett., 3 (1990), pp. 33–35.
[9] A. W. M. Dress and W. Wenzel, *Valuated matroids*, Adv. Math., 93 (1992), pp. 214–250.
[10] A. W. M. Dress and W. Wenzel, *A greedy-algorithm characterization of valuated Δ-matroids*, Appl. Math. Lett., 4 (1991), pp. 55–58.
[11] S. Fujishige, *Submodular Functions and Optimization*, 2nd ed., Ann. Discrete Math. 58, Elsevier, Amsterdam, 2005.
[12] J. F. Geelen, private communication, 1996.
[13] A. M. H. Gerards, *Matching*, 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, 1995, pp. 135–224.
[14] S. N. Kabadi and R. Sridhar, *Δ-matroid and jump system*, J. Appl. Math. Decis. Sci., 9 (2005), pp. 95–106.
[15] E. L. Lawler, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart and Winston, New York, 1976; Dover, Mineola, NY, 2001.
[16] L. Lovász, *The membership problem in jump systems*, J. Combin. Theory Ser. B, 70 (1997), pp. 45–66.
[17] K. Murota, *Two algorithms for valuated delta-matroids*, Appl. Math. Lett., 9 (1996), pp. 67–71.
[18] K. Murota, *Valuated matroid intersection* I: *Optimality criteria*, SIAM J. Discrete Math., 9 (1996), pp. 545–561.
[19] K. Murota, *Convexity and Steinitz's exchange property*, Adv. Math., 124 (1996), pp. 272–311.
[20] K. Murota, *Matrices and Matroids for Systems Analysis*, Springer-Verlag, Berlin, 2000.
[21] K. Murota, *Discrete Convex Analysis*, Monogr. Discrete Math. Appl. 10, SIAM, Philadelphia, 2003.
[22] K. Murota and K. Tanaka, *A steepest descent algorithm for M-convex functions on jump systems*, IEICE Trans. Fundamentals of Electr., Commun., Comput. Sci., E89-A (2006), to appear.
[23] W. R. Pulleyblank, *Matchings and extensions*, in Handbook of Combinatorics, Vol. 1, D. Graham, M. Grötschel, and L. Lovász, eds., Elsevier, Amsterdam, 1995, pp. 179–232.
[24] W. Wenzel, *Pfaffian forms and Δ-matroids*, Discrete Math., 115 (1993), pp. 253–266.
[25] W. Wenzel, *Δ-matroids with the strong exchange conditions*, Appl. Math. Lett., 6 (1993), pp. 67–70.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 227–239

ON STABILITY, ERROR CORRECTION, AND NOISE COMPENSATION IN DISCRETE TOMOGRAPHY*

ANDREAS ALPERS† AND PETER GRITZMANN†

Abstract. The task of reconstructing binary images from the knowledge of their line sums (discrete X-rays) in a given finite number m of directions is ill-posed. Even some small noise in the physical measurements can lead to dramatically different yet still unique solutions.

The present paper addresses in particular the following problems. Does discrete tomography have the power of error correction? Can noise be compensated by taking more X-ray images, and, if so, what is the quantitative effect of taking one more X-ray? Our main theorem gives the first nontrivial unconditioned (and best possible) stability result. In particular, we show that the Hamming distance between any two different sets of m X-ray images of the same cardinality is at least $2(m-1)$, and this is best possible. As a consequence, this result implies a Rényi-type theorem for denoising and shows that the noise compensating effect of X-rays is linear in their number.

Our theoretical results are complemented by determining the computational complexity of some underlying algorithmic tasks. In particular, we show that while there always is a certain inherent stability, the possibility of making (worst-case) efficient use of it is rather limited.

Key words. discrete tomography, stability, discrete inverse problems, computational complexity

AMS subject classifications. 90C31, 68R05, 11P05

DOI. 10.1137/040617443

1. Introduction.

Discrete tomography deals with the reconstruction of finite sets from knowledge about their interaction with certain query sets. The most prominent example is that of the reconstruction of a finite subset F of $\mathbb{Z}^d$ from its X-rays (i.e., line sums) in a small positive integer number m of directions. Applications of discrete tomography include quality control in semiconductor industry, image processing, graph theory, scheduling, statistical data security, game theory, etc. (see, e.g., [6], [8], [9], [13], [14], [17], [19]). The reconstruction task is an ill-posed discrete inverse problem, depicting (suitable variants of) all three Hadamard criteria [12] for ill-posedness. In fact, for general data there need not exist a solution, if the data is consistent, the solutions need not be uniquely determined, and even in the case of uniqueness, the solution may change dramatically with small changes of the data.

The papers [1] and [2] show just how unstable the reconstruction task really is: For arbitrarily large lattice sets even of the same cardinality, a total error of only $2(m-1)$ in the measurements can lead to unique but disjoint solutions. Clearly, this is an important issue for all practical applications where noise in the data cannot be avoided, particularly if the data stems from physical measurements.

The main theorem of the present paper shows that this number $2(m-1)$ is best possible in an ultimate sense. In Theorem 2.1 we prove that two finite sets of the same cardinality whose X-rays in a given set of m directions differ by a total of less than $2(m-1)$ are "tomographically equivalent." This means that either the X-rays differ by at least $2(m-1)$, or they do not differ at all. Note that the situation becomes trivial if the assumption on the equal cardinality of the lattice sets is omitted. Indeed, if the cardinalities of the two sets differ by k, then the total difference of the X-rays is

*Received by the editors October 21, 2004; accepted for publication (in revised form) August 8, 2005; published electronically March 15, 2006.
http://www.siam.org/journals/sidma/20-1/61744.html
†Zentrum Mathematik, Technische Universität München, Boltzmannstr. 3, D-85747 Garching bei München, Germany (alpers@ma.tum.de, gritzman@ma.tum.de).

at least km, and this is best possible (just delete k points of an arbitrary finite lattice set of cardinality at least k to obtain the second set).

Theorem 2.1 enables us to derive stability versions of all known uniqueness theorems, providing uniqueness even for somewhat noisy data. Complementing the theoretical results, we deal with the computational complexity of trying to take advantage of the inherent stability. The precise statements of our results will be given in the next section. Here we only summarize them qualitatively.

While it is clear that the total sum over all X-rays is a multiple of m and hence a small enough error in this number can be corrected, the problem of determining how the individual measurements should be corrected in order to provide consistency of the data is $\mathbb{NP}$-complete whenever $m \geq 3$ but easy for $m \leq 2$. Also, finding a set which best fits the data is $\mathbb{NP}$-hard for $m \geq 3$ but can be solved in polynomial time for $m \leq 2$.

The paper is organized as follows: After introducing some notation we state our main stability theorem, some of its corollaries, and the related algorithmic results in section 2. In sections 3 and 4 we give the proofs of our stability result and of the algorithmic results, respectively.

2. Main results: A stability theorem and some of its relatives. Let $d, m \in \mathbb{N}$, $d \geq 2$, and let $\mathbb{F}$ be a field with $\mathbb{Z} \subseteq \mathbb{F}$. Our underlying vector space will always be $\mathbb{F}^d$ but certain restrictions to the subring $\mathbb{Z}^d$ of all lattice points will also be relevant. Hence we will formulate some definitions and results in terms of $\mathbb{K} \in \{\mathbb{F}, \mathbb{Z}\}$. In particular, set

$$\mathcal{F}^d(\mathbb{K}) = \{F : F \subset \mathbb{K}^d \wedge F \text{ is finite}\}$$

and $\mathcal{F}^d = \mathcal{F}^d(\mathbb{Z})$. The elements of $\mathcal{F}^d$ are called *lattice sets*. Let $\mathcal{S}^d$ denote the set of all 1-dimensional linear subspaces of $\mathbb{F}^d$, and let $\mathcal{L}^d$ be the subset of $\mathcal{S}^d$ of all such subspaces that are spanned by vectors from $\mathbb{Z}^d$. The elements of $\mathcal{L}^d$ will be referred to as *lattice lines*. Further, for $S \in \mathcal{S}^d$ let $\mathcal{A}_{\mathbb{K}}(S) = \{v + S : v \in \mathbb{K}^d\}$.

Then, for $F \in \mathcal{F}^d(\mathbb{K})$ and $S \in \mathcal{S}^d$, the *(discrete 1-dimensional) X-ray of F parallel to S* is the function

$$X_S F : \mathcal{A}_{\mathbb{K}}(S) \to \mathbb{N}_0 = \mathbb{N} \cup \{0\}$$

defined by

$$X_S F(T) = |F \cap T| = \sum_{x \in T} \mathbf{1}_F(x)$$

for each $T \in \mathcal{A}_{\mathbb{K}}(S)$.

Two sets $F_1, F_2 \in \mathcal{F}^n(\mathbb{F})$ are called *tomographically equivalent* with respect to $S_1, \ldots, S_m \in \mathcal{S}^d$ if $X_{S_i} F_1 = X_{S_i} F_2$ for $i = 1, \ldots, m$.

Given m different lines $S_1, \ldots, S_m \in \mathcal{S}^d$, the basic questions in discrete tomography are as follows. What kind of information about a finite (lattice) set $F \in \mathbb{K}^d$ can be retrieved from its X-ray images $X_{S_1} F, \ldots, X_{S_m} F$? How difficult is the reconstruction algorithmically? How sensitive is the task to data errors? Here the data is given in terms of functions

$$f_i : \mathcal{A}_{\mathbb{K}}(S_i) \to \mathbb{N}_0, \qquad i = 1, \dots, m,$$

with finite support $\mathcal{T}_i \subseteq \mathcal{A}_{\mathbb{K}}(S_i)$ represented by appropriately chosen data structures; see [8]. Hence the difference of two data functions with respect to the same line $S \in \mathcal{S}^d$ is a function $h : \mathcal{A}_{\mathbb{K}}(S) \to \mathbb{Z}$; its size will be measured in terms of its ℓ_1-norm

$$\|h\|_1 = \sum_{T \in \mathcal{A}_{\mathbb{K}}(S)} |h(T)|.$$

For surveys on various aspects of discrete tomography see [10], [11], [13].

Our main stability result can now be formulated as follows.

THEOREM 2.1. *Let $S_1, \dots, S_m \in \mathcal{S}^d$ be different and $F_1, F_2 \in \mathcal{F}^d(\mathbb{K})$ with $|F_1| = |F_2|$. If*

$$\sum_{i=1}^{m} ||X_{S_i} F_1 - X_{S_i} F_2||_1 < 2(m-1),$$

then F_1 and F_2 are tomographically equivalent.

The proof will be given in section 3. Clearly, Theorem 2.1 is equivalent to the following theorem.

THEOREM 2.2. *Let $S_1, \dots, S_m \in \mathcal{S}^d$ be different. Then there do not exist $F_1, F_2 \in \mathcal{F}^d(\mathbb{K})$ with $|F_1| = |F_2|$ and $0 < \sum_{i=1}^{m} ||X_{S_i} F_1 - X_{S_i} F_2||_1 < 2(m-1)$.*

As corollaries to this stability result we may derive "noisy versions" of all known uniqueness theorems. In the following we give two such examples.

Rényi's well-known theorem [16] states that if we know the cardinality $|F|$ of a finite set F we can guarantee uniqueness from X-rays taken in any $m \geq |F| + 1$ different directions. Our first corollary shows that we can guarantee uniqueness, *even if the X-rays are not given precisely.*

COROLLARY 2.3. *Let $F_1, F_2 \in \mathcal{F}^d(\mathbb{K})$ with $|F_1| = |F_2|$, $m \in \mathbb{N}$ with $m \geq |F_1| + 1$, and let $S_1, \dots, S_m \in \mathcal{S}^d$ be different. If $\sum_{i=1}^{m} ||X_{S_i} F_1 - X_{S_i} F_2||_1 < 2|F_1|$, then $F_1 = F_2$.*

Proof. By Theorem 2.1, F_1 and F_2 are tomographically equivalent; hence the assertion follows from Rényi's theorem [16]. □

Corollary 2.3 shows the potential power of error correction in the setting of Rényi's theorem: A total error smaller than $2n$ can be compensated without increasing the number of X-rays taken if the cardinality n of the original set F is known. But even without knowing n precisely we can correct errors—at the expense, however, of taking more X-rays.

COROLLARY 2.4. *Let $F_1, F_2 \in \mathcal{F}^d(\mathbb{K})$ with $|F_1| \leq |F_2|$, $m \in \mathbb{N}$ with $m \geq 2|F_1|$, and let $S_1, \dots, S_m \in \mathcal{S}^d$ be different. Then $\sum_{i=1}^{m} ||X_{S_i} F_1 - X_{S_i} F_2||_1 < 2|F_1|$ implies $F_1 = F_2$.*

Proof. Clearly $\sum_{i=1}^{m} ||X_{S_i} F_1 - X_{S_i} F_2||_1 \geq m(|F_2| - |F_1|)$. Thus, $\sum_{i=1}^{m} ||X_{S_i} F_1 - X_{S_i} F_2||_1 < 2|F_1|$ implies $|F_1| = |F_2|$, and the assertion follows from Corollary 2.3. □

Next we give a stable version of a theorem of Gardner and Gritzmann [7] for the set $\mathcal{C}^d$ of *convex lattice sets*, i.e., of sets $F \in \mathcal{F}^d$ with $F = \text{conv}(F \cap \mathbb{Z}^d)$.

COROLLARY 2.5. *Let $F_1, F_2 \in \mathcal{C}^d$ with $|F_1| = |F_2|$.*

(i) *There are sets $\{S_1, S_2, S_3, S_4\} \subseteq \mathcal{L}^d$ of four lines such that $\sum_{i=1}^{4} ||X_{S_i} F_1 - X_{S_i} F_2||_1 < 6$ implies $F_1 = F_2$.*

(ii) *For any set $\{S_1, \dots, S_m\} \subseteq \mathcal{L}^d$ of $m \geq 7$ coplanar lattice lines, $\sum_{i=1}^{m} ||X_{S_i} F_1 - X_{S_i} F_2||_1 < 2(m-1)$ implies $F_1 = F_2$.*

Proof. By Theorem 2.1, F_1 and F_2 are tomographically equivalent in both parts of the statement; hence the assertion follows from the uniqueness theorems of [7]. □

Note that this theorem also holds for the somewhat more general class of *Q-convex lattice sets* because they are uniquely determined by the same sets of lattice lines as the convex lattice sets (see [5]).

Let us now turn to results on some algorithmic tasks related to stability and instability in discrete tomography. We concentrate on the case of finite lattice sets whose X-rays are taken in lattice directions. Thus, let $S_1, \ldots, S_m \in \mathcal{L}^d$. Proofs of the following statements will be given in section 4.

We begin with two examples of algorithmic consequences of Theorem 2.1, "noisy extensions" of known complexity results. It has been shown in [8] that the two problems

CONSISTENCY$_{\mathcal{F}^d}(S_1, \ldots, S_m)$
Input: For $i = 1, \ldots, m$ data functions $f_i : \mathcal{A}_{\mathbb{Z}}(S_i) \to \mathbb{N}_0$ with finite support.
Question: Does there exist a finite lattice set $F \in \mathcal{F}^d$ such that $X_{S_i}F = f_i$ for $i = 1, \ldots, m$?

and

UNIQUENESS$_{\mathcal{F}^d}(S_1, \ldots, S_m)$
Input: A set $F_1 \in \mathcal{F}^d$.
Question: Does there exist a set $F_2 \in \mathcal{F}^d$ with $F_1 \neq F_2$ such that $X_{S_i}F_1 = X_{S_i}F_2$ for $i = 1, \ldots, m$?

can be solved in polynomial time for $m \leq 2$ but are $\mathbb{NP}$-complete for $m \geq 3$.

With the aid of Theorem 2.1 these results can be extended as follows.

COROLLARY 2.6. *Let $S_1, \ldots, S_m \in \mathcal{L}^d$ be different. The two problems*

X-RAY-CORRECTION$_{\mathcal{F}^d}(S_1, \ldots, S_m)$
Input: *For every $i = 1, \ldots, m$ a data function $f_i : \mathcal{A}_{\mathbb{Z}}(S_i) \to \mathbb{N}_0$ with finite support.*
Question: *Does there exist a finite lattice set $F \in \mathcal{F}^d$ with $\sum_{i=1}^m ||X_{S_i}F - f_i||_1 \leq m - 1$?*

and

SIMILAR-SOLUTION$_{\mathcal{F}^d}(S_1, \ldots, S_m)$
Input: *A finite lattice set $F_1 \in \mathcal{F}^d$.*
Question: *Does there exist a finite lattice set $F_2 \in \mathcal{F}^d$ with $|F_1| = |F_2|$ and $F_1 \neq F_2$ such that $\sum_{i=1}^m ||X_{S_i}F_1 - X_{S_i}F_2||_1 \leq 2m - 3$?*

are in $\mathbb{P}$ for $m \leq 2$ but are $\mathbb{NP}$-complete for $m \geq 3$.

Note that X-RAY-CORRECTION$_{\mathcal{F}^d}(S_1, \ldots, S_m)$ can also be formulated as the task to decide, for given data functions $f_i : \mathcal{A}_{\mathbb{Z}}(S_i) \to \mathbb{N}_0$ $(i = 1, \ldots, m)$ with finite support, whether there exist "corrected" data functions $g_i : \mathcal{A}_{\mathbb{Z}}(S_i) \to \mathbb{N}_0$ $(i = 1, \ldots, m)$ with finite support that are consistent and do not differ from the given functions by more than a total of $m - 1$. Corollary 2.6 shows that this form of measurement correction is just as hard as checking consistency.

If the data is noisy it seems natural to try to find a finite lattice set that fits the measurements best. This task is studied in the following theorem.

THEOREM 2.7. *Let $S_1, \dots, S_m \in \mathcal{L}^d$ be different. The problem*

NEAREST-SOLUTION$_{\mathcal{F}^d}(S_1, \dots, S_m)$
Input: *For every $i = 1, \dots, m$, a data function $f_i : \mathcal{A}_{\mathbb{Z}}(S_i) \to \mathbb{N}_0$ with finite support.*
Task: *Determine a set $F^* \in \mathcal{F}^d$ such that $\sum_{i=1}^m ||X_{S_i} F^* - f_i||_1 = \min_{F \in \mathcal{F}^d} \sum_{i=1}^m ||X_{S_i} F - f_i||_1$*

is in $\mathbb{P}$ for $m \leq 2$ but is $\mathbb{NP}$-hard for $m \geq 3$.

From the $\mathbb{NP}$-hardness of CONSISTENCY$_{\mathcal{F}^d}(S_1, \dots, S_m)$ the statement for $m \geq 3$ follows easily. In fact, for a given instance $(f_1, \dots, f_m)$ of CONSISTENCY$_{\mathcal{F}^d}(S_1, \dots, S_m)$ let F^* denote a solution of NEAREST-SOLUTION$_{\mathcal{F}^d}(S_1, \dots, S_m)$ for the input $(f_1, \dots, f_m)$. Then $(f_1, \dots, f_m)$ is a yes-instance of CONSISTENCY$_{\mathcal{F}^d}(S_1, \dots, S_m)$ if and only if $X_{S_i} F^* = f_i$ for all $i = 1, \dots, m$. However, the proof of the polynomial-time solvability in the case $m = 2$ is more involved and will be given in section 4.

3. Proof of the main stability result. Note first that it is enough to prove Theorem 2.1 for $\mathbb{K} = \mathbb{F}$. The proof will be based on four lemmas. The first lemma is a simple combinatorial observation.

LEMMA 3.1. *Let $S \in \mathcal{S}^d$ and let $f, g : \mathcal{A}_{\mathbb{F}}(S) \to \mathbb{N}_0$ be data functions with finite support. Further, set $\mathcal{A}^+ = \{T \in \mathcal{A}_{\mathbb{F}}(S) : f(T) - g(T) > 0\}$ and $\mathcal{A}^- = \{T \in \mathcal{A}_{\mathbb{F}}(S) : f(T) - g(T) < 0\}$. Then*

$$||f - g||_1 = 2 \sum_{T \in \mathcal{A}^+} \big(f(T) - g(T)\big) - \|f\|_1 + \|g\|_1.$$

In particular, when $\|f\|_1 = \|g\|_1$ the number $||f - g||_1$ is even.

Proof. Since

$$\sum_{T \in \mathcal{A}_{\mathbb{F}}(S)} (f(T) - g(T)) = \sum_{T \in \mathcal{A}_{\mathbb{F}}(S)} f(T) - \sum_{T \in \mathcal{A}_{\mathbb{F}}(S)} g(T) = \|f\|_1 - \|g\|_1,$$

we have

$$\begin{aligned} ||f - g||_1 &= \sum_{T \in \mathcal{A}_{\mathbb{F}}(S)} |f(T) - g(T)| = \sum_{T \in \mathcal{A}^+} (f(T) - g(T)) - \sum_{T \in \mathcal{A}^-} (f(T) - g(T)) \\ &= \sum_{T \in \mathcal{A}^+} (f(T) - g(T)) - \sum_{T \in \mathcal{A}^-} (f(T) - g(T)) + \sum_{T \in \mathcal{A}^+} (f(T) - g(T)) \\ &\quad + \sum_{T \in \mathcal{A}^-} (f(T) - g(T)) - \|f\|_1 + \|g\|_1 \\ &= 2 \sum_{T \in \mathcal{A}^+} (f(T) - g(T)) - \|f\|_1 + \|g\|_1. \qquad \square \end{aligned}$$

In the present section we will apply Lemma 3.1 to the X-rays of sets $F_1, F_2 \in \mathcal{F}^d(\mathbb{F})$, i.e., to $f = X_S F_1$ and $g = X_S F_2$.

The next lemma is geometric in nature and will enable us to reduce the proof of Theorem 2.1 to the planar case.

LEMMA 3.2. *Let $d \geq 3$, $S_1, \dots, S_m \in \mathcal{S}^d$ be different and $F_1, F_2 \in \mathcal{F}^d(\mathbb{F})$. Then there exists a surjective linear map $\varphi : \mathbb{F}^d \to \mathbb{F}^2$ with the following properties.*

(i) *$\varphi(S_1), \dots, \varphi(S_m)$ are different lines in $\mathcal{S}^2$.*

(ii) *If $i \in \{1, \dots, m\}$ and $a, b \in F_1 \cup F_2$ satisfy $\varphi(b) \in \varphi(a) + \varphi(S_i)$, then $b \in a + S_i$.*

Proof. In order to satisfy the two properties the kernel $\ker(\varphi)$ will be chosen complementary to any plane spanned by two of the m lines, and also complementary to any plane spanned by one of the lines $S_1, \dots, S_m$ and a line generated by the difference of two of the vectors of $F_1 \cup F_2$. Let us denote the set of these exceptional planes by $\mathcal{P}$. Each of the planes $P \in \mathcal{P}$ can be described as the set of solutions of a homogeneous $(d-2) \times d$ system of linear equations; let A_P denote a corresponding coefficient matrix. Now, let $\pi_1, \dots, \pi_{2d}$ be different primes. Further, for $x \in \mathbb{F}$ let $B(x)$ be the $2 \times d$ matrix with row vectors $(x^{\pi_1}, x^{\pi_2}, \dots, x^{\pi_d})$ and $(x^{\pi_{d+1}}, x^{\pi_{d+2}}, \dots, x^{\pi_{2d}})$, and let $H(x)$ be the solution space of the corresponding homogeneous $2 \times d$ system. Then for each $P \in \mathcal{P}$ the determinant of the matrix composed of A_P and $B(x)$ is a nontrivial polynomial in x. (In fact, the coefficients are $(d-2) \times (d-2)$ subdeterminants of A_P, and by the choice of the exponents of x in $B(x)$ there is generically no cancellation.) Hence for all sufficiently large integers x, $H(x)$ is complementary to each plane $P \in \mathcal{P}$. Now taking a fixed such vector x, we define φ by choosing an arbitrary basis of $H(x)$, extend it to a basis of $\mathbb{F}^d$, and specify that φ maps the basis vectors of $H(x)$ to 0 and the remaining two to the standard basis vectors of $\mathbb{F}^2$. Then $\ker(\varphi) = H(x)$, whence φ has the desired properties. □

Note that a linear mapping φ with the properties of Lemma 3.2 is necessarily injective on $F_1 \cup F_2$.

The following two lemmas are more algebraic in nature. The next contains a well-known result on the elementary part of the Prouhet–Tarry–Escott Problem on solutions of a specific power system of polynomial equations. As a service to the reader we still outline the proof. For a survey on the Prouhet–Tarry–Escott Problem see [3] or [4].

LEMMA 3.3. *Let $x_1, \dots, x_q, y_1, \dots, y_q \in \mathbb{F}$ such that*

$$\sum_{i=1}^{q} x_i^j = \sum_{i=1}^{q} y_i^j$$

for $j = 1, \dots, q$. Then the multisets $\{x_1, \dots, x_q\}$ and $\{y_1, \dots, y_q\}$ coincide.

Proof. We show that $x_1, \dots, x_q$ and $y_1, \dots, y_q$ are the roots of the same polynomial of degree q.

For $i = 1, \dots, q$ let $p_i, s_i \in \mathbb{F}[X_1, \dots, X_q]$ be defined by

$$p_i = X_1^i + X_2^i + \cdots + X_q^i, \qquad s_i = \sum_{1 \le k_1 < \cdots < k_i \le q} X_{k_1} \cdots X_{k_i}.$$

The polynomials p_i and s_i are the well-known *power sums* and *elementary symmetric functions* of the indeterminates $X_1, \dots, X_q$, respectively. Clearly, for the indeterminates $X_1, \dots, X_q, Y$ we have

$$\prod_{i=1}^{q} (Y - X_i) = Y^q - s_1 Y^{q-1} + s_2 Y^{q-2} + \cdots + (-1)^q s_q.$$

Using the *Newton identities* (see, e.g., [15]) it follows inductively that for $i = 1, \dots, q$

$$s_i \in \mathbb{F}[p_1, \dots, p_q].$$

Since by assumption

$$p_i(x_1, \dots, x_q) = p_i(y_1, \dots, y_q) \quad \text{for } i = 1, \dots, q,$$

this implies

$$s_i(x_1,\dots,x_q) = s_i(y_1,\dots,y_q) \quad \text{for } i = 1,\dots,q.$$

Consequently,

$$\prod_{i=1}^{q}(Y - x_i) = \sum_{i=0}^{q}(-1)^i Y^{q-i} s_i(x_1,\dots,x_q) = \prod_{i=1}^{q}(Y - y_i);$$

i.e., the two polynomials $\prod_{i=1}^{q}(Y-x_i)$ and $\prod_{i=1}^{q}(Y-y_i)$ in $\mathbb{F}[Y]$ are identical. Hence $x_1,\dots,x_q$ is just a permutation of $y_1,\dots,y_q$. □

LEMMA 3.4. *Let $k \in \mathbb{N}$ and $\sigma_1,\dots,\sigma_{k+1},\tau_1,\dots,\tau_{k+1} \in \mathbb{F}$ such that $S_i = \operatorname{lin}\left\{(\sigma_i,\tau_i)^T\right\} \in \mathcal{S}^2$, $i = 1,\dots,k+1$, are different. Then*

$$(\tau_1 X - \sigma_1 Y)^k, \dots, (\tau_{k+1} X - \sigma_{k+1} Y)^k \in \mathbb{F}[X,Y]$$

form a basis of the $\mathbb{F}$-vector space V_k that is generated by the $k+1$ binomials Y^k, $X^1Y^{k-1},\dots,X^{k-1}Y^1$, $X^k \in \mathbb{F}[X,Y]$.

Proof. Every polynomial $(\tau_i X - \sigma_i Y)^k$ can be expressed in terms of its coefficient vector

$$\left(\binom{k}{0}\tau_i^0(-\sigma_i)^k, \dots, \binom{k}{k}\tau_i^k(-\sigma_i)^0\right)$$

with respect to the binomial basis $\{Y^k, X^1Y^{k-1},\dots,X^{k-1}Y^1, X^k\}$. Thus, we have to show only that these $k+1$ vectors are linearly independent, i.e., that the matrix

$$C = \left(\binom{k}{j-1}(\tau_i)^{j-1}(-\sigma_i)^{k-j+1}\right)_{i,j=1,\dots,k+1} \in \mathbb{F}^{(k+1)\times(k+1)}$$

is nonsingular.

Suppose first that $\sigma_1 \cdots \sigma_{k+1} \neq 0$. By setting $\rho_i = -\sigma_i^{-1}\tau_i$, and by denoting the Vandermonde matrix $(\rho_i^{j-1})_{i,j=1,\dots,k+1}$ by C', we obtain

$$\det(C) = \det(C') \cdot \prod_{i=1}^{k+1}\binom{k}{i-1}(-\sigma_i)^k = \prod_{i>j}(\rho_i - \rho_j) \cdot \prod_{i=1}^{k+1}\binom{k}{i-1}(-\sigma_i)^k.$$

Thus, if $\det(C) = 0$, then there exist indices i_0, j_0 in $\{1,\dots,k+1\}$ with $i_0 \neq j_0$ but $\rho_{i_0} = \rho_{j_0}$. This means that $\sigma_{i_0}^{-1}\tau_{i_0} = \sigma_{j_0}^{-1}\tau_{j_0}$, whence $S_{i_0} = S_{j_0}$, contrary to the assumption. Therefore $\det(C) \neq 0$.

Now suppose that one of the σ_i is zero. Without loss of generality we may assume that $\sigma_1 = 0$. Note that then $\sigma_i \neq 0$ for $i > 1$. The first row of C is now a nonzero multiple of $(0,\dots,0,1)$. By developing $\det(C)$ with respect to the first row, we see that the same argument as in the first case applies again. □

Now we are ready to prove our main stability result.

Proof of Theorem 2.1. Let $F_1, F_2 \in \mathcal{F}^d(\mathbb{F})$ with $|F_1| = |F_2|$ and $0 < \sum_{i=1}^{m} ||X_{S_i}F_1 - X_{S_i}F_2||_1 < 2(m-1)$. By Lemma 3.1, this implies that $m \geq 3$.

Suppose first that the error involves more than one direction; i.e., $X_{S_i}F_1 \neq X_{S_i}F_2$ for at least two different indices i_1 and i_2. By Lemma 3.1, $\|X_{S_i}F_1 - X_{S_i}F_2\|_1 \geq 2$ for $i = i_1, i_2$. Therefore, ignoring S_{i_1}, the sets F_1 and F_2 provide a counterexample already for $m-1$ directions. Hence we may in the following assume that $X_{S_i}F_1 =$

$X_{S_i}F_2$ for $i = 1, \dots, m-1$; i.e., the error occurs only for S_m. Similarly, we may assume that the error is exactly $2(m-2)$.

Next, we reduce the statement to the planar case. Let $d \geq 3$ and suppose that $F_1, F_2 \in \mathcal{F}^d(\mathbb{F})$ with $|F_1| = |F_2|$ and $0 < \sum_{i=1}^m ||X_{S_i}F_1 - X_{S_i}F_2||_1 < 2(m-1)$. Let φ be a linear mapping according to Lemma 3.2, and set $F_j' = \varphi(F_j)$ for $j = 1, 2$ and $S_i' = \varphi(S_i)$ for $i = 1, \dots, m$. Then $F_1', F_2' \in \mathcal{F}^2(\mathbb{F})$, $|F_1'| = |F_2'|$, $S_1', \dots, S_m' \in \mathcal{S}^2$ are different, and $X_{S_i'}F_j' = X_{S_i}F_j$ for $i = 1, \dots, m$ and $j = 1, 2$. Hence we obtain a counterexample already in dimension 2.

Finally we turn to the planar case. So, in the following let $d = 2$. The n points of F_1 and F_2 will be denoted by $(x_1, y_1), \dots, (x_n, y_n)$ and $(x_1', y_1'), \dots, (x_n', y_n')$, respectively.

Let $\sigma_1, \dots, \sigma_m, \tau_1, \dots, \tau_m \in \mathbb{F}$ be such that $S_i = \operatorname{lin}\left\{(\sigma_i, \tau_i)^T\right\}$ for $i = 1, \dots, m$. By Lemma 3.4 we know that for $k = 1, \dots, m-2$

$$(\tau_1 X - \sigma_1 Y)^k, \dots, (\tau_{k+1} X - \sigma_{k+1} Y)^k$$

form a basis of the $\mathbb{F}$-vector space V_k generated by the binomials $Y^k, X^1Y^{k-1}, \dots, X^{k-1}Y^1, X^k$. Since, of course, $(\tau_m X - \sigma_m Y)^k \in V_k$, there are coefficients $\alpha_{1,k}, \dots, \alpha_{m-1,k} \in \mathbb{F}$ such that

$$(\tau_m X - \sigma_m Y)^k = \sum_{i=1}^{m-1} \alpha_{i,k} (\tau_i X - \sigma_i Y)^k.$$

For every line T parallel to any of the lines $S_1, \dots, S_{m-1}$ we have $|F_1 \cap T| = |F_2 \cap T|$. Hence, as multisets the projections of F_1 and F_2 parallel to S_i (on any line complementary to S_i) coincide for $i = 1, \dots, m-1$. Thus

$$\{(\tau_i x_1 - \sigma_i y_1), \dots, (\tau_i x_n - \sigma_i y_n)\} = \{(\tau_i x_1' - \sigma_i y_1'), \dots, (\tau_i x_n' - \sigma_i y_n')\}$$

for $i = 1, \dots, m-1$. As a consequence we have

$$\begin{aligned} &\sum_{j=1}^n \left((\tau_m x_j - \sigma_m y_j)^k - (\tau_m x_j' - \sigma_m y_j')^k\right) \\ &= \sum_{j=1}^n \sum_{i=1}^{m-1} \alpha_{i,k} \left((\tau_i x_j - \sigma_i y_j)^k - (\tau_i x_j' - \sigma_i y_j')^k\right) = 0 \end{aligned}$$

for each $k = 1, \dots, m-2$.

Now we define the multiset differences

$$A = \{(\tau_m x_1 - \sigma_m y_1), \dots, (\tau_m x_n - \sigma_m y_n)\} \setminus \{(\tau_m x_1' - \sigma_m y_1'), \dots, (\tau_m x_n' - \sigma_m y_n')\}$$

and

$$B = \{(\tau_m x_1' - \sigma_m y_1'), \dots, (\tau_m x_n' - \sigma_m y_n')\} \setminus \{(\tau_m x_1 - \sigma_m y_1), \dots, (\tau_m x_n - \sigma_m y_n)\}.$$

Note that $|A|$ and $|B|$ count the positive excess of F_1 over F_2 and of F_2 over F_1, respectively, on lines parallel to S_m. To be more precise, let $\mathcal{A}^+ = \{T \in \mathcal{A}_{\mathbb{F}}(S_m) : X_{S_m}F_1(T) - X_{S_m}F_2(T) > 0\}$ and $\mathcal{A}^- = \{T \in \mathcal{A}_{\mathbb{F}}(S_m) : X_{S_m}F_1(T) - X_{S_m}F_2(T) < 0\}$. Then with the aid of Lemma 3.1

$$|A| = \sum_{T \in \mathcal{A}^+} \left(X_{S_m}F_1(T) - X_{S_m}F_2(T)\right) = \frac{1}{2}||X_{S_m}F_1 - X_{S_m}F_2||_1;$$

similarly,

$$|B| = \sum_{T \in \mathcal{A}^-} \big(X_{S_m} F_2(T) - X_{S_m} F_1(T)\big) = \frac{1}{2} ||X_{S_m} F_1 - X_{S_m} F_2||_1.$$

Hence

$$|A| = |B| = m - 2$$

and thus, particularly, $A \neq B$. Using the notation $A = \{a_1, \ldots, a_q\}$ and $B = \{b_1, \ldots, b_q\}$ with $q = m - 2$, we have for each $k = 1, \ldots, q$

$$\sum_{j=1}^{n} \big((\tau_m x_j - \sigma_m y_j)^k - (\tau_m x'_j - \sigma_m y'_j)^k\big) = \sum_{j=1}^{q} a_j^k - \sum_{j=1}^{q} b_j^k = 0,$$

a contradiction to Lemma 3.3. This completes the proof of Theorem 2.1. □

4. Proofs of the algorithmic results. In the following we give the proofs for the algorithmic results stated in section 2. We begin with the membership of X-Ray-Correction$_{\mathcal{F}^d}(S_1, \ldots, S_m)$ and Similar-Solution$_{\mathcal{F}^d}(S_1, \ldots, S_m)$ in the class $\mathbb{NP}$. Given an instance $(f_1, \ldots, f_m)$ or F_1, respectively, one would, of course, like to use as a certificate a corresponding set F or F_2, respectively. If the set is available and polynomial in the encoding length, the conditions can be checked efficiently. Let us call a set F *support consistent* if for each of the m directions the support of the X-ray $X_{S_i} F$ is a subset of the support of the data function f_i, i.e.,

$$\{T \in \mathcal{A}_{\mathbb{Z}}(S_i) : X_{S_i} F(T) \neq 0\} \subset \mathcal{T}_i \quad \text{for } i = 1, \ldots, m,$$

where

$$\mathcal{T}_i = \{T \in \mathcal{A}_{\mathbb{Z}}(S_i) \,:\, f_i(T) \neq 0\} \quad \text{for } i = 1, \ldots, m.$$

In fact, every support consistent solution is a subset of the *grid*

$$G = \mathbb{Z}^d \cap \bigcap_{i=1}^{m} \bigcup_{T \in \mathcal{T}_i} T,$$

and G contains only polynomially many points $v_1, \ldots, v_k$ of polynomially bounded size.

Since, in general, errors are allowed we cannot restrict ourselves to support consistent solutions. But then not every solution must consist of lattice points whose binary size is bounded by a polynomial in the input. The next lemma shows, however, that there always exist solutions of polynomial size.

Lemma 4.1. *Let $\gamma \in \mathbb{N}$ be a constant. Further, for $i = 1, \ldots, m$ let $f_i : \mathcal{A}_{\mathbb{Z}}(S_i) \to \mathbb{N}_0$ be a data function with finite support, and let $F \in \mathcal{F}^d$ be such that*

$$\sum_{i=1}^{m} ||X_{S_i} F - f_i||_1 \leq \gamma \sum_{i=1}^{m} ||f_i||_1.$$

Then there exists a finite lattice set $F^* \in \mathcal{F}^d$ of binary size that is bounded by a polynomial in the binary size of $(f_1, \ldots, f_m)$ with

$$|F| = |F^*| \qquad \text{and} \qquad \sum_{i=1}^{m} ||X_{S_i} F^* - f_i||_1 = \sum_{i=1}^{m} ||X_{S_i} F - f_i||_1 \quad \text{for } i = 1, \ldots, m.$$

Proof. Without loss of generality we may assume that the grid G contains the origin. Now, for $v_1, v_2 \in G$ and $i, j = 1, \dots, m$ with $i \neq j$, the point of intersection of the two lines $v_1 + S_i$ and $v_2 + S_j$ has binary size that is bounded by a polynomial in the binary size of $(f_1, \dots, f_m)$. Hence there is a constant λ of polynomial size such that $\lambda[-1,1]^d$ contains all such intersections and such that for every $v \in G$ and $i = 1, \dots, m$ the line $v + S_i$ contains at least two lattice points of $\lambda[-1,1]^d$. Let

$$\mathcal{T} = G + \{S_1, \dots, S_m\}, \qquad k = \max\left\{m\lambda, \gamma \sum_{i=1}^{m} ||f_i||_1\right\}$$

and

$$W = (1+k)\lambda[-1,1]^d, \qquad C = W \setminus \big(\lambda[-1,1]^d\big).$$

Then each line $v + S_i$ with $v \in G$ intersects the annulus C in at least $2k$ lattice points. Now, if $q \in F \setminus W$, then there is at most one line in $\mathcal{T}$ that passes through q. We will successively replace the points of $F \setminus W$ by points in C. Let us deal first with those points of $F \setminus W$ which are met by one of the X-ray lines in $\mathcal{T}$. We replace such points q one by one by the lattice point of C closest to q on that line with smallest ℓ_∞ norm among all such points which have not previously been inserted. By the choice of k there are always enough points of C on each line.

After having handled all such points we replace all points $q \in F \setminus W$ that are not met by any of the X-ray lines by a set of points of the same cardinality on the boundary of W that is disjoint from any line in $\mathcal{T}$. An elementary lattice point count shows that by the choice of k a set of appropriate cardinality always exists. This way we obtain a finite lattice set F^* with $|F| = |F^*|$. By construction, the X-ray images of F and F^* coincide on each line of $\mathcal{T}$. Also the total sums for F and F^* on all other lines are the same. This proves the assertion. □

It follows now directly from Lemma 4.1 that X-Ray-Correction$_{\mathcal{F}^d}(S_1, \dots, S_m)$ and Similar-Solution$_{\mathcal{F}^d}(S_1, \dots, S_m)$ are indeed in $\mathbb{NP}$.

For $m = 2$ the result of Lemma 4.1 can be sharpened. It is not just possible to avoid points "too far out" but it suffices to consider only instances and solutions "with no empty line in between." To be precise, we call a data function $f : \mathcal{A}_{\mathbb{Z}^d}(S) \to \mathbb{N}_0$ *consecutive* if for $v_1, v_2, v_3 \in \mathbb{Z}^d$ it is true that $f(v_2 + S) \neq 0$ whenever $f(v_1 + S) \neq 0$, $f(v_3 + S) \neq 0$, and $v_2 + S \subset \operatorname{conv}(v_1 + S) \cup (v_3 + S)$. Further, an m-tuple $(f_1, \dots, f_m)$ of data functions with respect to $S_1, \dots, S_m$ is called *consecutive* if $f_1, \dots, f_m$ are consecutive. Similarly, a finite lattice set F is called *consecutive* if and only if $(X_{S_1}F, \dots, X_{S_m}F)$ is consecutive. It is clear that for $m = 2$ we can always replace a given instance of any of our problems by an equivalent consecutive one.

Now we can give the proof of Corollary 2.6.

Proof of Corollary 2.6. Let first $m \geq 3$ and let us begin with X-Ray-Correction$_{\mathcal{F}^d}$ $(S_1, \dots, S_m)$.

Let $(f_1, \dots, f_m)$ be an instance of Consistency$_{\mathcal{F}^d}(S_1, \dots, S_m)$. Then $(f_1, \dots, f_m)$ is also an instance of X-Ray-Correction$_{\mathcal{F}^d}(S_1, \dots, S_m)$. Suppose first that no set $F \in \mathcal{F}^d$ exists with $\sum_{i=1}^m ||X_{S_i}F - f_i||_1 \leq m - 1$. Then, of course, $(f_1, \dots, f_m)$ is a no-instance of Consistency$_{\mathcal{F}^d}(S_1, \dots, S_m)$.

Thus, suppose there is a set $F \in \mathcal{F}^d$ with $\sum_{i=1}^m ||X_{S_i}F - f_i||_1 \leq m - 1$. Let $||f_1|| = \cdots = ||f_m||$. In polynomial time we can construct a line $T^* \in \mathcal{A}_{\mathbb{Z}}(S_1)$ with

$$T^* \cap \bigcup_{T \in \mathcal{T}_i} T \cap \bigcup_{T \in \mathcal{T}_j} T = \emptyset \qquad \text{for all } i \neq j.$$

Now let $f_1^*(T) = f_1(T)$ for $T \in \mathcal{A}_{\mathbb{Z}}(S_1) \setminus \{T^*\}$ and $f_1^*(T^*) = m-1$. Then, clearly, $(f_1^*, f_2, \ldots, f_m)$ is a yes-instance of X-RAY-CORRECTION$_{\mathcal{F}^d}(S_1, \ldots, S_m)$ if and only if $(f_1, f_2, \ldots, f_m)$ is a yes-instance of CONSISTENCY$_{\mathcal{F}^d}(S_1, \ldots, S_m)$. The result, therefore, is that CONSISTENCY$_{\mathcal{F}^d}(S_1, \ldots, S_m)$ reduces polynomially to X-RAY-CORRECTION$_{\mathcal{F}^d}(S_1, \ldots, S_m)$. Since by [8] the former is $\mathbb{NP}$-hard, so is the latter.

Next, let F_1 be an instance of UNIQUENESS$_{\mathcal{F}^d}(S_1, \ldots, S_m)$. Of course, F_1 is also an instance of SIMILAR-SOLUTION$_{\mathcal{F}^d}(S_1, \ldots, S_m)$. Let $F_2 \in \mathcal{F}^d$ with $|F_1| = |F_2|$ and $\sum_{i=1}^m ||X_{S_i}F_1 - X_{S_i}F_2||_1 < 2(m-1)$. Then by Theorem 2.1, F_2 is tomographically equivalent to F_1. Hence F_1 is a yes-instance of UNIQUENESS$_{\mathcal{F}^d}(S_1, \ldots, S_m)$ if and only if F_1 is a yes-instance of SIMILAR-SOLUTION$_{\mathcal{F}^d}(S_1, \ldots, S_m)$. Since UNIQUENESS$_{\mathcal{F}^d}(S_1, \ldots, S_m)$ is $\mathbb{NP}$-hard by [8] this concludes the proof for $m \geq 3$.

The case $m=1$ is trivial, so let $m=2$. The fact that SIMILAR-SOLUTION$_{\mathcal{F}^d}(S_1, S_2)$ is in $\mathbb{P}$ follows in conjunction with Theorem 2.1 directly from the polynomial-time solvability of UNIQUENESS$_{\mathcal{F}^d}(S_1, S_2)$.

Now let (f_1, f_2) be an instance of X-RAY-CORRECTION$_{\mathcal{F}^d}(S_1, S_2)$. Without loss of generality let (f_1, f_2) be consecutive. Clearly, (f_1, f_2) is a yes-instance if and only if there exist consecutive and consistent functions $g_i : \mathcal{A}_{\mathbb{Z}}(S_i) \to \mathbb{N}_0$ $i = 1, 2$ with $\sum_{i=1}^2 ||g_i - f_i||_1 \leq 1$. On the one hand, there are at most $\|f_1\|_1 + \|f_2\|_1 + 1$ many different choices of pairs (g_1, g_2) of such functions; hence all such pairs can be enumerated in polynomial time. On the other hand, for each choice (g_1, g_2) it can be checked in polynomial-time whether it is a yes-instance of CONSISTENCY$_{\mathcal{F}^d}$ (S_1, S_2). □

Finally we will show that NEAREST-SOLUTION$_{\mathcal{F}^d}(S_1, S_2)$ can be solved in polynomial time. (Again, the case $m = 1$ is trivial.)

Proof of the polynomial-time solvability of NEAREST-SOLUTION$_{\mathcal{F}^d}(S_1, S_2)$. Let (f_1, f_2) be an instance of NEAREST-SOLUTION$_{\mathcal{F}^d}(S_1, S_2)$. Without loss of generality we may assume that (f_1, f_2) is consecutive. Also, since the empty set is a feasible solution with error $\|f_1\|_1 + \|f_2\|_1$, we know that there is always a solution within the grid G' that is obtained from G by adding for $i = 1, 2$ to the support of f_i the next $\|f_1\|_1 + \|f_2\|_1$ lattice lines parallel to S_i and taking all intersections of any two of the extended two sets of parallel lines. Then G' contains at most $(2\|f_1\|_1 + \|f_2\|_1)(\|f_1\|_1 + 2\|f_2\|_1)$ lattice points which can all be determined in polynomial time. Let $N = |G'|$, and let M denote the number of different lines parallel to S_1 or S_2 that meet G. The points of G' will be the candidate points among which we will choose a solution.

Further, an optimal solution has at most $2\max\{\|f_1\|_1, \|f_2\|_1\}$ points. Therefore it suffices to solve at most that many instances with the same data but the additional constraint that the solution F has cardinality γ.

Let $F \in \mathcal{F}^d$ with $|F| = \gamma$. Then we have by Lemma 3.1

$$
\begin{aligned}
&||X_{S_1}F - f_1||_1 + ||X_{S_2}F - f_2||_1 \\
&\quad = 2\sum_{T \in \mathcal{A}_1^+} \big(X_{S_1}F(T) - f_1(T)\big) - |F| + \|f_1\|_1 + 2\sum_{T \in \mathcal{A}_2^+} \big(X_{S_2}F(T) - f_2(T)\big) - |F| + \|f_2\|_1 \\
&\quad = 2\left(\sum_{T \in \mathcal{A}_1^+} \big(X_{S_1}F(T) - f_1(T)\big) + \sum_{T \in \mathcal{A}_2^+} \big(X_{S_2}F(T) - f_2(T)\big)\right) - 2\gamma + \|f_1\|_1 + \|f_2\|_1,
\end{aligned}
$$

where $\mathcal{A}_i^+ = \{T \in \mathcal{A}_{\mathbb{Z}}(S_i) : X_{S_i}F(T) - f_i(T) > 0\}$ for $i = 1, 2$.

Hence it suffices to find a finite lattice set F with $|F| = \gamma$ that minimizes the sum of the excess of $X_{S_i}F(T)$ over $f_i(T)$.

Introducing one 0-1-variable for each candidate point of G', taking the incidence matrix $A \in \{0,1\}^{M\times N}$ whose rows correspond to the X-ray lines and whose columns correspond to the candidate points, collecting the X-ray data in a right-hand $b \in \mathbb{N}_0^M$, and using the notation $\mathbf{1}$ for a vector of ones of appropriate size, we can formulate this task as an integer linear programming problem.

$$\begin{aligned} & \mathbf{1}^T y \to \min \\ \text{s.t.} \quad & Ax \le b + y \\ & \mathbf{1}^T x = \gamma \\ & x \in \{0,1\}^N, y \in \mathbb{N}_0^M. \end{aligned}$$

Its linear programming relaxation can then be stated as the task to find a real vector solving

$$\begin{aligned} & \mathbf{1}^T y \to \min \\ \text{s.t.} \quad & C \begin{pmatrix} x \\ y \end{pmatrix} \le c, \end{aligned}$$

where

$$C = \begin{pmatrix} A^T & \mathbf{1} & -\mathbf{1} & -I_N & I_N & 0 \\ -I_M & 0 & 0 & 0 & 0 & -I_M \end{pmatrix}^T \quad \text{and} \quad c = (b, \gamma, -\gamma, 0, \mathbf{1}, 0)^T,$$

and where I_M and I_N denote the appropriately sized unit matrices.

We show that C is totally unimodular. Clearly it suffices to show that the submatrix

$$B = \begin{pmatrix} A \\ \mathbf{1}^T \end{pmatrix}$$

is totally unimodular. But this follows from the fact that each collection of rows from B can be split into two parts such that the difference of the sums of the rows in the first and in the second part is a vector with coefficients in $\{-1, 0, 1\}$ (see [18]). This is trivial if the collection does not involve the last row of B since the rows of A can be partitioned into two sets that correspond to the two directions and each column of A contains exactly two entries 1, one corresponding to S_1 and one corresponding to S_2. If, on the other hand, the last row is involved, take it as one part of the partition.

One can now use any polynomial-time linear programming algorithm to solve the task. □

REFERENCES

[1] A. Alpers and P. Gritzmann, *On the Degree of Ill-Posedness in Discrete Tomography*, preprint, 2004.

[2] A. Alpers, P. Gritzmann, and L. Thorens, *Stability and instability in discrete tomography*, in Digital and Image Geometry, Lecture Notes in Comput. Sci. 2243, Springer-Verlag, Berlin, 2001, pp. 175–186.

[3] P. Borwein, *Computational Excursions in Analysis and Number Theory*, Springer-Verlag, New York, 2002.

[4] P. Borwein and C. Ingalls, *The Prouhet-Tarry-Escott problem revisited*, Enseign. Math., 40 (1994), pp. 3–27.

[5] A. Daurat, *Determination of Q-convex sets by X-rays*, Theoret. Comput. Sci., 332 (2005), pp. 19–45.
[6] P.C. Fishburn, J.C. Lagarias, J.A. Reeds, and L.A. Shepp, *Sets uniquely determined by projections on axes* II: *Discrete case*, Discrete Math., 91 (1991), pp. 149–159.
[7] R.J. Gardner and P. Gritzmann, *Discrete tomography: Determination of finite sets by X-rays*, Trans. Amer. Math. Soc., 349 (1997), pp. 2271–2295.
[8] R.J. Gardner, P. Gritzmann, and D. Prangenberg, *On the computational complexity of reconstructing lattice sets from their X-rays*, Discrete Math., 202 (1999), pp. 45–71.
[9] R.J. Gardner, P. Gritzmann, and D. Prangenberg, *On the computational complexity of determining polyatomic structures by X-rays*, Theoret. Comput. Sci., 233 (2000), pp. 91–106.
[10] P. Gritzmann, *On the reconstruction of finite lattice sets from their X-rays*, in Discrete Geometry for Computer Imagery, E. Ahronovitz and C. Fiorio, eds., Springer-Verlag, Berlin, 1997, pp. 19–32.
[11] P. Gritzmann and S. de Vries, *Reconstructing crystalline structures from few images under high resolution transmission electron microscopy*, in Mathematics: Key Technology for the Future, W. Jäger, ed., Springer-Verlag, Berlin, 2003, pp. 441–459.
[12] J. Hadamard, *Lectures on Cauchy's Problem in Linear Partial Differential Equations*, Yale University Press, New Haven, CT, 1923.
[13] G.T. Herman and A. Kuba, eds., *Discrete Tomography: Foundations, Algorithms, and Applications*, Birkhäuser Boston, Cambridge, MA, 1999.
[14] R.W. Irving and M.R. Jerrum, *Three-dimensional statistical data security problems*, SIAM J. Comput., 23 (1994), pp. 170–184.
[15] M. Mignotte and D. Ştefănescu, *Polynomials: An Algorithmic Approach*, Springer-Verlag, Singapore, 1999.
[16] A. Rényi, *On projections of probability distributions*, Acta Math. Sci. Hungar., 3 (1952), pp. 131–142.
[17] H.J. Ryser, *Matrices of zeros and ones*, in Combinatorial Mathematics, Mathematical Association of America and Quinn & Boden, Rahway, NJ, 1963, pp. 61–78.
[18] A. Schrijver, *Theory of Linear and Integer Programming*, Wiley, Chichester, UK, 1986.
[19] C.H. Slump and J.J. Gerbrands, *A network flow approach to reconstruction of the left ventricle from two projections*, Comput. Graphics Image Processing, 18 (1982), pp. 18–36.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 240–260

COLLECTIVE TREE SPANNERS OF GRAPHS*

FEODOR F. DRAGAN†, CHENYU YAN†, AND IRINA LOMONOSOV‡

Abstract. In this paper we introduce a new notion of *collective tree spanners*. We say that a graph $G = (V, E)$ *admits a system of* μ *collective additive tree* r*-spanners* if there is a system $\mathcal{T}(G)$ of at most μ spanning trees of G such that for any two vertices x, y of G a spanning tree $T \in \mathcal{T}(G)$ exists such that $d_T(x, y) \leq d_G(x, y) + r$. Among other results, we show that any chordal graph, chordal bipartite graph or cocomparability graph admits a system of at most $\log_2 n$ collective additive tree 2-spanners. These results are complemented by lower bounds, which say that any system of collective additive tree 1-spanners must have $\Omega(\sqrt{n})$ spanning trees for some chordal graphs and $\Omega(n)$ spanning trees for some chordal bipartite graphs and some cocomparability graphs. Furthermore, we show that any c-chordal graph admits a system of at most $\log_2 n$ collective additive tree $(2\lfloor c/2 \rfloor)$-spanners, any circular-arc graph admits a system of two collective additive tree 2-spanners. Towards establishing these results, we present a general property for graphs, called (α, r)-*decomposition*, and show that any (α, r)-decomposable graph G with n vertices admits a system of at most $\log_{1/\alpha} n$ collective additive tree $2r$-spanners. We discuss also an application of the collective tree spanners to the problem of designing compact and efficient routing schemes in graphs. For any graph on n vertices admitting a system of at most μ collective additive tree r-spanners, there is a routing scheme of deviation r with addresses and routing tables of size $O(\mu \log^2 n / \log \log n)$ bits per vertex. This leads, for example, to a routing scheme of deviation $(2\lfloor c/2 \rfloor)$ with addresses and routing tables of size $O(\log^3 n / \log \log n)$ bits per vertex on the class of c-chordal graphs.

Key words. sparse spanners, tree spanners, graph distance, balanced separator, graph decomposition, chordal graphs, c-chordal graphs, message routing, efficient algorithms

AMS subject classifications. 05C05, 05C10, 05C12, 05C78, 05C85, 94C15, 68R10, 68Q25, 68W25

DOI. 10.1137/S089548010444167X

1. Introduction.

Many combinatorial and algorithmic problems are concerned with the distance d_G on the vertices of a possibly weighted graph $G = (V, E)$. Approximating d_G by a simpler distance (in particular, by tree-distance d_T) is useful in many areas such as communication networks, data analysis, motion planning, image processing, network design, and phylogenetic analysis (see [1, 8, 11, 19, 22, 52, 58, 59, 64, 66]). An arbitrary metric space (in particular a finite metric defined by a general graph) might not have enough structure to exploit algorithmically; on trees, since they have a simpler (acyclic) structure, many hard algorithmic problems have easy solutions. So, the general goal is, for a given graph G, to find a simpler (well-structured, sparse, etc.) graph $H = (V, E')$ with the same vertex-set such that the distance $d_H(u, v)$ in H between two vertices $u, v \in V$ is reasonably close to the corresponding distance $d_G(u, v)$ in the original graph G.

There are several ways to measure the quality of this approximation, two of them leading to the notion of a spanner. For $t \geq 1$, a spanning subgraph H of G is called a *multiplicative* t*-spanner* of G [22, 59, 58] if $d_H(u, v) \leq t \cdot d_G(u, v)$ for all $u, v \in V$. If $r \geq 0$ and $d_H(u, v) \leq d_G(u, v)+r$ for all $u, v \in V$, then H is called an *additive* r*-spanner*

*Received by the editors March 5, 2004; accepted for publication (in revised form) November 1, 2005; published electronically March 15, 2006. Results of this paper were partially presented at the SWAT '04 conference [30].

http://www.siam.org/journals/sidma/20-1/44167.html

†Department of Computer Science, Kent State University, Kent, OH 44242 (dragan@cs.kent.edu, cyan@cs.kent.edu).

‡Department of Computer Science, Hiram College, Hiram, OH 44234 (lomonosovi@hiram.edu).

of G [52]. The parameters t and r are called, respectively, the *multiplicative* and the *additive stretch factors*. Clearly, every additive r-spanner of G is a multiplicative $(r+1)$-spanner of G (but not vice versa). Note that the graphs considered in this paper are assumed to be unweighted (except in section 7 where we discuss how to extend our results to weighted graphs).

Graph spanners have applications in various areas, especially in distributed systems and communication networks. In [59], close relationships were established between the quality of spanners (in terms of stretch factor and the number of spanner edges $|E'|$), and the time and communication complexities of any synchronizer for the network based on this spanner. Also, sparse spanners are very useful in message routing in communication networks; in order to maintain succinct routing tables, efficient routing schemes can use only the edges of a sparse spanner [60]. Unfortunately, the problem of determining, for a given graph G and two integers $t \geq 2, m \geq 1$, whether G has a multiplicative t-spanner with m or fewer edges, is NP-complete (see [58]).

The sparsest spanners are tree spanners. Tree spanners occur in biology [5], and as it was shown in [57], they can be used as models for broadcast operations in communication networks. Tree spanners are favored also from the algorithmic point of view—many algorithmic problems are easily solvable on trees. Multiplicative tree t-spanners were studied in [19]. It was shown that, for a given graph G, the problem to decide whether G has a multiplicative tree t-spanner (the *multiplicative tree t-spanner problem*) is NP-complete for any fixed $t \geq 4$ and is linearly solvable for $t = 1, 2$. Recently, this NP-completeness result was improved—the multiplicative tree t-spanner problem is NP-complete for any fixed $t \geq 4$ even on some rather restricted graph classes: planar graphs [12], chordal graphs [14] and chordal bipartite graphs [15].

Nevertheless, some particular graph classes, such as cographs, complements of bipartite graphs, split graphs, regular bipartite graphs, interval graphs, permutation graphs, convex bipartite graphs, distance-hereditary graphs, directed path graphs, cocomparability graphs, AT-free graphs, strongly chordal graphs, and dually chordal graphs do admit additive tree r-spanners and/or multiplicative tree t-spanners for sufficiently small r and t (see [13, 18, 51, 55, 61, 62, 69]). We refer also to [1, 12, 14, 18, 19, 38, 52, 57, 58, 65] for more background information on tree and general sparse spanners.

Many graph classes (including hypercubes, planar graphs, chordal graphs, chordal bipartite graphs) do not admit any good tree spanner. For every fixed integer t there are planar chordal graphs and planar chordal bipartite graphs that do not admit tree t-spanners (additive as well as multiplicative) [21, 62]. However, as it was shown in [58], any chordal graph with n vertices admits a multiplicative 5-spanner with at most $2n-2$ edges and a multiplicative 3-spanner with at most $O(n \log n)$ edges (both spanners are constructable in polynomial time). Recently, the results were further improved. In [21], the authors show that every chordal graph admits an additive 4-spanner with at most $2n-2$ edges and an additive 3-spanner with at most $O(n \log n)$ edges. An additive 4-spanner can be constructed in linear time while an additive 3-spanner is constructable in $O(m \log n)$ time, where m is the number of edges of G. Even more, the method designed for chordal graph is extended to all c-chordal graphs. As a result, it was shown that any such graph admits an additive $(c+1)$-spanner with at most $2n-2$ edges which is constructable in $O(cn+m)$ time. Recall that a graph G is *chordal* if its largest induced (chordless) cycles are of length 3 and *c-chordal* if its largest induced cycles are of length c. Note also that [59] gives a method for constructing a multiplicative 3-spanner of the n-vertex hypercube with fewer than $7n$

edges and this construction was improved in [34] to give a multiplicative 3-spanner of the n-vertex hypercube with fewer than $4n$ edges.

1.1. Our results. In this paper we introduce a new notion of *collective tree spanners*, a notion slightly *weaker* than the one of a tree spanner and slightly *stronger* than the notion of a sparse spanner. We say that a graph $G = (V, E)$ *admits a system of μ collective additive tree r-spanners* if there is a system $\mathcal{T}(G)$ of at most μ spanning trees of G such that for any two vertices x, y of G a spanning tree $T \in \mathcal{T}(G)$ exists such that $d_T(x, y) \leq d_G(x, y)+r$ (a multiplicative variant of this notion can be defined analogously). Clearly, if G admits a system of μ collective additive tree r-spanners, then G admits an additive r-spanner with at most $\mu\times(n-1)$ edges (take the union of all those trees), and if $\mu = 1$ then G admits an additive tree r-spanner. Furthermore, any result on collective *additive* tree spanners can be translated into a result on collective *multiplicative* tree spanners since any graph, admitting a system of μ collective *additive* tree r-spanners, admits a system of μ collective *multiplicative* tree $(r + 1)$-spanners ($d_T(x, y) \leq d_G(x, y) + r$ implies $d_T(x, y)/d_G(x, y) \leq 1 + r/d_G(x, y) \leq r + 1$ for an unweighted graph G). Note also that any graph on n vertices admits a system of at most $n - 1$ collective additive tree 0-spanners (take $n - 1$ breadth-first-search–trees rooted at different vertices of G).

The introduction of this new notion was inspired by the works [6, 7] of Bartal and subsequent works [20, 37]. For example, motivated by Bartal's work on probabilistic approximation of general metrics with tree metrics, [20] gives a polynomial time algorithm that given a finite n point metric G, constructs $O(n \log n)$ trees and a probability distribution ψ on them such that the expected multiplicative stretch of any edge of G in a tree chosen according to ψ is at most $O(\log n \log \log n)$. These results led to approximation algorithms for a number of optimization problems including the group Steiner tree problem, the metric labeling problem, the buy-at-bulk network design problem and many others (see [6, 7, 20, 37] for more details).

In section 2 we define a large class of graphs, called (α, r)-*decomposable*, and show that any (α, r)-decomposable graph G with n vertices admits a system of at most $\log_{1/\alpha} n$ collective additive tree $2r$-spanners. Then, in sections 3 and 4, we show that chordal graphs, chordal bipartite graphs, and cocomparability graphs are all $(1/2, 1)$-decomposable graphs, implying that each graph from those families admits a system of at most $\log_2 n$ collective additive tree 2-spanners. These results are complemented by lower bounds, which say that any system of collective additive tree 1-spanners must have $\Omega(\sqrt{n})$ spanning trees for some chordal graphs and $\Omega(n)$ spanning trees for some chordal bipartite graphs and some cocomparability graphs. Furthermore, we show that any c-chordal graph is $(1/2, \lfloor c/2 \rfloor)$-decomposable, implying that each c-chordal graph admits a system of at most $\log_2 n$ collective additive tree $(2\lfloor c/2 \rfloor)$-spanners.

Thus, as a byproduct, we get that chordal graphs, chordal bipartite graphs, and cocomparability graphs admit additive 2-spanners with at most $(n - 1) \log_2 n$ edges and c-chordal graphs admit additive $(2\lfloor c/2 \rfloor)$-spanners with at most $(n - 1) \log_2 n$ edges. Our result for chordal graphs improves the known results from [58] and [21] on 3-spanners and answers the question posed in [21] whether chordal graphs admit additive 2-spanners with $O(n \log n)$ edges.

In section 5, we show that each circular-arc graph admits a system of two collective additive tree 2-spanners, and that for any constant $r \geq 0$ there is a circular-arc graph without any (one) additive tree r-spanner.

In section 6 we discuss an application of the collective tree spanners to the problem of designing compact and efficient routing schemes in graphs. For any graph

on n vertices admitting a system of at most μ collective additive tree r-spanners, there is a routing scheme of deviation r with addresses and routing tables of size $O(\mu \log^2 n / \log\log n)$ bits per vertex (for details see section 6). This leads, for example, to a routing scheme of deviation $(2\lfloor c/2 \rfloor)$ with addresses and routing tables of size $O(\log^3 n / \log\log n)$ bits per vertex on the class of c-chordal graphs. The latter improves the recent result on routing on c-chordal graphs obtained in [33] (see also [32] for the case of chordal graphs). We conclude the paper with section 7, where we discuss how to extend our results to weighted graphs, and section 8, where we discuss some further developments and future directions.

1.2. Basic notions and notations. All graphs occurring in this paper are connected, finite, undirected, loopless and without multiple edges. In a graph $G = (V, E)$ the *length* of a path from a vertex v to a vertex u is the number of edges in the path. The *distance* $d_G(u, v)$ between the vertices u and v is the length of a shortest path connecting u and v.

For a subset $S \subseteq V$, let $rad_G(S)$ and $diam_G(S)$ be the radius and the diameter, respectively, of S in G, i.e., $rad_G(S) = min_{v \in V}\{max_{u \in S}\{d_G(u, v)\}\}$, $diam_G(S) = max_{u,v \in S}\{d_G(u, v)\}$. A vertex $v \in V$ such that $d_G(u, v) \leq rad_G(S)$ for any $u \in S$ is called a central vertex for S. The value $rad_G(V)$ is called *the radius* of G. Let also $N(v)$ ($N[v]$) denote the open (closed) neighborhood of a vertex v in G, i.e., $N(v) = \{u \in V : uv \in E(G)\}$ and $N[v] = N(v) \cup \{v\}$.

2. (α, r)-decomposable graphs and their collective tree spanners. Different balanced separators in graphs were used by many authors in designing efficient graph algorithms (see [26, 27, 43, 44, 46, 50, 53, 54]). For example, bounded size balanced separators and bounded diameter balanced separators were recently employed in [43, 44, 50] for designing compact distance labeling schemes for different so-called well-separated families of graphs. We extend those ideas and apply them to our problem.

Let α be a positive real number smaller than 1 and r be a nonnegative integer. We say that an n-vertex graph $G = (V, E)$ is (α, r)-*decomposable* if there is a separator $S \subseteq V$ such that the following three conditions hold:

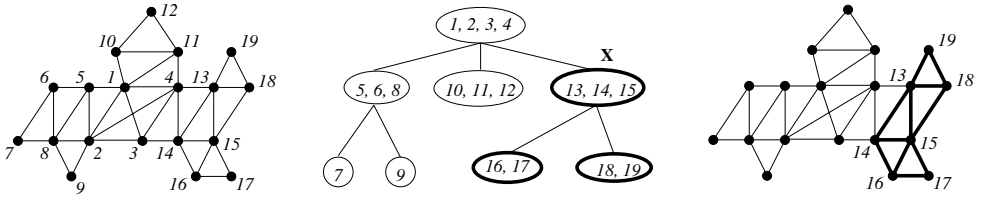
balanced separator condition—the removal of S leaves no connected component with more than αn vertices;

bounded separator-radius condition—$rad_G(S) \leq r$, i.e., there exists a vertex c in G (called *a central vertex for* S) such that $d_G(v, c) \leq r$ for any $v \in S$;

hereditary family condition—each connected component of the graph, obtained from G by removing vertices of S, is also an (α, r)-decomposable graph.

Note that, by definition, any graph of radius at most r is (α, r)-decomposable and that the size of S does not matter.

2.1. Collective tree spanners of (α, r)-decomposable graphs. Using the first and third conditions of the definition, one can construct for any (α, r)-decomposable graph G a *(rooted) balanced decomposition tree* $\mathcal{BT}(G)$ as follows. If G is of radius at most r, then $\mathcal{BT}(G)$ is a one-node tree. Otherwise, find a balanced separator S in G, which exists according to the balanced separator condition. Let $G_1, G_2, \ldots, G_p$ be the connected components of the graph $G - S$ obtained from G by removing vertices of S. For each graph G_i ($i = 1, \ldots, p$), which is (α, r)-decomposable by the hereditary family condition, construct a balanced decomposition tree $\mathcal{BT}(G_i)$ recursively, and build $\mathcal{BT}(G)$ by taking S to be the root and connecting the root of each tree $\mathcal{BT}(G_i)$ as a child of S. See Figure 1 for an illustration. Clearly, the nodes of $\mathcal{BT}(G)$ represent a partition of the vertex set V of G into *clusters* $S_1, S_2, \ldots, S_q$ of radius at most r



(a) (b) (c)

FIG. 1. (a) *A graph* G, (b) *its balanced decomposition tree* $\mathcal{BT}(G)$, *and* (c) *an induced subgraph* $G(\downarrow X)$ *of* G.

each. For a node X of $\mathcal{BT}(G)$, denote by $G(\downarrow X)$ the (connected) subgraph of G induced by vertices $\bigcup\{Y : Y$ is a descendent of X in $\mathcal{BT}(G)\}$ (here we assume that X is a descendent of itself).

It is easy to see that a balanced decomposition tree $\mathcal{BT}(G)$ of a graph G with n vertices and m edges has depth at most $\log_{1/\alpha} n$, which is $O(log_2 n)$ if α is a constant. Moreover, assuming that a balanced and bounded radius separator can be found in polynomial, say $p(n)$, time (for the special graph classes we consider later, $p(n)$ will be at most $O(n^3)$), the tree $\mathcal{BT}(G)$ can be constructed in $O((p(n)+m)\log_{1/\alpha} n)$ total time. Indeed, in each level of recursion we need to find balanced and bounded radius separators in current disjoint subgraphs and to construct the corresponding subgraphs of the next level. Also, since the graph sizes are reduced by a factor α, the recursion depth is at most $\log_{1/\alpha} n$.

Consider now two arbitrary vertices x and y of an (α, r)-decomposable graph G and let $S(x)$ and $S(y)$ be the nodes of $\mathcal{BT}(G)$ containing x and y, respectively. Let also $NCA_{\mathcal{BT}(G)}(S(x), S(y))$ be the nearest common ancestor of nodes $S(x)$ and $S(y)$ in $\mathcal{BT}(G)$ and $(X_0, X_1, \ldots, X_t)$ be the path of $\mathcal{BT}(G)$ connecting the root X_0 of $\mathcal{BT}(G)$ with $NCA_{\mathcal{BT}(G)}(S(x), S(y)) = X_t$ (in other words, $X_0, X_1, \ldots, X_t$ are the common ancestors of $S(x)$ and $S(y)$). The following lemmata are crucial to all our subsequent results.

LEMMA 2.1. *Any path* $P^G_{x,y}$, *connecting vertices* x *and* y *in* G, *contains a vertex from* $X_0 \cup X_1 \cup \cdots \cup X_t$.

Let $SP^G_{x,y}$ be a shortest path of G connecting vertices x and y, and let X_i be the node of the path $(X_0, X_1, \ldots, X_t)$ with the smallest index such that $SP^G_{x,y} \bigcap X_i \neq \emptyset$ in G. Then, the following lemma holds.

LEMMA 2.2. *We have* $d_G(x, y) = d_{G'}(x, y)$, *where* $G' := G(\downarrow X_i)$.

Proof. It is enough to show that the path $SP^G_{x,y}$ consists of only vertices of G'. Let us assume, by way of contradiction, that there is a vertex z of $SP^G_{x,y}$ that does not belong to G'. Let $SP^G_{x,z}$ be a subpath of $SP^G_{x,y}$ between x and z. Clearly, the node $S(z)$ of $\mathcal{BT}(G)$, containing vertex z, is not a descendent of X_i. Therefore, the nearest common ancestor of $S(x)$ and $S(z)$ in $\mathcal{BT}(G)$ is a node X_j from $\{X_0, X_1, \ldots, X_t\}$ with $j < i$. But then, by Lemma 2.1, the path $SP^G_{x,z}$ (and hence the path $SP^G_{x,y}$) must have a vertex in $X_0 \cup X_1 \cup \cdots \cup X_j$, contradicting the choice of X_i, $i > j$. □

For the graph $G' = G(\downarrow X_i)$, consider its arbitrary *breadth-first-search–tree (BFS-tree)* T' rooted at a central vertex c for X_i, i.e., a vertex c such that $d_{G'}(v, c) \leq r$ for any $v \in X_i$. Such a vertex exists in G' since G' is an (α, r)-decomposable graph and X_i is its balanced and bounded radius separator. The tree T' has the following distance property with respect to those vertices x and y.

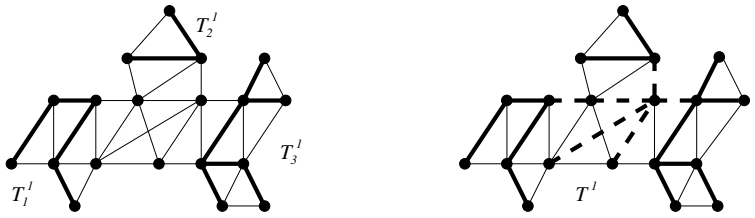


FIG. 2. (a) *Local subtrees* T_1^1, T_2^1, T_3^1 *of graph* G *from Figure* 1 *and* (b) *a corresponding spanning tree* T^1 *of* G *(dark solid edges are edges of local subtrees* T_1^1, T_2^1, T_3^1*, dashed edges are added to create one spanning tree* T^1 *on top of* T_1^1, T_2^1, T_3^1*).*

LEMMA 2.3. *We have* $d_{T'}(x,y) \le d_G(x,y) + 2r$.

Proof. We know, by Lemma 2.2, that a shortest path $SP^G_{x,y}$, intersecting X_i and not intersecting any X_l $(l < i)$, lies entirely in G'. Let x' be the vertex of $SP^G_{x,y} \cap X_i$ closest to x and y' be the vertex of $SP^G_{x,y} \cap X_i$ closest to y. Since T' is a BFS-tree of G' rooted at vertex c, we have

$$d_{T'}(x,c) = d_{G'}(x,c) \le d_{G'}(x,x') + d_{G'}(x',c) \le d_{G'}(x,x') + r = d_G(x,x') + r,$$
$$d_{T'}(y,c) = d_{G'}(y,c) \le d_{G'}(y,y') + d_{G'}(y',c) \le d_{G'}(y,y') + r = d_G(y,y') + r.$$

That is, $d_{T'}(x,y) \le d_{T'}(x,c) + d_{T'}(y,c) \le d_G(x,x') + d_G(y,y') + 2r$. Combining this with the fact that $d_G(x,y) \ge d_G(x,x') + d_G(y,y')$, we obtain $d_{T'}(x,y) \le d_G(x,y) + 2r$. □

Let now $B_1^i, \ldots, B_{p_i}^i$ be the nodes on depth i of the tree $\mathcal{BT}(G)$. For each subgraph $G_j^i := G(\downarrow B_j^i)$ of G $(i = 0, 1, \ldots, depth(\mathcal{BT}(G)),\ j = 1, 2, \ldots, p_i)$, denote by T_j^i a BFS-tree of graph G_j^i rooted at a central vertex c_j^i for B_j^i (see Figure 2 for an illustration). The trees T_j^i $(i = 0, 1, \ldots, depth(\mathcal{BT}(G)),\ j = 1, 2, \ldots, p_i)$ are called *local subtrees* of G, and, given the balanced decomposition tree $\mathcal{BT}(G)$, they can be constructed in $O((t(n)+m)\log_{1/\alpha} n)$ total time, where $t(n)$ is the time needed to find a central vertex c_j^i for B_j^i (a trivial upper bound for $t(n)$ is $O(n^3)$). From Lemma 2.3 the following general result can be deduced.

THEOREM 2.4. *Let* G *be an* (α, r)*-decomposable graph,* $\mathcal{BT}(G)$ *be its balanced decomposition tree and* $\mathcal{LT}(G) = \{T_j^i : i = 0, 1, \ldots, depth(\mathcal{BT}(G)),\ j = 1, 2, \ldots, p_i\}$ *be its local subtrees. Then, for any two vertices* x *and* y *of* G*, there exists a local subtree* $T_{j'}^{i'}$ *in* $\mathcal{LT}(G)$ *such that*

$$d_{T_{j'}^{i'}}(x,y) \le d_G(x,y) + 2r.$$

This theorem implies two important results for the class of (α, r)-decomposable graphs. Let G be an (α, r)-decomposable graph with n vertices and m edges, $\mathcal{BT}(G)$ be its balanced decomposition tree, and $\mathcal{LT}(G)$ be the family of its local subtrees (defined above). Consider a graph H obtained by taking the union of all local subtrees of G (by putting all of them together), i.e.,

$$H := \bigcup \{T_j^i : T_j^i \in \mathcal{LT}(G)\} = (V, \cup\{E(T_j^i) : T_j^i \in \mathcal{LT}(G)\}).$$

Clearly, H is a spanning subgraph of G, constructable in $O((p(n)+t(n)+m)\log_{1/\alpha} n)$ total time, and, for any two vertices x and y of G, $d_H(x,y) \le d_G(x,y)+2r$ holds. Also,

since for every level i ($i = 0, 1, \ldots, depth(\mathcal{BT}(G))$) of balanced decomposition tree $\mathcal{BT}(G)$, the corresponding local subtrees $T_1^i, \ldots, T_{p_i}^i$ are pairwise vertex-disjoint, their union has at most $n-1$ edges. Therefore, H cannot have more than $(n-1)\log_{1/\alpha} n$ edges in total. Thus, we have proven the following result.

THEOREM 2.5. *Any (α, r)-decomposable graph G with n vertices admits an additive $2r$-spanner with at most $(n-1)\log_{1/\alpha} n$ edges.*

Instead of taking the union of all local subtrees of G, one can fix i ($i \in \{0, 1, \ldots, depth(\mathcal{BT}(G))\}$) and consider separately the union of only local subtrees $T_1^i, \ldots, T_{p_i}^i$, corresponding to the level i of the decomposition tree $\mathcal{BT}(G)$, and then extend in linear $O(m)$ time that forest to a spanning tree T^i of G (using, for example, a variant of the Kruskal's spanning tree algorithm for the unweighted graphs). We call this tree T^i the *spanning tree of G corresponding to the level i of the balanced decomposition* $\mathcal{BT}(G)$. In this way we can obtain at most $\log_{1/\alpha} n$ spanning trees for G, one for each level i of $\mathcal{BT}(G)$. Denote the collection of those spanning trees by $\mathcal{T}(G)$. By Theorem 2.4, it is rather straightforward to show that for any two vertices x and y of G, there exists a spanning tree $T^{i'}$ in $\mathcal{T}(G)$ such that $d_{T^{i'}}(x, y) \leq d_G(x, y) + 2r$. Thus, we have the following theorem.

THEOREM 2.6. *Any (α, r)-decomposable graph G with n vertices admits a system $\mathcal{T}(G)$ of at most $\log_{1/\alpha} n$ collective additive tree $2r$-spanners.*

Note that such a system $\mathcal{T}(G)$ for an (α, r)-decomposable graph G with n vertices and m edges can be constructed in $O((p(n) + t(n) + m)\log_{1/\alpha} n)$ time, where $p(n)$ is the time needed to find a balanced and bounded radius separator S and $t(n)$ is the time needed to find a central vertex for S.

2.2. Extracting an appropriate tree from $\mathcal{T}(G)$. Now we will show that one can assign $O(\log_{1/\alpha} n \times \log n)$ bit labels to vertices of G such that, for any pair of vertices x and y, a tree $T^{i'}$ in $\mathcal{T}(G)$ with $d_{T^{i'}}(x, y) \leq d_G(x, y) + 2r$ can be identified in only $O(\log_{1/\alpha} n)$ time by merely inspecting the labels of x and y, without using any other information about the graph. This will be useful in an application of collective tree spanners, discussed in section 6.

Associate with each vertex x of G a $2 \times (depth(\mathcal{BT}(G)) + 1)$ array A_x such that, for each level i of $\mathcal{BT}(G)$, $A_x[1, i] = j$ and $A_x[2, i] = d_{T_j^i}(x, c_j^i)$ if there exists a local subtree T_j^i in $\mathcal{LT}(G)$ containing vertex x, and $A_x[1, i] = nil$ and $A_x[2, i] = \infty$, otherwise (i.e., the depth in $\mathcal{BT}(G)$ of node $S(x)$ containing x is smaller than i). Evidently, each label A_x ($x \in V$) can be encoded using $O(\log_{1/\alpha} n \times \log n)$ bits and a computation of all labels A_x, $x \in V$ can be performed together with the construction of system $\mathcal{T}(G)$.

Given labels A_x, A_y of vertices x and y, the following procedure will return in $O(\log_{1/\alpha} n)$ time an index $i' \in \{0, 1, \ldots, depth(\mathcal{BT}(G))\}$ such that, for tree $T^{i'} \in \mathcal{T}(G)$, $d_{T^{i'}}(x, y) \leq d_G(x, y) + 2r$ holds.

set $i' := 0$;
set $minsum := A_x[2, 0] + A_y[2, 0]$;
set $i := 1$;
while $(A_x[1, i] = A_y[1, i] \neq nil)$ and $(i \leq \log_{1/\alpha} n)$ do
  if $A_x[2, i] + A_y[2, i] < minsum$
    then set $i' := i$ and $minsum := A_x[2, i] + A_y[2, i]$;
  $i := i + 1$;
enddo
return i'.

This procedure simply finds, among all local subtrees containing both x and y, a subtree $T^{i'}_{j'}$ for which the sum $d_{T^{i'}_{j'}}(x, c^{i'}_{j'}) + d_{T^{i'}_{j'}}(y, c^{i'}_{j'})$ is minimum, and then returns its upper index i'.

To show that indeed $d_{T^{i'}}(x, y) \leq d_G(x, y) + 2r$, we will need to recall the proof of Lemma 2.3 (note that $d_{T^{i'}}(x, y) = d_{T^{i'}_{j'}}(x, y)$ by construction of $T^{i'}$). Let again $S(x)$ and $S(y)$ be the nodes of $\mathcal{BT}(G)$ containing vertices x and y, respectively, and let $(B^0, B^1_{j_1}, \ldots, B^t_{j_t})$ be the path of $\mathcal{BT}(G)$ connecting the root B^0 of $\mathcal{BT}(G)$ with $NCA_{\mathcal{BT}(G)}(S(x), S(y)) = B^t_{j_t}$. In Lemma 2.3 we proved that there exists an index $i \in \{0, 1, \ldots, t\}$ such that any BFS-tree T' of the graph $G(\downarrow B^i_{j_i})$ rooted at a center c for $B^i_{j_i}$ (including local subtree $T^i_{j_i}$ rooted at $c^i_{j_i}$) satisfies $d_{T'}(x, y) \leq d_{T'}(x, c) + d_{T'}(y, c) \leq d_G(x, y) + 2r$ (see inequalities (1) and (2) in that proof). Since, among local subtrees $T^0, T^1_{j_1}, \ldots, T^t_{j_t}$, the subtree $T^{i'}_{j'}$ has minimum sum $d_{T^{i'}_{j'}}(x, c^{i'}_{j'}) + d_{T^{i'}_{j'}}(y, c^{i'}_{j'})$, we conclude

$$\begin{aligned} d_{T^{i'}}(x, y) = d_{T^{i'}_{j'}}(x, y) &\leq d_{T^{i'}_{j'}}(x, c^{i'}_{j'}) + d_{T^{i'}_{j'}}(y, c^{i'}_{j'}) \\ &\leq d_{T^{i}_{j_i}}(x, c^{i}_{j_i}) + d_{T^{i}_{j_i}}(y, c^{i}_{j_i}) \leq d_G(x, y) + 2r. \end{aligned}$$

3. Acyclic hypergraphs, chordal graphs and (α, r)-decomposition. Let $H = (V, \mathcal{E})$ be a *hypergraph* with the vertex set V and the *hyperedge* set $\mathcal{E}$, i.e., $\mathcal{E}$ is a set of nonempty subsets of V. For every vertex $v \in V$, let $\mathcal{E}(v) = \{e \in \mathcal{E} : v \in e\}$. The *2-section graph* $2SEC(H)$ of a hypergraph H has V as its vertex-set and two distinct vertices are adjacent in $2SEC(H)$ if and only if they are contained in a common hyperedge of H. A hypergraph H is called *conformal* if every clique (a set of pairwise adjacent vertices) of $2SEC(H)$ is contained in a hyperedge $e \in \mathcal{E}$, and a hypergraph H is called *acyclic* if there is a tree T with node set $\mathcal{E}$ such that, for all vertices $v \in V$, $\mathcal{E}(v)$ induces a subtree T_v of T. For these and other hypergraph notions see [10].

The following theorem represents two well-known characterizations of acyclic hypergraphs. Let $\mathcal{C}(G)$ be the set of all maximal (by inclusion) cliques of a graph $G = (V, E)$. The hypergraph $(V, \mathcal{C}(G))$ is called the clique-hypergraph of G. Recall that a graph G is *chordal* if it does not contain any induced cycles of length greater than 3.

THEOREM 3.1 (see [2, 9, 10, 17, 36, 67]). *Let $H = (V, \mathcal{E})$ be a hypergraph. Then the following conditions are equivalent:*

(i) *H is an acyclic hypergraph;*
(ii) *H is conformal and $2SEC(H)$ of H is a chordal graph;*
(iii) *H is the clique hypergraph $(V, \mathcal{C}(G))$ of some chordal graph $G = (V, E)$.*

Later we will need also the following known result. A vertex v of a graph G is called *simplicial* if its neighborhood $N(v)$ forms a clique in G.

THEOREM 3.2 (see [17, 25]). *Let $G = (V, E)$ be a graph. Then the following conditions are equivalent:*

(i) *G is a chordal graph;*
(ii) *the clique hypergraph $(V, \mathcal{C}(G))$ of G is acyclic (in other words, G is the intersection graph of a family of subtrees of a tree);*
(iii) *G has a perfect elimination ordering. i.e., an ordering $v_1, v_2, \ldots, v_n$ of vertices of G such that, for any i, $i \in \{1, 2, \ldots, n\}$, vertex v_i is simplicial in graph $G(v_i, \ldots, v_n)$, the subgraph of G induced by vertices $v_i, \ldots, v_n$.*

Let now $G = (V, E)$ be an arbitrary graph and r be a positive integer. We say that G *admits a radius r acyclic covering* if there is a family $\mathcal{S}(G) = \{S_1, \ldots, S_k\}$ of

subsets of V such that

(1) $\bigcup_{i=1}^{k} S_i = V$;
(2) for any edge xy of G there is a subset S_i ($i \in \{1, \ldots, k\}$) with $x, y \in S_i$;
(3) $H = (V, \mathcal{S}(G))$ is an acyclic hypergraph;
(4) $rad_G(S_i) \leq r$ for each $i = 1, \ldots, k$.

A class of graphs $\mathcal{F}$ is called *hereditary* if every induced subgraph of a graph G belongs to $\mathcal{F}$ whenever G is in $\mathcal{F}$. A class of graphs $\mathcal{F}$ is called (α, r)-*decomposable* if every graph G from $\mathcal{F}$ is (α, r)-decomposable.

THEOREM 3.3. *Let $\mathcal{F}$ be a hereditary class of graphs such that any $G \in \mathcal{F}$ admits a radius r acyclic covering. Then $\mathcal{F}$ is a $(1/2, r)$-decomposable class of graphs.*

Proof. Consider a graph $G \in \mathcal{F}$ and let $\mathcal{S}(G) = \{S_1, \ldots, S_k\}$ be its radius r acyclic covering. Since $H = (V, \mathcal{S}(G))$ is an acyclic hypergraph, $2SEC(H)$ is chordal and H is conformal. It is well known [47], that every n-vertex chordal graph Γ contains a maximal clique C such that if the vertices in C are deleted from Γ, every connected component in the graph induced by any remaining vertices is of size at most $n/2$. Moreover, according to [47], for any chordal graph on n vertices and m edges, such a separating clique C can be found in $O(n+m)$ time. Applying this result to an n-vertex chordal graph $2SEC(H)$, we will get in at most $O(n^2)$ time a maximal clique S of $2SEC(H)$ such that any connected component of the graph $2SEC(H) - S$ (obtained from $2SEC(H)$ by deleting vertices of S) has at most $n/2$ vertices. Since $2SEC(H)$ is obtained from G by adding some new edges, removing vertices of S from the original graph G will leave no connected component (in $G - S$) with more than $n/2$ vertices. Furthermore, since $\mathcal{F}$ is a hereditary class of graphs, all connected components of $G - S$ induce graphs from $\mathcal{F}$ (and they can be assumed by induction to be $(1/2, r)$-decomposable graphs). It remains to note that, from conformality of H, there must exist a set S_i in $\mathcal{S}(G)$ which contains S, that is, $rad_G(S) \leq rad_G(S_i) \leq r$ must hold. □

Since for a chordal graph $G = (V, E)$ the clique hypergraph $(V, \mathcal{C}(G))$ is acyclic and chordal graphs form a hereditary class of graphs, from Theorem 3.3 and Theorems 2.5 and 2.6, we immediately conclude the following corollaries.

COROLLARY 3.4. *Any chordal graph G with n vertices and m edges admits an additive 2-spanner with at most $(n-1)\log_2 n$ edges, and such a sparse spanner can be constructed in $O(m \log_2 n)$ time.*

COROLLARY 3.5. *Any chordal graph G with n vertices and m edges admits a system $\mathcal{T}(G)$ of at most $\log_2 n$ collective additive tree 2-spanners, and such a system of spanning trees can be constructed in $O(m \log_2 n)$ time.*

Note that, since any additive r-spanner is a multiplicative $(r+1)$-spanner, Corollary 3.4 improves a known result of Peleg and Schäffer on sparse spanners of chordal graphs. In [58], they proved that any chordal graph with n vertices admits a multiplicative 3-spanner with at most $O(n \log_2 n)$ edges and a multiplicative 5-spanner with at most $2n - 2$ edges. Both spanners can be constructed in polynomial time. Note also that their result on multiplicative 5-spanners was earlier improved in [21], where the authors showed that any chordal graph with n vertices admits an additive 4-spanner with at most $2n - 2$ edges, constructable in linear time. Motivated by this and Corollary 3.5, it is natural to ask whether a system of constant number of collective additive tree 4-spanners exists for a chordal graph (or, generally, for which r, a system of constant number of collective additive tree r-spanners exists for any chordal graph). Recall that the problem whether a chordal graph admits a (one) multiplicative tree t-spanner is NP-complete for any $t > 3$ [14].

Peleg and Schäffer showed also in [58] that there are n-vertex chordal graphs for which any multiplicative 2-spanner will need to have at least $\Omega(n^{3/2})$ edges. This result leads to the following observation on collective additive tree 1-spanners of chordal graphs.

OBSERVATION 3.6. *There are n-vertex chordal graphs for which any system of collective additive tree 1-spanners will need to have at least $\Omega(\sqrt{n})$ spanning trees.*

Proof. Indeed, the existence of a system of $o(\sqrt{n})$ collective additive tree 1-spanners for a chordal graph will lead to the existence of an additive 1-spanner (and hence, of a multiplicative 2-spanner) with $o(n^{3/2})$ edges. □

4. Collective tree spanners in c-chordal graphs. A graph G is *c-chordal* if it does not contain any induced cycles of length greater than c; c-chordal graphs naturally generalize the class of chordal graphs. Chordal graphs are precisely the 3-chordal graphs.

THEOREM 4.1. *The class of c-chordal graphs is $(1/2, \lfloor c/2 \rfloor)$-decomposable.*

Proof. By Theorem 3.3 and since c-chordal graphs form a hereditary class of graphs, we need only to show that any c-chordal graph G admits a radius $\lfloor c/2 \rfloor$ acyclic covering. The existence of a radius $\lfloor c/2 \rfloor$ acyclic covering for G easily follows from a famous result of [43], which states that any c-chordal graph $G = (V, E)$ admits a special kind of *Robertson and Seymour tree-decomposition* [63]. That is, a tree $\mathcal{DT}(G)$, whose nodes are subsets of V, exists such that

- $(1')$ $\bigcup\{S : S \text{ is a node of } \mathcal{DT}(G)\} = V$;
- $(2')$ for any edge xy of G there is a node S of $\mathcal{DT}(G)$ with $x, y \in S$;
- $(3')$ for any tree nodes X, Y, Z of $\mathcal{DT}(G)$, if Y is on the path from X to Z in $\mathcal{DT}(G)$, then $X \cap Z \subseteq Y$;
- $(4')$ $diam_G(S) \leq \lfloor c/2 \rfloor$ for each node S of $\mathcal{DT}(G)$.

The reader might notice a close similarity between these four properties and the four properties from the definition of a radius r acyclic covering. In fact, they are almost equivalent. Note that $diam_G(S) \leq \lfloor c/2 \rfloor$ implies $rad_G(S) \leq \lfloor c/2 \rfloor$. Let $\mathcal{S}(G) = \{S : S \text{ is a node of } \mathcal{DT}(G)\}$ and consider a hypergraph $H = (V, \mathcal{S}(G))$. We claim that for a family $\mathcal{S}(G)$ of subsets of V, properties (1), (2) and (3) are equivalent to properties $(1')$, $(2')$ and $(3')$. Indeed, since, by property $(3')$, $v \in X \cap Z$ implies v belongs to any Y on the path of $\mathcal{DT}(G)$ from X to Z, for any vertex $v \in V$ the elements of $\mathcal{S}(G)$ containing vertex v induce a subtree in $\mathcal{DT}(G)$. Hence, by definition, $H = (V, \mathcal{S}(G))$ is an acyclic hypergraph. Conversely, let that for a graph G, a family $\mathcal{S}(G)$ of subsets of V satisfies properties (1), (2) and (3). Then, the acyclicity of the hypergraph $H = (V, \mathcal{S}(G))$ implies the existence of a tree T with node set $\mathcal{S}(G)$ such that for any vertex $v \in V$, the elements of $\mathcal{S}(G)$ containing v induce a subtree in T. Therefore, if two nodes X and Z of the tree T contain a vertex v then any node Y of T between X and Z must contain v, too. □

A balanced separator of radius at most $\lfloor c/2 \rfloor$ of a c-chordal graph G on n vertices and m edges can be found in $O(n^3)$ time as follows. Use an $O(nm)$ time algorithm from [33] to construct a Robertson–Seymour tree-decomposition $\mathcal{DT}(G)$ of G (it will have at most n nodes [33]). Then define the family $\mathcal{S}(G) = \{S : S \text{ is a node of } \mathcal{DT}(G)\}$ and consider the 2-section graph $2SEC(H)$ of an acyclic hypergraph $H = (V, \mathcal{S}(G))$. $2SEC(H)$ can be constructed in at most $O(n^3)$ time. Using an algorithm from [47], find a balanced separator C of a chordal graph $2SEC(H)$ in $O(n^2)$ time. We know that C is a maximal clique of $2SEC(H)$ and there must exist a set $S \in \mathcal{S}(G)$ which coincides with C (by conformality of H). As we showed earlier (see the proof of Theorem 3.3), $C = S$ is a balanced separator of radius at most $\lfloor c/2 \rfloor$ of G.

Thus, from Theorems 2.5 and 2.6, we conclude the following corollaries.

COROLLARY 4.2. *Any c-chordal graph G with n vertices admits an additive $(2\lfloor c/2\rfloor)$-spanner with at most $(n-1)\log_2 n$ edges, and such a sparse spanner can be constructed in $O(n^3 \log_2 n)$ time.*

COROLLARY 4.3. *Any c-chordal graph G with n vertices admits a system $\mathcal{T}(G)$ of at most $\log_2 n$ collective additive tree $(2\lfloor c/2\rfloor)$-spanners, and such a system of spanning trees can be constructed in $O(n^3 \log_2 n)$ time.*

Note that there are c-chordal graphs which do not admit any radius r acyclic covering with $r < \lfloor c/2\rfloor$. Consider, for example, the complement $\overline{C_6}$ of an induced cycle $C_6 = (a-b-c-d-e-f-a)$, which is a 4-chordal graph. A family $\mathcal{S}(\overline{C_6})$ consisting of one set $\{a,b,c,d,e,f\}$ gives a trivial radius $2 = \lfloor 4/2\rfloor$ acyclic covering of $\overline{C_6}$, and a simple consideration shows that no radius 1 acyclic covering can exist for $\overline{C_6}$ (it is impossible, by simply adding new edges to $\overline{C_6}$, to get a chordal graph in which each maximal clique induces a radius one subgraph of $\overline{C_6}$). In the next subsection we will show that yet an interesting subclass of 4-chordal graphs, namely, the class of chordal bipartite graphs, does admit radius 1 acyclic coverings.

4.1. Collective tree spanners in chordal bipartite graphs. A bipartite graph $G = (X \cup Y, E)$ is *chordal bipartite* if it does not contain any induced cycles of length greater than 4 [48].

For a chordal bipartite graph G, consider a hypergraph $H = (X \cup Y, \{N[y] : y \in Y\})$. In what follows we show that H is an acyclic hypergraph.

LEMMA 4.4. *The 2-section graph $2SEC(H)$ of H is chordal.*

Proof. First notice that any $y \in Y$ is simplicial in $2SEC(H)$ by construction of H and definition of $2SEC(H)$. Assume now, by way of contradiction, that there is an induced cycle C_p of length p, $p \geq 4$, in $2SEC(H)$. Necessarily, all vertices of C_p are from part X of G, since C_p is induced and all vertices from Y are simplicial in $2SEC(H)$. Let $C_p = (x_1, x_2, \ldots, x_p, x_1)$. For any edge $x_i x_{i+1}$ of C_p (including the edge $x_p x_1$), since it is not an edge of G, there must exist a vertex y_i in Y such that both x_i and x_{i+1} are adjacent to y_i in G. Also, since C_p is induced in $2SEC(H)$, y_i is not adjacent to any other vertex of C_p. Therefore, a cycle $(x_1, y_1, x_2, y_2, \ldots, x_p, y_p, x_1)$ of G must be induced. But, since its length is $2p \geq 8$, a contradiction with G being a chordal bipartite graph arises. □

LEMMA 4.5. *The hypergraph $H = (X \cup Y, \{N[y] : y \in Y\})$ is conformal.*

Proof. Let C be a clique of $2SEC(H)$ consisting of p vertices. First, note that, by definitions of H and $2SEC(H)$, the clique C can contain at most one vertex from Y. If C contains a vertex from Y (say $y \in C \cap Y$) then for all $v \in C \setminus \{y\}$, vy is an edge of G, and therefore $C \subseteq N[y]$ must hold. Let now $C \cap Y = \emptyset$. By induction on p we will show that there exists a vertex $y \in Y$ such that $C \subset N[y]$. Since G is connected, any vertex $x \in C \subseteq X$ has a neighbor in Y. Also, by definition of $2SEC(H)$, for any edge uv of $2SEC(H)$ with $u, v \in X$ there must exist a vertex y in Y adjacent to both u and v. Assume now, by induction, that each $p-1$ vertice of C has a common neighbor y in Y. Consider three different vertices a, b and c in C and three corresponding vertices a', b' and c' in Y such that $C \setminus \{a\} \subset N[a']$, $C \setminus \{b\} \subset N[b']$ and $C \setminus \{c\} \subset N[c']$. Since graph G cannot have any induced cycles of length 6, the cycle $(a - b' - c - a' - b - c' - a)$ of G cannot be induced. Without loss of generality, assume that a is adjacent to a' in G. But then, all p vertices of C are contained in $N[a']$. □

Since chordal bipartite graphs form a hereditary class of graphs and, for any chordal bipartite graph $G = (X \cup Y, E)$, a family $\{N[y] : y \in Y\}$ of subsets of $X \cup Y$

satisfies all four conditions of radius 1 acyclic covering, by Theorem 3.3 we have the following theorem.

THEOREM 4.6. *The class of chordal bipartite graphs is* $(1/2, 1)$*-decomposable.*

Hence, by Theorems 2.5 and 2.6, we immediately conclude the following corollaries.

COROLLARY 4.7. *Any chordal bipartite graph* G *with* n *vertices and* m *edges admits an additive* 2*-spanner with at most* $(n-1)\log_2 n$ *edges, and such a sparse spanner can be constructed in* $O(nm \log_2 n)$ *time.*

COROLLARY 4.8. *Any chordal bipartite graph* G *with* n *vertices and* m *edges admits a system* $\mathcal{T}(G)$ *of at most* $\log_2 n$ *collective additive tree* 2*-spanners, and such a system of spanning trees can be constructed in* $O(nm \log_2 n)$ *time.*

Recall that the problem whether a chordal bipartite graph admits a (one) multiplicative tree t-spanner is NP-complete for any $t > 3$ [15]. Also, any chordal bipartite graph G with n vertices admits an additive 4-spanner with at most $2n-2$ edges which is constructable in linear time [21]. Again, it is interesting to know whether a system of constant number of collective additive tree 4-spanners exists for a chordal bipartite graph. We have the following observation on collective additive tree 1-spanners for chordal bipartite graphs.

OBSERVATION 4.9. *There are chordal bipartite graphs on* $2n$ *vertices for which any system of collective additive tree* 1*-spanners will need to have at least* $\Omega(n)$ *spanning trees.*

Proof. Consider the complete bipartite graph $G = K_{n,n}$ on $2n$ vertices (which is clearly a chordal bipartite graph), and let $\mathcal{T}(G)$ be a system of μ collective additive tree 1-spanners of G. Then, for any two adjacent vertices x and y of G there must exist a spanning tree T in $\mathcal{T}(G)$ such that $d_T(x, y) \leq 2$. If $d_T(x, y) = 2$, then a common neighbor z of x and y in G would form a triangle with vertices x and y, which is impossible for $G = K_{n,n}$. Hence, $d_T(x, y) = 1$ must hold. Thus, any edge xy of G is an edge of some tree $T \in \mathcal{T}(G)$. Since there are n^2 graph edges to cover by spanning trees from $\mathcal{T}(G)$, we conclude $\mu \geq n^2/(2n-1) > n/2$. □

4.2. Collective tree spanners in cocomparability graphs. We will use the following definition of cocomparability graphs (see [16, 48, 56]). A graph G is a *cocomparability graph* if it admits a vertex ordering $\sigma = [v_1, v_2, \ldots, v_n]$, called a *cocomparability ordering*, such that, for any $i < j < k$, if v_i is adjacent to v_k, then v_j must be adjacent to v_i or to v_k. According to [56], such an ordering of a cocomparability graph can be constructed in linear time. It is well known also that cocomparability graphs are 4-chordal and they contain all interval graphs, all permutation graphs, and all trapezoid graphs (see, e.g., [16, 48] for the definitions).

Since $\overline{C_6}$ is a cocomparability graph, cocomparability graphs generally do not admit radius 1 acyclic coverings (although, we can show that both the class of permutation graphs and the class of trapezoid graphs do admit radius 1 acyclic coverings [28]). Here we will present a very simple direct proof for the statement that the class of cocomparability graphs is $(1/2, 1)$-decomposable.

THEOREM 4.10. *The class of cocomparability graphs is* $(1/2, 1)$*-decomposable. Moreover, for a given cocomparability graph* G *with* n *vertices and* m *edges a decomposition tree* $\mathcal{BT}(G)$ *can be constructed in* $O(m \log_2 n)$ *time.*

Proof. Let G be a cocomparability graph with a cocomparability ordering $\sigma = [v_1, v_2, \ldots, v_n]$. Consider the closed neighborhood of the vertex $v_{\lceil n/2 \rceil}$. We claim that the graph G' obtained from G by removing vertices of $N[v_{\lceil n/2 \rceil}]$ has no connected components with more that $n/2$ vertices. Indeed, there are no more than $n/2$ vertices

in G which are on the left (analogously, on the right) side of $v_{\lceil n/2 \rceil}$ with respect to σ. Also, if there is an edge connecting vertices v_i and v_j with $i < \lceil n/2 \rceil < j$, then at least one of these vertices must belong to $N[v_{\lceil n/2 \rceil}]$ as σ is a cocomparability ordering. Therefore, each connected component G_s of G' has at most $n/2$ vertices since it consists of vertices which are only on one side of $v_{\lceil n/2 \rceil}$. It is clear also that the ordering σ projected to the vertices of G_s gives a cocomparability ordering of G_s. Hence we can assume by induction that G_s is a $(1/2, 1)$-decomposable graph. □

Hence, we have the following corollaries.

COROLLARY 4.11. *Any cocomparability graph G with n vertices and m edges admits an additive 2-spanner with at most $(n-1)\log_2 n$ edges, and such a sparse spanner can be constructed in $O(m \log_2 n)$ time.*

COROLLARY 4.12. *Any cocomparability graph G with n vertices and m edges admits a system $\mathcal{T}(G)$ of at most $\log_2 n$ collective additive tree 2-spanners, and such a system of spanning trees can be constructed in $O(m \log_2 n)$ time.*

It is known [62] that any cocomparability graph admits a (one) additive tree 3-spanner. In a forthcoming paper [31], using different technique, we show that the result stated in Corollary 4.12 can further be improved. One can show that any cocomparability graph admits a system of two collective additive tree 2-spanners and there are cocomparability graphs which do not have any (one) additive tree 2-spanners. Since the complete bipartite graph $K_{n,n}$ is a cocomparability graph, from the proof of Observation 4.9, we also have the following observation.

OBSERVATION 4.13. *There are cocomparability graphs on n vertices for which any system of collective additive tree 1-spanners will need to have at least $\Omega(n)$ spanning trees.*

5. Collective tree spanners in circular-arc graphs. In this section we describe another way of obtaining a system of few collective additive tree spanners. We demonstrate it on the class of circular-arc graphs.

The *intersection graph* of a family of n sets is the graph where the vertices are the sets, and the edges are the pairs of sets that intersect. Every graph is the intersection graph of some family of sets. A graph $G = (V, E)$ is an *interval graph* if it is the intersection graph of a finite set of intervals (line segments) on a line. A graph G is a *circular-arc graph* if it is the intersection graph of a finite set of arcs on a circle. An interval graph is a special case of a circular-arc graph; it is a circular-arc graph that can be represented with arcs that do not cover the entire circle. Hence, if we remove from a circular-arc graph $G = (V, E)$ a vertex $v \in V$ together with its neighbors, the resulting graph will be interval [48] (see Figure 3 for an illustration).

It is well known that any interval graph admits an additive tree 2-spanner, and such a tree spanner is computable in linear time [61]. On the other hand, for any constant $r \geq 0$, there is a circular-arc graph without any additive tree r-spanner. Indeed, consider an induced cycle C_q on $q \geq 3$ vertices. Clearly, it is a circular-arc graph. Let P be an arbitrary spanning path of C_q and x and y be the end vertices of P. Then, trivially, $d_{C_q}(x, y) = 1$, $d_P(x, y) = q - 1$, i.e., a circular-arc graph C_q does not admit any additive tree $(q-3)$-spanner. In what follows we show that two spanning trees are enough to collectively additively 2-span any circular-arc graph.

Let $G = (V, E)$ be a circular-arc graph, u be its arbitrary vertex, and T_u be a BFS-tree of G rooted at u. Consider an interval graph G^- obtained from G by removing vertices of $N[u]$. For each connected component of G^-, compute its additive tree 2-spanner using a linear time algorithm from [61]. Extend obtained forest to a spanning tree T of the original graph G (see Figure 3 for an illustration).

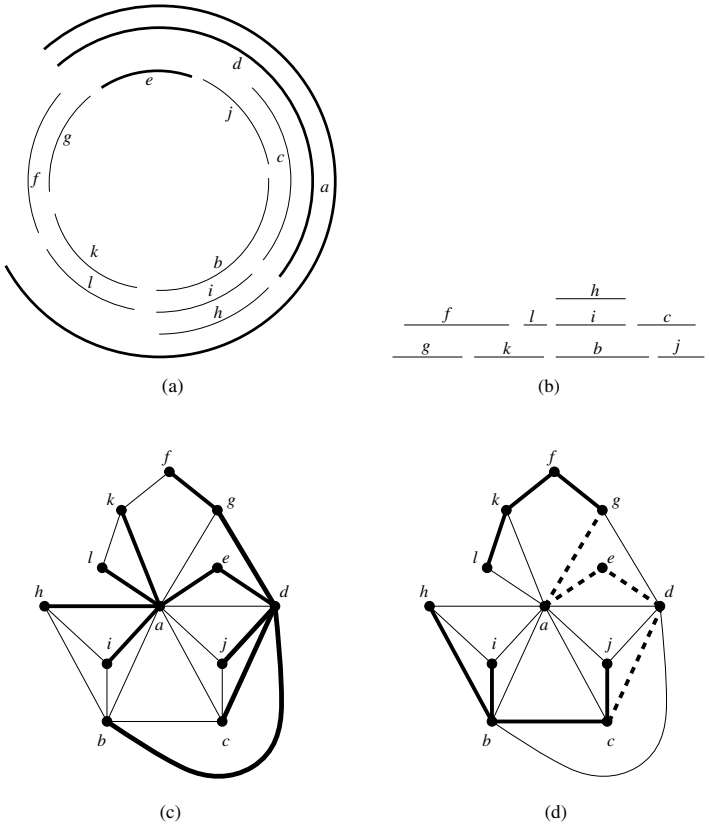


FIG. 3. (a) *A set S of circular-arcs (with arcs from $N[e]$ in bold);* (b) *the set of intervals corresponding to $S \setminus N[e]$;* (c) *the corresponding to S circular-arc graph G with a spanning tree T_e in bold;* (d) *a spanning tree T of G obtained from two local tree 2-spanners.*

LEMMA 5.1. *Spanning trees $\{T, T_u\}$ are collective additive tree 2-spanners of a circular-arc graph G.*

Proof. Let x and y be two arbitrary vertices of G. If there is a shortest path in G connecting vertices x and y and avoiding the neighborhood $N[u]$ of u, then $d_G(x,y) = d_{G^-}(x,y)$ and, by construction of T, $d_T(x,y) \leq d_{G^-}(x,y)+2 = d_G(x,y)+2$ holds. Let now a shortest path P connecting x and y in G intersect $N[u]$, and let v be a vertex from $N[u] \cap P$. Since T_u is a shortest-path tree of G rooted at u, we have $d_G(x,u) = d_{T_u}(x,u)$ and $d_G(u,y) = d_{T_u}(u,y)$. Hence, $d_G(x,y) = d_G(x,v)+d_G(v,y) \geq d_G(x,u) - 1 + d_G(u,y) - 1 = d_{T_u}(x,u) + d_{T_u}(u,y) - 2 \geq d_{T_u}(x,y) - 2$. ∎

Hence, we conclude the following theorem and corollary.

THEOREM 5.2. *Any circular-arc graph G admits a system of two collective additive tree 2-spanners, and such a system of spanning trees can be constructed in linear time.*

COROLLARY 5.3. *Any circular-arc graph G with n vertices and m edges admits an additive 2-spanner with at most $2n-2$ edges, and such a sparse spanner can be constructed in $O(m+n)$ time.*

6. Collective tree spanners and routing labeling schemes. Routing is one of the basic tasks that a distributed network of processors must be able to perform.

A *routing scheme* is a mechanism that can deliver packets of information from any vertex of the network to any other vertex. More specifically, a routing scheme is a distributed algorithm. Each processor in the network has a routing daemon running on it. This daemon receives packets of information and has to decide whether these packets have already reached their destination, and if not, how to forward them towards their destination. Each packet of information has a *header* attached to it. This header contains the address of the destination of the packet, and in some cases, some additional information that can be used to guide the routing of this message towards its destination. Each routing daemon has a local *routing table* at its disposal. It has to decide, based on this table and on the packet header only, whether to pass the packet to its host, or whether to forward the packet to one of its neighbors in the network.

The efficiency of a routing scheme is measured in terms of its *multiplicative stretch*, called *delay*, (or *additive stretch*, called *deviation*), namely, the maximum ratio (or surplus) between the length of a route, produced by the scheme for some pair of vertices, and their distance.

A straightforward approach for achieving the goal of guaranteeing optimal routes is to store a complete routing table in each vertex v in the network, specifying for each destination u the first edge (or an identifier of that edge, indicating the output port) along some shortest path from v to u. However, this approach may be too expensive for large systems since it requires a total of $O(n^2 \log d)$ memory bits in an n-processor network with maximum degree d [41]. Thus, an important problem in large-scale communication networks is the design of routing schemes that produce efficient routes and have relatively low memory requirements (see [3, 24, 35, 49, 57, 60, 68]).

This problem can be approached via localized techniques based on *labeling schemes* [57]. Informally speaking, the routing problem can be presented as requiring us to assign a label to every vertex of a graph. This label can contain the address of the vertex as well as the local routing table. The labels are assigned in such a way that at every source vertex v and given the address of any destination vertex u, one can decide the output port of an outgoing edge of v that leads to u. The decision must be taken locally in v, based solely on the label of v and the address of u.

Following [57], one can give the following formal definition. A family $\Re$ of graphs is said to *have an* $l(n)$ *routing labeling scheme* if there is a *function* L labeling the vertices of each n-vertex graph in $\Re$ with distinct labels of up to $l(n)$ bits, and there exists an efficient algorithm, called the *routing decision*, that given the label of a source vertex v and the label of the destination vertex (the header of the packet), decides in time polynomial in the length of the given labels and using only those two labels, whether this packet has already reached its destination, and if not, to which neighbor of v to forward the packet. Thus, the goal is, for a family of graphs, to find routing labeling schemes with small stretch factor, relatively short labels, and fast routing decision.

To obtain routing schemes for general graphs that use $o(n)$-bit label for each vertex, one has to abandon the requirement that packets are always routed on shortest paths, and settle instead for the requirement that packets are routed on paths with relatively small stretch [3, 4, 24, 35, 60, 68]. A delay-3 scheme that uses labels of size $\tilde{O}(n^{2/3})$ was obtained in [24], and a delay-5 scheme that uses labels of size $\tilde{O}(n^{1/2})$ was obtained in [35].[1] Recently, authors of [68] further improved these results. They presented a routing scheme that uses only $\tilde{O}(n^{1/2})$ bits of memory at each vertex of

[1]Here, $\tilde{O}(f)$ means $O(f \text{ polylog } n)$.

an n-vertex graph and has delay 3. Note that each routing decision takes constant time in their scheme, and the space is optimal, up to logarithmic factors, in the sense that every routing scheme with delay < 3 must use, on some graphs, routing tables of total size $\Omega(n^2)$, and hence $\Omega(n)$ at some vertex (see [39, 42, 45]).

There are many results on optimal (with delay 1) routing schemes for particular graph classes, including complete graphs, grids (alias meshes), hypercubes, complete bipartite graphs, unit interval and interval graphs, trees and 2-trees, rings, tori, unit circular-arc graphs, outerplanar graphs, and squaregraphs. All those graph families admit optimal routing schemes with $O(d \log n)$ labels and $O(\log d)$ routing decision. These results follow from the existence of special so-called *interval routing* schemes for those graphs. We will not discuss details of this scheme here; for precise definitions and an overview of this area, we refer the reader to [41].

Observe that in interval routing schemes the local memory requirement increases with the degree of the vertex. Routing labeling schemes aim at overcoming the problem of large degree vertices. In [40], a shortest-path routing labeling scheme for trees of arbitrary degree and diameter is described that assigns each vertex of an n-vertex tree a $O(\log^2 n/\log\log n)$-bit label. Given the label of a source vertex and the label of a destination it is possible to compute, in constant time, the neighbor of the source that heads in the direction of the destination. A similar result was independently obtained also in [68]. This result for trees was recently used in [32, 33] to design interesting low-deviation routing schemes for chordal graphs and general c-chordal graphs. Reference [32] describes a routing labeling scheme of deviation 2 with labels of size $O(\log^3 n/\log\log n)$ bits per vertex and $O(1)$ routing decision for chordal graphs. Reference [33] describes a routing labeling scheme of deviation $2\lfloor c/2 \rfloor$ with labels of size $O(\log^3 n)$ bits per vertex and $O(\log\log n)$ routing decision for the class of c-chordal graphs.

Our collective additive tree spanners give much simpler and easier to understand means of constructing compact and efficient routing labeling schemes for all (α, r)-decomposable graphs. We simply reduce the original problem to the problem on trees.

Let G be an (α, r)-decomposable graph and let $\mathcal{T}(G) = \{T^1, T^2, \ldots, T^\mu\}$ ($\mu \leq O(\log_2 n)$) be a system of μ collective additive tree $2r$-spanners of G. We can preprocess each tree T^i using the $O(n \log_2 n)$ algorithm from [40] and assign to each vertex v of G a tree label $L^i(v)$ of size $O(\log^2 n/\log\log n)$ bits associated with the tree T^i. Then we can form a label $L(v)$ of v of size $O(\log^3 n/\log\log n)$ bits by concatenating the μ tree labels. We store in $L(v)$ also the string A_v of length $O(\log^2 n)$ bits described in subsection 2.2. Thus, $L(v) := A_v \circ L^1(v) \circ \cdots \circ L^\mu(v)$.

Now assume that a vertex v wants to send a message to a vertex u. Given the labels $L(v)$ and $L(u)$, v first uses their substrings A_v and A_u to find in $\log_2 n$ time an index i such that for tree $T^i \in \mathcal{T}(G)$, $d_{T^i}(v, u) \leq d_G(v, u) + 2r$ holds. Then, v extracts from $L(u)$ the substring $L^i(u)$ and forms a header of the message $H(u) := i \circ L^i(u)$. Now, the initiated message with the header $H(u) = i \circ L^i(u)$ is routed to the destination using the tree T^i: when the message arrives at an intermediate vertex x, vertex x using own substring $L^i(x)$ and the string $L^i(u)$ from the header makes a constant time routing decision.

Thus, the following result is true.

THEOREM 6.1. *Each (α, r)-decomposable graph with n vertices and m edges admits a routing labeling scheme of deviation $2r$ with addresses and routing tables of size $O(\log^3 n/\log\log n)$ bits per vertex. Once computed by the sender in $\log_2 n$ time, headers never change. Moreover, the scheme is computable in $O((p(n) + t(n) + m + n\log_2 n)\log_2 n)$ time, and the routing decision is made in constant time per vertex,*

TABLE 1
Routing labeling schemes obtained for special graph classes via collective additive tree spanners.

Graph class	Scheme construction time	Addresses and routing tables (bits per vertex)	Message initiation time	Routing decision time	Deviation
Chordal	$O(m\log_2 n + n\log_2^2 n)$	$O(\log^3 n/\log\log n)$	$\log_2 n$	$O(1)$	2
Chordal bipartite	$O(nm\log_2 n)$	$O(\log^3 n/\log\log n)$	$\log_2 n$	$O(1)$	2
Cocomparability	$O(m\log_2 n + n\log_2^2 n)$	$O(\log^3 n/\log\log n)$	$\log_2 n$	$O(1)$	2
c-Chordal	$O(n^3\log_2 n)$	$O(\log^3 n/\log\log n)$	$\log_2 n$	$O(1)$	$2\lfloor c/2\rfloor$
Circular-arc	$O(n\log_2 n + m)$	$O(\log^2 n)$	$O(1)$	$O(1)$	2

where $p(n)$ is the time needed to find a balanced and bounded radius separator S and $t(n)$ is the time needed to find a central vertex for S.

Projecting this theorem to the particular graph classes considered in this paper, we obtain the following results summarized in Table 1. For circular-arc graphs, the labels are of size $O(\log^2 n)$ bits per vertex since this size labels are needed to decide in constant time which tree T or T_u is good for routing for given source x and destination y. We will choose tree $T' \in \{T, T_u\}$ such that $d_{T'}(x,y) = \min\{d_T(x,y), d_{T_u}(x,y)\}$. According to [57], in $O(n\log_2 n)$ total time the vertices of an n-vertex tree T can be labeled with labels of up to $O(\log^2 n)$ bits such that, given two labels of two vertices x, y of T, it is possible to compute in constant time the distance $d_T(x,y)$, by merely inspecting the labels of x and y.

7. Extension to the weighted graphs. Although in our previous discussions graph G is assumed (for simplicity) to be unweighted, the obtained results, in slightly modified form, are true even for weighted graphs.

Let $G = (V, E, w)$ be a *weighted graph* with the weight function $w : E \to R^+$. In a weighted graph G, the *length of a path* is the sum of the weights of edges participating in the path. The *distance* $d_G(x,y)$ between vertices x and y is the length of a shortest-length path connecting vertices x and y.

It is easy to see that, if in sections 2–4 we consider shortest path trees instead of BFS-trees, interpret r as an upper bound on the weighted radius of a balanced separator $S \subseteq V$, and denote the maximum edge weight by w, then the following corollaries from the previous results are true.

- Any weighted (α, r)-decomposable graph with n vertices, where r is an upper bound on the weighted radius of a balanced separator, admits a system of at most $\log_{1/\alpha} n$ collective additive tree $2r$-spanners.
- Any weighted c-chordal graph with n vertices admits a system of at most $\log_2 n$ collective additive tree $(2\lfloor c/2\rfloor\mathsf{w})$-spanners.
- Any weighted chordal, chordal bipartite, or cocomparability graph with n vertices admits a system of at most $\log_2 n$ collective additive tree $2\mathsf{w}$-spanners.

8. Conclusion and further developments. In this paper, we introduced a new notion of *collective tree spanners*, and showed that any (α, r)-decomposable graph G with n vertices admits a system of at most $\log_{1/\alpha} n$ collective additive tree $2r$-spanners. As a consequence, we got that any chordal graph, chordal bipartite graph

or cocomparability graph admits a system of at most $\log_2 n$ collective additive tree 2-spanners. We complemented these results by lower bounds, which say that any system of collective additive tree 1-spanners must have $\Omega(\sqrt{n})$ spanning trees for some chordal graphs and $\Omega(n)$ spanning trees for some chordal bipartite graphs and some cocomparability graphs. We also showed that every c-chordal graph admits a system of at most $\log_2 n$ collective additive tree $(2\lfloor c/2 \rfloor)$-spanners and every circular-arc graph admits a system of two collective additive tree 2-spanners. Furthermore, we discussed an application of the collective tree spanners to the problem of designing compact and efficient routing schemes in graphs.

Collective tree spanners can find applications also in designing compact and efficient distance labeling schemes for graphs, defined in [57]. As shown in [57], the vertices of any n-vertex tree T can be labeled with labels of up to $O(\log^2 n)$ bits such that, given two labels of two vertices x, y of T, it is possible to compute in constant time the distance $d_T(x, y)$ by merely inspecting the labels of x and y. Hence, any n-vertex graph G, admitting a system of μ collective additive tree r-spanners, admits a labeling that assigns $O(\mu \log^2 n)$ bit labels to vertices of G such that, given two labels of two vertices x, y of G, it is possible to compute in $O(\mu)$ time an additive r-approximation to the distance $d_G(x, y)$ by merely inspecting the labels of x and y, without using any other information about the graph.

In forthcoming papers [23, 29, 31], we investigate the collective tree spanners problem in other special families of graphs such as homogeneously orderable graphs, AT-free graphs, House–Hole–Domino-free graphs, graphs of bounded tree-width (including series-parallel graphs, outerplanar graphs), graphs of bounded asteroidal number, and others. We show that

- any homogeneously orderable graph admits a system of at most $\log_2 n$ collective additive tree 2-spanners and (one) additive tree 3-spanner,
- any House–Hole–Domino-free graph admits a system of at most $2 \log_2 n$ collective additive tree 2-spanners,
- any AT-free graph admits a system of two collective additive tree 2-spanners,
- any graph whose asteroidal number is bounded by a constant admits a system of a constant number of collective additive tree 3-spanners,
- any graph whose tree-width is bounded by a constant admits a system of at most $O(\log_2 n)$ collective additive tree 0-spanners,
- any graph whose clique-width is bounded by a constant admits a system of at most $O(\log_2 n)$ collective additive tree 2-spanners.

We conclude this paper with a few open questions/problems:

1. What is the complexity of the problem, "Given a graph G and integers μ, r, decide whether G has a system of at most μ collective additive tree r-spanners" for different $\mu \geq 1$, $r \geq 0$ on general graphs and on different restricted families of graphs?
2. What is the best trade-off between the number of trees μ and the additive stretch factor r on planar graphs? (So far, we can state only that any planar graph admits a system of $O(\sqrt{n})$ collective additive tree 0-spanners.)
3. What would be some more applications where collective tree spanners could be useful? The fact that collective tree spanners give a collection of (good) trees might make it easy to adapt many tree algorithms for the graphs that have collective tree r-spanners.

When this paper was already under review for this journal, we learned from A. Gupta that they introduced in [49] a notion of tree covers of graphs which is identical to our notion of collective multiplicative tree spanners. They additionally showed there

that any planar graph admits a system of at most $2\log_2 n$ collective multiplicative tree 3-spanners. This result makes question 2 even more intriguing.

Acknowledgments. We are very grateful to anonymous referees for many useful suggestions.

REFERENCES

[1] I. Althöfer, G. Das, D. Dobkin, D. Joseph, and J. Soares, *On sparse spanners of weighted graphs*, Discrete Comput. Geom., 9 (1993), pp. 81–100.
[2] G. Ausiello, A. D'Arti, and M. Moscarini, *Chordality properties on graphs and minimal conceptual connections in sematic data models*, J. Comput. System Sci., 33 (1986), pp. 179–202.
[3] B. Awerbuch, A. Bar-Noy, N. Linial, and D. Peleg, *Improved routing strategies with succinct tables*, J. Algorithms, 11 (1990), pp. 307–341.
[4] B. Awerbuch and D. Peleg, *Routing with polynomial communication-space tradeoff*, SIAM J. Discrete Math., 5 (1992), pp. 151–162.
[5] H.-J. Bandelt and A. Dress, *Reconstructing the shape of a tree from observed dissimilarity data*, Adv. in Appl. Math., 7 (1986), pp. 309-343.
[6] Y. Bartal, *Probabilistic approximations of metric spaces and its algorithmic applications*, in Proceedings of the 37th Annual Symposium on Foundations of Computer Science, IEEE, 1996, pp. 184–193.
[7] Y. Bartal, *On approximating arbitrary metrices by tree metrics*, Proceedings of the 13th Annual ACM Symposium on Theory of Computing, 198, pp. 161-168.
[8] J.-P. Barthélemy and A. Guénoche, *Trees and Proximity Representations*, Wiley, New York, 1991.
[9] C. Beeri, R. Fagin, D. Maier, and M. Yannakakis, *On the desirability of acyclic database schemes*, J. ACM, 30 (1983), pp. 479–513.
[10] C. Berge, *Hypergraphs*, North-Holland, Amsterdam, 1989.
[11] S. Bhatt, F. Chung, F. Leighton, and A. Rosenberg, *Optimal simulations of tree machines*, in Proceedings of the 27th Annual Symposium on Foundations of Computer Science, IEEE, 1986, pp. 274–282.
[12] U. Brandes and D. Handke, *NP–Completeness Results for Minimum Planar Spanners*, preprint, University of Konstanz, Konstanzer Schriften in Mathematik und Informatik, Nr. 16, Germany, 1996.
[13] A. Brandstädt, V. Chepoi, and F. F. Dragan, *Distance approximating trees for chordal and dually chordal graphs*, J. Algorithms, 30 (1999), pp. 166-184.
[14] A. Brandstädt, F. F. Dragan, H.-O. Le, and V. B. Le, *Tree spanners on chordal graphs: Complexity, algorithms, open problems*, in Proceedings of the 13th International Symposium on Algorithms and Computation, Lecture Notes in Comput. Sci. 2518, Springer, Berlin, 2002, pp. 163–174.
[15] A. Brandstädt, F. F. Dragan, H.-O. Le, V. B. Le, and R. Uehara, *Tree spanners for bipartite graphs and probe interval graphs*, in Graph-Theoretic Concepts in Computer Science, Lecture Notes in Comput. Sci. 2880, Springer, Berlin, 2003, pp. 106–118.
[16] A. Brandstädt, V. B. Le, and J. Spinrad, *Graph Classes: A Survey*, SIAM, Philadelphia, 1999.
[17] P. Buneman, *A characterization of rigid circuit graphs*, Discrete Math., 9 (1974), pp. 205–212.
[18] L. Cai, *Tree Spanners: Spanning Trees that Approximate the Distances*, Ph.D. thesis, University of Toronto, 1992.
[19] L. Cai and D. G. Corneil, *Tree spanners*, SIAM J. Discrete Math., 8 (1995), pp. 359–387.
[20] M. Charikar, C. Chekuri, A. Goel, S. Guha, and S. Plotkin, *Approximating a finite metric by a small number of tree metrics*, in Proceedings of the 39th Annual Symposium on Foundations of Computer Science, IEEE, 1998, pp. 379–388.
[21] V. D. Chepoi, F. F. Dragan, and C. Yan, *Additive spanners for k-chordal graphs*, Proceedings of the 5th Conference on Algorithms and Complexity, Lecture Notes in Comput. Sci. 2653, Springer, Berlin, 2003, pp. 96–107.
[22] L. P. Chew, *There are planar graphs almost as good as the complete graph*, J. Comput. System Sci., 39 (1989), pp. 205–219.
[23] D. G. Corneil, F. F. Dragan, E. Köhler, and C. Yan, *Collective tree 1-spanners for interval graphs*, in Graph-Theoretic Concepts in Computer Science, Lecture Notes in Comput. Sci. 3787, Springer, Berlin, 2005, pp. 151–162.

[24] L. Cowen, *Compact routing with minimum stretch*, in Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms, 1999, pp. 255–260.
[25] G. A. Dirac, *On rigid circuit graphs*, Abh. Math. Sem. Univ. Hamburg, 25 (1961), pp. 71–76.
[26] H. N. Djidjev, *On the problem of partitioning planar graphs*, SIAM J. Alg. Discrete Meth., 3 (1982), pp. 229–240.
[27] H. N. Djidjev, *A separator theorem for graphs of fixed genus*, Serdica, 11 (1985), pp. 319–329.
[28] F. F. Dragan and I. Lomonosov, *On compact and efficient routing in certain graph classes*, in Proceedings of the 15th Annual International Symposium on Algorithms and Computation, Lecture Notes in Comput. Sci. 3341, Springer, Berlin, 2004, pp. 402–414.
[29] F. F. Dragan and C. Yan, *Collective Tree Spanners of Homogeneously Orderable Graphs*, in preparation.
[30] F. F. Dragan, C. Yan, and I. Lomonosov, *Collective tree spanners of graphs*, in Proceedings of the 9th Scandinavian Workshop on Algorithm Theory, Lecture Notes in Comput. Sci. 3111, Springer, Berlin, 2004, pp. 64–76.
[31] F. F. Dragan, C. Yan, and D. G. Corneil, *Collective tree spanners and routing in AT-free related graphs*, in Graph-Theoretic Concepts in Computer Science, Lecture Notes in Comput. Sci. 3353, Springer, Berlin, 2004, pp. 68–80.
[32] Y. Dourisboure and C. Gavoille, *Improved compact routing scheme for chordal graphs*, in Proceedings of the 16th International Conference on Distributed Computing, Lecture Notes in Comput. Sci. 2508, Springer, Berlin, 2002, pp. 252–264.
[33] Y. Dourisboure and C. Gavoille, *Tree-Decompositions with Bags of Small Diameter*, Discrete Math., 2003, to appear.
[34] W. Duckwortha and M. Zito, *Sparse hypercube 3-spanners*, Discrete Appl. Math., 103 (2000), pp. 289–295.
[35] T. Eilam, C. Gavoille, and D. Peleg, *Compact routing schemes with low stretch factor*, in Proceedings of the 17th Annual ACM Symposium Prin. Distr. Comput., 1998, pp. 11–20.
[36] R. Fagin, *Degrees of acyclicity for hypergraphs and relational database schemes*, J. ACM, 30 (1983), pp. 514–550.
[37] J. Fakcharoenphol, S. Rao, and K. Talwar, *A tight bound on approximating arbitrary metrics by tree metrics*, in Proceedings of the 35th ACM Symposium on Theory of Computing, 2003, pp. 448–455.
[38] S. P. Fekete and J. Kremer, *Tree spanners in planar graphs*, Discrete Appl. Math., 108 (2001), pp. 85–103.
[39] P. Fraigniaud and C. Gavoille, *Memory requirements for universal routing schemes*, in Proceedings of the 14th Annual ACM Symposium Prin. Distr. Comput., 1995, pp. 223–230.
[40] P. Fraigniaud and C. Gavoille, *Routing in trees*, in Proceedings of the 28th Int. Colloquium on Automata, Languages and Programming, Lecture Notes in Comput. Sci. 2076, Springer, Berlin, 2001, pp. 757–772.
[41] C. Gavoille, *A survey on interval routing schemes*, Theoret. Comput. Sci., 245 (1999), pp. 217–253.
[42] C. Gavoille and M. Gengler, *Space-efficiency of routing schemes of stretch factor three*, J. Parallel and Distr. Comput., 61 (2001), pp. 679–687.
[43] C. Gavoille, M. Katz, N. A. Katz, C. Paul, and D. Peleg, *Approximate distance labeling schemes*, in Proceedings of the 9th Annual European Symposium on Algorithms, Lecture Notes in Comput. Sci. 2161, Springer, Berlin, 2001, pp. 476–487.
[44] C. Gavoille, D. Peleg, S. Pérennes, and R. Raz, *Distance labeling in graphs*, in Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms, 2001, pp. 210–219.
[45] C. Gavoille and S. Pérennès, *Memory requirements for routing in distributed networks*, in Proceedings of the 15th Ann. ACM Symposium on Prin. Distr. Comput., 1996, pp. 125–133.
[46] J. R. Gilbert, J. P. Hutchinson, and R. E. Tarjan, *A separator theorem for graphs of bounded genus*, J. Algorithms, 5 (1984), pp. 391–407.
[47] J. R. Gilbert, D. J. Rose, and A. Edenbrandt, *A separator theorem for chordal graphs*, SIAM J. Alg. Discrete Meth., 5 (1984), pp. 306–313.
[48] M. C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.
[49] A. Gupta, A. Kumar, and R. Rastogi, *Traveling with a Pez dispenser (or, routing issues in MPLS)*, SIAM J. Comput., 34 (2005), pp. 453–474. Appeared also in FOCS IEEE, 2001.
[50] M. Katz, N. A. Katz, and D. Peleg, *Distance labeling schemes for well-separated graph classes*, in Proceedings of the 17th Annual Symposium on Theoretical Aspects of Computer Science, Lecture Notes in Comput. Sci. 1770, Springer, Berlin, 2000, pp. 516–528.
[51] H.-O Le and V. B. Le, *Optimal tree 3-spanners in directed path graphs*, Networks, 34 (1999), pp. 81–87.

[52] A. L. Liestman and T. Shermer, *Additive graph spanners*, Networks, 23 (1993), pp. 343–364.
[53] R. J. Lipton and R. E. Tarjan, *A separator theorem for planar graphs*, SIAM J. Appl. Math., 36 (1979), pp. 177–189.
[54] R. J. Lipton and R. E. Tarjan, *Applications of a planar separator theorem*, SIAM J. Comput., 9 (1980), pp. 615–627.
[55] M. S. Madanlal, G. Venkatesan, and C. Pandu Rangan, *Tree* 3*-spanners on interval, permutation and regular bipartite graphs*, Inform. Process. Lett., 59 (1996), pp. 97–102.
[56] R. M. McConnell and J. P. Spinrad, *Linear-time transitive orientation*, in Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms, 1997, pp. 19–25.
[57] D. Peleg, *Distributed Computing: A Locality-Sensitive Approach*, SIAM Monogr. Discrete Math. Appl., SIAM, Philadelphia, 2000.
[58] D. Peleg and A. A. Schäffer, *Graph spanners*, J. Graph Theory, 13 (1989), pp. 99–116.
[59] D. Peleg and J. D. Ullman, *An optimal synchronizer for the hypercube*, in Proceedings of the 6th ACM Symposium on Prin. of Distr. Comput., 1987, pp. 77–85.
[60] D. Peleg and E. Upfal, *A tradeoff between space and efficiency for routing tables*, in Proceedings of the 20th ACM Symposium on the Theory of Computing, 1988, pp. 43–52.
[61] E. Prisner, *Distance approximating spanning trees*, in Proceedings of the 14th Annual Symposium on Theoretical Aspects of Computer Science, Lecture Notes in Comput. Sci. 1200, Springer, Berlin, 1997, pp. 499–510.
[62] E. Prisner, D. Kratsch, H.-O. Le, H. Müller, and D. Wagner, *Additive tree spanners*, SIAM J. Discrete Math., 17 (2003), pp. 332–340.
[63] N. Robertson and P. D. Seymour, *Graph minors. Algorithmic aspects of tree-width*, J. Algorithms, 7 (1986), pp. 309–322.
[64] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*, W. H. Freeman, San Francisco, 1973.
[65] J. Soares, *Graph spanners: A survey*, Congr. Numer., 89 (1992), pp. 225–238.
[66] D. L. Swofford and G. J. Olsen, *Phylogeny reconstruction*, in Molecular Systematics, D. M. Hillis and C. Moritz, eds., Sinauer Associates, Sunderland, MA, 1990, pp. 411–501.
[67] R. E. Tarjan and M. Yannakakis, *Simple linear time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs*, SIAM J. Comput., 13 (1984), pp. 566–579.
[68] M. Thorup and U. Zwick, *Compact routing schemes*, Proceedings of the 13th Annual ACM Symposium on Par. Alg. and Arch., 2001, pp. 1–10.
[69] G. Venkatesan, U. Rotics, M. S. Madanlal, J. A. Makowsky, and C. Pandu Rangan, *Restrictions of minimum spanner problems*, Inform. and Comput., 136 (1997), pp. 143–164.

SIAM J. DISCRETE MATH.
Vol. 20, No. 1, pp. 261–271

THE STEINER k-CUT PROBLEM*

CHANDRA CHEKURI†, SUDIPTO GUHA‡, AND JOSEPH (SEFFI) NAOR§

Abstract. We consider the Steiner k-cut problem which generalizes both the k-cut problem and the multiway cut problem. The Steiner k-cut problem is defined as follows. Given an edge-weighted undirected graph $G = (V, E)$, a subset of vertices $X \subseteq V$ called *terminals*, and an integer $k \leq |X|$, the objective is to find a minimum weight set of edges whose removal results in k disconnected components, each of which contains at least one terminal. We give two approximation algorithms for the problem: a greedy $(2 - \frac{2}{k})$-approximation based on Gomory–Hu trees, and a $(2 - \frac{2}{|X|})$-approximation based on rounding a linear program. We use the insight from the rounding to develop an exact bidirected formulation for the global minimum cut problem (the k-cut problem with $k = 2$).

Key words. multiway cut, k-cut, Steiner tree, minimum cut, linear program, approximation algorithm

AMS subject classifications. 68Q25, 68W25, 90C27, 90C59

DOI. 10.1137/S0895480104445095

1. Introduction.

The k-cut problem and the multiway cut problem are fundamental graph partitioning problems. In both problems we are given an undirected edge-weighted graph $G = (V, E)$ with $w(e)$ denoting the weight of edge $e \in E$. In the k-cut problem the goal is to find a minimum weight set of edges whose removal separates the graph into k disconnected components. In the multiway cut problem we are given a set of k terminals, $X \subseteq V$, and the goal is to find a minimum weight set of edges whose removal separates the graph into components such that each terminal is in a different connected component. In this paper we consider a generalization of the two problems, namely, the Steiner k-cut problem. In this problem, we are given an undirected weighted graph G, a set of *terminals* $X \subseteq V$, and an integer $k \leq |X|$. The goal is to find a minimum weight set of edges whose removal separates the graph into k components with vertex sets $V_1, V_2, \ldots, V_k$, such that $V_i \cap X \neq \emptyset$ for $1 \leq i \leq k$. If $X = V$, we obtain the k-cut problem. If $|X| = k$, we obtain the multiway cut problem.

The k-cut problem can be solved in polynomial time for fixed k [5, 6], but it is NP-complete when k is part of the input [5]. In contrast, the multiway cut problem is NP-complete for all $k \geq 3$ and is also APX-hard for all $k \geq 3$ [2]. It follows that the Steiner k-cut problem is NP-complete and APX-hard for all $k \geq 3$. For the multiway cut problem Calinescu, Karloff, and Rabani [1] gave a $1.5 - 1/k$ approximation using an interesting geometric relaxation. Karger et al. [7] improved the analysis of the

*Received by the editors July 12, 2004; accepted for publication (in revised form) September 26, 2005; published electronically March 24, 2006. A preliminary version appeared in Proceedings of the 30th International Colloquium on Automata, Languages, and Programming (ICALP), 2003, pp. 189–199.

http://www.siam.org/journals/sidma/20-1/44509.html

†Bell Labs, Lucent Technologies, 600 Mountain Ave., Murray Hill, NJ 07974 (chekuri@research.bell-labs.com). This author's research was supported in part by US-Israel BSF grant 2002276.

‡Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104 (sudipto@cis.upenn.edu). This author's research was supported in part by an Alfred P. Sloan Research Fellowship and by NSF Award CCF-0430376.

§Computer Science Dept., Technion, Haifa 32000, Israel (naor@cs.technion.ac.il). This author's research was supported in part by US-Israel BSF grant 2002276 and by EU contract IST-1999-14084 (APPOL II).

integrality gap of this relaxation and obtained an approximation ratio of $1.3438 - \epsilon_k$, where $\epsilon_k \to 0$ as $k \to \infty$. For the k-cut problem Saran and Vazirani [11] gave a $2 - \frac{2}{k}$ approximation algorithm using a greedy algorithm. This result was improved by [13] to $2 - \frac{3}{k}$ for odd k and to $2 - \frac{3k-4}{k^2-k}$ for even k. Recently, two different 2-approximations for the k-cut problem were obtained. The algorithm of Naor and Rabani [9] is based on rounding a linear programming (LP) formulation of the problem, and the algorithm of Ravi and Sinha [10] is based on the notion of network strength and Lagrangian relaxation.

The authors have learned of related independent work of Maeda, Nagamochi, and Ibaraki [8] (in Japanese) and Zhao, Nagamochi, and Ibaraki [14]. The Steiner k-cut was considered in [8], where it is shown that a greedy algorithm similar to the one we describe in this paper has an approximation ratio of $2 - 2/k$. In [14], the authors define a generalization of the Steiner k-cut problem which they refer to as the multiway partition problem (MPP). MPP is defined as follows. We are given a finite set V, a set of terminals $X \subseteq V$, and an integer k such that $|X| \geq k$. We are also given a submodular function f on V that assigns a real value $f(S)$ to each subset $S \subseteq V$. The function f is provided as an oracle. The goal is to partition V into k sets $V_1, V_2, \ldots, V_k$ such that $V_i \cap X \neq \emptyset$ for $1 \leq i \leq k$ and minimize $f(V_1) + f(V_2) + \cdots + f(V_k)$. It is shown in [14] that the greedy algorithm that iteratively increases the size of the partition yields a $(2 - \frac{2}{k})$-approximation for MPP. The Steiner k-cut problem can be seen to be a special case of MPP: given an edge-weighted graph $G = (V, E)$, we can define a submodular function f where $f(S) = \frac{1}{2} \sum_{e \in \delta_G(S)} w_e$.

1.1. Results. We provide two approximation algorithms for the Steiner k-cut problem. The first algorithm we present is combinatorial and has an approximation ratio of $(2 - \frac{2}{k})$. This algorithm is based on choosing cuts from the Gomory–Hu tree of the given graph and is similar to approximation algorithms developed for the k-cut problem and the multiway cut problem [12]. Maeda, Nagamochi, and Ibaraki [8] obtained the same result earlier, but our proof is considerably simpler. Also, as we mentioned earlier, Zhao, Nagamochi, and Ibaraki [14] show that the greedy algorithm yields a $(2 - \frac{2}{k})$ approximation for MPP. Our main result is a 2-approximation algorithm for the Steiner k-cut problem which is based on rounding a LP formulation. Although our formulation is a straightforward generalization of the formulation in [9] (for the k-cut problem), our rounding scheme differs substantially. The rounding in [9] exploits the properties of optimal solutions to the LP relaxation. These properties do not hold for the relaxation of the Steiner k-cut problem. Instead, we rely on the primal dual algorithm and the analysis of Goemans and Williamson [4] for the Steiner tree problem. As a consequence, our rounding algorithm extends to any feasible solution of the LP formulation. This interesting new connection might have future applications.

We conclude with a bidirected formulation for the global minimum cut problem and prove that the linear relaxation of this formulation is exact. The formulation and analysis are inspired by our analysis for the Steiner k-cut problem. This formulation and its integrality gap may have been known previously; however, we could not find a published reference and hence include it here.

2. Combinatorial $(2 - \frac{2}{k})$-approximation algorithm. We assume without loss of generality that the given graph G is connected. A natural greedy algorithm for the Steiner k-cut problem is the following iterative algorithm. In each iteration, find a minimum weight cut that increases the number of distinct components that contain a terminal. This algorithm has been shown to achieve a $(2 - \frac{2}{k})$-approximation algorithm

for both the k-cut problem and the multiway cut problem (see, e.g., [12]) and for MPP [14]. However, the analysis of this algorithm is nontrivial. As in [11, 12], we consider an alternative algorithm that is based on the Gomory–Hu tree representation of the minimum cuts in a graph. Recall that a Gomory–Hu tree for an edge-weighted undirected graph $G = (V, E)$ is an edge-weighted tree $T = (V, E_T)$ with weight function c that has the following property: for all $u, v \in V$, the weight of a minimum cut separating u and v in G is equal to the smallest edge weight on the unique path between u and v in T. In particular, for $(u, v) \in E_T$, $c(u, v)$ is the weight of the minimum cut separating u and v in G, and the partition of V induced by the removal of (u, v) from T induces such a minimum cut. We run the natural greedy algorithm mentioned above on the tree T: *Iteratively, pick the smallest weight edge in T separating a pair of terminals that are not already separated until k components, each of which contains a terminal, are generated.*

It is easy to see that we pick $k - 1$ edges in T. We take the union of the cuts associated with these edges and this defines our solution for the Steiner k-cut problem in G.

PROPOSITION 2.1. *The algorithm produces a feasible solution to the Steiner k-cut problem.*

We need a simple proposition about Gomory–Hu trees.

PROPOSITION 2.2. *Let $T = (V, E_T)$ be a Gomory–Hu tree for a connected graph $G = (V, E)$. For any pair of vertices (s, t) in G and an $s - t$ cut $(S, V - S)$ in G, there is an edge $(u, v) \in E_T$ such that $u \in S$, $v \in V - S$, and (u, v) lies on the path between s and t in T.*

Now we argue about the cost of the solution produced by the Gomory–Hu tree based algorithm. Our analysis is similar to that of the analysis for the Gomory–Hu tree based algorithm for the k-cut problem (see Theorem 4.8 in [12, page 42]). However, the analysis is not a straightforward extension; in the Steiner k-cut problem, the terminals constrain the choice of cuts, and we need to identify a mapping to the optimal set of cuts in a careful manner.

LEMMA 2.3. *The cost of the $(k - 1)$ edges picked by the algorithm is at most $(2 - 2/k)$ times the cost of the optimal solution.*

Proof. Fix an optimal solution A to the Steiner k-cut problem. Let $V_1, V_2, \ldots, V_k$ be the partitioning of V defined by A. Clearly, each set V_i $(i = 1, \ldots, k)$ contains at least one terminal from X. From each set V_i we arbitrarily choose a terminal t_i contained in V_i. Define cuts $A_i = (V_i, V \setminus V_i)$ for $i = 1, \ldots, k$, and let $w(A_i)$ denote the weight of cut A_i. Assume without loss of generality that $w(A_1) \leq w(A_2) \leq \cdots \leq w(A_k)$. Observe that each edge in the optimum solution A participates in exactly two of the cuts $A_1, \ldots, A_k$; hence the weight of the optimal solution A is $w(A) = \sum_{i=1}^{k} w(A_i)/2$. Let $B_1, B_2, \ldots, B_{k-1}$ denote the $k - 1$ cuts chosen by the above Gomory–Hu tree based algorithm. We claim that

$$w(B_i) \leq w(A_i), \qquad 1 \leq i \leq k - 1. \tag{1}$$

Assuming the claim, we have that

$$\sum_{i=1}^{k-1} w(B_i) \leq \left(1 - \frac{1}{k}\right) \sum_{i=1}^{k} w(A_i) \leq 2\left(1 - \frac{1}{k}\right) w(A),$$

which proves the desired bound on the performance of the algorithm.

To prove (1), we identify a set of edges $e_1, e_2, \ldots, e_{k-1}$ of the Gomory–Hu tree T with the following properties:

1. $w(A_i) \geq c(e_i)$, for $1 \leq i \leq k-1$, and since $w(A_1) \leq w(A_2) \leq \cdots \leq w(A_k)$, it follows that $w(A_i) \geq \max_{1 \leq j \leq i} c(e_i)$.
2. The removal of $e_1, e_2, \ldots, e_i$ creates $i+1$ components in T, each containing a terminal.

Assuming the existence of $e_1, e_2, \ldots, e_{k-1}$ as above, let $f_1, f_2, \ldots, f_{k-1}$ be the edges of T picked by the algorithm. We claim that $c(f_i) \leq \max_{1 \leq j \leq i} c(e_i)$; this follows by observing that there is some edge in $\{e_1, e_2, \ldots, e_i\}$ that when added to $\{f_1, \ldots, f_{i-1}\}$ would yield a new component containing a terminal. If not, removing the edges in $\{f_1, f_2, \ldots, f_{i-1}\} \cup \{e_1, \ldots, e_i\}$ would result in at most i components each containing a terminal which contradicts the definition of the e_i. Therefore,

$$w(B_i) = c(f_i) \leq \max_{1 \leq j \leq i} c(e_i) \leq w(A_i).$$

We obtain $e_1, \ldots, e_{k-1}$ as follows. Let $E' \subseteq E_T$ be the set of edges of T that cross the partition of V induced by the optimum solution $V_1, V_2, \ldots, V_k$. In other words, $(u,v) \in E'$ if and only if $(u,v) \in E_T$, $u \in V_i$, $v \in V_j$, and $i \neq j$; root the tree at t_k. For each t_i, $1 \leq i \leq k-1$, we let e_i be first edge in the directed path from t_i to the root t_k that is in E'; by Proposition 2.2, e_i exists. Also, for $i \neq j$, e_i and e_j are distinct; otherwise, the path between t_i and t_j in T would not have any edges in E' and this contradicts Proposition 2.2. Further, since e_i crosses the partition V_i, from the Gomory–Hu tree property, $w(A_i) \geq c(e_i)$. We claim that removing $e_1, e_2, \ldots, e_i$ from T will disconnect the set $\{t_1, t_2, \ldots, t_i, t_k\}$ in T. Suppose that this is not the case. Clearly, t_k is separated from $t_1, \ldots, t_i$; therefore for some $h, \ell \leq i$, t_h and t_ℓ are connected by a path P after removing $e_1, \ldots, e_i$. Let v be the least common ancestor of t_h and t_ℓ in T rooted at t_k. From our assumption e_h and e_ℓ are both above v. This implies that no edge in P is in E', and therefore P connects t_h and t_ℓ even after $e_1, \ldots, e_{k-1}$ are removed, contradicting Proposition 2.2. □

Given a Gomory–Hu tree for the input graph, the iterative greedy algorithm that we described can be easily implemented in $O(n^2)$ time. This potentially could be improved, but we do not attempt it since the running time to build a Gomory–Hu tree is currently $\Omega(n^2)$ even for sparse graphs. We conclude with the following theorem.

THEOREM 2.4. *There is a $(2 - \frac{2}{k})$-approximation algorithm for the Steiner k-cut problem that runs in $O(n^2 + \tau)$ time, where τ is the time required to build a Gomory–Hu tree for the input graph.*

3. LP formulation and a 2-approximation. We consider the following integer programming formulation for the Steiner k-cut problem. For each edge e we have a binary variable $d(e)$ which is 1 if the edge e belongs to the cut and 0 otherwise. Let T be a Steiner tree on the terminal set X in G. In any feasible Steiner k-cut, at least $k-1$ edges of T have to be cut. Based on this we obtain the following integer program for the Steiner k-cut problem:

$$\begin{aligned}
(K) \qquad \min \sum_{e \in E} w(e) \cdot d(e) \quad & \textit{subject to:} \\
\sum_{e \in T} d(e) \;&\geq\; k-1 \qquad \forall\, T : T \text{ Steiner tree on } X \\
d(e) \;&\in\; \{0,1\} \qquad \forall\, e \in E.
\end{aligned}$$

A relaxation of this integer program is obtained by allowing the variables $d(e)$ to assume values in $[0,1]$. The variables $d(e)$ are to be interpreted as inducing a semimet-

ric[1] on V. Our formulation above is a straightforward extension of the formulation of Naor and Rabani [9] for the k-cut problem. In the k-cut problem $X = V$, and hence [9] considers only spanning trees of G.

Unfortunately, we do not know how to solve the LP (K) in polynomial time. Consider, for example, the separation oracle required for running the Ellipsoid algorithm. Given edge weights $d(e)$, the separation oracle has to check that the minimum weight Steiner tree on X in G is of weight at least $k-1$. However, this problem is NP-hard. Note that for the k-cut problem, a polynomial time separation oracle is available because the minimum spanning tree (MST) of a graph can be computed in polynomial time.

We can use an approximate separation oracle based on the MST heuristic for the Steiner tree problem. Given edge weights $d(e)$, $e \in E$, we define the metric completion. For an unordered pair of vertices uv we let $d(uv)$ denote the shortest path distance from u to v in G with edge weights defined by d. Let G_X be the complete graph on the terminal set X. The oracle computes the MST on G_X where for each pair uv in G_X the weight of the edge uv is $d(uv)$. If the MST is of weight at least $k-1$, the oracle concludes that d is feasible. If the weight of the MST is less than $k-1$, it is easy to find a corresponding Steiner tree on X in G whose weight is less than $k-1$. In other words, we are solving the following relaxation:

$$(K') \qquad \min \sum_{uv \in E(G)} w(uv) \cdot d(uv) \quad \textit{subject to:}$$

$$\sum_{uv \in E(T)} d(uv) \geq k-1 \qquad T \text{ spanning tree in } G_X \tag{2}$$

$$d(uv) + d(vw) \geq d(uw) \qquad u, v, w \in V \tag{3}$$

$$d(uv) \in [0,1] \qquad u, v \in V. \tag{4}$$

For an edge $e \in E(G)$ with $e = uv$, we use $d(e)$ and $d(uv)$ interchangeably. The next lemma follows from the discussion.

LEMMA 3.1. *The LP (K') is a valid relaxation for the Steiner k-cut problem and it can be solved optimally in polynomial time.*

For the multiway cut problem we note that the LP (K') is equivalent to a LP that constrains the terminals to be at a distance of at least 1 from each other. This latter LP has been shown to have an integrality gap of $2(1-1/k)$ [2]. We will obtain the same result as well for the Steiner k-cut problem. We now prove a property of feasible solutions to (K') that will be useful later.

LEMMA 3.2. *In any feasible solution to (K') there is $X' \subseteq X$ such that $|X'| \geq k$, and for any two distinct vertices u and v in X', $d(uv) > 0$.*

Proof. For any two, not necessarily distinct, vertices u and v in X, define a relation R as follows: uRv if and only if $d(uv) = 0$. Since d is symmetric and satisfies triangle inequality (hence the relation is transitive), R defines an equivalence relation on X. We need to prove that the number of equivalence classes in R is at least k. Suppose this is not the case. For any two vertices a and b in V, $d_{ab} \leq 1$. Hence, there is a spanning tree on X of cost at most $\ell - 1$, where ℓ is the number of distinct equivalence classes. If $\ell < k$, we get a contradiction to the feasibility of the solution to (K'). ∎

[1] A semimetric is a distance function that is symmetric and satisfies triangle inequality. It differs from a metric in that it need not satisfy reflexivity, that is, distinct points can be at distance 0 from each other.

Note that the above proof is constructive and a set X' satisfying the required properties can be easily computed. In the rest of the paper it is convenient to assume that $X' = X$ and that for each $u, v \in X$, $d(uv) > 0$.

3.1. A strategy to round the LP. We show how to round a solution to (K') to yield a 2-approximation to the Steiner k-cut problem. To this end, we use the Goemans and Williamson primal-dual approximation algorithm for the Steiner tree problem [4] (henceforth referred to as the GW algorithm) to find a family of cuts.

Let $\bar{d}$ be any feasible solution to the LP (K'). Then, $\bar{d}$ defines a weight function on the edges of G. Let $G_{\bar{d}}$ denote the resulting edge-weighted graph. We run the GW primal-dual algorithm on the graph $G_{\bar{d}}$ to create a Steiner tree on X. To find a minimum Steiner tree on X in $G_{\bar{d}}$, the GW algorithm uses the following cut based LP relaxation of the Steiner tree problem. Let $x(e)$ be 1 if e is in the Steiner tree and 0 otherwise: every cut that separates the terminal set has to be covered by at least one edge. This yields the following LP where the variables are relaxed to be in $[0, 1]$. Note that the variables $\bar{d}(e)$ in the formulations below are treated as constants obtained from a solution to (K').

Each subset of vertices $S \subset V$ defines a cut which we denote by $\delta(S)$:

$$(STP) \qquad \min \sum_e \bar{d}(e) \cdot x(e) \quad \textit{subject to:}$$

$$\sum_{e \in \delta(S)} x(e) \geq 1 \qquad \forall\, S : S \text{ separates } X \tag{5}$$

$$x(e) \in [0, 1] \qquad \forall\, e. \tag{6}$$

The dual of this LP is the following:

$$(STD) \qquad \max \sum_S y(S) \quad \textit{subject to:}$$

$$\sum_{S : e \in \delta(S)} y(S) \leq \bar{d}(e) \qquad \forall \text{ e} \tag{7}$$

$$y(S) \geq 0 \qquad \forall\, S : S \text{ separates } X. \tag{8}$$

The GW algorithm is a primal-dual algorithm that incrementally grows a dual solution while maintaining feasibility and computes a corresponding feasible primal Steiner tree such that the cost of the Steiner tree computed is at most twice the value of the dual solution found. Let y' be the dual solution produced by the GW algorithm upon termination and let T be the tree returned by the algorithm. Then the following properties hold for y' and T [4].

1. y' is a feasible solution to (STD).
2. T is a tree that spans the terminal set X.
3. Sets S (representing cuts) with $y'(S) > 0$ form a laminar family. Let $\mathcal{S}$ denote this family of sets.
4. $\sum_{e \in T} \bar{d}(e) \leq 2(1 - 1/|X|) \sum_{S \in \mathcal{S}} y'(S)$.
5. For any $u \in X$, $\sum_{S : u \in S} y'(S) \leq \frac{1}{2} \cdot \max_{v \in X, v \neq u} \bar{d}(uv) \leq \frac{1}{2}$.
6. For any $u, v \in X$ such that $\bar{d}(uv) > 0$, there exists a cut S such that $y'(S) > 0$ and $|S \cap \{u, v\}| = 1$.

With the above discussion in place, we are ready to describe our rounding procedure. For a cut S, let $w(S) = \sum_{e \in \delta(S)} w(e)$ denote the weight of S in G; we observe the following claim.

CLAIM 3.3. $\sum_{S\in\mathcal{S}} y'(S)w(S) \le \sum_e w(e)\bar{d}(e)$.

Proof. We have the following:

$$\begin{aligned}\sum_{S\in\mathcal{S}} y'(S)w(S) &= \sum_{S\in\mathcal{S}} y'(S) \sum_{e\in\delta(S)} w(e)\\ &= \sum_e w(e) \sum_{S:e\in\delta(S)} y'(S)\\ &\le \sum_e w(e)\bar{d}(e).\end{aligned}$$

The final inequality follows from constraint (7) since y' is a feasible solution to (STD). □

CLAIM 3.4. $2(1-1/|X|)\sum_{S\in\mathcal{S}} y'(S) \ge (k-1)$.

Proof. The GW algorithm guarantees that $2(1-1/|X|)\sum_S y'(S) \ge \sum_{e\in T}\bar{d}(e)$. Since T is a spanning tree on X, from the feasibility of $\bar{d}$ for (K'), $\sum_{e\in T}\bar{d}(e) \ge k-1$ by (2). The claim follows by combining the two equalities. □

3.2. Choosing the cuts. We describe how we choose the cuts from $\mathcal{S}$. We partition $\mathcal{S}$ into classes $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_\ell$ such that two cuts S and S' are in the same class $\mathcal{S}_i$ if and only if $S\cap X = S'\cap X$. Clearly, the number of classes is at least $|X| \ge k$. For a class $\mathcal{S}_i$, let C_i be a least weight cut in $\mathcal{S}_i$. Let $\mathcal{C}$ be the collection of C_i, $1 \le i \le \ell$. Without loss of generality assume that the classes are ordered such that $w(C_1) \le w(C_2) \le \ldots \le w(C_\ell)$.

A solution to the problem consists of a set of edges. Our algorithm outputs a collection of cuts from $\mathcal{C}$ with the solution consisting of all edges that belong to one of the chosen cuts; we therefore think of the cuts as defining the solution. The algorithm considers classes in increasing order of their index and while considering class $\mathcal{S}_i$, adds C_i to the solution if adding the cut produces a new component containing a terminal from X. The process stops when $k-1$ cuts are chosen. This procedure is well defined and yields a feasible solution for the following reason. From Lemma 3.2 and property 6 of the GW algorithm, if all the cuts $C_1, C_2, \ldots, C_\ell$ are chosen, we obtain k (or more) components, each containing a terminal from X. We now upper bound the value of the solution output by the algorithm. Let $1 = i_1 < i_2 < \cdots < i_{k-1} < \ell$ denote the indices of the $k-1$ classes chosen by the algorithm. We let $y'(\mathcal{S}_i)$ denote $\sum_{S\in\mathcal{S}_i} y'(S)$.

DEFINITION 3.5. *Given a collection of distinct cuts $\mathcal{B}$, we say that a cut $C\in\mathcal{B}$ is* basic *with respect to $\mathcal{B}$ if there is no cut $C'\in\mathcal{B}$ such that $C'\subsetneq C$.*

From the laminarity of $\mathcal{S}$ and hence of $\mathcal{B}$, the set of basic cuts in $\mathcal{B}$ is well defined and disjoint. Let $\mathcal{A}_j$ denote the set of cuts $C_1, C_2, \ldots, C_j$.

LEMMA 3.6. *Let q_j be the number of basic cuts in $\mathcal{A}_j$ and let p_j be the number of components created by the algorithm after the first j cuts have been considered. Then*

- $\sum_{1\le h\le j} y'(\mathcal{S}_h) \le q_j/2$,
- *$p_j \ge q_j$, and if $p_j = q_j$, then the components are induced by the basic cuts in $\mathcal{A}_j$ and $X \subset \cup_{h=1}^{j} C_j$.*

Proof. From the analysis of the GW algorithm we have that for any cut S, $\sum_{S'\supseteq S} y'(S') \le \Delta/2$, where Δ is the diameter of G. In our case $\Delta = 1$. Since every cut in $\mathcal{A}_j$ is a superset of some basic cut in $\mathcal{A}_j$, we have that $\sum_{1\le h\le j} y'(\mathcal{S}_h) \le q_j/2$.

Let $r_1 < r_2 < \cdots < r_{q_j}$ be the indices of the basic cuts in $\mathcal{A}_j$. Note that the cuts in $\mathcal{S}$ are laminar and hence these basic cuts are disjoint. We now argue that $p_j \ge q_j$. Let $X_h = X \cap C_{r_h}$, $1 \le h \le q_j$, and let $X' = X - \uplus_{h=1}^{q_j} X_h$. We claim that for $h < h'$, X_h and $X_{h'}$ are in separate components; otherwise, the algorithm when processing

C_{r_h} would add it to the solution and separate X_h and $X_{h'}$; therefore $p_j \geq q_j$. By the same argument, it follows that if X' is not empty, X_h and X' are in separate components as well and in this case $p_j \geq q_j + 1$. Thus, if $p_j = q_j$, $X' = \emptyset$, and each X_h is in a separate component. ∎

Let $\alpha = 1/(1 - 1/|X|)$. From the analysis of the GW algorithm we have that $\sum_{h=1}^{\ell} y'(\mathcal{S}_h) = \sum_S y'(S) \geq \alpha(k-1)/2$. The main tool in our analysis is the following lemma.

LEMMA 3.7. *For $1 \leq r \leq k-1$, $\sum_{j \geq i_r} y'(\mathcal{S}_j) \geq \alpha(k-r)/2$.*

Proof. Let $f = i_r - 1$, then $p_f = r$. We consider two cases based on q_f.

If $p_f > q_f$, we have that $q_f \leq r-1$, and by Lemma 3.6, $\sum_{1 \leq h \leq f} y'(\mathcal{S}_h) \leq (r-1)/2$. Since $\sum_{1 \leq h \leq \ell} y'(\mathcal{S}_h) \geq \alpha(k-1)/2$ it follows that $\sum_{i_r \leq j \leq \ell} y'(\mathcal{S}_j) \geq \alpha(k-r)/2$.

Now we consider the case $p_f = q_f$. From Lemma 3.6, the components at this stage are induced by the basic cuts in $\mathcal{A}_f$. Let the basic cuts be $C_{j_1}, C_{j_2}, \ldots, C_{j_r}$. Let X_h denote the terminals in C_{j_h}. Recall that $X = \uplus_h X_h$ and hence $\sum_{1 \leq h \leq r} |X_h| = |X|$. The tree T created by the GW algorithm is of cost $k-1$. We note that the part of the tree that connects the components $C_{j_1}, C_{j_2}, \ldots, C_{j_r}$ costs at most $r-1$ since the diameter of the graph is at most 1. For $1 \leq h \leq r$, let T_h be the minimal subtree of T that connects X_h. It follows that $\sum_{1 \leq h \leq r} \sum_{e \in T_h} \bar{d}_e \geq k - 1 - (r-1) \geq k - r$. Let $L_h = \{i \mid (C_i \cap X) \subsetneq X_h\}$ be the indices of classes that contain a proper subset of terminals from X_h. From the analysis of the GW algorithm applied to tree T_h and terminals set X_h, we obtain that

$$\sum_{i \in L_h} y'(\mathcal{S}_i) \geq \frac{1}{2(1 - 1/|X_h|)} \sum_{e \in T_h} \bar{d}_e;$$

therefore

$$\sum_{1 \leq h \leq r} \sum_{i \in L_h} y'(\mathcal{S}_i) \;\geq\; \sum_{1 \leq h \leq r} \frac{1}{2(1 - 1/|X_h|)} \sum_{e \in T_h} \bar{d}_e \;\geq\; \frac{1}{2(1 - 1/|X|)}(k - r).$$

We now claim that if $i \in \uplus_h L_h$, then $i > f = i_r - 1$. For if $i \in L_h$, then C_{j_h} would not be basic in $C_1, C_2, \ldots, C_f$; therefore

$$\sum_{j \geq i_r} y'(\mathcal{S}_h) \;\geq\; \sum_{1 \leq h \leq r} \sum_{i \in L_h} y'(\mathcal{S}_i).$$

This finishes the proof of the lemma. ∎

COROLLARY 3.8. $\sum_{r=1}^{k-1} w(C_{i_r}) \leq 2(1 - 1/|X|) \leq \sum_S y'(S) w(S)$.

Proof. For $1 \leq h \leq \ell$ let $z_h = \sum_{j \geq h} y'(\mathcal{S}_j)$. Recall that $1 = i_1 < i_2 < \cdots < i_{k-1} < \ell$ are the indices of the cuts chosen by the algorithm and that $w(C_1) \leq w(C_2) \leq \cdots \leq w(C_\ell)$; hence,

$$\begin{aligned}
\sum_S y'(S) w(S) &= \sum_{h=1}^{\ell} \sum_{S \in \mathcal{S}_h} y'(S) w(S) \\
&\geq \sum_{h=1}^{\ell} y'(\mathcal{S}_h) w(C_h) \\
&\geq w(C_{i_{k-1}}) z_{i_{k-1}} + \sum_{r=1}^{k-2} w(C_{i_r})(z_{i_r} - z_{i_{r+1}}).
\end{aligned}$$

From Lemma 3.7 we have that $z_{i_r} \geq \alpha(k-r)/2$. The right-hand side of the last inequality above is minimized when $z_{i_r} = \alpha(k-r)/2$ for $1 \leq r \leq k-1$. Therefore,

$$\sum_S y'(S)w(S) \geq \frac{1}{2}\alpha \sum_{r=1}^{k-1} w(C_{i_r}).$$

This yields the desired inequality. □

From Corollary 3.8 and Claim 3.3 we obtain that

$$\sum_{r=1}^{k-1} w(C_{i_r}) \;\leq\; 2(1-1/|X|)\sum_S y'(S)w(S) \;\leq\; 2(1-1/|X|)\sum_e w_e \bar{d}_e.$$

Thus the integrality gap of (K') is upper bounded by $2(1-1/|X|)$.

Lower bound on the integrality gap. The integrality gap of (K') (and (K)) is no better than $2(1-1/|X|)$ even when $k=2$ and $X=V$ (the global minimum cut problem). Consider the unit weight cycle on n vertices. Clearly, an integral solution has to cut at least two edges to separate the cycle into two components. Consider the following feasible solution to the relaxation. We set $d(e) = 1/(n-1)$ on each edge of the cycle; for all other edges, $d(e)$ is the shortest path distance induced by the distances on the cycle edges. The value of this solution is $n/(n-1)$. Hence, the integrality gap is $2(1-1/n)$.

THEOREM 3.9. *The integrality gap of the LP (K') is $2(1-1/|X|)$.*

4. An exact formulation for the global minimum cut problem. In the previous section we saw that LP (K') has an integrality gap of $2(1-1/n)$ for the 2-cut problem, i.e., for the global minimum cut problem. Here we give a bidirected formulation of the global minimum cut problem. Given an undirected weighted graph $G=(V,E)$, let $G^b=(V,A)$ be the directed graph obtained by replacing each edge $e \in E$ between u and v by two directed arcs (u,v) and (v,u). The weights of both (u,v) and (v,u) in G^b are set to $w(e)$. Let r be any vertex in $V(G)$. An *arborescence* in a directed graph rooted at a vertex r is a spanning out-tree from r (also known as a *branching*). Our formulation is based on G^b. For an arc $a \in A$, let $d(a)=1$ if a is chosen to the cut, and let $d(a)=0$ otherwise. The following is a valid integer program for the global minimum cut problem:

$$\begin{aligned}
(B) \qquad \min \sum_{a\in A} w(a)\cdot d(a) \quad & \textit{subject to:} \\
\sum_{a\in T} d(a) \;&\geq\; 1 \qquad\qquad T \text{ arborescence rooted at } r \text{ in } G^b \\
d(a) \;&\in\; \{0,1\} \qquad\qquad a \in A.
\end{aligned}$$

Although the above integer program is similar to integer program (K), we remark that for $k>2$ we do not obtain a valid formulation for the k-cut problem if we replace the right-hand side of the constraint above by $k-1$.

We obtain a LP by relaxing each variable $d(a)$ to be in $[0,1]$. We show that the value of the LP is exactly equal to the global minimum cut of the graph G. The separation oracle needed to solve (B) in polynomial time by the Ellipsoid algorithm is the minimum cost arborescence problem in directed graphs. We can use the algorithm of Edmonds [3] for this purpose. In fact, Edmonds [3] showed that the arborescence polytope is integral and we use this to show that (B) is exact for the minimum cut

problem. The proof is similar in outline to the one in section 3, but we use arborescences in place of spanning trees, and the result of Edmonds [3] on the integrality of the arborescence polytope in place of the GW algorithm. Let $\bar{d}$ be an optimal solution to (B). Let G_d^b be the graph G^b equipped with $\bar{d}$ as costs on the edges of G^b. We find a minimum cost arborescence in G_d^b using the following formulation. For each arc a, variable $x(a) = 1$ if a belongs to the arborescence and 0 otherwise:

$$
\begin{aligned}
(AP) \qquad \min \sum_{a \in A} d(a) \cdot x(a) \quad & \textit{subject to:} \\
\sum_{a \in \delta(S)} x(a) \;&\geq\; 1 && \forall\, S : S \neq V \text{ and } r \in S \\
x(a) \;&\in\; [0,1] && \forall\, a.
\end{aligned}
$$

The dual of the above LP is the following:

$$
\begin{aligned}
(AD) \qquad \max \sum_{S} y(S) \quad & \textit{subject to:} \\
\sum_{S : a \in \delta(S)} y(S) \;&\leq\; d(a) && \forall \text{ a} \\
y(S) \;&\geq\; 0 && \forall\, S : S \neq V \text{ and } r \in S.
\end{aligned}
$$

Let $\bar{x}^*$ and $\bar{y}^*$ be optimal primal and dual solutions to (AP) and (AD) on the graph G_d^b. From the feasibility of $\bar{d}$, it follows that $\sum_a d(a)x^*(a) \geq 1$. From weak duality we therefore also obtain that $\sum_S y^*(S) \geq 1$. Let $\mathcal{S} = \{S \mid y^*(S) > 0\}$ be the set of all cuts with strictly positive dual values. Let $C \in \mathcal{S}$ be a cut such that $w(S)$ is the cheapest cut. We pick C as our solution. We now show that $w(C) \leq \sum_a w(a)d(a)$, which shows that the weight of the cut is at most the value of the optimal solution to (B). We see that

$$
\begin{aligned}
\sum_S y^*(S)w(S) &= \sum_S y^*(S) \sum_{a \in \delta(S)} w(a) \\
&= \sum_a w(a) \sum_{S : a \in \delta(S)} y^*(S) \\
&\leq \sum_a w(a)d(a).
\end{aligned}
$$

The last inequality follows from the feasibility of y^*. We have that $\sum_S y^*(S)w(S) \leq \sum_a w(a)d(a)$ and $\sum_S y^*(S) \geq 1$. Therefore, the weight of the cheapest cut is no more than $\sum_a w(a)d(a)$.

THEOREM 4.1. *The LP relaxation of (B) can be solved in polynomial time and is an exact formulation for the global minimum cut problem.*

5. Conclusions. Our study of LP relaxations for the Steiner k-cut problem was partly motivated by the goal of obtaining an approximation algorithm for the k-cut problem with a ratio better than 2. This has been accomplished for the multiway cut problem by a strengthened LP relaxation [1]. Our results show that the available approximation techniques for the k-cut problem extend to the Steiner k-cut problem. In the process we have shown an interesting connection between laminar cut families obtained from the primal-dual algorithm of Goemans and Williamson [4] and their use in analyzing the LP relaxation for the Steiner k-cut problem. Several interesting questions are open.

- Is the k-cut problem APX-hard?
- Is there an approximation algorithm for the k-cut problem with ratio better than 2?
- What is the integrality gap of the geometric relaxation in [1] for the multiway cut problem?

Acknowledgments. We thank David Shmoys and Zoya Svitkina for pointing out an erroneous proof in a previous version of the paper. We thank two anonymous referees for comments which helped improve the clarity of the proofs in section 3.2.

REFERENCES

[1] G. Călinescu, H. Karloff, and Y. Rabani, *An improved approximation algorithm for MULTIWAY CUT*, J. Comput. System Sci., 60 (2000), pp. 564–574.
[2] E. Dahlhaus, D. S. Johnson, C. H. Papadimitriou, P. D. Seymour, and M. Yannakakis, *The complexity of multiterminal cuts*, SIAM J. Comput., 23 (1994), pp. 864–894.
[3] J. Edmonds, *Optimum branchings*, J. Res. Nat. Bur. Standards, 71B (1967), pp. 233–240.
[4] M. Goemans and D. P. Williamson, *A general approximation technique for constrained forest problems*, SIAM J. Comput., 24 (1995), pp. 296–317.
[5] O. Goldschmidt and D. S. Hochbaum, *Polynomial algorithm for the k-cut problem*, Math. Oper. Res., 19 (1994), pp. 24–37.
[6] D. R. Karger and C. Stein, *A new approach to the minimum cut problem*, J. ACM, 43 (1996), pp. 601–640.
[7] D. R. Karger, P. Klein, C. Stein, M. Thorup, and N. E. Young, *Rounding algorithms for a geometric embedding of minimum multiway cut*, Math. Oper. Res., 29 (2004), pp. 436–461.
[8] N. Maeda, H. Nagamochi, and T. Ibaraki, *Approximate algorithms for multiway objective point split problems of graphs (in Japanese)*, Computing devices and algorithms (in Japanese) (Kyoto, 1993). Surikaisekikenkyusho Kokyuroku, 833 (1993), pp. 98–109.
[9] J. Naor and Y. Rabani, *Approximating k-cuts*, in Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms, Washington, DC, 2001, pp. 26–27.
[10] R. Ravi and A. Sinha, *Approximating k-cuts via network strength*, in Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, 2002, pp. 621–622.
[11] H. Saran and V. Vazirani, *Finding k cuts within twice the optimal*, SIAM J. Comput., 24 (1995), pp. 101–108.
[12] V. Vazirani, *Approximation Algorithms*, Springer, Berlin, 2001.
[13] L. Zhao, H. Nagamochi, and T. Ibaraki, *Approximating the minimum k-way cut in a graph via minimum 3-way cuts*, J. Combin. Optim., 5 (2001), pp. 397–410.
[14] L. Zhao, H. Nagamochi, and T. Ibaraki, *Greedy splitting algorithms for approximating multiway partition problems*, Math. Program. A, 102 (2005), pp. 167–183.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 273–286

MULTICOLORED HAMILTON CYCLES AND PERFECT MATCHINGS IN PSEUDORANDOM GRAPHS*

DANIELA KÜHN† AND DERYK OSTHUS†

Abstract. Given $0 < p < 1$, we prove that a pseudorandom graph G with edge density p and sufficiently large order has the following property: Consider any red/blue-coloring of the edges of G and let r denote the proportion of edges which have the color red. Then there is a Hamilton cycle C so that the proportion of red edges of C is close to r. The analogue also holds for perfect matchings instead of Hamilton cycles. We also prove a bipartite version which is used elsewhere to give a minimum-degree condition for the existence of a Hamilton cycle in a 3-uniform hypergraph.

Key words. Hamilton cycles, perfect matchings, random graphs, pseudorandom graphs

AMS subject classifications. 05C45, 05C70, 05C80

DOI. 10.1137/050627010

1. Introduction.

1.1. Overview. It is well known that random graphs, pseudorandom graphs, and ε-superregular graphs have some strong Hamiltonicity properties in common. For instance, a recent result of Frieze and Krivelevich [10] states that, for every constant $0 < p < 1$, with high probability almost all edges of the random graph $G_{n,p}$ can be packed into edge-disjoint Hamilton cycles. (They derive this from a similar result about ε-superregular graphs.)

Hamiltonicity has also been investigated from the viewpoint of (anti-)Ramsey theory. For example, Albert, Frieze, and Reed [1] proved that there is a linear function $k = k(n)$ such that for every edge-coloring of the complete graph K_n on n vertices which uses each color at most k times there is a Hamilton cycle where each edge has a different color. This improves bounds by previous authors. A related problem for random graphs was also considered by Cooper and Frieze [6].

Here, we prove a related result about colorings of bipartite ε-superregular graphs (which will imply analogous statements for pseudorandom and random graphs). Roughly speaking, we prove that given a k-coloring of a sufficiently large ε-superregular graph G (where ε is sufficiently small) there is a Hamilton cycle C in G which is strongly multicolored (or well balanced) in the following sense: for all colors i, the proportion of edges in C of color i is close to the proportion of edges in G which have color i. We derive this from a related result about random perfect matchings (Theorem 1.1) which is also a crucial tool in [12]; see section 1.3.

This paper is organized as follows. In sections 2 and 3.1 we collect some tools which we will need in our proofs. In section 3.2 we then use these tools to deduce some simple properties of random perfect matchings in ε-superregular graphs. The core result of this paper is Lemma 3.8 in section 3.3, which proves Theorem 1.1 for special graphs H. In the final section, the remaining results in this paper are easily deduced from Lemma 3.8 and the results in section 3.2.

*Received by the editors March 17, 2005; accepted for publication (in revised form) September 14, 2005; published electronically March 24, 2006.
http://www.siam.org/journals/sidma/20-2/62701.html
†School of Mathematics, Birmingham University, Edgbaston, Birmingham B15 2TT, UK (kuehn@maths.bham.ac.uk, osthus@maths.bham.ac.uk).

1.2. Statement of results. Given a bipartite graph $G = (A, B)$ with vertex classes A and B, we denote the edge set of G by $E(A, B)$ and let $e(G) = e(A, B) = |E(A, B)|$. The *density* of a bipartite graph $G = (A, B)$ is defined to be

$$d(A, B) := \frac{e(A, B)}{|A||B|}.$$

Given $0 < \varepsilon < 1$ and $d \in [0, 1]$, we say that G is (d, ε)-*regular* if for all sets $X \subseteq A$ and $Y \subseteq B$ with $|X| \geq \varepsilon|A|$ and $|Y| \geq \varepsilon|B|$ we have $(1-\varepsilon)d < d(X, Y) < (1+\varepsilon)d$. We say that G is (d, ε)-*superregular* if it is (d, ε)-regular and, furthermore, if $(1 - \varepsilon)d|B| < d_G(a) < (1+\varepsilon)d|B|$ for all vertices $a \in A$ and $(1-\varepsilon)d|A| < d_G(b) < (1+\varepsilon)d|A|$ for all $b \in B$. This is more or less equivalent to the traditional notions of ε-regularity and ε-superregularity—see section 2.

THEOREM 1.1. *For all positive constants $d, \nu_0, \eta \leq 1$ there is a positive $\varepsilon = \varepsilon(d, \nu_0, \eta)$ and an integer $N_0 = N_0(d, \nu_0, \eta)$ such that the following holds for all $n \geq N_0$ and all $\nu \geq \nu_0$. Let $G = (A, B)$ be a (d, ε)-superregular bipartite graph whose vertex classes both have size n and let H be a subgraph of G with $e(H) = \nu e(G)$. Choose a perfect matching M uniformly at random in G. Then with probability at least $1 - \mathrm{e}^{-\varepsilon n}$ we have*

$$(1 - \eta)\nu n \leq |M \cap E(H)| \leq (1 + \eta)\nu n.$$

At first sight it may seem surprising that the only parameter of H that is relevant here is the number of its edges. However, this is quite natural in view of the fact that the assertion would be trivial if instead of a perfect matching one would choose n edges independently and uniformly at random.

The case when H is a sufficiently large induced subgraph of G was proved earlier by Rödl and Ruciński [13] as a tool in their alternative proof of the blow-up lemma of Komlós, Sárközy, and Szemerédi.

From Theorem 1.1 we will also deduce a (weaker) analogue for Hamilton cycles.

THEOREM 1.2. *For all integers k and all positive constants $d, \nu, \eta \leq 1$ there is a positive $\varepsilon = \varepsilon(d, \nu, \eta)$ and an integer $N_1 = N_1(k, d, \nu, \eta)$ such that the following holds for all $n \geq N_1$. Let $G = (A, B)$ be a (d, ε)-superregular bipartite graph whose vertex classes both have size n. For each $1 \leq i \leq k$ let H_i be a subgraph of G with $e(H_i) = \nu_i e(G)$, where $\nu_i \geq \nu$. Then G contains a Hamilton cycle C such that for all $1 \leq i \leq k$*

$$(1 - \eta)2\nu_i n \leq |C \cap E(H_i)| \leq (1 + \eta)2\nu_i n.$$

Theorems 1.1 and 1.2 can in turn be used to deduce analogues for nonbipartite graphs (see the final section for details). For this, we need to modify the notion of (d, ε)-superregularity as follows. Given $0 < \varepsilon < 1$ and $d \in [0, 1]$, we say that a graph G with n vertices is (d, ε)-*regular* if for all disjoint sets $X, Y \subseteq V(G)$ with $|X|, |Y| \geq \varepsilon n$ we have $(1 - \varepsilon)d < d(X, Y) < (1 + \varepsilon)d$. We say that G is (d, ε)-*superregular* if it is (d, ε)-regular and, furthermore, if $(1 - \varepsilon)dn < d_G(x) < (1 + \varepsilon)dn$ for all vertices x of G.

THEOREM 1.3. *For all integers k and all positive constants $d, \nu, \eta \leq 1$ there is a positive $\varepsilon = \varepsilon(d, \nu, \eta)$ and an integer $N_2 = N_2(k, d, \nu, \eta)$ such that the following holds for all $n \geq N_2$. Let G be a (d, ε)-regular graph with n vertices. For each $1 \leq i \leq k$, let H_i be a subgraph of G with $e(H) = \nu_i e(G)$, where $\nu_i \geq \nu$ for all $i \geq k$. Then*

(i) *G contains a Hamilton cycle C such that for all i* $(1-\eta)\nu_i n \le |C \cap E(H_i)| \le (1+\eta)\nu_i n$;
(ii) *if n is even then G contains a perfect matching M such that for all i* $(1-\eta)\nu_i n/2 \le |M \cap E(H_i)| \le (1+\eta)\nu_i n/2$.

Note that the assertion is not even trivial (but much easier to prove) in the special case where G is the complete graph K_n. Moreover, let $G_{n,p}$ be a random graph on n vertices obtained by connecting each pair of vertices with probability p (independently of all the other pairs). For given $0 < p < 1$ and n sufficiently large, $G_{n,p}$ is (p,ε)-superregular with high probability (in fact the probability that this is not the case is easily seen to decrease exponentially in n). Thus the assertion of Theorem 1.3 holds with high probability in this case. Also, if G is dn-regular and the second eigenvalue of the adjacency matrix is at most λdn for sufficiently small λ, then G is (d,ε)-superregular (see, e.g., Chung [7, Theorem 5.1]) so the result applies in this case, too (such graphs are often called pseudorandom graphs).

1.3. Application: Loose Hamilton cycles in 3-uniform hypergraphs. A fundamental theorem of Dirac states that every graph on n vertices with minimum degree at least $n/2$ contains a Hamilton cycle. In [12], we prove an analogue of this for 3-uniform hypergraphs, which we describe below. All the results proved in this paper except Theorems 1.2 and 1.3 and Lemma 3.8 are used as a tool in [12].

One way to extend the notion of the minimum degree of a graph to that of a 3-uniform hypergraph $\mathcal{H}$ is as follows. Given two distinct vertices x and y of $\mathcal{H}$, the *neighborhood* $N(x,y)$ *of* (x,y) in $\mathcal{H}$ is the set of all those vertices z which form a hyperedge together with x and y. The *minimum degree* $\delta(\mathcal{H})$ is defined to be the minimum $|N(x,y)|$ over all pairs of vertices of $\mathcal{H}$.

We say that a 3-uniform hypergraph $\mathcal{C}$ is a *cycle of order* n if there exists a cyclic ordering $v_1,\dots,v_n$ of its vertices such that every consecutive pair v_iv_{i+1} lies in a hyperedge of $\mathcal{C}$ and such that every hyperedge of $\mathcal{C}$ consists of 3 consecutive vertices. A cycle is *tight* if every three consecutive vertices form a hyperedge. A cycle of order n is *loose* if it has the minimum possible number of hyperedges among all cycles on n vertices. A *Hamilton cycle* of a 3-uniform hypergraph $\mathcal{H}$ is a subhypergraph of $\mathcal{H}$ which is a cycle containing all its vertices. The following result is proved in [12].

THEOREM 1.4. *For each $\varepsilon > 0$ there is an $n_0 = n_0(\varepsilon)$ such that every 3-uniform hypergraph $\mathcal{H}$ with $n \ge n_0$ vertices and minimum degree at least $n/4 + \varepsilon n$ contains a loose Hamilton cycle.*

The bound on the minimum degree is essentially best possible in the sense that there are hypergraphs with minimum degree $\lceil n/4\rceil - 1$ which do not even contain some (not necessarily loose) Hamilton cycle. Recently, Rödl, Ruciński, and Szemerédi [14] proved that if the minimum degree is at least $n/2+\varepsilon n$ and n is sufficiently large, then one can even guarantee a tight Hamilton cycle. This is also best possible up to the error term (they announced in [14] that the error term εn can in fact be omitted).

2. Notation and a probabilistic estimate. Given a graph G, we write $N_G(x)$ for the neighborhood of a vertex x in G and let $d_G(x) := |N_G(x)|$. Given $\varepsilon > 0$, we say that G is *ε-regular* if for all sets $X \subseteq A$ and $Y \subseteq B$ with $|X| \ge \varepsilon|A|$ and $|Y| \ge \varepsilon|B|$ we have $|d(A,B) - d(X,Y)| < \varepsilon$. This (more traditional) notion of regularity is more or less equivalent to the one defined in the introduction. Indeed, clearly every (d,ε)-regular graph is also $2\varepsilon d$-regular (and thus 2ε-regular). Conversely, if $d = d(A,B) \ge \sqrt{\varepsilon}$ then every ε-regular bipartite graph (A,B) is $(d,\sqrt{\varepsilon})$-regular.

Given a positive number ε and sets $A, Q \subseteq T$, we say that *A is split ε-fairly by*

Q if

$$\left| \frac{|A \cap Q|}{|Q|} - \frac{|A|}{|T|} \right| \leq \varepsilon.$$

Thus, if ε is small and A is split ε-fairly by Q, then the proportion of all those elements of T which lie in A is almost equal to the proportion of all those elements of Q which lie in A. We will use the following version of the well-known fact that if Q is random then it tends to split large sets ε-fairly. It is an easy consequence of standard large deviation bounds for the hypergeometric distribution; see, e.g., [12] for a proof.

PROPOSITION 2.1. *For each $0 < \varepsilon < 1$ there exists an integer $q_0 = q_0(\varepsilon)$ such that the following holds. Given $t \geq q \geq q_0$ and a set T of size t, let Q be a subset of T which is obtained by successively selecting q elements uniformly at random without repetitions. Let $\mathcal{A}$ be a family of at most q^{10} subsets of T such that $|A| \geq \varepsilon t$ for each $A \in \mathcal{A}$. Then with probability at least $1/2$ every set in $\mathcal{A}$ is split ε-fairly by Q.*

3. Perfect matchings in superregular graphs. In this section, we collect and prove several results about (random) perfect matchings in bipartite superregular graphs G which will all be needed to prove Theorems 1.1 and 1.2. Moreover, Lemmas 3.6 and 3.7 will also be used in [12]. The main result of this section is Lemma 3.8. Given a reasonably regular small subgraph H of G, it gives precise bounds on the likely number of all those edges of H that are contained in a random perfect matching M of G. This is proved in the third subsection. In the first subsection, we collect some tools which we will need in the other two subsections. In the second subsection, we give likely upper bounds on the number of all those edges of an arbitrary sparse subgraph H of G that are contained in a random perfect matching and on the number of cycles in the union of two random perfect matchings in G.

3.1. Known results on counting perfect matchings. We use the following version of Stirling's inequality (the bound is a weak form of a result of Robbins; see, e.g., [4]).

PROPOSITION 3.1. *For all integers $n \geq 1$ we have*

$$\left(\frac{n}{\mathrm{e}}\right)^n \leq n! \leq 3\sqrt{n}\left(\frac{n}{\mathrm{e}}\right)^n. \tag{1}$$

We will frequently use the following immediate consequence of the lower bound in Stirling's inequality:

$$\binom{n}{k} \leq \left(\frac{\mathrm{e}n}{k}\right)^k. \tag{2}$$

We will also use that

$$1 - x \geq \mathrm{e}^{-x-x^2} \text{ for all } 0 < x < 0.45 \tag{3}$$

(see, e.g., [4, section 1.1]).

We also need the following result of Brégman [5] which settles a conjecture of Minc on the permanent of a 0-1 matrix. (A short proof of it was given by Schrijver [15]; see also [3].) We state this result in terms of an upper bound on the number of perfect matchings of a bipartite graph.

THEOREM 3.2. *The number of perfect matchings in a bipartite graph $G = (A, B)$ is at most*

$$\prod_{a \in A} (d_G(a)!)^{1/d_G(a)}.$$

An application of Stirling's inequality (Proposition 3.1) to Theorem 3.2 immediately yields the following.

COROLLARY 3.3. *For all $\varepsilon > 0$ there is an integer $d = d_0(\varepsilon)$ so that the following holds: Let $G = (A, B)$ be a bipartite graph with $|A| = |B| = n$ and let $m(G)$ denote the number of perfect matchings in G. Then*

$$m(G) \le (1+\varepsilon)^n \prod_{a \in A} \frac{\max\{d_G(a), d_0\}}{\mathrm{e}}.$$

A very useful lower bound on the number of perfect matchings in a k-regular bipartite graph is provided by the following result by Egorichev [8] and Falikman [9], which was formerly known as the van der Waerden conjecture.

THEOREM 3.4. *Let G be a k-regular bipartite graph whose vertex classes have size n. Then the number of perfect matchings in G is at least $(k/n)^n n!$.*

To bound the number of perfect matchings in superregular graphs, we will use the following theorem of Alon, Rödl, and Ruciński [2]. (Actually, we will only apply the lower bound, which is based on Theorem 3.4. The upper bound in Theorem 3.5 is an easy consequence of Corollary 3.3.) Note that their result is stated slightly differently in [2] as the definition of (d, ε)-superregularity in [2] is slightly different.

THEOREM 3.5. *For every $0 < \varepsilon < 1/4$ there exists an integer $n_1 = n_1(\varepsilon)$ such that whenever $d > 0$ and G is a (d, ε)-superregular bipartite graph whose vertex classes both have size $n \ge n_1$, then the number $m(G)$ of perfect matchings in G satisfies*

$$(d(1-4\varepsilon))^n n! \le m(G) \le (d(1+4\varepsilon))^n n!.$$

3.2. Simple properties of random perfect matchings. Based on the results in section 3.1, we can easily deduce the next lemma, which implies that if we are given a (super)regular graph G and a "bad" subgraph F of G which is comparatively sparse, then a random perfect matching of G will probably contain only a few bad edges. The "moreover" part will only be used in [12]; the assertion about (d, ε)-regular graphs will be used in [12] and the proof of Theorem 1.1.

LEMMA 3.6. *For all positive constants ε and d with $d \le 1$ and $\varepsilon \le 1/6$ there exists an integer $n_0 = n_0(\varepsilon, d)$ such that the following holds. Let G be a (d, ε)-superregular graph whose vertex classes A and B satisfy $|A| = |B| =: n \ge n_0$. Let M be a perfect matching chosen uniformly at random from the set of all perfect matchings of G. Let F be a subgraph of G such that all but at most $\Delta' n$ vertices in F have degree most $\Delta' dn$, where $1/2 \ge \Delta' \ge 18\varepsilon$. Then the probability that M contains at least $9\Delta' n$ edges of F is at most $\mathrm{e}^{-2\varepsilon n}$. Moreover, the statement also holds if we assume that G is dn-regular, where $dn \in \mathbb{N}$.*

Proof. First suppose that F has maximum degree at most $\Delta' dn$. Let $F' \supseteq F$ be a subgraph of G such that $d_{F'}(a) = \Delta' dn$ for each vertex $a \in A$. (Such an F' exists since $d_G(a) \ge (1-\varepsilon)dn \ge \Delta' dn$ as G is (d, ε)-superregular.) Given a set $A' \subseteq A$, we denote by $F'_{A'}$ the bipartite graph with vertex classes A and B in which every vertex $a \in A'$ is joined to all the vertices $b \in N_{F'}(a)$, while every vertex $a \in A \setminus A'$ is joined to all the vertices $b \in N_G(a) \setminus N_{F'}(a)$. For an integer $q \ge \mathrm{e}^2 \Delta' n$, let $m(q)$ denote the number of perfect matchings in G which contain precisely q edges from F'. Every such matching M' can be obtained by first fixing a q-element set $A' \subseteq A$ and then choosing a perfect matching in the graph $F'_{A'}$. (So the elements of A' correspond to the q endvertices of the edges in $M' \cap E(F')$.) If we apply Corollary 3.3 to $F'_{A'}$ we

now obtain

$$m(q) \le \binom{n}{q}(1+\varepsilon)^n \left(\frac{\Delta' dn}{\mathrm{e}}\right)^q \left(\frac{(1+\varepsilon)dn}{\mathrm{e}}\right)^{n-q} \overset{(2)}{\le} \left(\frac{\mathrm{e}n}{q}\right)^q \left(\frac{dn}{\mathrm{e}}\right)^n (\Delta')^q (1+\varepsilon)^{2n}.$$

Let $m(G)$ denote the number of perfect matchings in G. Then the lower bound in Theorem 3.5 implies that

$$m(G) \ge (d(1-4\varepsilon))^n n! \overset{(1)}{\ge} \left(\frac{(1-4\varepsilon)dn}{\mathrm{e}}\right)^n.$$

Thus the probability $m(q)/m(G)$ that M contains exactly q edges from F' is at most

$$\left(\frac{\mathrm{e}n\Delta'}{q}\right)^q (1+5\varepsilon)^{3n} \le \mathrm{e}^{-q}(1+5\varepsilon)^{3n} \le \mathrm{e}^{(15\varepsilon-\Delta')n} \le \mathrm{e}^{-3\varepsilon n}.$$

(To see the first inequality, use that $q \ge \mathrm{e}^2\Delta' n$.) By summing this bound over all $q \ge \mathrm{e}^2\Delta' n$, we find that the probability that M contains at least $\mathrm{e}^2\Delta' n$ edges of F' is at most $n\mathrm{e}^{-3\varepsilon n} \le \mathrm{e}^{-2\varepsilon n}$. Since $F \subseteq F'$, this implies that with probability at most $\mathrm{e}^{-2\varepsilon n}$ the matching M contains at least $\mathrm{e}^2\Delta' n$ edges of F. If F is now allowed to have up to $\Delta' n$ vertices whose degree is larger than $\Delta' dn$, this can increase the number of edges of F in M by at most $\Delta' n$, which implies the result.

The same proof also works in the case where G is dn-regular. We now use the lower bound $m(G) \ge d^n n! \ge (dn/\mathrm{e})^n$ which follows from Theorem 3.4 and inequality (1). □

In the following lemma we will use Theorems 3.4 and 3.5 to show that a randomly chosen 2-factor in a (super)regular graph G will typically contain only a few cycles. We will need this fact in the proof of Theorem 1.2 (and in [12] again, as mentioned earlier). A similar observation was also used in Frieze and Krivelevich [10]. The "moreover" part will only be used in [12].

LEMMA 3.7. *For all positive constants $\varepsilon < 1/64$ and $d \le 1$ there exists an integer $n_0 = n_0(\varepsilon, d)$ such that the following holds. Let G be a (d,ε)-superregular graph whose vertex classes A and B satisfy $|A| = |B| =: n \ge n_0$. Let M_1 be any perfect matching in G. Let M_2 be a perfect matching chosen uniformly at random from the set of all perfect matchings in $G - M_1$. Let $R = M_1 \cup M_2$ be the resulting 2-factor. Then the probability that R contains more than $n/(\log n)^{1/5}$ cycles is at most e^{-n}. Moreover, the statement also holds if we assume that G is dn-regular, where $dn \in \mathbb{N}$, and that G and M_1 are disjoint.*

Proof. Let $G' := G - M_1$. Let $m(G')$ denote the number of perfect matchings in G'. Since the deletion of a perfect matching from G still leaves a $(d, 2\varepsilon)$-superregular graph, Theorem 3.5 implies that

$$m(G') \ge ((1-8\varepsilon)d)^n n! \overset{(1),(3)}{\ge} \mathrm{e}^{-9\varepsilon n}\left(\frac{dn}{\mathrm{e}}\right)^n.$$

Let $k := n/(\log n)^{1/2}$ and $\ell' := (\log n)^{1/4}$. Given an integer $\ell \le \ell'$, let $f_{k,\ell}$ denote the number of 2-factors of G which contain M_1 and have at least k cycles of length 2ℓ. We will now find an upper bound on $f_{k,\ell}$. For this, note that the number of possibilities

for choosing a set $C_{k,\ell}$ of k disjoint cycles of length 2ℓ in G where every second edge is contained in M_1 is at most

$$\frac{1}{k!}n^{\ell k} \overset{(1)}{\leq} \left(\frac{\mathrm{e}}{k}n^{\ell}\right)^k =: c_{k,\ell}.$$

(Indeed, each such cycle of length 2ℓ is determined by an ordered choice of ℓ edges in M_1.) By Corollary 3.3, given some $C_{k,\ell}$ as above, the number of matchings on the remaining vertices of $G - M_1$ is at most

$$(1+\varepsilon)^{2n}\left(\frac{dn}{\mathrm{e}}\right)^{n-k\ell} \leq \mathrm{e}^{2\varepsilon n}\left(\frac{dn}{\mathrm{e}}\right)^{n-k\ell} =: d_{k,\ell}.$$

Hence we have that $f_{k,\ell} \leq c_{k,\ell}d_{k,\ell}$. Altogether, this implies that the probability $f_{k,\ell}/m(G')$ that a random 2-factor R (chosen as in the statement of the lemma) contains at least k cycles of length 2ℓ can be bounded as follows.

$$\frac{f_{k,\ell}}{m(G')} \leq \mathrm{e}^{11\varepsilon n}\left(\frac{\mathrm{e}}{k}n^{\ell}\right)^k\left(\frac{\mathrm{e}}{dn}\right)^{k\ell} = \mathrm{e}^{11\varepsilon n}\left(\frac{\mathrm{e}^{\ell+1}}{kd^{\ell}}\right)^k \leq \mathrm{e}^{11\varepsilon n}k^{-k/2} \leq \mathrm{e}^{-2n}.$$

To derive the third inequality, we used the fact that $(\mathrm{e}/d)^{\ell'}$ (and thus $(\mathrm{e}/d)^{\ell}$) is small compared to k. For the final one, we used that $k\log k$ is large compared to n.

Hence the probability that there is an $\ell \leq \ell'$ such that the random 2-factor R contains at least k cycles of length 2ℓ is at most $\ell'\mathrm{e}^{-2n} \leq \mathrm{e}^{-n}$. Note that the number of cycles of length at least $2\ell'$ in R is at most $2n/(2\ell')$. Thus with probability at least $1-\mathrm{e}^{-n}$ the number of cycles in R is at most $k\ell' + n/\ell' = 2n/(\log n)^{1/4}$, which implies the first part of the lemma.

The proof of the "moreover" part of Lemma 3.7 is almost the same, except that we use the lower bound $m(G) \geq (dn/\mathrm{e})^n$ on the number of perfect matchings in G which follows from Theorem 3.4 by an application of (1). □

3.3. Counting perfect matchings which contain a given number of edges of an almost regular subgraph.

LEMMA 3.8. *For each positive constant $\beta \neq 1$ there is a constant $f(\beta)$ with $0 < f(\beta) \leq 1$ such that the following holds. Suppose that $\alpha, \varepsilon, \xi, c'$, and d are positive constants with $\varepsilon \ll \alpha, c', d \leq 1$ and $\alpha, c' \ll \xi \ll f(\beta) \leq 1$. There exists an integer $n_0 = n_0(\alpha, \varepsilon, \xi, c', d)$ for which the following is true. Let G be a bipartite (d,ε)-superregular graph whose vertex classes V and W satisfy $|V| = |W| =: n \geq n_0$. Let H be a subgraph of G with vertex classes $C \subseteq V$ and $D \subseteq W$, where $c'n \leq |C| = cn \leq 2c'n$ and*

$$\alpha dn \leq d_H(v) \leq (1+\xi)\alpha dn \text{ for all vertices } v \in C.$$

Let M be a perfect matching chosen uniformly at random from the set of all perfect matchings in G. Then

(i) $\mathbb{P}(|M \cap E(H)| \leq \beta\alpha cn) \leq \mathrm{e}^{-f(\beta)\alpha cn}$ *if* $\beta < 1$,

(ii) $\mathbb{P}(|M \cap E(H)| \geq \beta\alpha cn) \leq \mathrm{e}^{-f(\beta)\alpha cn}$ *if* $\beta > 1$.

The intuition behind this result is the following (see also the remark after Theorem 1.1): If the inclusion of the edges of G into the random perfect matching M would be mutually independent and equally likely, then the probability that a given edge e is contained in M would be close to $|M|/e(G)$. Thus the expected value of $|M \cap E(H)|$ would be close to $ne(H)/e(G)$ which in turn is close to $n(\alpha dn)(cn)/(dn^2) = \alpha cn$. The

above result would thus immediately follow by an application of some large deviation bound on the tail of the binomial distribution.

The basic strategy of the proof is similar to that of [13], where the authors assume that H is a sufficiently large induced subgraph of G. The main difficulty of our proof is due to the fact that H is assumed to be rather small compared to G.

Proof. Let $m(G)$ denote the total number of perfect matchings in G. If we apply Stirling's formula (1) to the lower bound in Theorem 3.5, we obtain

$$m(G) \geq \left(\frac{(1-4\varepsilon)dn}{\mathrm{e}}\right)^n \overset{(3)}{\geq} \left(\frac{dn}{\mathrm{e}}\right)^n \mathrm{e}^{-5\varepsilon n}. \tag{4}$$

Given $a \leq cn$, let $m(a)$ be the number of perfect matchings in G which meet $E(H)$ in precisely a edges. Our aim is to show that $m(a)$ is much smaller than $m(G)$ if a is significantly smaller or larger than αcn. Let $\sum_J$ denote the summation over all matchings J in H of cardinality a. Given such a matching J, let $m(J)$ denote the number of perfect matchings M' in $G(J) := G - V(J) - E(H)$. Thus M' together with J forms a perfect matching of G which intersects H in exactly a edges and so $m(a) = \sum_J m(J)$. We claim that for all matchings J as above, we have

$$m(J) \leq \left(\frac{dn}{\mathrm{e}}\right)^{n-a} \mathrm{e}^{-\alpha cn-a}\mathrm{e}^{\xi a+5\varepsilon n}. \tag{5}$$

The first term is the roughly the bound we would get if we would use only the fact that $G(J)$ has maximum degree $(1+\varepsilon)dn$. The second term is a small but crucial improvement on this estimate. The third term is an insignificant error term.

We now prove (5). By Corollary 3.3, we have

$$m(J) \leq (1+\varepsilon)^{n-a} \prod_{v \in V \setminus V(J)} \frac{\max\{d_{G(J)}(v), d_0(\varepsilon)\}}{\mathrm{e}}, \tag{6}$$

where $d_0(\varepsilon)$ is the integer defined in Corollary 3.3. Thus we have reduced the problem of bounding $m(J)$ to that of finding accurate upper bounds on the degrees of the vertices in $G(J)$. Recall that the vertex classes of H are C and D and that $\Delta(G) \leq (1+\varepsilon)dn$ since G is (d,ε)-superregular. For a vertex $v \in C \setminus V(J)$ we have

$$d_{G(J)}(v) \leq dn(1+\varepsilon-\alpha) =: q_H.$$

We say that a vertex $v \in V \setminus V(H)$ is *average for* J if in the graph G it has at least $(1-\varepsilon)d(a-\varepsilon n)$ neighbors in $W \cap V(J)$. Let V^{av} be the set of such vertices. For $v \in V^{av}$, we have

$$d_{G(J)}(v) \leq dn(1+\varepsilon-(1-\varepsilon)(a/n-\varepsilon)) =: q_J.$$

Since G is (d,ε)-superregular, we have that $|V^{av}| \geq n - cn - \varepsilon n$ if $a \geq \varepsilon n$. If $a \leq \varepsilon n$, then trivially every vertex in $v \in V \setminus V(H)$ is average for J, so the above bound on $|V^{av}|$ holds in this case, too. Moreover, note that both $q_H \geq d_0(\varepsilon)$ and $q_J \geq d_0(\varepsilon)$ since n is sufficiently large compared to ε. Thus, inserting all these bounds into (6) gives

$$m(J) \leq (1+\varepsilon)^n \mathrm{e}^{a-n}(q_H)^{|C\setminus V(J)|}(q_J)^{|V^{av}|}((1+2\varepsilon)dn)^{n-a-|C\setminus V(J)|-|V^{av}|}.$$

Now note that $q_J \le (1+2\varepsilon)dn$ to deduce that the right-hand side is maximized if $|V^{av}|$ is minimized. Thus

$$\begin{aligned} m(J) &\le \mathrm{e}^{\varepsilon n}\mathrm{e}^{a-n}(q_H)^{cn-a}(q_J)^{(1-c-\varepsilon)n}((1+2\varepsilon)dn)^{\varepsilon n} \\ &\le \left(\frac{dn}{\mathrm{e}}\right)^{n-a}\exp Q, \end{aligned} \tag{7}$$

where $Q := \varepsilon n + Q_H + Q_J + 2\varepsilon(\varepsilon n)$ and

$$\begin{aligned} Q_H :=& (\varepsilon-\alpha)(cn-a), \\ Q_J :=& [\varepsilon-(1-\varepsilon)(a/n-\varepsilon)][(1-c-\varepsilon)n]. \end{aligned}$$

Note that, we made use of the fact that $1+x \le \mathrm{e}^x$ three times in order to obtain (7). Now observe that

$$\begin{aligned} Q_H &\le -\alpha cn + \alpha a + \varepsilon n, \\ Q_J &\le \varepsilon n - a(1-c-\varepsilon)(1-\varepsilon) + \varepsilon n \le -a(1-2c) + 2\varepsilon n. \end{aligned}$$

Altogether, we thus have

$$Q \le \varepsilon n - \alpha cn + \alpha a + \varepsilon n - a + 2ac + 2\varepsilon n + \varepsilon n \le -\alpha cn - a + \xi a + 5\varepsilon n,$$

which proves (5).

Let p_a denote the probability that a perfect matching which is chosen uniformly at random in the set of all perfect matchings in G contains exactly a edges of H. Thus $p_a = m(a)/m(G) = \sum_J m(J)/m(G)$. Let $|\sum_J|$ denote the number of summands, i.e., the number of matchings in H of cardinality a. Each matching of cardinality a in H can be obtained by first choosing a subset of a vertices in C and then choosing one neighbor in H for each vertex in this subset. Thus, writing $(x/0)^0 := 1$ for all $x > 0$, it follows that

$$\left|\sum_J\right| \le \binom{cn}{a}((1+\xi)\alpha dn)^a \overset{(2)}{\le} \left(\frac{\mathrm{e}^{1+\xi}\alpha cdn^2}{a}\right)^a. \tag{8}$$

Since the bound (5) on $m(J)$ is independent of J, we can now combine (4) and (5) to obtain

$$\begin{aligned} p_a = \sum_J \frac{m(J)}{m(G)} &\le \left|\sum_J\right|\left(\frac{\mathrm{e}}{dn}\right)^a \mathrm{e}^{5\varepsilon n}\mathrm{e}^{-\alpha cn-a}\mathrm{e}^{\xi a+5\varepsilon n} \\ &\overset{(8)}{\le} \left(\frac{\mathrm{e}\alpha cn}{a}\right)^a \mathrm{e}^{-\alpha cn}\mathrm{e}^{2\xi a+10\varepsilon n}. \end{aligned}$$

Now define β' by $a = \beta'\alpha cn$ and let $g(\beta') := \log\{(\mathrm{e}/\beta')^{\beta'}/\mathrm{e}\}$. Then

$$p_a \le \left(\left(\frac{\mathrm{e}}{\beta'}\right)^{\beta'}\mathrm{e}^{-1}\right)^{\alpha cn}\mathrm{e}^{2\xi a+10\varepsilon n} \le \exp\left\{\alpha cn(g(\beta')+2\xi\beta'+\xi)\right\}.$$

Now set $\mu := \alpha cn$ to obtain

$$p_a \le \exp\{\mu(g(\beta')+\xi(1+2\beta'))\}.$$

(Note that if $\xi = 0$ and $\beta' < 1$, this would be exactly the standard Chernoff bound on the probability that $X \le \beta'\mu$, where X has a binomial distribution with mean μ; see, e.g., Theorem A.12 in [3].) It is easy to check that $g(\beta') < 0$ if $\beta' \neq 1$.

The assertion (i) (i.e. the case $\beta < 1$) of the lemma now follows with $f(\beta) := -g(\beta)/4$ by summing over all values of a between 1 and $\beta\mu$. Indeed, as $g(\beta')$ is negative and increasing for $\beta' < 1$, we have

$$\mathbb{P}\left(|M \cap E(H)| \le \beta\alpha cn\right) \le \beta\mu \exp\{\mu g(\beta) + 3\xi\} \le \beta\mu \exp\{\mu g(\beta)/2\},$$

as required. To prove the assertion (ii) of the lemma, we first consider the case $1 < \beta \le \beta' \le \mathrm{e}^2$. As $g(\beta')$ is negative and decreasing for $\beta' > 1$, it follows that

$$p_a \le \exp\{\mu(g(\beta) + 17\xi)\} \le \exp\{\mu g(\beta)/2\}.$$

Next consider the case that $\beta' \ge \mathrm{e}^2$. It is easy to check that $g(\beta') \le -\beta'$. Thus

$$p_a \le \exp\{\mu(-\beta' + \xi(1 + 2\beta'))\} \le \exp\{-\mu\beta'/2\}.$$

Similarly to the case (i), the assertion of the lemma in case (ii) now follows by summing the bounds on p_a over all values of a between $\beta\mu$ and cn. ∎

4. Proof of Theorems 1.1–1.3. We will prove Theorem 1.1 by decomposing H into small "almost regular" subgraphs H_{ij} and a small remainder F. We will apply Lemma 3.8 to each of the H_{ij} separately and then use Lemma 3.6 to show that a random perfect matching contains only a negligible number of edges of F.

Proof of Theorem 1.1. By adding all the vertices in $V(G) \setminus V(H)$ to H, we may assume that H is a spanning subgraph of G. Set $\beta := 1 + \eta/4$, define $f(\beta)$ as in the statement of Lemma 3.8, and choose parameters $\alpha, \varepsilon, \xi, c'$ so that $0 < \varepsilon \ll \alpha, c', d \le 1$ and $c' \ll \alpha \ll \xi \ll \nu, \eta, f(\beta)$. Thus the restrictions in the statement of Lemma 3.8 are satisfied. Choose N_0 to be sufficiently large compared to both $1/\varepsilon$ and the integer $n_0(\alpha, \varepsilon, \xi, c', d)$ defined in Lemma 3.8. Finally, fix a constant c such that $cn \in \mathbb{N}$ and $c' \le c \le 2c'$.

First, we prove the upper bound in Theorem 1.1. Let ℓ be the smallest integer so that $\mathrm{e}^{\xi\ell/2}\alpha > 1 + \varepsilon$. Thus

$$\ell \le \frac{2}{\xi}\log(2/\alpha) \le 1/\sqrt{c}. \tag{9}$$

Let A_0 be the set of vertices in A with $d_H(a) < \alpha dn$. For all $i \ge 1$, let $\alpha_i := \mathrm{e}^{\xi(i-1)/2}\alpha$. Thus

$$\alpha_{i+1} \le (1 + \xi)\alpha_i \tag{10}$$

since $\mathrm{e}^{\xi/2} \le 1 + \xi$ (see, e.g., [4, section 1.1]). Moreover,

$$1 + \varepsilon < \alpha_{\ell+1} \le 2. \tag{11}$$

For all i with $1 \le i \le \ell$, let A_i be the set of vertices in $a \in A$ with $\alpha_i dn \le d_H(a) < \alpha_{i+1}dn$. Since G is (d, ε)-superregular and thus $d_H(a) \le d_G(a) \le (1 + \varepsilon)dn$ for each $a \in A$, it follows that the A_i with $0 \le i \le \ell$ give a partition of A.

We now define a partition of the edge set of H into graphs H_{ij}. Given $1 \le i \le \ell$, define q_i by $|A_i| = q_i cn$ and let $q(i) := \lfloor q_i \rfloor$. We partition the vertices in A_i into

$q(i)+1$ parts A_{ij} with $0 \le j \le q(i)$ as follows: the partition is arbitrary except that we require that $|A_{ij}| = cn$ for all $j \ge 1$. Thus $|A_{i0}| < cn$ and so

$$\sum_{i=1}^{\ell} |A_{i0}| \le \ell cn \le \sqrt{c}n \le \alpha n. \tag{12}$$

Let H_{ij} be the subgraph of H induced by A_{ij} and B. Then for all $a \in A_{ij}$, we have

$$\alpha_i dn \le d_{H_{ij}}(a) < \alpha_{i+1} dn \overset{(10)}{\le} (1+\xi)\alpha_i dn. \tag{13}$$

Let H_{00} be the subgraph of H which is induced by A_0 and B. Given $1 \le i \le \ell$, let H_{i0} be the subgraph of H which is induced by A_{i0} and B. Let F denote the union of all the H_{i0} with $0 \le i \le \ell$. Then

(14)
$$e(F) \le \alpha dn|A_0| + \sum_{i=1}^{\ell} |A_{i0}|\alpha_{i+1} dn \overset{(11),(12)}{\le} \alpha dn^2 + 2\alpha dn^2 \le 4\alpha e(G) \le \eta e(H)/4.$$

Let M be a perfect matching chosen uniformly at random from the set of all perfect matchings in G. Let $X_{ij} := |M \cap E(H_{ij})|$ and $\mu_i := \alpha_i cn$. (Note that μ_i can be thought of as roughly the expected value of X_{ij}.) Then for all i, j with $i, j \ge 1$ we can apply Lemma 3.8(ii) to H_{ij} to see that with probability at least $1 - \mathrm{e}^{-f(\beta)\mu_i}$ we have $X_{ij} \le \beta\mu_i$ (apply the lemma with α_i taking on the role of the parameter α there). Moreover, we can apply Lemma 3.6 to F as follows: Let $\Delta' := \alpha$. Then (12) implies that at most $\Delta' n$ vertices of F have degree more than $\Delta' dn$. Thus Lemma 3.6 implies that with probability at least $1 - \mathrm{e}^{-2\varepsilon n}$ we have

$$|M \cap E(F)| \le 9\alpha n \le \eta\nu n/2.$$

But F and the sets $E(H_{ij})$ with $i, j \ge 1$ form a partition of $E(H)$, and so with probability at least $1 - \mathrm{e}^{-2\varepsilon n} - \sum_{i=1}^{\ell} q(i)\mathrm{e}^{-f(\beta)\mu_i} \ge 1 - \mathrm{e}^{-\varepsilon n}$ we have

$$|M \cap E(H)| \le \eta\nu n/2 + \beta \sum_{i=1}^{\ell} q(i)\mu_i \le \eta\nu n/2 + \beta \sum_{i=1}^{\ell} |A_i|\alpha_i.$$

Now use the fact that $\sum_{i=1}^{\ell} |A_i|\alpha_i dn \le e(H) \le (1+\varepsilon)\nu dn^2$ to see that $|M \cap E(H)| \le \eta\nu n/2 + \beta(1+\varepsilon)\nu n \le (1+\eta)\nu n$, as required.

The proof of the lower bound is almost exactly the same: in this case, we let $\beta = 1 - \eta/4$. The graphs H_{ij} are defined as before. We now apply Lemma 3.8(i) to H_{ij} to see that with probability at least $1 - \sum_{i=1}^{\ell} q(i)\mathrm{e}^{-f(\beta)\mu_i} \ge 1 - \mathrm{e}^{-\varepsilon n}$ we have $X_{ij} \ge \beta\mu_i$ for all i, j with $i \ge 1$. Thus with probability at least $1 - \mathrm{e}^{-\varepsilon n}$, we have

$$\begin{aligned} |M \cap E(H)| &\ge \beta \sum_{i=1}^{\ell} q(i)\mu_i \ge \beta \sum_{i=1}^{\ell} (|A_i| - cn)\alpha_i \overset{(11)}{\ge} \beta \sum_{i=1}^{\ell} |A_i|\alpha_i - 2\beta\ell cn \\ &\overset{(9)}{\ge} \beta \sum_{i=1}^{\ell} |A_i|\alpha_i - 4\sqrt{c}n \ge \beta \sum_{i=1}^{\ell} |A_i|\alpha_i - \eta\nu dn/2. \end{aligned} \tag{15}$$

But

$$\sum_{i=1}^{\ell} |A_i|\alpha_i dn \overset{(10)}{\geq} (1-2\xi)\sum_{i=1}^{\ell} |A_i| dn\alpha_{i+1} \geq (1-2\xi)(e(H)-e(F))$$
$$\overset{(14)}{\geq} (1-2\xi)(1-\eta/4)e(H) \geq (1-\eta/2)\nu dn^2,$$

which implies the result together with (15). □

We can now easily deduce Theorem 1.2 from Theorem 1.1 and Lemma 3.7.

Proof of Theorem 1.2. Put $\varepsilon := \min\{1/64, d/5, \varepsilon(d,\nu,\eta/2)/2\}$, where $\varepsilon(d,\nu,\eta/2)$ is as defined in Theorem 1.1. Let N_1 be sufficiently large compared to $1/\eta$, $1/\nu$, and k as well as larger than $n_0(\varepsilon,d)$ and $N_0(d,\nu,\eta/2)$ defined in Lemma 3.7 and Theorem 1.1, respectively.

Choose a perfect matching M_1 uniformly at random in G and then choose a perfect matching M_2 uniformly at random in $G - M_1$. Lemma 3.7 implies that with probability at least $1 - \mathrm{e}^{-n}$ the resulting 2-factor $R = M_1 \cup M_2$ contains at most $n/(\log n)^{1/5}$ cycles. Moreover, Theorem 1.1 implies that we may assume that

$$(1-\eta/2)2\nu_i n \leq |R \cap E(H_i)| \leq (1+\eta/2)2\nu_i n \tag{16}$$

for all $i \leq k$. Thus it suffices to prove that there is a Hamilton cycle C in G which has sufficiently many edges in common with R. This is achieved using a standard argument based on expansion properties of G.

Let C' be any cycle in R with the property that there are adjacent vertices x and y on C' such that x has a neighbor z outside C'. (Using that G is (d,ε)-superregular, it is easy to see that such a cycle always exists unless R is already a Hamilton cycle. Indeed, since $\delta(G) \geq (1-\varepsilon)dn$, each cycle in R of length at most $(1-\varepsilon)dn$ will have a neighbor outside and thus can be taken to be C'. On the other hand, $|N_G(X)| \geq (1-\varepsilon)n$ for any set X of size at least $(1-\varepsilon)dn/2 \geq \varepsilon n$ which lies in one of the vertex classes of G. This implies that if all the cycles in R have length at least $(1-\varepsilon)dn$ and R is not a Hamilton cycle, then we can take for C' any cycle of R.)

Let C'' denote the cycle in R which contains z. Let P denote the path obtained from $C' \cup C''$ by adding the edge xz and deleting xy as well as one of the edges on C'' adjacent to z. Note that the length of P is odd. If one of the endpoints of P has a neighbor outside P, we can further enlarge P in a similar way. So suppose we can no longer enlarge P in this way and view P as a directed path whose first vertex is denoted by x and whose final vertex is denoted by y. Thus all the neighbors of x and y lie on P. Moreover, since P is odd, x and y lie in different vertex classes of G.

We claim that there is a cycle C^* which has the same vertex set as P. Let X_1 be the set consisting of the first $\lfloor d_G(x)/2 \rfloor$ neighbors of x on P and let X_2 consist of all other neighbors. Define Y_1 and Y_2 similarly. It is easily seen that either (i) all vertices in Y_1 come before all those in X_2 or (ii) all vertices in X_1 come before those in Y_2. Suppose first that (i) holds. Note that $|X_i|, |Y_j| \geq \delta(G)/4 \geq (1-\varepsilon)dn/4 \geq \varepsilon n$ and so the (d,ε)-superregularity of G implies that there is an edge $e \in E(G)$ between a predecessor p of some vertex $y_1 \in Y_1$ and a successor s of some vertex $x_2 \in X_2$. We thus obtain a cycle C^* whose vertex set is $V(P)$ by removing the edges py_1 and x_2s from P and adding the three edges e, xx_2, and yy_1. The case (ii) is identical except that we now consider the predecessors of the vertices in X_1 and the successors of the vertices in Y_2.

Altogether, we have now constructed a 2-factor where the number of cycles has decreased. Continuing in this way, we eventually arrive at a Hamilton cycle C. It is easy to check that the symmetric difference of C and R contains only at most $5n/(\log n)^{1/5} \leq \eta\nu n/2$ edges. Together with (16) this shows that C is as required in the theorem. □

It remains to deduce Theorem 1.3 from Theorems 1.1 and 1.2.

Proof of Theorem 1.3. First suppose that n is even. Set $n' := n/2$. Consider a random partition of the vertex set of G into two sets A and B of equal size. Let G' be the bipartite subgraph of G between A and B. Lemma 2.1 implies that we may assume that the graph G' is $(d, 2\varepsilon)$-superregular (in the bipartite sense) if n is sufficiently large compared to ε. Also, Lemma 2.1 implies that we may assume that the density of the bipartite subgraph of H_i between A and B is still close to $\nu_i d$ for all $i \leq k$. Thus we can apply Theorems 1.1 and 1.2 in this case.

Now suppose that n is odd and set $n' := \lfloor n/2 \rfloor$. Delete any vertex x from the vertex set of G. Again, Lemma 2.1 implies that we may assume that the bipartite graph $G' = (A, B)$ constructed as above on the remaining $2n'$ vertices is $(d, 3\varepsilon)$-superregular if n is sufficiently large compared to ε. Moreover, we may assume that for all $i \leq k$ the density of the bipartite subgraph of H_i between A and B is still very close to that of H_i, i.e., close to $\nu_i d$. Thus we may apply Theorem 1.2 to obtain a Hamilton cycle C' which satisfies

$$(1 - \eta/2)2\nu_i n' \leq |C' \cap E(H_i - x)| \leq (1 + \eta/2)2\nu_i n'.$$

Let P be a Hamilton path obtained from C' by adding an edge between x and some vertex $y \in C'$ and deleting one of the two edges on C' incident to y. As in the proof of Theorem 1.2, one can easily show that one can transform P into a Hamilton cycle C by deleting two and adding three edges. Then C is as required in Theorem 1.3(i). □

REFERENCES

[1] M. J. Albert, A. Frieze, and B. Reed, *Multicoloured Hamilton cycles*, Electron. J. Combin., 2 (1995), #R10.

[2] N. Alon, V. Rödl, and A. Ruciński, *Perfect matchings in ε-regular graphs*, Electron. J. Combin., 5 (1998), #R13.

[3] N. Alon and J. Spencer, *The Probabilistic Method*, 2nd edition, Wiley-Interscience, New York, 2000.

[4] B. Bollobás, *Random Graphs*, 2nd edition, Cambridge Studies in Advanced Mathematics 73, Cambridge University Press, Cambridge, UK, 2001.

[5] L. M. Brégman, *Some properties of nonnegative matrices and their permanents*, Soviet Mathematics Doklady, 14 (1973), pp. 945–949.

[6] C. Cooper and A. Frieze, *Multicoloured Hamilton cycles in random graphs; an anti-Ramsey threshold*, Electron. J. Combin., 2 (1995), #R19.

[7] F. R. K. Chung, *Spectral Graph Theory*, CBMS Reg. Conf. Ser. Math. 92, AMS, Providence, RI, 1997.

[8] G. P. Egorichev, *The solution of the van der Waerden problem for permanents*, Dokl. Akad. Nauk SSSR, 258 (1981), pp. 1041–1044.

[9] D. I. Falikman, *A proof of the van der Waerden's conjecture on the permanent of a doubly stochastic matrix*, Mat. Zametki, 28 (1981), pp. 931–938.

[10] A. Frieze and M. Krivelevich, *On packing Hamilton cycles in ε-regular graphs*, J. Combin. Theory Ser. B, 94 (2005), pp. 159–172.

[11] S. Janson, T. Łuczak, and A. Ruciński, *Random Graphs*, Wiley-Interscience, New York, 2000.

[12] D. Kühn and D. Osthus, *Loose Hamilton cycles in 3-uniform hypergraphs of large minimum degree*, J. Combin. Theory Ser. B, to appear.

[13] V. Rödl and A. Rucíński, *Perfect matchings in ε-regular graphs and the blow-up lemma*, Combinatorica, 19 (1999), pp. 437–452.
[14] V. Rödl, A. Rucíński, and E. Szemerédi, *A Dirac-type theorem for 3-uniform hypergraphs*, Combin. Probab. Comput., 15 (2006), pp. 229–251.
[15] A. Schrijver, *A short proof of Minc's conjecture*, J. Combin. Theory Ser. A, 25 (1978), pp. 80–83.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 287–290

ON MULTICAST REARRANGEABLE 3-STAGE CLOS NETWORKS WITHOUT FIRST-STAGE FAN-OUT*

HONG-BIN CHEN† AND FRANK K. HWANG†

Abstract. For the multicast rearrangeable 3-stage Clos networks where input crossbars do not have fan-out capability, Kirkpatrick, Klawe, and Pippenger gave a sufficient condition and also a necessary condition which differs from the sufficient condition by a factor of 2. In this paper, we first tighten their conditions. Then we propose a new necessary condition based on the affine plane such that the necessary condition matches the sufficient condition for an infinite class of 3-stage Clos networks.

Key words. rearrange, Clos networks, multicast, affine plane

AMS subject classifications. 94C15, 05B05

DOI. 10.1137/05062336X

1. Introduction.

Consider a 3-stage Clos network $C(n_1, r_1, m, n_2, r_2)$, where the input stage consists of r_1 $n_1 \times m$ crossbars, the middle stage m $r_1 \times r_2$ crossbars, the output stage r_2 $m \times n_2$ crossbars, and where there exists one link between every pair of crossbars between two adjacent stages (see Figure 1).

The inlets of the input crossbars are the inputs of the network, and the outlets of the output crossbars are the outputs of the network. In the multicast traffic network, an input can appear in a request more than once. If the appearance is restricted to at most f times, the traffic is called an f-cast traffic. If there is no restriction, then it is called the *broadcast* traffic. A network is rearrangeable if any set of disjoint pairs of inputs and outputs can be simultaneously connected. If the calls come sequentially, rearrangeability means we can disconnect all existing connections and reroute them together with the new call simultaneously.

A crossbar is said to have the fan-out capability if the crossbar itself can route multicast traffic without blocking, i.e., any inlet can be connected to any number of idle outlets regardless of other connections. If the crossbars in a given stage perform only point-to-point connections, then we say the stage has no fan-out capability. Four models have been studied [3] on 3-stage Clos networks:

Model 0. no restriction on fan-out capability,
Model 1. input stage has no fan-out capability,
Model 2. middle stage has no fan-out capability,
Model 3. output stage has no fan-out capability.

Masson and Jordan [7] proved that $C(n_1, r_1, m, n_2, r_2)$ under model 2 is multicast rearrangeable if and only if $m \geq \max\{\min\{n_1 f, N_2\}, \min\{n_2, N1\}\}$. However, necessary and sufficient conditions are not known under the other models. Under model 1, Kirkpatrick, Klawe, and Pippenger [6] gave a sufficient condition for $C(n_1, r_1, m, n_2, r_2)$ to be multicast rearrangeable, and also a necessary condition which differs from the

*Received by the editors January 27, 2005; accepted for publication (in revised form) October 12, 2005; published electronically March 24, 2006.
http://www.siam.org/journals/sidma/20-2/62336.html
†Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan (andan.am92g@nctu.edu.tw, fhwang@math.nctu.edu.tw). The work of the second author was partially supported by Republic of China, National Science Council grant NSC 92-2115-M-009-014.

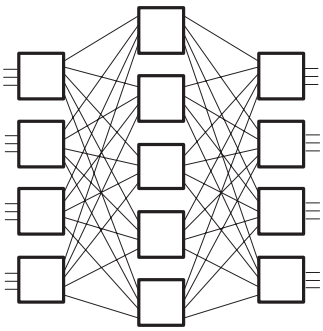


FIG. 1. $C(3,4,5,3,4)$.

sufficient condition by a factor of 2. In this paper, we tighten their conditions, then propose a new necessary condition which matches the sufficient condition for an infinite class of networks.

2. Main results. Kirkpatrick, Klawe, and Pippenger proved the following theorem.

THEOREM 2.1. (i) $C(n_1, r_1, m, n_2, r_2)$ *is broadcast rearrangeable under model* 1 *if* $m \geq n_1 + (n_2(n_2-1)r_2)^{1/2}$, (ii) $C(n_1, r_1, m, n_2, r_2)$ *is broadcast rearrangeable under model* 1 *only if* $m \geq \frac{n_1+(n_2(n_2-1)r_2)^{1/2}}{2}$.

In the following theorem we modify Theorem 2.1(i) by considering integrality of the number of crossbars. We also improve Theorem 2.1(ii).

THEOREM 2.2. (i) $C(n_1, r_1, m, n_2, r_2)$ *is broadcast rearrangeable under model* 1 *if* $m \geq \lceil (\frac{n_2 r_2}{n_2-1})^{1/2} \rceil (n_2-1) + (n_1 - n_2 + 1)^+$, *where* $x^+ = \max\{x, 0\}$, (ii) $C(n_1, r_1, m, n_2, r_2)$ *is broadcast rearrangeable under model* 1 *only if* $m \geq \max\{n_1, \lfloor n_2/2 \rfloor \lfloor (2r_2)^{1/2} \rfloor\}$.

Proof. (i) The set L of requests each asking to connect to at least $\lceil (\frac{n_2 r_2}{n_2-1})^{1/2} \rceil$ outputs has size of at most

$$\frac{n_2 r_2}{\lceil (\frac{n_2 r_2}{n_2-1})^{1/2} \rceil} \leq \frac{n_2 r_2}{(\frac{n_2 r_2}{n_2-1})^{1/2}} = [n_2 r_2 (n_2-1)]^{1/2} \leq \left\lceil \left(\frac{n_2 r_2}{n_2-1} \right)^{1/2} \right\rceil (n_2-1).$$

Route each of these requests through a distinct middle crossbar. A request g other than these can ask for connections to a set O_g of at most $\lceil (\frac{n_2 r_2}{n_2-1})^{1/2} \rceil - 1$ output crossbars. Such a request has to be routed through a middle crossbar not taken by any of the at most $(\lceil (\frac{n_2 r_2}{n_2-1})^{1/2} \rceil - 1)(n_2-1)$ outputs on crossbars in O_g, nor by the $n_1 - 1$ inputs on the same input crossbar as g. Therefore

$$\left(\left\lceil \left(\frac{n_2 r_2}{n_2-1} \right)^{1/2} \right\rceil - 1 \right) (n_2-1) + (n_1-1) + 1 = \left\lceil \left(\frac{n_2 r_2}{n_2-1} \right)^{1/2} \right\rceil (n_2-1) + n_1 - n_2 + 1$$

middle crossbars are sufficient to route g. However, if $n_1 - n_2 + 1 < 0$, the number of middle crossbars still cannot be less than the number required to route L.

(ii) Construct a complete graph K_v with $v = \lfloor (2r_2)^{1/2} \rfloor$ vertices and $e = \binom{v}{2} \leq r_2$ edges. Label each edge by a distinct output crossbar. Take $c = \lfloor n_2/2 \rfloor$ copies of K_v, keeping the edge-labels intact, and label the vc vertices by the set $\{1, 2, \ldots, vc\}$. Identify each vertex u as a request $(O_{u1}, \ldots, O_{u(v-1)})$, where $O_{u1}, \ldots, O_{u(v-1)}$ are the labels of the $v-1$ edges incident to u. Note that each output crossbar appears in $2c \leq n_2$ requests.

Since every pair of requests intersect in at least one output crossbar, each of the vc requests must be routed through a distinct middle crossbar. □

Our construction in (ii) improves over that of [6] by increasing the number of edges in K_v. The corresponding graph in [6] contains only $r_2^{1/2}$ vertices, hence roughly $r_2/2$ edges.

COROLLARY 2.3. *Part* (ii) *is valid for f-cast traffic with $f \geq \lfloor (2r_2)^{1/2} \rfloor - 1$.*

Next we give a stronger necessary condition which is based on the observation that requests from the same input crossbar cannot use the same middle crossbar, even if the requests do not intersect (in output crossbars). Then the request graph we need to construct is no longer a complete graph, but a graph whose vertices can be partitioned into r_1 subsets such that an edge exists between every pair of vertices from different subsets. Such a graph corresponds to a resolvable block design.

A block design $B(v, b, r, k, 1)$ is a collection of b k-subsets (called blocks) of a v-set S, $k < v$, such that each pair of elements of S appears together in exactly one block and each element of S appears in exactly r blocks. $B(v, b, r, k, 1)$ is resolvable if the blocks can be partitioned into r orbits such that each element appears once in each orbit. For example, the following 12 blocks form a resolvable $B(9, 12, 4, 3, 1)$ which can be grouped into 4 orbits, each of 3 blocks, so that the blocks in each orbit together contain each element exactly once:

$$(\{1,2,3\},\{4,8,9\},\{5,6,7\}),$$
$$(\{1,5,8\},\{3,4,6\},\{2,7,9\}),$$
$$(\{1,4,7\},\{2,6,8\},\{3,5,9\}),$$
$$(\{1,6,9\},\{2,4,5\},\{3,7,8\}).$$

A block design $B(n^2, n^2 + n, n + 1, n, 1)$ is called an affine plane of order n. It is well known that every affine plane is resolvable and that an affine plane of order q exists whenever q is a prime power; see [1].

THEOREM 2.4. *For every prime power q, and every M in the range $1 \leq M \leq r$, there exist $n_1 \geq q, r_1 \geq M, n_2 \geq M$, and $r_2 \geq q^2$ such that $C(n_1, r_1, m, n_2, r_2)$ is broadcast rearrangeable only if $m \geq q(q + 1)$.*

Proof. For a prime power q we know that there exists an affine plane of order q, i.e., a resolvable block design $B(q^2, q^2 + q, q + 1, q, 1)$. Identify the elements as the output crossbars, the orbits as the input crossbars, and the blocks as inputs, while the elements in a block i represent the output crossbars input i requests to connect. Note that each output crossbar appears in $M \leq n_2$ requests. Hence the given set of requests are legitimate.

By our construction, two requests from different input crossbars (orbits) intersect in one output crossbar and hence must be routed through different middle crossbars. Requests from the same input crossbar do not intersect in any output crossbar, but still have to be routed through different middle crossbars since they share the input crossbar. Therefore the total number of middle crossbars required is at least the number of requests constructed above, which is $q(q + 1)$. □

Introduction of the parameter M is just to broaden the applicability of Theorem 2.4, i.e., n_2 does not have to be equal r, but can be less.

Now we show that the necessary condition of Theorem 2.4 matches the sufficient condition in Theorem 2.2(i). Theorem 2.4 shows the necessary condition is $m \geq q(q+1)$. From Theorem 2.2(i), setting $n_1 = \frac{v}{k} = q, n_2 = M = q+1$, and $r_2 = v = q^2$,

the sufficient condition is

$$m \geq \left\lceil \left(\frac{Mq^2}{q} \right)^{1/2} \right\rceil q + (q - (q+1) + 1)^+ = \lceil ((q+1)q)^{1/2} \rceil q = q(q+1),$$

same as the necessary condition.

COROLLARY 2.5. *Theorem* 2.4 *holds for* f*-cast traffic with* $f \geq q$.

3. Conclusions. The current necessary condition for broadcast rearrangeable 3-stage Clos networks differs from the sufficient condition by a factor of 2. We tightened these conditions such that they match for an infinite class of 3-stage Clos networks. This shows that our tightened conditions cannot be further improved for general parameters.

While the main results obtained for multicast rearrangeable 3-stage Clos networks are for broadcast networks so far, our arguments for necessary conditions are valid also for f-cast networks for some specific f as shown in the corollaries, thus starting the study of f-cast rearrangeable 3-stage Clos networks, which has been a vacuum so far.

The model-1 model can be interpreted in two ways. One is that the input crossbars do not have the fan-out capability, and thus perhaps can be obtained with a cheaper cost. The other is that they do have the fan-out capability, but our routing algorithm chooses not to use it. This type of routing algorithm has been used in the mixed-requirement model where all point-to-point requests meet the strictly nonblocking requirement and all f-cast requests for $f \geq 2$ meet the rearrangeable requirement [2, 4, 5].

REFERENCES

[1] I. ANDERSON, *Combinatorial designs: Construction methods*, Ellis Horwood Limited, New York, 1990.

[2] D.-Z. DU AND H. Q. NGO, *An extension of DHH-Erdös conjecture on cycle-plus-triangle graphs*, Taiwanese J. Math., 6 (2002), pp. 261–267.

[3] F. K. HWANG, *The mathematical theory of nonblocking switching networks*, 2nd ed., World Scientific, Hackensack, NJ, 2004.

[4] F. K. HWANG, S. C. LIAW, AND L. D. TONG, *Strictly nonblocking* 3*-stage Clos networks with some rearrangeable multicast capability*, IEEE Trans. Commun., 51 (2003), pp. 1765–1767.

[5] F. K. HWANG AND C. H. LIN, *Broadcasting in a* 3*-stage point-to-point nonblocking networks*, Inter. J. Rel. Qual. Safety Engin., 2 (1995), pp. 299–307.

[6] D. G. KIRKPATRICK, M. KLAWE, AND N. PIPPENGER, *Some graph-coloring theorems with applications to generalized connection networks*, SIAM J. Algebraic Discrete Methods, 6 (1985), pp. 576–582.

[7] G. M. MASSON AND B. W. JORDAN, JR., *Generalized multi-stage connection networks*, Networks, 2 (1972), pp. 191–209.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 291–301

A DICHOTOMY THEOREM ON FIXED POINTS OF SEVERAL NONEXPANSIVE MAPPINGS*

TOMÁS FEDER†

Abstract. The problem of finding a fixed point of a nonexpansive mapping on a hypercube is that it has a polynomial time algorithm. In fact, it is known that one can find a 2-satisfiability characterization of the set of all fixed points in polynomial time. This implies that the problem of finding a vertex that is a common fixed point of several given nonexpansive mappings on a hypercube is that it has a polynomial time algorithm.

We consider the problem of finding a vertex that is a common fixed point of several given nonexpansive mappings on a more general Cartesian product of graphs. For a single nonexpansive mapping, a known polynomial time algorithm finds a fixed point and a 2-satisfiability-like characterization of all fixed points. We introduce graphs with a farthest point property (also called *apiculate graphs* in [H. J. Bandelt and V. Chepoi, *The Algebra of Metric Betweenness: Subdirect Representations, Retracts, and Axiomatics*, manuscript]), and show that finding a common fixed point of several nonexpansive mappings on Cartesian products of such graphs involves using a polynomial time algorithm. We generalize this result to any family of graphs having a majority function.

By contrast, the smallest graph (in the sense of having the fewest vertices, and the fewest edges of those having the fewest vertices) without the farthest point property is $K_{2,3}$, and finding a vertex that is a fixed point of two given nonexpansive mappings (retractions) on a Cartesian product of graphs isomorphic to $K_{2,3}$ is NP-complete. More generally, we exhibit an infinite family of graphs without the farthest point property giving NP-completeness. We show that for any family of graphs not having a majority function, the existence of a common fixed point of two nonexpansive mappings on Cartesian products of such graphs is NP-complete. This proves a dichotomy for the problem based on the existence of a majority function; a similar dichotomy is obtained for the special case of nonexpansive mappings that are retractions. Finally we characterize the families of chordal graphs corresponding to both dichotomies.

Key words. fixed points, nonexpansive mappings, product graphs, apiculate graphs, retractions

AMS subject classifications. 68R10

DOI. 10.1137/S0895480103427734

1. Introduction.

Many results in computational complexity take the form of a dichotomy theorem, where every problem in a given class is shown to be either polynomial time solvable or NP-complete. An early result is Schaefer's dichotomy of Boolean constraint satisfaction problems [7]. Beyond the Boolean domain, an approach to general constraint satisfaction was proposed by Feder and Vardi [5]. Recently, Bulatov has classified 3-element and conservative constraint satisfaction problems, with two dichotomy theorems [2, 3].

A similar dichotomy and classification project, for network stability problems, was initiated by Mayr and Subramanian [6] and Subramanian [8, 9]. They showed that every Boolean network stability problem is either monotone, linear, adjacency-preserving, or NP-complete. They also showed that the monotone and linear cases are polynomial, and that a family of adjacency-preserving problems, the scatter-free case (containing stable matching as a special case), is polynomial as well. The general adjacency-preserving case was studied as the problem of finding a fixed point of a nonexpansive mapping on a hypercube, and was shown polynomial by Feder [4], also

*Received by the editors May 11, 2003; accepted for publication (in revised form) October 25, 2005; published electronically March 24, 2006.
http://www.siam.org/journals/sidma/20-2/42773.html
†Stanford University, 268 Waverley St., Palo Alto, CA 94301 (tomas@theory.stanford.edu).

generalizing the nonexpansive case to other Cartesian products of graphs, and thus extending the results beyond the Boolean domain.

The n-dimensional hypercube (or n-cube) is the graph $G = (V, E)$, where V consists of all n-bit vectors $x = x_1x_2 \cdots x_n$, $x_i \in \{0, 1\}$, and two vertices x, y in V are joined by an edge in E if there exists a $1 \leq i \leq n$ such that $x_i \neq y_i$, and $x_j = y_j$ for all $j \neq i$. The distance $d(x, y)$ between two vertices x, y in the n-cube equals the number of positions $1 \leq i \leq n$ such that $x_i \neq y_i$.

A mapping $f : V(G) \to V(G)$ on the n-cube G is *nonexpansive* if $d(f(x), f(y)) \leq d(x, y)$ for all vertices x, y in G. A *fixed point* of a nonexpansive mapping f is a vertex x such that $f(x) = x$. Assuming that a nonexpansive f on the n-cube is given by a black box that can be queried in polynomial time, we specify an input x to f and the black box gives the image $f(x)$. Feder [4] gave a polynomial time algorithm for finding a fixed point x of a nonexpansive f on the n-cube, if one exists, and a second polynomial time algorithm that finds a 2-satisfiability instance on the variables x_i whose set of solutions characterizes the set of fixed points $x = x_1x_2 \cdots x_n$ of the nonexpansive mapping f.

Suppose we are given a collection of nonexpansive mappings $f_i : V(G) \to V(G)$ on the n-cube G, $1 \leq i \leq m$, and that we wish to find a common fixed point x with $f_i(x) = x$ for all i. We may then combine the m corresponding 2-satisfiability instances to obtain a single 2-satisfiability instance characterizing the set of all common fixed points, which can be solved in polynomial time to find a common fixed point, if one exists.

We study a generalization of the problem of finding common fixed points of a collection of nonexpansive mappings to Cartesian products of graphs, which include the hypercube as the simplest special case. Given n graphs $G_1, G_2, \ldots, G_n$, their *Cartesian product* $G = G_1 \square G_2 \square \cdots \square G_n$ has vertices $V(G)$ given by $x = x_1x_2 \cdots x_n$ with $x_i \in V(G_i)$, and two vertices x, y in $V(G)$ are joined by an edge in $E(G)$ if there exists an $1 \leq i \leq n$ such that (x_i, y_i) is an edge in $E(G_i)$, and $x_j = y_j$ for all $j \neq i$. The distance function d on G satisfies $d(x, y) = \sum_{1 \leq i \leq n} d_i(x_i, y_i)$, where d_i is the distance function on G_i.

A mapping $f : V(G) \to V(G)$ on a Cartesian product $G = G_1 \square G_2 \square \cdots \square G_n$ is *nonexpansive* if $d(f(x), f(y)) \leq d(x, y)$ for all vertices x, y in $V(G)$. A *fixed point* of a nonexpansive mapping f on G is a vertex x such that $f(x) = x$. Assuming again that f is given by a black box that can be queried in polynomial time, Feder [4] showed the following theorem.

THEOREM 1.1. *There is a polynomial time algorithm (polynomial in the sum of the sizes of the G_i) that finds sets $S_{ij} \subseteq V(G_i \square G_j)$ for all $1 \leq i < j \leq n$ such that, given a partial assignment of values $a_i \in V(G_i)$ for $i \in S \subseteq \{1, 2, \ldots, n\}$ with $|S| \geq 2$, there exists a fixed point x of the nonexpansive mapping f such that $x_i = a_i$ for all $i \in S$ if and only if $a_ia_j \in S_{ij}$ for all $1 \leq i < j \leq n$ with $i, j \in S$. The partial assignment of values a_i can thus be extended to a fixed point x in polynomial time by considering the sets S_{ij}.*

Suppose now we are given a collection of nonexpansive mappings $f_i : V(G) \to V(G)$ on a Cartesian product $G = G_1 \square G_2 \square \cdots \square G_n$, with $1 \leq i \leq m$. We wish to determine whether the question of the existence of a common fixed point x with $f_i(x) = x$ for all $1 \leq i \leq m$ can be solved in polynomial time; we obtain both positive and negative answers to this question.

Given vertices x_i, y_i in $V(G_i)$, the *interval* $I(x_i, y_i)$ is the set of vertices t_i in G_i such that $d(x_i, y_i) = d(x_i, t_i) + d(t_i, y_i)$. We say that G_i satisfies the *farthest point*

property if for all vertices x_i, y_i, z_i there is a unique vertex t_i in $I(x_i, y_i) \cap I(x_i, z_i)$ that maximizes $d(x_i, t_i)$ over all t_i in $I(x_i, y_i) \cap I(x_i, z_i)$. Note that, in particular, cliques and cycles satisfy the farthest point property, and that the Cartesian products and the retracts of graphs satisfying the farthest point property also satisfy the farthest point property.

We give a polynomial time algorithm for finding a common fixed point x with $f_i(x) = x$ for all $1 \leq i \leq m$ in the Cartesian product $G = G_1 \square G_2 \square \cdots \square G_n$, when all G_i satisfy the farthest point property. We generalize this result to any family $\mathcal{G}$ of graphs G_i having a majority function. The results are shown in a more general form analogous to Theorem 1.1 involving a structural property that we show holds only for families $\mathcal{G}$ of graphs G_i having a majority function.

The smallest graph not satisfying the farthest point property is the complete bipartite graph $K_{2,3}$. We define a family of graphs $H_{a,b,c}$ for integers $a, b, c \geq 1$ that do not satisfy the farthest point property; in particular $a = b = c = 1$ gives $H_{1,1,1} = K_{2,3}$. We consider the Cartesian product $G = G_1 \square G_2 \square \cdots \square G_n$, in the case where all G_i are isomorphic to a given $H_{a,b,c}$. We then define pairs of nonexpansive mappings $f_1, f_2 : V(G) \to V(G)$ that are *retractions*, i.e., these mappings satisfy $f_i(f_i(x)) = f_i(x)$ for all vertices x in G. We show that for all choices of $H_{a,b,c}$ with $a, b, c \geq 1$, the question of whether two such mappings have a common fixed point $f_1(x) = f_2(x) = x$ is NP-complete.

We show that if a family $\mathcal{G}$ of graphs G_i does not have a majority function, then the existence of a common fixed point of two nonexpansive mappings f_i on Cartesian products of graphs in $\mathcal{G}$ is NP-complete. This proves the dichotomy and classification of the problem of finding common fixed points of several nonexpansive mappings on Cartesian products of graphs in $\mathcal{G}$ as polynomial or NP-complete depending on whether $\mathcal{G}$ has a majority function or not. A similar dichotomy holds for the special case of retraction mappings.

Finally we characterize the families of chordal graphs that are polynomial or NP-complete for the problems of fixed points of multiple nonexpansive mappings and of fixed points of multiple retraction mappings.

2. Common fixed points with farthest point property and with majority function. Let $f_i : V(G) \to V(G)$ be nonexpansive mappings on a Cartesian product $G = G_1 \square G_2 \square \cdots \square G_n$, with $1 \leq i \leq m$. Assume again that f_i is given by a black box that can be queried in polynomial time. We generalize Theorem 1.1 in the case where each G_i satisfies the farthest point property.

THEOREM 2.1. *Suppose each G_i satisfies the farthest point property. There is a polynomial time algorithm that finds sets $S_{ij} \subseteq V(G_i \square G_j)$ for all $1 \leq i < j \leq n$ such that, given a partial assignment of values $a_i \in V(G_i)$ for $i \in S \subseteq \{1, 2, \ldots, n\}$ with $|S| \geq 2$, there exists a common fixed point x of the nonexpansive mappings f_i such that $x_i = a_i$ for all $i \in S$ if and only if $a_i a_j \in S_{ij}$ for all $1 \leq i < j \leq n$ with $i, j \in S$. The partial assignment of values a_i can thus be extended to a common fixed point x with $f_i(x) = x$ for all $1 \leq i \leq m$ in polynomial time by considering the sets S_{ij}.*

Proof. We can apply Theorem 1.1 to find the corresponding sets S_{ij}^k for each f_k. We can then define sets T_{ij} that are the intersection of all S_{ij}^k over all $1 \leq k \leq m$. Define a mapping $g_i(x_i, y_i, z_i) = t_i$ on $V(G_i)$ to be the unique t_i maximizing $d(x_i, t_i)$ for $t_i \in I(x_i, y_i) \cap I(x_i, z_i)$ by the definition of the farthest point property. Note that if x, y, z are fixed points of f_k, so that $f_k(x) = x, f_k(y) = y, f_k(z) = z$, then $t = t_1 t_2 \cdots t_n$ with $t_i = g_i(x_i, y_i, z_i)$ satisfies $t \in I(x, y) \cap I(x, z)$, $f_k(t) \in I(x, y) \cap I(x, z)$, and $d(x, t) = d(x, f_k(t))$, so $f_k(t)_i \in I(x_i, y_i) \cap I(x_i, z_i)$, and $d(x_i, t_i) = d(x_i, f_k(t)_i)$,

therefore $f_k(t)_i = t_i$ and $f_k(t) = t$. That is, the vertex t is also a fixed point of f_k.

Thus the sets S^k_{ij} are such that if $x_ix_j, y_iy_j, z_iz_j \in S^k_{ij}$, then $g_i(x_i, y_i, z_i)g_i(x_j, y_j, z_j) \in S^k_{ij}$. Furthermore, we have $g_i(x_i, x_i, y_i) = g_i(x_i, y_i, x_i) = g_i(y_i, x_i, x_i) = x_i$. Such a function g_i is called a majority function. We may assume that the graphs G_i are disjoint, and extend the mappings g_i on each G_i to a single mapping g on $V(G_1) \cup V(G_2) \cup \cdots \cup V(G_n)$ satisfying $g(x, x, y) = g(x, y, x) = g(y, x, x) = x$. Furthermore, the sets T_{ij} are closed under g, that is, if $x_ix_j, y_iy_j, z_iz_j \in T_{ij}$, then $g(x_i, y_i, z_i)g(x_j, y_j, z_j) \in T_{ij}$. The mapping g is thus a *majority function* for the sets T_{ij}. The problem of simultaneously satisfying the conditions $x_ix_j \in T_{ij}$ when there exists a majority function is a constraint satisfaction that can be solved in polynomial time, as shown by Feder and Vardi [5]. In fact the algorithm in [5] infers smaller sets $S_{ij} \subseteq T_{ij}$ from the sets T_{ij} by enforcing the condition that for all $l \neq i, j$, if $x_ix_j \in S_{ij}$, then there exists an x_l such that $x_ix_l \in S_{il}$ and $x_jx_l \in S_{jl}$ (otherwise x_ix_j is removed from S_{ij}). As shown in [5], these inferred sets S_{ij} satisfy the condition that any partial solution can be extended to a full solution, as in the statement of the Theorem, completing the proof. □

A family of graphs $\mathcal{G}$ has a *majority function* if each graph G_i in $\mathcal{G}$ has a function g_i such that $g_i(x_i, x_i, y_i) = g_i(x_i, y_i, x_i) = g_i(y_i, x_i, x_i) = x_i$ for all x_i, y_i in $V(G_i)$, and for every pair of graphs G_i, G_j in $\mathcal{G}$, if f is a nonexpansive mapping on $G_i \square G_j$, and x_ix_j, y_iy_j, z_iz_j are fixed points of f, then $g_i(x_i, y_i, z_i)g_j(x_j, y_j, z_j)$ is a fixed point of f as well.

Note from Theorem 2.1 that the family of graphs with the farthest point property has a majority function. Another example of graphs with a majority function is the family of graphs G_i such that for all x_i, y_i, z_i in $V(G_i)$ there is a unique vertex t_i in $V(G_i)$ minimizing $d(x_i, t_i) + d(y_i, t_i) + d(z_i, t_i)$.

THEOREM 2.2. *The statement of Theorem* 2.1 *holds for a family $\mathcal{G}$ of graphs G_i if and only if $\mathcal{G}$ has a majority function.*

Proof. Suppose $\mathcal{G}$ has a majority function. Obtain the sets S^k_{ij} as in Theorem 2.1. The nonexpansive mapping f_k on G can be projected to a nonexpansive mapping $f_{k,ij}$ on $G_i \square G_j$ whose fixed points are given by S^k_{ij} as shown in [4]. The projection $f_{k,ij}$ is obtained by fixing inputs x_i, x_j, finding a periodic point p for the resulting function on the remaining positions other than i, j, and obtaining the outputs y_i, y_j, independently of the choice of periodic point p. Thus the S^k_{ij} are closed under the majority functions g_i, g_j, and the result follows as in Theorem 2.1.

If $\mathcal{G}$ does not have a majority function, then by the compactness theorem, the family $\mathcal{G}$ has a finite subfamily $\mathcal{G}'$ that does not have a majority function. Then the sets S^k_{ij} that can be defined as fixed points of nonexpansive mappings $f_{k,ij}$ do not satisfy the 2-*Helly property*, since the 2-Helly property is equivalent to the existence of a majority function, as shown by Feder and Vardi [5]. The 2-Helly property is the statement that a partial assignment of a_i can be extended to a full solution if and only if partial assignments consisting of just pairs a_ia_j can be extended to a full solution, as in the statement of Theorem 2.1. □

3. Common fixed points without farthest point property and without majority function. Given integers $a, b, c \geq 1$, let $H_{a,b,c}$ be the graph consisting of five particular vertices x, y, z, u, v, and six paths joining them, with two paths of length a, from x to u and from x to v, two paths of length b, from y to u and from y to v, and two paths of length c, from z to u and from z to v. Note that in particular, for $a = b = c = 1$, the graph $H_{1,1,1}$ is a complete bipartite graph $K_{2,3}$ with u, v on one side and x, y, z on the other. The graphs $H_{a,b,c}$ do not satisfy the farthest point

property, since $d(x,t)$ is maximized for $t \in I(x,y) \cap I(x,z)$ at both $t = u$ and $t = v$.

For a particular choice of integers $a, b, c \geq 1$, consider the Cartesian product $G = G_1 \square G_2 \square \cdots \square G_n$, where each G_i is isomorphic to $H_{a,b,c}$, so that vertices x, y, z, u, v in $H_{a,b,c}$ correspond to vertices x_i, y_i, z_i, u_i, v_i in G_i.

A retraction on G is a nonexpansive mapping $f : V(G) \to V(G)$ such that $f(f(x)) = f(x)$ for all $x \in V(G)$. We define several retractions on G. The retractions f_{u_i} map r to s in such a way that $s_j = r_j$ for $j \neq i$, $s_i = r_i$ for r_i on the paths joining u_i to x_i, y_i, z_i, respectively, and if r_i is on a path joining v_i to one of x_i, y_i, z_i, then s_i is the corresponding vertex on the corresponding path joining u_i to one of x_i, y_i, z_i, respectively. Similarly, the retractions f_{v_i} map r to s in such a way that $s_j = r_j$ for $j \neq i$, $s_i = r_i$ for r_i on the paths joining v_i to x_i, y_i, z_i, respectively, and if r_i is on a path joining u_i to one of x_i, y_i, z_i, then s_i is the corresponding vertex on the corresponding path joining v_i to one of x_i, y_i, z_i, respectively.

For $t_i \in \{x_i, y_i, z_i\}$ and $w_j \in \{x_j, y_j, z_j\}$, the retractions f_{t_i,w_j} compose f_{u_i} with f_{u_j}, and then with the mapping taking r to s in such a way that $s_k = r_k$ for $k \neq i, j$, and $s_i = r_i, s_j = r_j$ unless $r_i = t_i$ and $r_j = w_j$, in which case s_i and s_j are the neighbors of t_i and w_j on paths to u_i and u_j, respectively.

THEOREM 3.1. *For each choice of integers $a, b, c \geq 1$, one can define retractions f_1, f_2 as compositions of the retractions $f_{u_i}, f_{v_i}, f_{t_i,w_j}$ such that the existence of a common fixed point $f_1(x) = f_2(x) = x$ is NP-complete.*

Proof. An instance of the one-in-three satisfiability problem has a collection of variables and triples of variables (X, Y, Z), and asks whether there exists an assignment of values in $\{0, 1\}$ to each variable in such a way that for each triple (X, Y, Z) in this instance, one of X, Y, Z has value 1 and the other two have value 0. The one-in-three satisfiability problem is NP-complete [7].

Let n be the number of triples (X, Y, Z) in a one-in-three satisfiability problem, so that each G_i corresponds to one such triple, with a correspondence between X, Y, Z and x_i, y_i, z_i in G_i. Let f_1 be the composition of the retractions f_{v_i} for $1 \leq i \leq n$. Let f_2 be the composition of the retractions f_{u_i} for $1 \leq i \leq n$, composed with the retractions f_{t_i,w_j} such that t_i and w_j correspond to two distinct variables T and W in a triple (X, Y, Z) in the one-in-three satisfiability instance.

The mappings f_1 and f_2 are retractions. Since the range of f_1 contains only vertices s such that s_i is on a path from v_i to one of x_i, y_i, z_i, and since the range of f_2 contains only vertices s such that s_i is on a path from u_i to one of x_i, y_i, z_i, it follows that the only candidate fixed points $f_1(s) = f_2(s) = s$ have all $s_i \in \{x_i, y_i, z_i\}$. A choice of such an s with $s_i \in \{x_i, y_i, z_i\}$ thus chooses a variable from the corresponding triple (X, Y, Z). Furthermore, if X is chosen for one such triple, then X will also be chosen for all other triples containing X; otherwise $f_2(s) \neq s$, since the retractions f_{t_i,w_j} used in the definition of f_2 guarantee that we cannot choose two distinct variables T and W that share a triple. Thus the choice of s_i satisfies $f_1(s) = f_2(s) = s$ if and only if the s_i chooses an element from $\{x_i, y_i, z_i\}$ corresponding to a solution to the one-in-three satisfiability problem. □

We next show the following theorem.

THEOREM 3.2. *Let $\mathcal{G}$ be a family of graphs that does not have a majority function. Then the question of whether several nonexpansive mappings f_i on a Cartesian product G of graphs G_i in $\mathcal{G}$ have a common x such that $f_i(x) = x$ for all f_i is NP-complete.*

Proof. As in the proof of Theorem 2.2, if $\mathcal{G}$ does not have a majority function, then $\mathcal{G}$ has a finite subfamily $\mathcal{G}'$ that does not have a majority function, and then the sets S_{ij}^k that can be defined as fixed points of nonexpansive mappings $f_{k,ij}$ do

not satisfy the 2-Helly property. This means that one can define an instance of the problem with variables x_i constrained by sets S^k_{ij} of fixed points of $f_{k,ij}$ in such a way that for some $r \geq 3$, there exist values $a_1, a_2, \ldots, a_r$ such that no solution x has $x_i = a_i$ for all $1 \leq i \leq r$, but for each $1 \leq i \leq r$, there exists a solution x^i that has $x^i_j = a_j$ for all $1 \leq j \leq r$ with $j \neq i$.

Let b_i be a possible value for x^i_i which is closest to a_i in G_i. Let P_i denote a shortest path from a_i to b_i in G_i. We may restrict each variable x_i with $1 \leq i \leq r$ ranging over G_i to the path P_i by considering fixed points of the retraction that maps G_i to P_i by mapping vertices in G_i at distance l from a_i to the point on P_i at distance $\min(l, d(a_i, b_i))$ from a_i.

Let H_i for $1 \leq i \leq r$ be a graph consisting of an edge joining two vertices 0 and 1. Considering a retraction on $H_i \square P_i$ that leaves $0a_i$ and all $1x$ fixed by mapping all $0x$ with $x \neq a_i$ to $1y$, where y is the neighbor of x on P_i toward a_i. If $y_i x_i$ is the vertex of $H_i \square G_i$ so constrained, then there is no solution with $y_i = 0$ for all $1 \leq i \leq r$, since such a solution would have $x_i = a_i$ for all $1 \leq i \leq r$. However, for each $1 \leq i \leq r$, there exists a solution with $y_i = 1$ and $y_j = 0$ for all $1 \leq j \leq r$ with $j \neq i$.

In particular, if we require $y_j = 0$ for all $3 \leq j \leq r$ with an appropriate retraction on H_j, we have that (y_1, y_2, y_3) cannot be $(0, 0, 0)$ but can be $(0, 0, 1)$, $(0, 1, 0)$, or $(1, 0, 0)$. Furthermore, for $1 \leq i < j \leq 3$, we can define a retraction on $H_i \square H_j$ that maps 11 to 00, while leaving $00, 01, 10$ fixed. This enforces $(y_i, y_j) \neq (1, 1)$, so that the only possible values for (y_1, y_2, y_3) are $(0, 0, 1)$, $(0, 1, 0)$, and $(1, 0, 0)$. Therefore, by using fixed points of several nonexpansive mappings, we have defined the one-in-three satisfiability relation on a Boolean domain, and the result follows from the NP-completeness of one-in-three satisfiability [7]. □

COROLLARY 3.3. *Let $\mathcal{G}$ be a family of graphs. Then the problem of finding a common fixed point of several nonexpansive mappings f_i on a product $G = G_1 \square G_2 \square \cdots \square G_n$ of graphs G_i from $\mathcal{G}$ is either polynomial time solvable or NP-complete, even with just two nonexpansive mappings f_1 and f_2 such that f_1 is composed of nonexpansive mappings on $G_{2i-1} \square G_{2i}$, and f_2 is given by a permutation σ on $1 \leq i \leq n$ and isomorphisms from G_i to $G_{\sigma(i)}$.*

Proof. If $\mathcal{G}$ has a majority function, then polynomiality is shown in Theorem 2.2.

If $\mathcal{G}$ does not have a majority function, then in the NP-completeness proof of Theorem 3.2 we have only nonexpansive mappings f^k_{ij} on $G_i \square G_j$ whose fixed points give S^k_{ij}. We may then make r copies of each G_i, one for each f^k, so that we now have a single nonexpansive mapping f_1 so composed. It remains to force equality between the r copies of G_i given by $G_{i1}, \ldots, G_{ir}$, which is achieved by letting $\sigma(il) = i(l+1)$, with $l + 1$ obtained modulo r, and f_2 mapping G_{il} to $G_{\sigma(il)}$, giving this equality correspondence for fixed points of f_2. □

COROLLARY 3.4. *Let $\mathcal{G}$ be a family of graphs. Then the problem of finding a common fixed point of several retractions f_i on a product $G = G_1 \square G_2 \square \cdots \square G_n$ of graphs G_i from $\mathcal{G}$ is either polynomial time solvable or NP-complete, even with just retractions on products $G_i \square G_j$.*

Proof. Modify the definition of a majority function for $\mathcal{G}$ by considering only nonexpansive mappings f on $G_i \square G_j$ that are retractions. The polynomiality in the presence of a majority function follows as in Theorem 2.2, while the NP-completeness in the absence of a majority follows as in Theorem 3.2 since all nonexpansive mappings explicitly defined in the proof are retractions. □

4. Chordal graphs. In this section we characterize the families of chordal graphs giving rise to polynomiality or NP-completeness for the problems of fixed

points of multiple nonexpansive mappings and of fixed points of multiple retractions. A chordal graph is a graph that does not contain any chordless cycles of length at least four, where a chordless cycle is a cycle where no two nonconsecutive vertices in the cycle are joined by an edge.

Consider a choice of integers $l_1, l_2, l_3 \geq 0$. The graph $R^0 = R^0_{l_1,l_2,l_3}$ has a core S^0 consisting of a single vertex y; in addition, there are three vertices x_1, x_2, x_3 in R^0, where vertex x_i is joined to the core S^0 by a path p_i of length l_i from x_i to y.

The graph $R^1 = R^1_{l_1,l_2,l_3}$ has a core S^1 consisting of a triangle $y_1y_2y_3$; in addition, there are three vertices x_1, x_2, x_3 in R^1, where vertex x_i is joined to the core S^0 by a path p_i of length l_i from x_i to y_i.

The graph $R^2 = R^2_{l_1,l_2,l_3}$ has a core S^2 consisting of six vertices $y_1, y_2, y_3, z_{12}, z_{23}$, z_{31} with a triangle z_{12}, z_{23}, z_{31} and edges joining y_1 to both z_{12} and z_{31}, joining y_2 to both z_{12} and z_{23}, and joining y_3 to both z_{23} and z_{31}; in addition, there are three vertices x_1, x_2, x_3 in R^2, where vertex x_i is joined to the core S^0 by a path p_i of length l_i from x_i to y_i.

A subgraph H of a graph G is an *isometric subgraph* if for every pair of vertices a, b in H, the distance between a and b is the same in H as in G.

Let G be a chordal graph and consider a triple of vertices a_1, a_2, a_3 in G. The triple is of type 0 if G has an isometric subgraph isomorphic to $R^0_{l_1,l_2,l_3}$ with the vertices x_1, x_2, x_3 coinciding with a_1, a_2, a_3, respectively. The triple is of type 1 if it is not of type 0 and G has an isometric subgraph isomorphic to $R^1_{l_1,l_2,l_3}$ with the vertices x_1, x_2, x_3 coinciding with a_1, a_2, a_3, respectively. The triple is of type 2 if it is not of type 0 or 1 and G has an isometric subgraph isomorphic to $R^2_{l_1,l_2,l_3}$ with the vertices x_1, x_2, x_3 coinciding with a_1, a_2, a_3, respectively.

LEMMA 4.1. *Let G be a chordal graph. Then every triple of vertices a_1, a_2, a_3 in G is of type* 0*, type* 1*, or type* 2*.*

Proof. Let b_1, b_2, b_3 be three vertices in G satisfying $d(a_i, a_j) = d(a_i, b_i)+d(b_i, b_j)+d(b_j, a_j)$ for $i \neq j$. Note that we may always choose $(b_1, b_2, b_3) = (a_1, a_2, a_3)$; choose the vertices b_i so as to minimize $D = d(b_1, b_2)+d(b_2, b_3)+d(b_1, b_3)$. Assume $d(b_1, b_2) \leq d(b_2, b_3) \leq d(b_1, b_3)$. We consider three cases.

If $d(b_1, b_2) = 0$, then $d(b_2, b_3) = d(b_1, b_3) = 0$, since in this case $D = 0$ can be achieved by setting b_3 to the common vertex $b_1 = b_2$. The three paths from a_1, a_2, a_3 to $b_1 = b_2 = b_3$ then form an $R^0_{l_1,l_2,l_3}$ isometric subgraph.

If $d(b_1, b_2) = 1$, then $d(b_2, b_3) = d(b_1, b_3) = 1$. Otherwise, if $d(b_1, b_3) = d(b_2, b_3) + 1$, then in this case $D = 0$ can be achieved by setting b_1 and b_3 to the vertex b_2; and if $d(b_2, b_3) = d(b_1, b_3) \geq 2$, then the two paths from b_2 to b_3 and from b_1 to b_3 must be disjoint except for b_3, since otherwise we could move b_3 to the next common vertex and reduce D; furthermore, the ith vertex on the path from b_2 to b_3 can only have an edge to the ith vertex on the path from b_1 to b_3, since otherwise we could use an edge from the ith vertex in one path to the $(i + 1)$th vertex on the other path to modify the second path to go through the ith vertex in the first path, so that the two paths share a vertex other than b_3 as before. The fact that these are the only additional edges for the cycle through b_1, b_2, b_3 implies that the graph has a chordless cycle of even length at least four, contrary to chordality. Thus b_1, b_2, b_3 form a triangle and the three paths from a_1, a_2, a_3 to b_1, b_2, b_3 form an $R^1_{l_1,l_2,l_3}$ isometric subgraph.

Finally, if $d(b_1, b_2) \geq 2$, then $d(b_1, b_2) = d(b_2, b_3) = d(b_1, b_3) = 2$. Otherwise, the only edges joining the path from b_1 to b_2 to the path from b_1 to b_3 must join the ith vertex in one path to the ith vertex in the other path, otherwise we could reduce D by replacing one of the two paths by a path that shares a vertex other than b_1 with

the other path; furthermore, only one such value i is possible, otherwise we would have a chordless cycle of even length at least four, contrary to chordality. Thus there is only one edge joining each pair of the three paths from b_1 to b_2, from b_1 to b_3, and from b_2 to b_3, and adding three edges to the cycle going through b_1, b_2, b_3 can only add enough chords to forbid chordless cycles if $d(b_1, b_2) = d(b_2, b_3) = d(b_1, b_3) = 2$. In that case, we have paths $b_1c_{12}b_2$, $b_2c_{23}b_3$, and $b_3c_{31}b_1$, with chords forming a triangle $c_{12}c_{23}c_{31}$. These six vertices form the core S^2, and the three paths from a_1, a_2, a_3 to b_1, b_2, b_3 form an $R^2_{l_1,l_2,l_3}$ isometric subgraph, since a shorter distance $d(a_1, c_{23}) = l_1+1$ would allow achieving $D = 0$ by setting b_1, b_2, b_3 to the vertex c_{23}, and similarly for $d(a_2, c_{31}) = l_2 + 1$ and $d(a_3, c_{12}) = l_3 + 1$. □

We strengthen this result from isometric subgraphs to retracts.

LEMMA 4.2. *Let G be a chordal graph, with a triple of vertices a_1, a_2, a_3 of type j giving an $R^j_{l_1,l_2,l_3}$ isometric subgraph. Then the $R^j_{l_1,l_2,l_3}$ isometric subgraph is a retract of G.*

Proof. Map the vertices at distance $d \leq l_i$ from a_i in G to the corresponding vertex at distance d from a_i on the path from a_i to the core S^j. If $j = 0, 1$, then the remaining vertices of G can be mapped to any vertex in the core S^j. If $j = 2$, then the remaining vertices can be mapped to vertices in the triangle z_{12}, z_{23}, z_{31} inside the core S^2, since no vertex will be required to map to a vertex adjacent to all three vertices y_1, y_2, y_3 in the core, otherwise such a vertex would give type 0 to the triple a_1, a_2, a_3. □

The following result characterizes the images of products of cores.

LEMMA 4.3. *Let f be a nonexpansive mapping on a product $G \square G'$ of two chordal graphs, with three vertices $x_1x'_1$, $x_2x'_2$, $x_3x'_3$ that are fixed points of f. Suppose the triple x_1, x_2, x_3 is of type j in G and gives an $R^j_{l_1,l_2,l_3}$ isometric subgraph with core S^j; suppose the triple x'_1, x'_2, x'_3 is of type j' in G' and gives an $R^{j'}_{l_1,l_2,l_3}$ isometric subgraph with core $S^{j'}$. Then the image of $S^j \square S^{j'}$ under f is a product $U^j \square U^{j'}$ for two choices of cores U^j and $U^{j'}$ isomorphic to S^j and $S^{j'}$, respectively.*

Proof. The vertex y_i (or y if $j = 0$) maximizes $d(x_i, y_i)$ subject to $d(x_i, x_{i'}) = d(x_i, y_i) + d(y_i, x_{i'})$ for $i' \neq i$. Similarly, the vertex y'_i (or y' if $j = 0$) maximizes $d(x'_i, y'_i)$ subject to $d(x'_i, x'_{i'}) = d(x'_i, y'_i) + d(y'_i, x'_{i'})$ for $i' \neq i$. It follows that if $f(y_iy'_i) = t_it'_i$, then $d(t_i, t_{i'}) = d(y_i, y_{i'}) = j$ and $d(t'_i, t'_{i'}) = d(y'_i, y'_{i'}) = j'$.

If $j = 0$ and $j' = 0$, then $S^j \square S^{j'}$ is a single vertex and the result follows. If $j = 1$ and $j' = 0$, then the three vertices t_1, t_2, t_3 form a triangle and the result follows. If $j = 2$ and $j' = 0$, then the three vertices t_1, t_2, t_3 are at pairwise distance 2, and if we consider the vertices $u_{ij}y' = f(z_{ij}y')$, then the three u_{ij} must be different, otherwise we could set the type j to 0 if two coincide. Thus the six vertices t_i and u_{ij} form a subgraph isomorphic to S^j and the result follows.

If $j = 1$ and $j' = 1$, then the three vertices t_i form a triangle and the three vertices t'_i also form a triangle. It follows that $f(y_iy'_{i'})$ is $t_it'_{i'}$ or $t_{i'}t'_i$. Furthermore, the six-cycle $y_1y'_2, y_1y'_3, y_2y'_3, y_2y'_1, y_3y'_1, y_3y'_2, y_1y'_2$ implies that either all $f(y_iy'_{i'}) = t_it'_{i'}$ or all $f(y_iy'_{i'}) = t_{i'}t'_i$, so the result follows.

If $j = 2$ and $j' = 1, 2$, then consider the S^2 given by $y_1y'_i, y_2y'_i, y_3y'_i, z_{12}y'_i, z_{23}y'_i, z_{31}y'_i$; it must map to $t_1t'_i, t_2t'_i, t_3t'_i, u_{12}t'_i, u_{23}t'_i, u_{31}t'_i$ or if $j' = 2$, then possibly also to $t_it'_1, t_it'_2, t_it'_3, t_iu'_{12}, t_iu'_{23}, t_iu'_{31}$. To see this, say if $i = 1$, note that the two paths of length two $y_1y'_1, z_{12}y'_1, y_2y'_1$ and $y_1y'_1, z_{31}y'_1, y_3y'_1$ must map to paths of length two because they approach $y_2y'_2$ and $y_3y'_3$, respectively. Furthermore, $z_{12}y'_1$ and $z_{31}y'_1$ must map to different vertices, since $f(y_2y'_2)$ and $f(y_3y'_3)$ cannot be approached simultaneously from $f(y_1y'_1)$. This gives the two possible ways of mapping the triangle

$y_1y'_1, z_{12}y'_1, z_{31}y'_1$, along G or along G'. If $j' = 1$, then in the second case we would have to switch to proceeding along G for the two paths of length two, and the triangles could not be completed. Note then that the triangle $z_{12}y'_1, z_{31}y'_1, z_{23}y'_1$ cannot be mapped to a single edge, else again $f(y_2y'_2)$ and $f(y_3y'_3)$ could be approached simultaneously from $f(y_1y'_1)$. The same argument implies that the triangles $y_2y'_1, z_{23}y'_1, y_{12}y'_1$ and $y_3y'_1, z_{23}y'_1, y_{31}y'_1$ must remain as triangles, proving the claim.

It follows again that we have either all $f(y_iy'_{i'}) = t_it'_{i'}$ or alternatively if $j' = 2$ all $f(y_iy'_{i'}) = t_{i'}t'_i$ by the same six-cycle argument from the case $j = 1$ and $j' = 1$. Thus all $f(y_iy'_{i'}) = t_it'_{i'}$, $f(y_iz'_{i'i''}) = t_iu'_{i'i''}$, $f(z_{ii'}y'_{i''}) = u_{ii'}t'_{i''}$, $f(z_{ii'}z'_{i''i'''}) = u_{ii'}u'_{i''i'''}$, or alternatively if $j' = 2$ we have all $f(y_iy'_{i'}) = t_{i'}t'_i$, $f(y_iz'_{i'i''}) = u_{i'i''}t'_i$, $f(z_{ii'}y'_{i''}) = t_{i''}u'_{ii'}$, $f(z_{ii'}z'_{i''i'''}) = u_{i''i'''}u'_{ii'}$.

There is thus only one possible isomorphism from $S^j \square S^{j'}$ to $U^j \square U^{j'}$, unless $j = j' = 1$ or $j = j' = 2$, in which case a second isomorphism first exchanges S^j and $S^{j'}$. □

Let G be a chordal graph and consider a triple a_1, a_2, a_3 of type j giving possibly several different $R^j_{l_1,l_2,l_3}$. If the corresponding possibly different cores S^j share a vertex, then this vertex in the core is said to be a *uniquely determined* vertex.

THEOREM 4.4. *Given a family $\mathcal{G}$ of chordal graphs, the problem of finding a common fixed point of several retractions on products of graphs from $\mathcal{G}$ can be solved in polynomial time if all triples of vertices a_1, a_2, a_3 of type j from a graph G in $\mathcal{G}$ give at least one uniquely determined vertex in the corresponding cores S^j. Otherwise the problem is NP-complete.*

Proof. If the property in the statement of the theorem holds, then for every triple a_1, a_2, a_3 of type j from a graph G_i in $\mathcal{G}$, define the majority function $g_i(a_1, a_2, a_3) = b$, where b is a uniquely determined vertex in the corresponding core S^j. Given a retraction mapping f on $G_i \square G_{i'}$ with fixed points $a_1a'_1, a_2a'_2, a_3a'_3$, the image under f of a product of cores $S^j \square S^{j'}$ gives by Lemma 4.3 such a product of cores which is fixed under f by the definition of a retraction. The fixed $S^j \square S^{j'}$ must contain $bb' = g_i(a_1, a_2, a_3)g_{i'}(a'_1, a'_2, a'_3)$ since b and b' are uniquely determined vertices and thus must belong to S^j and $S^{j'}$, respectively. Thus bb' is a fixed point of f. Furthermore, if two of a_1, a_2, a_3 are the same vertex a, then this triple is of type 0 with core a and so $g_i(a_1, a_2, a_3) = a$. This proves that the g_i are majority functions, and a polynomial time algorithm follows from Corollary 3.4.

In the other case, suppose three vertices a_1, a_2, a_3 of type j from a graph G in $\mathcal{G}$ do not give a uniquely determined vertex in the corresponding cores S^j. We show that G does not have a majority function. First, we observe that G retracts to an $R^j_{l_1,l_2,l_3}$ by Lemma 4.2, so $b = g(a_1, a_2, a_3)$ must belong to this $R^j_{l_1,l_2,l_3}$. Second, the vertex b cannot be a vertex in a core S^j, since every such vertex can be avoided by a different choice of S^j, since there is no uniquely determined core vertex. So the only possibility left is that b is on the path from some a_i to a core S^j, say $a_i = a_1$. Consider, however, a graph P with only two adjacent vertices c and d, so that on P we have $g'(c, d, d) = d$. We can define a retraction f on $R^j_{l_1,l_2,l_3} \square P$ that leaves every vertex fixed except for the vertices ud, where u is on the path from a_1 to the core S^j (but not in S^j); here we define $f(ud) = vc$, where v follows u on the path from a_1 to the core S_j. It follows that a_1c, a_2d, a_3d are fixed points of f, but bd is not a fixed point of f, so we do not have a majority function. By Corollary 3.4 the problem is NP-complete. □

We derive a similar result for general nonexpansive mappings.

THEOREM 4.5. *Given a family $\mathcal{G}$ of chordal graphs, the problem of finding a*

common fixed point of several nonexpansive mappings on products of graphs from $\mathcal{G}$ can be solved in polynomial time if all triples of vertices a_1, a_2, a_3 of type j from a graph G in $\mathcal{G}$ have all vertices in the corresponding cores S^j uniquely determined, except for the case $j = 2$, where there is either never a y_i that is not uniquely determined or never a $z_{ii'}$ that is not uniquely determined.

Proof. Again we define a majority function $g_i(a_1, a_2, a_3) = b$ for triples a_1, a_2, a_3 of type j by letting b be the uniquely determined vertex y in the core S^0 if $j = 0$, letting b be the uniquely determined vertex y_1 in the core S^1 if $j = 1$, and letting b be the uniquely determined vertex y_1 in the core S^2 if $j = 2$ and there is never a y_i that is not uniquely determined in this case, or letting b be the uniquely determined vertex z_{12} in the core S^2 if $j = 2$ and there is never a $z_{ii'}$ that is not uniquely determined. If $g_i(a_1, a_2, a_3) = b$ and $g_{i'}(a'_1, a'_2, a'_3) = b'$, and $a_1a'_1, a_2a'_2, a_3a'_3$ are fixed points of a nonexpansive f, then bb' is in the product of cores $S^j \square S^{j'}$ and must be left fixed by f by Lemma 4.3. It follows that the g_i are majority functions and a polynomial time algorithm follows from Theorem 2.2.

If the conditions of the theorem are not met, we show that there is no majority function. As in Theorem 4.4, we are left with setting $g_i(a_1, a_2, a_3) = b$, where b is uniquely determined in the corresponding core S^j. However, if, say, $g_i(a_1, a_2, a_3) = y_1$ in the core S^1, then we must now set, say, $g_{i'}(a'_1, a'_2, a'_3) = y'_1$ in the core $S^{j'}$ when $j' = 1$, since there is a nonexpansive mapping that exchanges the two cores S^1 and $S^{j'}$ as suggested by Lemma 4.3 so that $f(y_iy'_{i'}) = y_{i'}y'_i$, forcing $i = i'$ to get a fixed point. This thus requires all y_i in S^1 to be uniquely determined. The case with $j = 2$ and the S^2 similary requires all y_i in the S^2 to be uniquely determined, or all $z_{ii'}$ in the S^2 to be uniquely determined, otherwise by exchanging the two cores we would get a vertex that is not a fixed point. The exchanging of the two cores can be performed after retracting $R^j_{l_1,l_2,l_3} \square R^j_{l'_1,l'_2,l'_3}$ to the product of cores $S^j \square S^j$ joined by paths of length $l_1 + l'_1, l_2 + l'_2, l_3 + l'_3$ to $a_1a'_1, a_2a'_2, a_3a'_3$, respectively. This retraction is obtained by first mapping a vertex uv on a path from $y_iy'_{i'}$ to $x_ix'_{i'}$ with $i \neq i'$ to a vertex $u'v'$ on such a path with $d(u, u') = d(v, v')$ and either $u' = y_i$ or $v' = y'_{i'}$. Subsequently, every vertex not on the product of cores $S^j \square S^j$ is on a shortest path from this product to a vertex $x_ix'_i$, and can be mapped according to the distance to $x_ix'_i$ along the path of length $l_i + l'_i$ from $x_ix'_i$ and proceeding if necessary into the product of cores. Consequently, the absence of a majority function gives NP-completeness by Theorem 3.2. □

5. Conclusion. We have studied the problem of finding a vertex in a Cartesian product of graphs that is a simultaneous fixed point of several given nonexpansive mappings. We found that the problem is polynomial time solvable if the graphs satisfy a farthest point property, and otherwise the simplest counter examples to the farthest point property with two retraction mappings give NP-completeness. The polynomial cases for several given nonexpansive mappings extend to any family of graphs that has a majority function, while families of graphs without a majority function give NP-completeness using just two nonexpansive mappings, thus establishing a dichotomy in the complexity of the problem giving a classification depending on the existence of a majority function for any given class of graphs. A similar classification is obtained for the case of nonexpansive mappings that are retraction mappings. The families of graphs having a majority function for nonexpansive mappings or retractions can be simply characterized in the case of chordal graphs.

Finally we note that following the work in [4], one may find in polynomial time a characterization for the set of periodic points of a nonexpansive mapping f given by a

black box, which constitutes a retract to the given product graph. One may similarly obtain as in [4] such a characterization for vertices belonging to fixed subproducts of f. The results in this paper may thus be extended from fixed points and retractions to periodic points and fixed subproducts.

REFERENCES

[1] H.-J. Bandelt and V. Chepoi, *The Algebra of Metric Betweenness: Subdirect Representations, Retracts, and Axiomatics of Weakly Median Graphs*, manuscript.

[2] A. A. Bulatov, *A dichotomy constraint on a three-element set*, in Proceedings of the 43rd Annual ACM Symposium on Theory of Computing, Montreal, 2002, pp. 649–658.

[3] A. A. Bulatov, *Tractable conservative constraint satisfaction problems*, in Proceedings of the 18th Annual IEEE Symposium on Logic in Computer Science, Ottawa, 2003, pp. 321–330.

[4] T. Feder, *Stable Networks and Product Graphs*, Doctoral dissertation, Stanford University, Palo Alto, CA, 1990. Mem. Amer. Math. Soc. 116 (1995), pp. 1–223.

[5] 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 (1999), pp. 57–104.

[6] E. Mayr and A. Subramanian, *The complexity of circuit value and network stability*, in Fourth Annual Conference on Structure in Complexity Theory, Eugene, OR, (1989), pp. 114–123. Full version in J. Comput. System Sci., 44 (1992), pp. 302–323.

[7] T. J. Schaefer, *The complexity of satisfiability problems*, in Proceedings of the Tenth Annual ACM Symposium on Theory of Computing, San Diego, ACM, New York, 1978, pp. 216–226.

[8] A. Subramanian, *A new approach to stable matching problems*, SIAM J. Comput., 23 (1994), pp. 671–701.

[9] A. Subramanian, *The Complexity of Circuit Value and Network Stability*, Doctoral dissertation, Stanford University, Palo Alto, CA, 1989.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 302–327

REAL NUMBER GRAPH LABELLINGS WITH DISTANCE CONDITIONS*

JERROLD R. GRIGGS† AND XIAOHUA TERESA JIN†

Abstract. The theory of integer λ-labellings of a graph, introduced by Griggs and Yeh [J. R. Griggs and R. K.-C. Yeh, *SIAM J. Discrete Math.*, 5 (1992), pp. 586–595], seeks to model efficient channel assignments for a network of transmitters. To prevent interference, labels for nearby vertices must be separated by specified amounts k_i depending on the distance i, $1 \le i \le p$. Here we expand the model to allow real number labels and separations. The main finding ("D-Set Theorem") is that for any graph, possibly infinite, with maximum degree at most Δ, there is a labelling of minimum span in which all of the labels have the form $\sum_{i=1}^{p} a_i k_i$, where the a_i's are integers ≥ 0. We show that the minimum span is a continuous function of the k_i's, and we conjecture that it is piecewise linear with finitely many pieces. Our stronger conjecture is that the coefficients a_i can be bounded by a constant depending only on Δ and p. We offer results in strong support of the conjectures, and we give formulas for the minimum spans of several graphs with general conditions at distance two.

Key words. channel assignment, graph labelling, generalized coloring

AMS subject classifications. 05C78, 05C15, 90B18

DOI. 10.1137/S0895480105446708

1. Integer labellings with distance conditions.

A steadily growing body of literature has evolved in the past 15 years on efficient integer labellings of the vertices of a finite simple graph with restrictions not only on adjacent vertices—as is the case with traditional graph coloring—but also on vertices at distance two.

In the traditional *Channel Assignment Problem*, introduced by Hale [17] and studied by Cozzens and Roberts [6] and many others, vertices of a graph $G = (V, E)$ correspond to transmitter locations, and their labels represent transmission channels. Adjacent vertices correspond to pairs of transmitters that interfere with each other due to their proximity. There is a given finite set T of integers ≥ 0, with $0 \in T$, of forbidden differences in channels for adjacent vertices. A vertex labelling $f : V \to \mathbb{Z}$ is a *T-coloring* provided that $|f(v) - f(w)| \notin T$ whenever vertices v and w are adjacent. Of course, one can select channels $f(v)$ that are very far apart, but this would require allocating a very large band of the frequency spectrum to the network. To optimize the assignment f we seek to minimize the *span*

$$\mathrm{sp}(f) := \max_{v \in V} f(v) - \min_{v \in V} f(v).$$

Note that labels need not be distinct. The set of labels used may contain gaps in the interval between the smallest and largest labels. It is the width of the interval, given by $\mathrm{sp}(f)$, that we seek to minimize.

In 1988 Roberts [27] described a new channel assignment problem, suggested by Tim Lanfear at NATO. This time we consider a given network of transmitters in the

*Received by the editors December 20, 2004; accepted for publication (in revised form) August 26, 2005; published electronically April 21, 2006. This research was supported in part by NSF grants DMS-0072187 and DMS-0302307. This research was also described in the second author's dissertation [19].

http://www.siam.org/journals/sidma/20-2/44670.html

†Mathematics Department, University of South Carolina, Columbia, SC 29208 (griggs,jin2@math.sc.edu). New address for second author is Department of Mathematics and Statistics, University of Vermont, Burlington, VT 05405 (xiaohua.jin@uvm.edu).

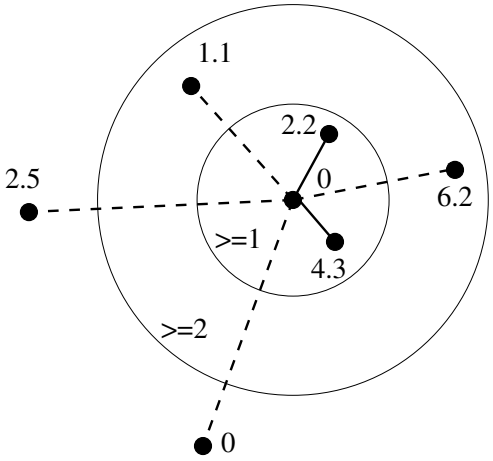


FIG. 1.1. *Labels for a planar transmitter network.*

plane, with two different levels of interference. An integer channel is to be assigned to each transmitter such that channels for nearby transmitters (within, say, 100 miles) are distinct, and for very close transmitters (within, say, 50 miles) they differ by at least two. There is some spectral spreading of transmitters that decreases with distance between transmitters. Again, the goal is to construct a feasible labelling with minimum span.

For instance, Figure 1.1 shows a transmitter location in the plane with some other transmitters around it. The small circle is at 50 miles, while the large circle shows points at 100 miles from the center. A possible real-number channel assignment is shown, in which the central transmitter is assigned 0, two other very close transmitters have labels at least two, and the two nearby transmitters that are not very close have labels at least one. This labelling satisfies the distance-labelling conditions for every pair of vertices, not just pairs involving the central transmitter. In the example, the label 0 is repeated, but at distance more than 100 miles from the center.

Griggs quickly discovered that this problem of labelling a planar transmitter network is quite challenging. In order to develop some heuristics for the real problem, he decided to investigate the natural graph analogue of the distance-labelling problem above, in which the vertices are the transmitters and adjacent vertices correspond to transmitters that are very close [16].

Specifically, for a finite simple graph $G = (V, E)$, consider a labelling $f : V \to \mathbb{Z}$ such that for all vertices $v, w \in V$,

$$|f(v) - f(w)| \geq \begin{cases} 2, & \text{if} \quad d(v,w) = 1; \\ 1, & \text{if} \quad d(v,w) = 2, \end{cases}$$

where $d(v, w)$ is the distance between v and w in G (the minimum number of edges in any path from v to w). For such labellings f, called λ-labellings of G, we seek to determine the minimum span, denoted $\lambda(G)$.

For instance, the 4-cycle can be labelled by nonnegative integers by assigning 0 to one vertex and moving around the cycle, assigning each vertex the smallest label satisfying the conditions above. We end up using $0, 2, 4, 6$ at consecutive vertices, but this "greedy first-fit" labelling is *not* optimal. We find that $\lambda(C_4) = 4$ is achieved by

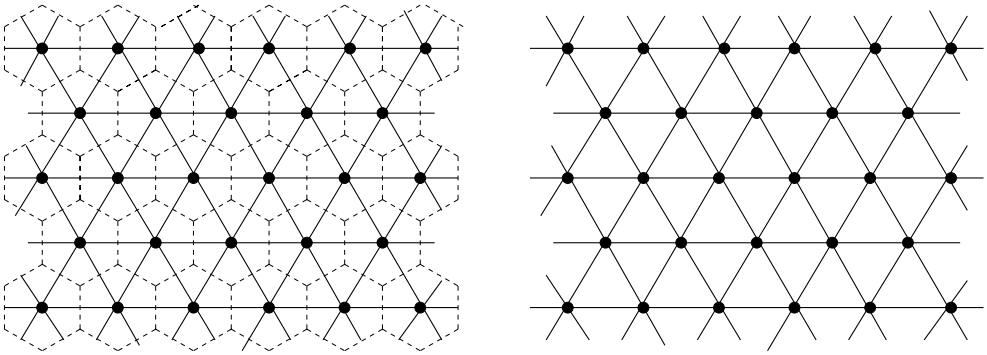


FIG. 1.2. *The honeycomb of hexagonal coverage regions of the plane (left); the corresponding triangular lattice* $\Gamma_\triangle$ *of transmitters (right).*

labelling successive vertices by $0, 4, 1, 3$.

By converting a planar network of transmitters—the problem of Roberts—to the graph problem, it is true that a pair of vertices at distance two in the graph corresponds to a pair of transmitters that are nearby, but not very close (at distance between 50 and 100 miles). However, a pair of transmitters in the plane can be close, but not very close, while their corresponding vertices in the graph are more than distance two apart. Their vertices need not even belong to the same component of the graph! For instance, in Figure 1.1 the transmitter labelled 1.1 is not very close to any other transmitter, so it would be isolated in the corresponding graph, although it is actually close to the transmitter at the center and to the one labelled 2.2.

Nonetheless, the study of $\lambda(G)$ for graphs G should lead to good bounds and heuristics for efficiently labelling planar networks. Also, for some natural arrays the problems are equivalent. Especially, one particular array often used in practice for mobile communication networks assigns a hexagonal coverage region to each transmitter, with the hexagons fitting together in a honeycomb tiling. This is efficient in the sense of using a small number of transmitters. The graph corresponding to this example is called the *triangular lattice*, $\Gamma_\triangle$ (see Figure 1.2). In this case the graph problem does properly represent the real problem in the plane. Besides, the λ-labelling problem has turned out to be quite interesting on its own as a generalized graph coloring problem.

The natural generalization of $\lambda(G)$ to deal with multiple levels of interference was introduced by Griggs in the original paper with Yeh [16]: Let $\mathbb{N}$ denote the set of natural numbers, $\{0, 1, 2, \ldots\}$ (note that 0 is included). Given integers $k_1, \ldots, k_p \in \mathbb{N}$, let $L(k_1, \ldots, k_p)$ denote the set of labellings $f : V \to \mathbb{Z}$ such that for all $v, w \in V$

$$|f(v) - f(w)| \geq k_i, \quad \text{if } d(v, w) = i \leq p.$$

We may abbreviate this by $L(\mathbf{k})$, where $\mathbf{k} = (k_1, k_2, \ldots, k_p)$. Note that $L(\mathbf{k})$ always depends on the graph G being considered. We seek the optimal span of any labelling f, denoted by

$$\lambda(G; k_1, \ldots, k_p) := \min_{f \in L(\mathbf{k})} \mathrm{sp}(f).$$

We also denote this lambda number of G by $\lambda(G; \mathbf{k})$. Alternate notation, also used by many authors, is $\lambda_{k_1,\ldots,k_p}(G)$. Another notation is in use for $p = 2$ [8], where

$\lambda_k^j(G)$ represents our $\lambda(G; j, k)$. However, we recommend against it, as it cannot be extended to model conditions at distances more than two. Also, we allow $k_1 < k_2$, and their notation does not make it easy to tell which of j and k refers to k_1 or k_2.

Translating a labelling f, by adding the same element to all labels $f(v)$, preserves the span. Hence, to determine $\lambda(G; \mathbf{k})$, it suffices to consider labellings with smallest label 0. For such labellings f, $\mathrm{sp}(f) = \max_{v \in V} f(v)$.

Ordinary graph coloring corresponds to the case $k_1 = 1$ and $k_i = 0$ otherwise. More precisely, the chromatic number $\chi(G)$ of a graph G is expressed in our theory by

$$\chi(G) = \lambda(G; 1) + 1,$$

where the difference of one arises due to our allowing 0 to be a label. While it is convenient that λ is one off from χ, allowing 0 to be a label gives us the nice Scaling Property, in which if all separations k_i are multiplied by the same constant c, then so is the optimal span λ. This principle is stated explicitly later when we formulate the theory of *real number* labellings.

An interesting special case, more general than above, is for $\mathbf{k}$ consisting of p ones. Here we have

$$\lambda(G; 1, \ldots, 1) = \chi(G^p) - 1,$$

where G^p is the graph that has the same vertex set V, and vertices v, w are joined by an edge whenever their distance in G is at most p.

Recent literature has expanded beyond the basic case of $L(2,1)$-labellings. Numerous papers consider $\lambda(G; k, 1)$ for arbitrary integers $k > 0$, or, more generally, $\lambda(G; p, q)$ for integer separations $p \geq q$.

Here is an overview of the rest of the paper. In the next section we introduce our model of labellings by real numbers. This more general setting is natural for many of the models in which labellings with distance conditions arise, since it would seem that the labels, e.g., frequencies, can actually be real numbers. Besides, we shall see that real number labellings offer greater insight into the behavior of the lambda number $\lambda(G; \mathbf{k})$ when it is viewed as a function of the separations $\mathbf{k} = (k_1, \ldots, k_p)$ with fixed graph G and separation distance p.

We also extend our model to allow *infinite* graphs G, such as the triangular lattice mentioned earlier. We will usually restrict our attention to infinite graphs with bounded degrees, to be assured of dealing with graphs of finite span. We denote by $\mathcal{G}_\Delta$ the class of simple graphs, possibly infinite, with maximum degree at most Δ. In particular, $\Gamma_\Delta \in \mathcal{G}_6$.

Section 3 presents (without proof) our formulas for real number labellings with conditions at distance two for the triangular lattice, paths, cycles, and the square lattice. Their behavior is instructive and serves to motivate the results and conjectures in the rest of the paper.

Our main general discovery about real number labellings, the D-Set Theorem, is presented in section 4. It applies to the class of graphs $\mathcal{G}_\Delta$. It shows that $\lambda(G; \mathbf{k})$ must be a sum of k_i's, repetitions allowed. Moreover, there is a labelling achieving $\lambda(G; \mathbf{k})$ in which every label has the form $\sum_i a_i k_i$, where the coefficients a_i are nonnegative integers, the smallest label is zero, and the largest label is the span. Further, if G is finite, we can restrict the possible labels to those sums with $\sum_i a_i < n$, where $n = |V|$.

Among the consequences of the D-Set Theorem is that when all k_i are integers, our problem reduces to the familiar integer lambda labelling problem of section 1:

There is always an integer labelling that is optimal for the real number labelling problem.

In section 5 we show that $\lambda(G;\mathbf{k})$ is continuous as a function of the separations $\mathbf{k}$, and we conjecture that it is piecewise linear with only finitely many linear pieces. This is verified for separations at distance two ($p = 2$) and for finite graphs. Section 6 poses a stronger conjecture that there is always a labelling as in the D-Set Theorem in which the coefficients are bounded above by a constant that depends only on p and the degree bound Δ. We can prove this for conditions out to distance two ($p = 2$).

Section 7 recalls the conjecture of [16] concerning the largest possible value of $\lambda(G;2,1)$ for graphs G of maximum degree $\Delta \geq 2$. We survey the progress on this longstanding conjecture. We present upper bounds on $\lambda(G;\mathbf{k})$ in terms of the separations k_i and the maximum degree.

We describe work that is closely related to our project in section 8. Besides the fascinating conjectures that remain open, the model of real number labellings has opened up interesting new lines of research, which we describe in the final section.

2. From integer to real number labellings. When the paper of Griggs and Yeh [16] introduced λ-labellings, it actually began with a special case of real number labellings, in which the transmitters were assigned real number labels and the separations were $\mathbf{k} = (2d, d)$, where d is some real number, not necessarily integer. It was shown that for any (finite) graph G, there is an optimal labelling in which all labels are multiples of d. But by a change of scale, dividing the separations and the labels by d, the problem was transformed into that of determining the optimal integer labelling span $\lambda(G;2,1)$. The same method applies to real labellings whenever the separations k_i are multiples of the same real number d.

It has always seemed overly restrictive that channel assignment models assumed the channels—the vertex labels—are integers (or, equivalently, all are multiples of the same number). This is certainly the case for familiar VHF television, with its integer channels ranging from 2 to 13, but for FM radio and possibly even for UHF television, the channels appear to have a continuous range of possibilities. The efficient allocation of bands of the radio frequency spectrum is a subject receiving considerable publicity.

Reviewing the progress on λ-labellings over the years, it is now sensible to consider the generalization in which the separations k_i and the labels $f(v)$ are arbitrary real numbers. As we would demand, this more general problem reduces to the familiar integer labellings when the separations and labels are integers. However, a rich variety of interesting new problems has been exposed by considering real number labellings, and their solution does not generally follow from their integer restrictions. Further, we have gained valuable new insights into the original integer λ-labellings by thinking in this more general context—see the D-Set Theorem.

This study also widens the class of graphs to include infinite graphs, in order to be able to deal properly with infinite arrays of transmitters. For instance, in cellular communications, a very large flat area is partitioned honeycomb-style into hexagonal cells, with a transmitter located in the center of each cell (its coverage area). This transmitter placement is most efficient (minimizes the number of transmitters). The channel assignment for the transmitter network is equivalent to the lambda labelling of the vertices of the dual graph, where each vertex corresponds to a transmitter. Extending the cellular network over the whole plane, the dual graph is a planar graph in which every vertex has six neighbors, which form a cycle around the vertex. The regions of the embedding of the dual graph are all triangles. This infinite 6-regular

graph is called the *triangular lattice*, which we denote by Γ_Δ. Because of its potentially practical implications, the lambda labelling of Γ_Δ is of particular interest to us.

Let $G = (V, E)$ be any graph, possibly infinite. A *real number labelling* of G is a function $f : V \to \mathbb{R}$, and its *span* is

$$\mathrm{sp}(f) := \sup_{v \in V} f(v) - \inf_{v \in V} f(v).$$

We consider *real* separations $k_1, \ldots, k_p \in [0, \infty)$. Define $L(G; \mathbf{k}) = L(G; k_1, \ldots, k_p)$ to be the set of real labellings $f : V \to \mathbb{R}$ such that

$$|f(v) - f(w)| \geq k_i \quad \text{if } d(v, w) = i \leq p.$$

We may write $L(\mathbf{k})$ or $L(k_1, \ldots, k_p)$ if G is understood.

We then define

$$\lambda(G; \mathbf{k}) := \inf_{f \in L(\mathbf{k})} \mathrm{sp}(f).$$

We will usually restrict our attention to the class $\mathcal{G}_\Delta$ of simple graphs with maximum degree at most Δ. We shall see that there are labellings with bounded span in $L(G; \mathbf{k})$ for $G \in \mathcal{G}_\Delta$, so $\lambda(G; \mathbf{k}) < \infty$ for all such G. Compactness arguments may be used to show that for $G \in \mathcal{G}_\Delta$, there is a labelling f that actually achieves $\lambda(G; \mathbf{k})$, meaning that its span is $\lambda(G; \mathbf{k})$, and there are vertices on which f assumes its minimum and maximum values, which are $\lambda(G; \mathbf{k})$ apart (recall this is true for infinite graphs). However, we will derive this information by a simpler approach in the next section. Our approach will also provide information about possible values of $\lambda(G; \mathbf{k})$.

As a simple introduction to real labellings, let us determine $\lambda(P_3; k, 1)$, where P_3 is the path on three vertices and k is any real ≥ 0. One valid labelling is $(k, 0, k + 1)$, and there is no way to improve on it among labellings where the middle label is the smallest of the three. If we put the smallest label on an end, and the largest is in the middle, we can do no better than $(0, k + 1, 1)$. If the smallest and largest labels are at the ends, we can either use $(0, k, 2k)$, provided that $k \geq 1/2$, or $(0, k, 1)$, when $k \leq 1/2$.

It follows that

$$\lambda(P_3; k, 1) = \begin{cases} 1, & \text{if} \quad 0 \leq k \leq 1/2; \\ 2k, & \text{if} \quad 1/2 \leq k \leq 1; \\ k + 1, & \text{if} \quad 1 \leq k. \end{cases}$$

Notice that we are allowing $k_1 = k$ to be less than $k_2 = 1$, which seems strange at first. Indeed, in models of interference between nearby transmitters, one expects the interference to decrease with distance, so that the required separations k_i would be nonincreasing as i grows. However, some recent papers have considered situations where the k_i may increase with i, and our model allows arbitrary $k_i \geq 0$ (which is mathematically interesting regardless of its usefulness).

We mention how it can arise in practice that $k_1 < k_2$. Jin and Yeh [20] cite a packet communication model of Bertossi and Bonuccelli [2] that considers such a case. Message packets are being sent throughout a wireless network of computer stations (computers and transceivers). The computer stations are the vertices and wireless connections between them are the edges if they can hear each other due to their proximity. Using the Code Division Multiple Access protocol (CDMA), each computer

station is assigned a control code, and packets are sent along the edges simultaneously, using the control codes of the computer stations sending them. These codes correspond to channels in our model. A problem arises whenever a computer station receives packets simultaneously from two different adjacent computer stations that cannot hear each other and use the same control code, and the receiving computer station has to ask for the packets to be resent. Avoiding this interference then requires that no two computer stations at distance two in the network use the same control code. Minimizing the number of different control codes used is then the $L(0,1)$ problem for the corresponding graph.

Of course, this is really just a standard graph coloring problem in disguise, but for a different graph: Given $G = (V, E)$, we form the graph $G' = (V, E') = (G^2 - G)$, in which $E' = E(G^2) - E(G)$ contains pairs of vertices that are at distance two in G. Then $\lambda(G; 0, 1) = \chi(G') - 1$. It would be interesting to find other situations that require lambda labellings $L(k_1, \ldots, k_p)$ in which $k_i < k_j$ for some $i < j$.

Returning to the path P_3, we have already given the values $\lambda(P_3; k, 1)$. Let us note also that $\lambda(P_3; 1, 0) = \chi(P_3) - 1 = 1$. We can now obtain all lambda numbers for the path P_3 with conditions at distance two from the following principle.

PROPERTY 2.1 (Scaling Property). *For all reals $k_1, \ldots, k_p, c \geq 0$ and all graphs G,*

$$\lambda(G; ck_1, \ldots, ck_p) = c\lambda(G; k_1, \ldots, k_p).$$

This property is an immediate consequence of the definition of λ, since if any labelling of G with separation conditions $\mathbf{k}$ has its labels each multiplied by c, it gives a labelling with separation conditions $c\mathbf{k}$, and vice versa.

For conditions at distance two, the Scaling Property gives us that

$$\lambda(G; k_1, k_2) = k_2\lambda(G; k, 1)$$

for $k = k_1/k_2, k_2 > 0$. So we can derive all values of λ with $k_2 > 0$ from the one-parameter values $\lambda(G; k, 1)$, $k \geq 0$, such as we gave above for $G = P_3$. We can also obtain the values $\lambda(G; k_1, 0)$, which is given by $k_1 \lim_{k\to\infty}(\lambda(G; k, 1)/k)$.

We next give a simple general upper bound on $\lambda(G; \mathbf{k})$ in terms of the maximum degree and the separations k_i for use in the proofs. Better bounds are given later in section 8.

LEMMA 2.2. *Let G be a graph, possibly infinite, of maximum degree at most Δ. Let $p \in \mathbb{Z}^+$, $\mathbf{k} = (k_1, \ldots, k_p)$, and $k = \max_i\{k_i\}$. Then $\lambda(G; \mathbf{k}) \leq k\Delta^p$.*

Proof. Let $G \in \mathcal{G}_\Delta$. For such p and $\mathbf{k}$ we have that

$$\lambda(G; \mathbf{k}) \leq \lambda(G; k, \ldots, k) = k\lambda(G; 1, \ldots, 1) = k(\chi(G^p) - 1),$$

which is, in turn, at most k times the maximum degree of graph G^p. Since G has at most $\Delta(\Delta - 1)^{i-1}$ vertices at distance i from any given vertex, we get that the maximum degree of G^p is at most

$$\Delta \sum_{i=1}^{p} (\Delta - 1)^{i-1} \leq \Delta^p. \qquad \square$$

Note that a labelling that satisfies the bound of Lemma 2.2 can be obtained by first arbitrarily ordering the vertices in some component of G^p. One can then greedily color the vertices in the component one-by-one in order by nonnegative integers,

always selecting the least color not already assigned to any neighboring vertex. Do this for each component. (This is a so-called *first-fit labelling*.) Then multiply all the labels by k to obtain a suitable labelling in $L(G;\mathbf{k})$.

3. Optimal spans with conditions at distance two for special graphs. In order to motivate the D-Set Theorem and other general results in later sections concerning the general behavior of $\lambda(G;\mathbf{k})$ viewed as a function of $\mathbf{k}$, we present some of our findings for particular graphs. Some of the results in this section were obtained, in part, by using the D-Set Theorem. But many of them were obtained independently of it, before the discovery of the D-Set Theorem, and played a role in its discovery. We postpone the long and intricate proofs with many cases to later papers [15, 14].

We consider only conditions at distance at most two. As noted above, it suffices to determine $\lambda(G;k,1)$. There are various results in the literature concerning $\lambda(G;k,1)$ when k is a positive integer. In the 2000 Mathematical Competition in Modeling (MCM), a problem of this kind (written by Griggs for the contest) was presented. The problem can be found in the special journal issue for the contest that includes a survey article by Griggs [12], or on the Web at www.comap.com. The problem was selected and reported on by 271 teams, each consisting of three undergraduates, from universities worldwide. Each team had a long weekend (less than four days) to research the problem, write and run programs, and put together a paper. They had access to libraries, computers, and the Web, but no human assistance was permitted.

Teams were asked in this problem to investigate distance labellings of the triangular lattice graph Γ_Δ. In some cellular communication networks [22] a large planar region is partitioned into hexagonal cells with a transmitter at the center of each cell. This method gives efficient coverage (minimizes the number of transmitters needed). Strong interference occurs between transmitters in adjacent cells, while lighter interference occurs between transmitters in cells with just one cell in between. We may form a graph, with a vertex for each cell and an edge between each two vertices that represent adjacent cells. In this case, we are fortunate in that the graph labelling problem with conditions at distance two is actually equivalent to the original transmitter problem in the plane. When the planar coverage region is the entire plane, the corresponding graph is an infinite 6-regular graph, the *triangular lattice*, which we denote by Γ_Δ. MCM teams were asked to determine $\lambda(G;2,1)$ for G corresponding to a certain large region and then for the entire plane (for $G=\Gamma_\Delta$). While experts in the subject already knew the (unpublished) answer, it was pleasing to see how many teams succeeded. MCM teams were asked to determine what they could about $\lambda(\Gamma_\Delta;k,1)$ for integers $k>1$. Several teams devised labellings that turned out to be optimal, though no team came up with a valid proof for general k; their lower bound proofs were not adequate. Condensed versions of the winning papers are collected in the special UMAP journal issue mentioned above [12].

A subsequent manuscript of Zhu and Shi [30] considers $\lambda(\Gamma_\Delta;k_1,k_2)$ for general integers $k_1\ge k_2\ge 1$. It provided more impetus to undertake the study contained in this paper. Note that by scaling, we shall find that to determine $\lambda(\Gamma_\Delta;k_1,k_2)$ for such integers is equivalent to determining $\lambda(\Gamma_\Delta;k,1)$ for rationals $k\ge 1$.

With considerable effort, the present authors have completely determined the values $\lambda(\Gamma_\Delta;k,1)$ for all reals $k\ge 1$. For reals $0\le k<1$, we have been chipping away, determining the exact value for small k and for some other values k, and bounds otherwise.

THEOREM 3.1 (see [14]). *For the triangular lattice Γ_Δ, we have the following values (or bounds, where it is not yet determined) for optimal spans of labellings with*

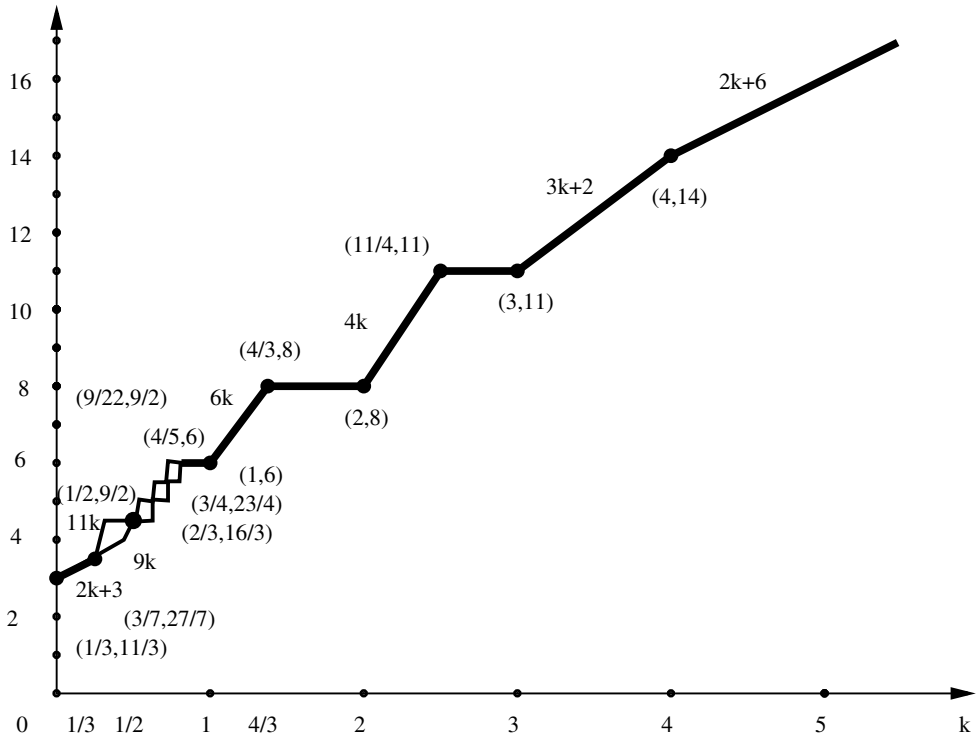


FIG. 3.1. *A graph of* $\lambda(\Gamma_{\triangle}; k, 1)$ *as a function of* k.

conditions at distance two:

$$\lambda(\Gamma_{\triangle}; k, 1) = \begin{cases} 2k+3 & \text{if} \quad 0 \le k \le 1/3, \\ \in [2k+3, 11k] & \text{if} \quad 1/3 \le k \le 9/22, \\ \in [2k+3, 9/2] & \text{if} \quad 9/22 \le k \le 3/7, \\ \in [9k, 9/2] & \text{if} \quad 3/7 \le k \le 1/2, \\ \in [9/2, 16/3] & \text{if} \quad 1/2 \le k \le 2/3, \\ \in [16/3, 23/4] & \text{if} \quad 2/3 \le k \le 3/4, \\ \in [23/4, 6] & \text{if} \quad 3/4 \le k \le 4/5, \\ 6 & \text{if} \quad 4/5 \le k \le 1, \\ 6k & \text{if} \quad 1 \le k \le 4/3, \\ 8 & \text{if} \quad 4/3 \le k \le 2, \\ 4k & \text{if} \quad 2 \le k \le 11/4, \\ 11 & \text{if} \quad 11/4 \le k \le 3, \\ 3k+2 & \text{if} \quad 3 \le k \le 4, \\ 2k+6 & \text{if} \quad k \ge 4. \end{cases}$$

A graph of $\lambda(\Gamma_{\Delta}; k, 1)$ as a function of k is presented in Figure 3.1. Coordinates are given for endpoints and isolated points that are known precisely. For $k \ge 1$ the graph is seen to be nondecreasing, continuous, and piecewise linear, and the same appears likely for $k \le 1$. Curiously, it is neither convex nor concave, nor is it even strictly increasing (at least three sections are flat). It will follow from Theorem 5.6

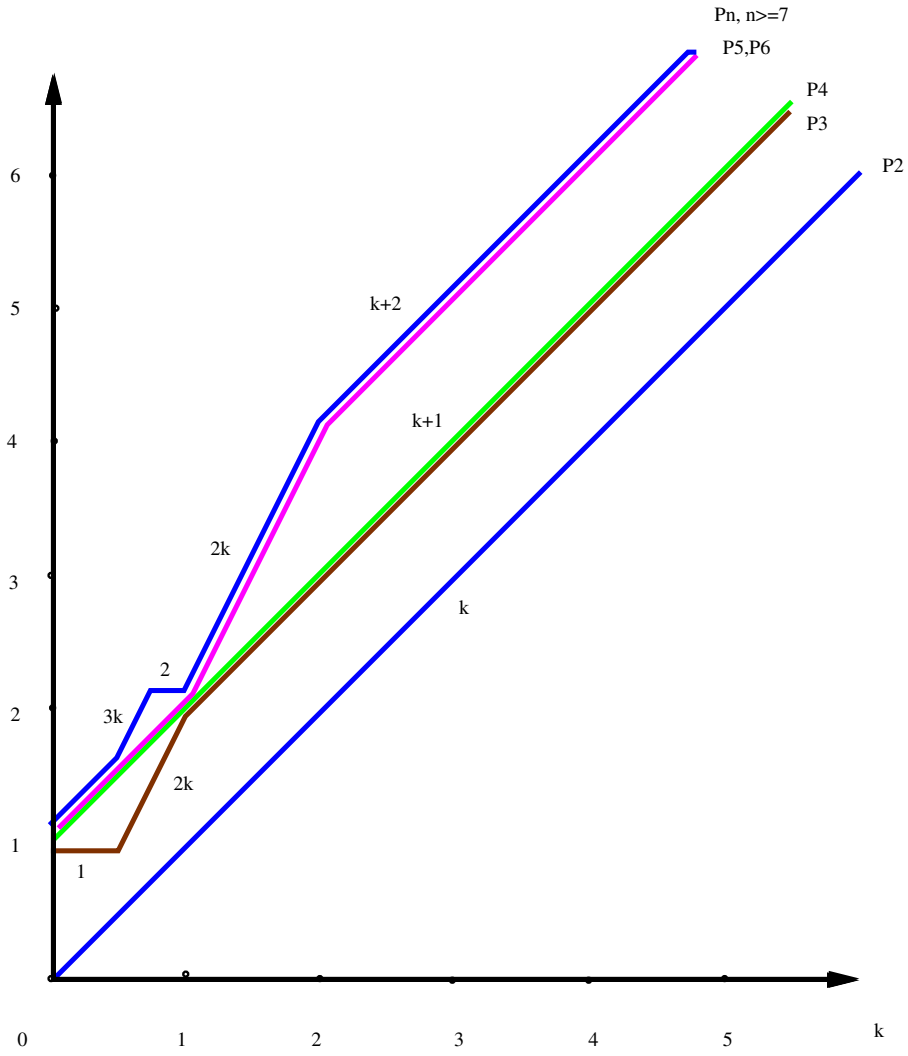


FIG. 3.2. *The functions* $\lambda(P_n; k, 1)$ *for the paths* P_n *on* n *vertices.*

that the graph is piecewise linear over its whole domain, $[0, \infty)$, even though we cannot yet give it on $[0, 1]$.

We have completely determined $\lambda(G; k, 1)$ for paths P_n and cycles C_n on n vertices.

THEOREM 3.2 (see [15], [19]). *For the paths* P_n*, the optimal span* $\lambda(P_n; k, 1)$ *with conditions at distance two, for* $k \geq 0$*, is shown in Figure* 3.2*. In particular, the optimal span is the same for all* k *for* $n \geq 7$*.*

THEOREM 3.3 (see [15], [19]). *For the cycles* C_n*, the optimal span* $\lambda(C_n; k, 1)$ *with conditions at distance two, for* $k \geq 0$*, is shown in Figure* 3.3 *for* $n = 3, 4, 5$ *and in Figure* 3.4 *for* $n \geq 6$*.*

Note. For $n \geq 6$, the choice of curve to follow in Figure 3.4 depends on the value of n modulo 12. For instance, in the interval $[2/3, 2]$, one follows the lower piece when n is 0 (mod 12) and the upper piece when n is 1 (mod 12).

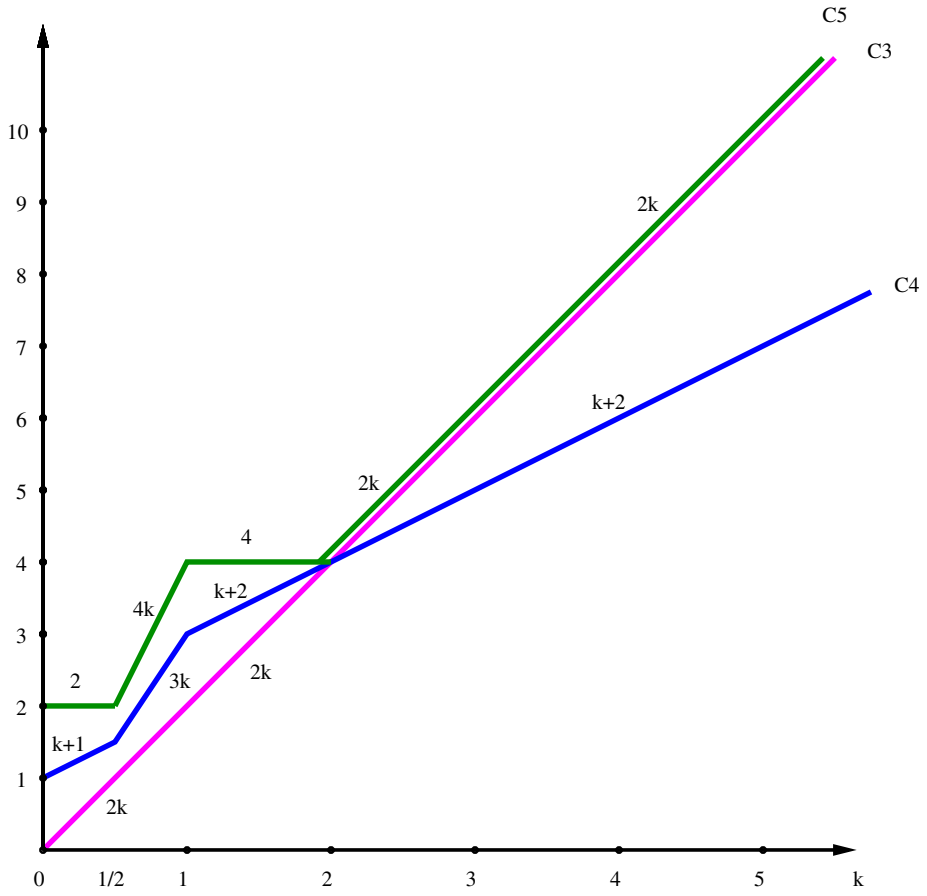


FIG. 3.3. *The functions* $\lambda(C_n; k, 1)$ *for the cycles* C_n, $n = 3, 4, 5$.

After obtaining these path and cycle lambda number formulas, we realized that they were already known in part: Georges and Mauro [8] determined them for integer separations $k_1 \geq k_2$. By the Scaling Property their formulas give $\lambda(G; k, 1)$ for rationals $k \geq 1$ when G is a path or cycle (see section 8 for related remarks).

Notice that for each path and cycle, the graph is again a continuous nondecreasing piecewise linear function. Also, the linear formulas for the straight sections of the graphs above are always of the form $ak + b$, where a and b are nonnegative integers. For the graph Γ_Δ, one of the winning teams in the modeling contest, from Washington University [11], claimed that this should be the case for the triangular lattice for all (integers) k. They turned out to be correct. Indeed, we shall see there is a piecewise linearity result, where the pieces are nonnegative integer linear functions of k_1 and k_2, for general graphs of bounded degree (Theorem 5.6).

We wish to state one more important example, the square lattice $\Gamma_\square$, which is used in some applications. Here, the vertices correspond to the integer lattice points in the plane, and edges join pairs of vertices that are equal in one coordinate and are consecutive in the other coordinate. It is possible to give the complete formula for labellings with conditions at distance two, and, as expected, it is piecewise linear with finitely many pieces. The graph of $\lambda(\Gamma_\square; k, 1)$ is shown in Figure 3.5.

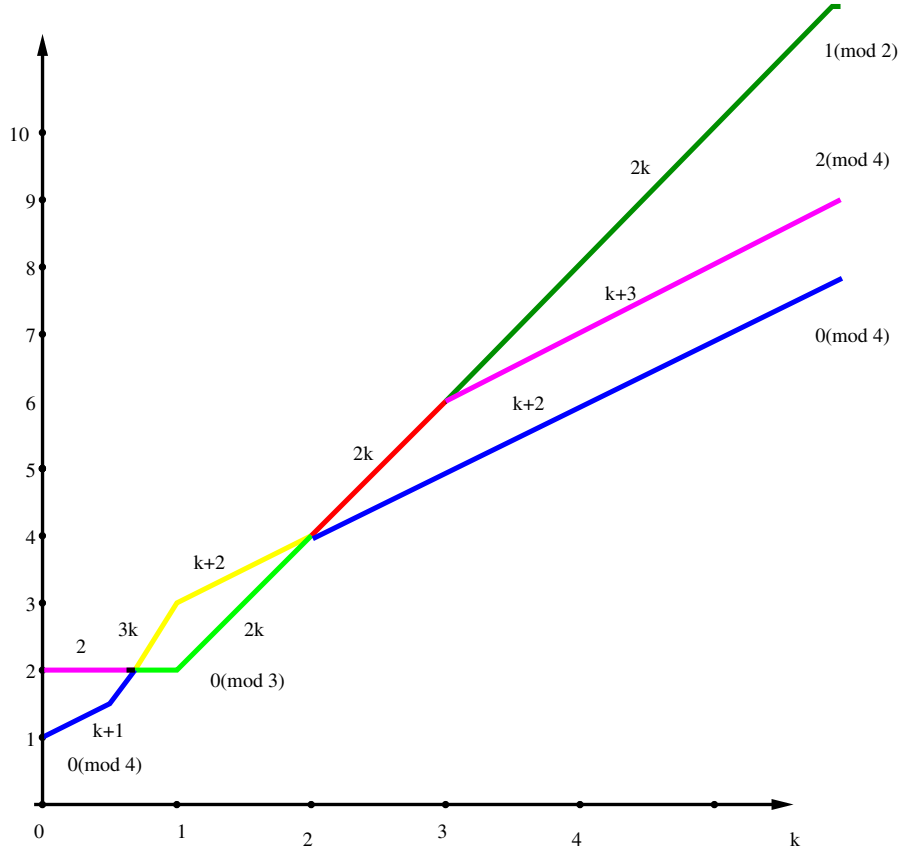


FIG. 3.4. *The functions* $\lambda(C_n; k, 1)$ *for the cycles* C_n, $n \geq 6$.

THEOREM 3.4 (see [14]). *For the square lattice* $\Gamma_\square$ *we have the following values for optimal spans of labellings with conditions at distance two:*

$$\lambda(\Gamma_\square; k, 1) = \begin{cases} k+3 & \text{if} \quad 0 \leq k \leq 1/2, \\ 7k & \text{if} \quad 1/2 \leq k \leq 4/7, \\ 4 & \text{if} \quad 4/7 \leq k \leq 1, \\ 4k & \text{if} \quad 1 \leq k \leq 4/3, \\ k+4 & \text{if} \quad 4/3 \leq k \leq 3/2, \\ 3k+1 & \text{if} \quad 3/2 \leq k \leq 5/3, \\ 6 & \text{if} \quad 5/3 \leq k \leq 2, \\ 3k & \text{if} \quad 2 \leq k \leq 8/3, \\ 8 & \text{if} \quad 8/3 \leq k \leq 3, \\ 2k+2 & \text{if} \quad 3 \leq k \leq 4, \\ k+6 & \text{if} \quad k \geq 4. \end{cases}$$

4. *D*-Set Theorem on real labellings. Motivated by partial results for paths and cycles (our complete solution came later) and by observations described in the previous section, we came to realize that for any finite graph G and any fixed separation vector $\mathbf{k} = (k_1, \ldots, k_p)$, $\lambda(G; \mathbf{k})$ must be a sum of the separations k_i (repeats allowed). Indeed, this holds as well for the infinite graph Γ_Δ (as claimed by the MCM

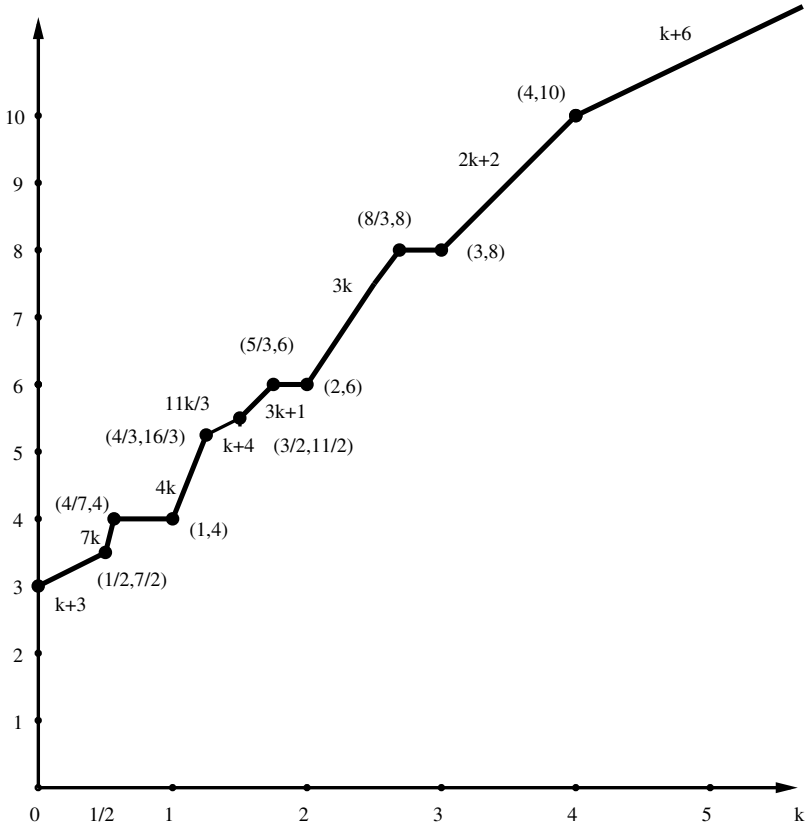


FIG. 3.5. *A graph of* $\lambda(\Gamma_{\square}; k, 1)$ *as a function of* k.

team from Washington University), and, in general, for any graph with bounded degrees.

Given $\mathbf{k}$, let us define the "D-set"

$$D(k_1, \ldots, k_p) := \left\{ \sum_{i=1}^{p} a_i k_i : a_i \in \mathbb{N}, 1 \le i \le p \right\},$$

abbreviated by $D(\mathbf{k})$. It turns out that it suffices to consider labellings with labels in $D(\mathbf{k})$ in order to prove the existence of optimal λ-labellings.

THEOREM 4.1 (D-Set Theorem). *Let $G = (V, E)$ be a graph, possibly infinite, of bounded maximum degree. Let $k_1, \ldots, k_p$ be real numbers ≥ 0. Then there is an optimal labelling $f \in L(\mathbf{k})$ with all labels $f(v) \in D(\mathbf{k})$ in which the smallest label is 0, the largest label is $\lambda(G; \mathbf{k})$. In particular, $\lambda(G; \mathbf{k}) \in D(\mathbf{k})$. Moreover, if G is finite, each label $f(v)$ and $\lambda(G; \mathbf{k})$ can be expressed in the form $\sum_i a_i k_i$, where the nonnegative integer coefficients a_i satisfy $\sum_i a_i < |V|$.*

Proof. Let $\mathbf{k} = (k_1, \ldots, k_p)$, and suppose G is a graph in $\mathcal{G}_\Delta$. Let $f \in L(\mathbf{k})$ be any labelling of span at most $\Delta^p k$, where $k = \max_i\{k_i\}$; such labellings exist by Lemma 2.2. By translating the labelling f, if necessary, we may assume that $\inf_v f(v) = 0$. We modify f to get another labelling f^*, with span no larger than for f, such that all labels for f^* belong to $D(\mathbf{k})$. Also, f^* will have smallest label 0.

Let us define the *D-floor* of a real number $x \geq 0$ by

$$\lfloor x \rfloor_D := \max\{y \in D(\mathbf{k}) : y \leq x\}.$$

Note that $D(\mathbf{k})$ contains only a finite number of elements $\leq x$, so this is a maximum, not a supremum. Then we define the new labelling f^* at each vertex v by

$$f^*(v) := \lfloor f(v) \rfloor_D.$$

Because $D(\mathbf{k})$ has only finitely many elements in $[0, \Delta^p k]$, and $\inf_v f(v) = 0$, it follows that f^* has minimum value 0. By design, all values $f^*(v)$ belong to $D(\mathbf{k})$. It suffices to show that $f^* \in L(\mathbf{k})$, which requires checking the separation conditions. Suppose that $u, v \in V$ are at distance $i \leq p$ in G. Without loss of generality, suppose $f(u) \geq f(v)$. Since $f \in L(\mathbf{k})$, we find that

$$f(u) \geq f(v) + k_i \geq \lfloor f(v) \rfloor_D + k_i = f^*(v) + k_i.$$

Since $f^*(v) + k_i \in D(\mathbf{k})$, it follows from the definition of the D-floor that $\lfloor f(u) \rfloor_D \geq f^*(v) + k_i$, so that

$$f^*(u) \geq f^*(v) + k_i,$$

and the separation conditions are satisfied.

We have shown that for every $f \in L(\mathbf{k})$, there exists $f^* \in L(\mathbf{k})$ with $\mathrm{sp}(f^*) \leq \mathrm{sp}(f)$ such that $\mathrm{sp}(f^*) \leq C$ and $\mathrm{sp}(f^*) \in D(\mathbf{k})$. Since $D(\mathbf{k}) \cap [0, C]$ is finite, it follows that λ exists, belongs to $D(\mathbf{k})$, and is at most C. Also, there exists an optimal labelling where the labels are in $D(\mathbf{k})$, with smallest label 0 and largest label $\lambda(G; \mathbf{k})$.

If G has $n < \infty$ vertices, let f be an optimal labelling of G as above. The minimum label $f(v)$ is 0; say it occurs at $v_1, \ldots, v_s$. Consider the smallest label > 0; say it occurs at v_{s+1}. We may decrease $f(v_{s+1})$ to 0 without any problem, and repeat this process, unless there is some $t < s+1$ such that v_t and v_{s+1} are at some distance d with $k_d > 0$. But we may at least still decrease $f(v_{s+1})$ until it is some value $k_i \geq k_d > 0$, where some vertex v_r is at distance i from v_{s+1}, with $r < s+1$.

Then we decrease the next smallest label until it is a sum of at most two k_i's, not necessarily distinct, and so on, until all labels are sums of fewer than n k_i's. In doing this, labels only get smaller or remain unchanged. Thus, in the end we still have an optimal labelling, and its span is a sum of fewer than n k_i's, not necessarily distinct. □

We now confirm that real number labellings include integer labellings as a special case:

COROLLARY 4.2. *Let $G = (V, E)$ be a graph, possibly infinite, of bounded maximum degree. Let $k_1, \ldots, k_p$ be integers ≥ 0. Then $\lambda(G; \mathbf{k})$ is an integer, and there is an optimal labelling with smallest label* 0 *and every label integer.*

Example 4.3. Suppose G is a graph with bounded maximum degree, and suppose $\mathbf{k} = (5, 3)$. Then there is an optimal labelling with smallest label 0 and all labels, and $\lambda(G; 5, 3)$, belonging to $D(5, 3) = \{0, 3, 5, 6, 8, 9, 10, \ldots\}$. In particular, to search for $\lambda(G; 5, 3)$, it suffices to try nonnegative integer labellings, with smallest label 0, that do not use 1, 2, 4, or 7. This could speed up computing $\lambda(G; 5, 3)$.

The D-Set Theorem is particularly useful for proving lower bounds on $\lambda(G; \mathbf{k})$. (Explicit labellings are used to prove upper bounds.) Here are two proofs of (sharp) lower bounds for paths that rely on the D-Set Theorem.

Example 4.4. For $1/2 \le k \le 1$, we have $\lambda(P_3; k, 1) \ge 2k$.

Proof. Suppose not; say $\lambda(P_3; k, 1) < 2k$. This forces $k > 1/2$. By the D-Set Theorem, there is an optimal labelling f using labels only in $[0, 2k) \cap D(k, 1) = \{0, k, 1\}$. For P_3, the labels must be distinct by the separation conditions. But even that is impossible, since label k is too close to 1: their difference, $1 - k$ is less than k and 1. So no such f exists. □

Example 4.5. For $1/2 \le k \le 2/3$ (resp., $2/3 \le k \le 1$), $\lambda(P_7; k, 1) \ge 3k$ (resp., ≥ 2).

Proof. For this range of k values we have that the smallest elements of $D(k, 1)$ are, in order, $0, k, 1, 2k, k + 1$, followed by $3k$ (resp., 2), if $1/2 \le k \le 2/3$ (resp., $2/3 \le k \le 1$). By the D-Set Theorem, it suffices to show that $\lambda(P_7; k, 1) > k + 1$ for $1/2 < k \le 1$. Suppose not; say f is a $L(k, 1)$-labelling of P_7 using labels in the set $\{0, k, 1, 2k, k + 1\}$.

Vertices not labelled with 0 have labels in $[k, k + 1]$. Since P_3 has minimum span $2k > 1$ by Example 4.4, f cannot assign nonzero labels to three consecutive vertices. But f cannot assign 0 to two vertices at distance two or less. Hence, 0 labels must be used at some vertices at distance a multiple of three. If three labels are zeros, they are at the middle vertex and both endpoints of P_7; then all other labels are at least one, and one of them is at least two, which is larger than the span, a contradiction. If two labels are zeros, neither of them is at the endpoints of P_7. The two nonzero labels in between are both at least one, and at least k apart from each other, so they must be 1 and $k + 1$. Then the nonzero label at distance two from label 1 is at least 2, which is larger than the span, a contradiction. □

5. Piecewise linearity. For a fixed graph G and a fixed integer p, we wish to understand the behavior of the optimal span λ as a function of the minimum separations $k_1, \ldots, k_p$ in the set

$$T^p = \{(k_1, \ldots, k_p) \in \mathbb{R}^p : k_i \ge 0 \ \forall i\}.$$

In particular, we want to see why it is piecewise linear in all examples we have studied.

We first obtain the continuity of λ on T^p, which ends up being unexpectedly tricky to prove near the boundary.

THEOREM 5.1. *Let $G = (V, E)$ be a graph, possibly infinite, of bounded maximum degree. Let $p \in \mathbb{Z}^+$. Then $\lambda(G; \mathbf{k})$ is continuous and nondecreasing as a function of $\mathbf{k}$ on T^p.*

Proof. Let $G \in \mathcal{G}_\Delta$. For any $\mathbf{k} = (k_1, \ldots, k_p)$, $\mathbf{k}' = (k'_1, \ldots, k'_p) \in T_p$, we write $\mathbf{k} \le \mathbf{k}'$ if $k_i \le k'_i$ for all i. The function λ is nondecreasing, since if $\mathbf{k} \le \mathbf{k}'$, the separations $\mathbf{k}'$ are more restrictive than $\mathbf{k}$, so that $L(\mathbf{k}') \subseteq L(\mathbf{k})$, and thus $\lambda(G; \mathbf{k}) \le \lambda(G; \mathbf{k}')$.

We show that $\lambda(G; \mathbf{k})$ is continuous at an arbitrary $\mathbf{k} \in T^p$. Let I be the set of indices i where $k_i > 0$. Let $\mathbf{k}^*$ be any element of T_p that is distance at most $\varepsilon > 0$ from $\mathbf{k}$. We need to show that $|\lambda(G; \mathbf{k}^*) - \lambda(G; \mathbf{k})|$ can be made arbitrarily small by selecting ε small enough. Assume $\varepsilon < (\min_{i \in I} k_i)/2$. Define vectors $\mathbf{k}', \mathbf{k}'' \in T^p$, where $\mathbf{k}'' \le \mathbf{k} \le \mathbf{k}'$, as follows. Let $k'_i = k_i + \varepsilon$ for all i, while $k''_i = k_i - \varepsilon$ for $i \in I$ and $k''_i = k_i = 0$ otherwise. By design, $\mathbf{k}'' \le \mathbf{k}^* \le \mathbf{k}'$. As λ is nondecreasing, we have that

$$|\lambda(G; \mathbf{k}^*) - \lambda(G; \mathbf{k})| \le \lambda(G; \mathbf{k}') - \lambda(G; \mathbf{k}''),$$

and it suffices to show that $\lambda(G; \mathbf{k}') - \lambda(G; \mathbf{k}'')$ can be made arbitrarily small as $\varepsilon \to 0$.

Let f'' be an optimal labelling as in the D-Set Theorem achieving $\lambda(G;\mathbf{k}'')$. We will modify f'' to obtain a labelling $f' \in L(\mathbf{k}')$ with span only slightly larger. Specifically, since $\lambda(G;\mathbf{k}') - \lambda(G;\mathbf{k}'') \le \mathrm{sp}(f') - \mathrm{sp}(f'')$, it will suffice that $\mathrm{sp}(f') - \mathrm{sp}(f'') \to 0$ as $\varepsilon \to 0$. Since $0 \le k_i' - k_i'' \le 2\varepsilon$ for all i, f' will be feasible for separations $\mathbf{k}'$ if, for each pair of vertices at distance at most p in G, f' increases the separation between their labels by at least 2ε.

By the D-Set Theorem, all labels used by f'' belong to the set $D(\mathbf{k}'') \cap [0, \lambda(G;\mathbf{k}'')]$. Let us denote these labels by $0 = r_1 < r_2 < \cdots r_A$. Each of these labels r_j has the form $\sum_{i\in I} a_i k_i''$, where the coefficients a_i are integers ≥ 0. Now we have

$$a_i(k_i/2) \le a_i k_i'' \le \lambda(G;\mathbf{k}'') \le \lambda(G;\mathbf{k}) \le \Delta^p k,$$

where we use Lemma 2.2. Hence, for all $i \in I$, $a_i \le 2C/k_i$, so that a_i is bounded in terms of $\mathbf{k}$ and Δ. It follows that the number of labels used by f'', A, is bounded in terms of $\mathbf{k}$ and Δ.

Now we modify the labels r_i two different ways. Let δ be a small number, depending on ε. First, increase each label r_i by $(i-1)\delta$, which increases the separation between each pair of distinct labels by at least δ. Secondly, take an optimal vertex coloring g of graph G^p using colors that are integers in the interval $[0, B]$, where $B = \Delta^p$ (see proof of Lemma 2.2). Then increase the labels $f''(v)$ again, this time by $2g(v)\varepsilon$.

The labelling obtained after the two augmentations is what we call f', and it depends on both ε and δ. Consider any pair of vertices $v, w \in V$ which are at some distance $i \le p$ in G. If $f''(v) = f''(w)$ (which can happen only if $k_i = 0$), only the second augmentation changes their difference, and we get that

$$|f'(v) - f'(w)| = |2(g(v) - g(w))\varepsilon| \ge 2\varepsilon = |f''(v) - f''(w)| + 2\varepsilon,$$

which is what we claimed. On the other hand, suppose that $f''(v) \ne f''(w)$, say, $f''(v) < f''(w)$. The first operation must moves their labels at least δ farther apart, while the second operation may move them closer together, but by at most $2B\varepsilon$. Let us then specify that $\delta = (2B+2)\varepsilon$, so that in f' the separation between labels for such v, w increases by at least $\delta - 2B\varepsilon = 2\varepsilon$ over f''. We have that $f' \in L(\mathbf{k}')$.

Now we compare the span of f' with that of f''. The smallest label in f' is at least 0, while the largest label may increase over that in f'' due to the two operations, by at most $A\delta = A(2B+2)\varepsilon$ from the first operation, and by at most another $2B\varepsilon$ from the second operation. Thus, $\mathrm{sp}(f') - \mathrm{sp}(f'')$ is at most a constant times ε, the constant depending only on $\mathbf{k}$ and Δ, and it goes to 0 with ε. □

Next we consider the piecewise linearity of $\lambda(G;\mathbf{k})$. We say that a function f defined on domain $A \subseteq T^p$ is PL on A if it is piecewise linear on A with only finitely many pieces. More specifically, we mean that A can be split by finitely many hyperplanes such that, on each of the closed (polyhedral) regions, f is linear. Further, f is continuous, that is, the linear formulas for two adjacent regions agree on the boundary between them.

We begin with the piecewise linearity for finite graphs, and then consider infinite graphs with bounded degrees.

THEOREM 5.2. *Let $G = (V, E)$ be a finite graph. Let $p \in \mathbb{Z}^+$. Then $\lambda(G;\mathbf{k})$ is PL as a function of $\mathbf{k}$ on T^p. Specifically, the domain T^p can be split by finitely many hyperplanes through the origin into closed convex polyhedral cones, such that $\lambda(G;\mathbf{k})$ is given by a linear function of the k_i's on each cone.*

Proof. Let $G = (V, E)$ be a finite graph on n vertices. Let us partition T^p into polyhedral cones by taking all hyperplanes with equations of the form

$$\sum_{i=1}^{p} b_i k_i = 0, \text{ where } \sum_i |b_i| < 2n.$$

By the D-Set Theorem, for any point $\mathbf{k} \in T^p$, $\lambda(G; \mathbf{k})$ is the minimum, over feasible labellings $f : V \to D(\mathbf{k})$, of the maximum label $f(v), v \in V$. Further, it suffices to consider such f in which for all vertices v, $f(v)$ has the form $\sum_{i=1}^{p} a_i(v)k_i$, where the nonnegative integer coefficients $a_i(v)$ satisfy $\sum_{i=1}^{p} a_i(v) < n$. (Note that for every $\mathbf{k} \in T^p$, there is some feasible labelling, hence some feasible labelling of this form.)

Now we turn things around and fix a labelling f of this form and consider the feasible region for f, meaning the set of values $\mathbf{k} \in T^p$ for which $f \in L(G; \mathbf{k})$. We claim that it is a union of convex cones with vertex at the origin.

To see this, note that $\mathbf{k} \in T^p$ is feasible for such f whenever it is feasible for each pair of vertices u, v at distance between 1 and p. If u and v are at distance d, say, this means that $f(u) - f(v)$ is either $\geq k_d$ or $\leq -k_d$. For the pair u, v the two constraints are bounded by the hyperplanes through the origin with equations $\sum_i (a_i(u) - a_i(v))k_i = k_d$ or $= -k_d$, which both have the stated form. Then the feasible region for f is the intersection, over such pairs u, v, of these sets in T^p, each a union (possibly empty) of at most two closed half-spaces. So it is a union of polyhedral cones of the form stated in the theorem.

Within the feasible region of f, the maximum label $\max_{v \in V} \sum_{i=1}^{p} a_i(v)k_i$ depends on $\mathbf{k}$. Chopping the feasible region by all possible comparisons between values of f, we get that the feasible region is refined into a union of closed polyhedral cones, bounded by the hyperplanes above plus the additional hyperplanes $f(u) = f(v)$, over distinct $u, v \in V$, that is, $\sum_i (a_i(u) - a_i(v))k_i = 0$, which also has the stated form.

By taking all possible hyperplanes of the stated form T^p is divided into polyhedral cones through the origin such that in each such cone (cell) K some nonempty collection of our labellings f is feasible on all of K, and for each such feasible f, the maximum label $f(v)$ is achieved at a single vertex v (by a single linear formula in $\mathbf{k}$). Similarly, the minimum of these maximum labels, over all feasible f on K, will be given by a single linear formula (some label $f(v)$) throughout K.

To summarize, cutting T^p by all of the finitely many hyperplanes described above divides it into a finite number of convex polyhedral cones such that $\lambda(G; \mathbf{k})$ is given by a linear formula of $\mathbf{k}$ in each (closed) cone, and we see $\lambda(G; \mathbf{k})$ is PL. □

We remark that since the formulas for adjacent cells K and K' agree on their boundaries, the continuity of $\lambda(G; \mathbf{k})$ follows for finite graphs G. The strength of Theorem 5.1 is evidently that continuity holds as well for infinite graphs $G \in \mathcal{G}_\Delta$.

Now consider infinite graphs with bounded maximum degree, say, $G \in \mathcal{G}_\Delta$. The same arguments above extend, but now the number of hyperplanes cutting through the origin is infinite, so the convex cones in the feasible region are not necessarily polyhedral. Nonetheless, we conjecture that Theorem 5.2 can be extended to such infinite graphs.

Let $G = (V, E)$ be a graph, possibly infinite, of bounded maximum degree. Let $p \in \mathbb{Z}^+$. Then $\lambda(G; \mathbf{k})$ is continuous and nondecreasing as a function of $\mathbf{k}$ on T^p.

CONJECTURE 5.3 (PL Conjecture). *Let $G = (V, E)$ be a graph, possibly infinite, of bounded maximum degree. Let $p \in \mathbb{Z}^+$. Then $\lambda(G; \mathbf{k})$ is PL as a function of $\mathbf{k}$ on T^p. Specifically, the domain T^p can be split by finitely many hyperplanes through*

the origin into closed convex polyhedral cones such that $\lambda(G;\mathbf{k})$ is given by a linear function of the k_i's on each cone.

Despite considerable effort we have not yet succeeded in proving this conjecture. We can give weaker, though still quite strong, results in support of it. One strategy is to restrict the domain by staying away from the coordinate planes (avoiding very small values of the separations k_i). For a number $\varepsilon > 0$, let us consider the region $T^p(\varepsilon)$ of all points $\mathbf{k}$ with all $k_i \geq \varepsilon(\sum_i k_i)$. Consider any point $\mathbf{k} \in T^p(\varepsilon)$. By Lemma 2.2, $\lambda(G;\mathbf{k}) \leq \Delta^p \sum_i k_i$. By the D-Set Theorem there is an optimal labelling in which each label has the form $\sum a_i k_i$, so that, for all i, $a_i k_i \leq \Delta^p \sum_i k_i$, from which our assumption on $\mathbf{k}$ implies each coefficient is at most a constant, Δ^p/ε. We can then proceed as for Theorem 5.2 and derive the PL property.

THEOREM 5.4. *Let $G = (V, E)$ be a graph, possibly infinite, of bounded maximum degree. Let $p \in \mathbb{Z}^+$. Then for any $\varepsilon > 0$ the function $h(\mathbf{k}) = \lambda(G;\mathbf{k})$ is PL on $T^p(\varepsilon)$.*

Our other supporting result for the PL Conjecture 5.3 is to prove it for conditions out to distance two, that is, for $p = 2$. This explains why we obtained PL graphs for $\lambda(G; k, 1)$ for the graphs we considered. It depends on a special sort of argument that we have been unable to extend to larger p. We can derive the PL Theorem for $p = 2$ (Theorem 5.6) by a different argument in section 8. We first require some simple bounds on $\lambda(G; k_1, k_2)$ depending on the chromatic number. Note that the upper bound here may be either better or worse than in Lemma 2.2, depending on k_1, k_2.

LEMMA 5.5. *Let $G = (V, E)$ be a graph, possibly infinite, of maximum degree at most $\Delta > 0$. Then*

$$(\chi - 1)k_1 \leq \lambda(G; k_1, k_2) \leq (\chi - 1)k_1 + \chi\Delta^2 k_2,$$

where χ is the chromatic number of G. Also,

$$(\chi(G^2 - G) - 1)k_2 \leq \lambda(G; k_1, k_2) \leq \chi(G^2 - G)\Delta k_1 + (\chi(G^2 - G) - 1)k_2.$$

Proof. We prove the first display; the proof of the second is similar. The lower bound follows easily from

$$\lambda(G; k_1, k_2) \geq \lambda(G; k_1, 0) = \lambda(G; k_1) = k_1\lambda(G; 1) = k_1(\chi - 1).$$

For the upper bound, let us employ two labellings of G. First, take any optimal coloring f_1 of G, where the colors are integer labels in $[0, \chi - 1]$. (Note that if we instead use a greedy first-fit coloring of G, we might not get an optimal coloring; the labels would be in $[0, \Delta]$.) Second, take labelling f_2 to be a greedy $L(0, 1)$-labelling, as in the proof of Lemma 2.2, so that the labels in f_2 are integers in the interval $[0, \Delta^2]$. Then we define the labelling

$$f = (k_1 + \Delta^2 k_2)f_1 + k_2 f_2.$$

By design, $f \in L(k_1, k_2)$, and its span is at most

$$(k_1 + \Delta^2 k_2)(\chi - 1) + k_2\Delta^2 = (\chi - 1)k_1 + \chi\Delta^2 k_2. \qquad \square$$

THEOREM 5.6. *Let $G = (V, E)$ be a graph, possibly infinite, of bounded maximum degree. Then $\lambda(G; k_1, k_2)$ is PL as a function of (k_1, k_2) on T^2. Specifically, the domain T^2 can be partitioned by finitely many lines through the origin into closed*

convex polyhedral cones such that $\lambda(G; k_1, k_2)$ is given by a linear function of the k_i's on each cone.

Proof. Let $G \in \mathcal{G}_\Delta$. By the Scaling Property, $\lambda(G; k_1, k_2) = k_2\lambda(G; k, 1)$ for $k_2 > 0$ by setting $k = k_1/k_2$, so that it suffices to show that $\lambda(G; k, 1)$, denote this by $g(k)$, is piecewise linear as a function of k. The proof of Theorem 5.2 gives us the piecewise linearity we want, except there may be infinitely many linear pieces when G is infinite. It is enough to prove that $g(k)$ is eventually linear for sufficiently large k and also for sufficiently small $k > 0$. Theorem 5.4 guarantees that it is piecewise linear with integer coefficients, with only finitely many pieces, between the linear pieces at the ends.

Let us deal with the case of large k. Consider some $k_o > \chi\Delta^2$. As in the proof of Theorem 5.2, $g(k_o)$ is on a linear segment of the graph of g with a formula of the form $\alpha k + \beta$, where $\alpha, \beta \in \mathbb{N}$. Moreover, the upper bound in Lemma 5.5 gives us that

$$\alpha k_o + \beta \leq (\chi - 1)k_o + \chi\Delta^2,$$

which forces

$$(\alpha - (\chi - 1))k_o \leq \chi\Delta^2.$$

But since $k_o > \chi\Delta^2$, it must be that $\alpha \leq \chi - 1$, as α is integral.

There must be a largest integer coefficient α over the values $k > \chi\Delta^2$; say it is α_o at k_o so that $g(k_o) = \alpha_o k_o + \beta_o$. None of the linear formulas for k larger than k_o can be of the form $\alpha_o k + \beta$ with $\beta > \beta_o$, because there is no way to get above the linear function $\alpha_o k + \beta$ without some piece having a slope $\alpha > \alpha_o$, contradicting the maximality of α_o, using the fact that $g(k)$ is continuous. Hence, for all $k \geq k_o$, $g(k) \leq \alpha_o k + \beta_o$. Then by the lower bound of Lemma 5.5, $(\chi - 1)k \leq \alpha_o k + \beta_o$ for all large k. It follows that $\alpha_o = \chi - 1$.

Linear pieces of the graph of $g(k)$ for $k > k_o$ with slope α_o must have decreasing values of β, so that each subsequent linear piece with formula $\alpha_o k + \beta$ will have a lower value of $\beta \in \mathbb{N}$. There then must be a last linear piece with $\alpha = \alpha_o$. Then this piece never ends, because if it did, then after that the slopes would all be at most $\chi - 2$, and eventually the graph of g would drop by the lower bound of Lemma 5.5.

It remains to go the other way and show that, for sufficiently small k, $g(k)$ is linear. This is equivalent to showing that $\lambda(G; 1, k)$ is eventually linear as k grows. The same method used before now works, except the roles of k_1 and k_2 are reversed. We conclude that $g(k)$ is eventually linear as $k \to 0$, with a formula of the form $\alpha k + \chi(G^2 - G)$, and it is piecewise linear overall with only finitely many pieces. ∎

6. Bounds on the coefficients. We made a special effort to determine whether the piecewise linearity theorem (Theorem 5.2) holds more generally than for finite graphs. Does it hold for infinite graphs of bounded degree, that is, for the class $\mathcal{G}_\Delta$? We verified our PL Conjecture 5.3 in cases where we could bound the coefficients a_i independent of $\mathbf{k}$. The full PL Conjecture would follow (by the arguments used to prove Theorem 5.2) if one can prove this strengthening of the D-Set Theorem.

Conjecture 6.1 (Coefficient Bound Conjecture). *Let $G = (V, E)$ be a graph, possibly infinite, of bounded maximum degree. Let $p \in \mathbb{Z}^+$. Then there exists a constant $c_1 = c_1(G, p)$ such that, for all $\mathbf{k} \in T^p$, there is an optimal labelling $f \in L(\mathbf{k})$ with all labels $f(v) \in D(\mathbf{k})$ in which the smallest label is 0, the largest label is $\lambda(G; \mathbf{k})$, and each of the labels $f(v)$ and $\lambda(G; \mathbf{k})$ can be expressed in the form $\sum_i a_i k_i$, where the nonnegative integer coefficients a_i are at most c_1.*

We cannot see how to derive the conjecture above from the PL Conjecture. We would need to know more, such as the domain can be split into finitely many regions such that in each region there is a single labelling of G that is optimal. Of course then there would be a collection of just finitely many labellings f_j such that, for any $\mathbf{k}$, some labelling in the collection is optimal (and feasible, of course) for $\mathbf{k}$.

We suspect that coefficient bounds can be given that work for all graphs with given maximum degree. Specifically, we propose this strengthening of the Coefficient Bound Conjecture.

CONJECTURE 6.2 (Delta Bound Conjecture). *Let $\Delta, p \in \mathbb{Z}^+$. Then there exists a constant $c_2 = c_2(\Delta, p)$ such that, for all $\mathbf{k} \in T^p$ and all graphs $G = (V, E)$, possibly infinite, of maximum degree at most Δ, there is an optimal labelling $f \in L(G; \mathbf{k})$ with all labels $f(v) \in D(\mathbf{k})$ in which the smallest label is 0, the largest label is $\lambda(G; \mathbf{k})$, and each of the labels $f(v)$ and $\lambda(G; \mathbf{k})$ can be expressed in the form $\sum_i a_i k_i$, where the nonnegative integer coefficients a_i are at most c_2.*

We have not even established this Delta Bound Conjecture yet for general finite graphs G, since the bound on the coefficient sums $\sum_i a_i$ in the D-Set Theorem, $n - 1$, is not restricted by Δ. It does hold trivially for $p = 1$ (with coefficient bound $a_1 = \chi(G) - 1 \leq \Delta$). It is not clear to us how the proof of the PL Conjecture for $p = 2$ (Theorem 5.6) can be used to obtain coefficient bounds to verify the Delta Bound Conjecture for $p > 2$. However, we can present another approach, which then gives a different proof of Theorem 5.6, one that may be useful in trying to prove the Delta Bound Conjecture (and, hence, the PL Conjecture) for general p.

THEOREM 6.3. *Let $\Delta \in \mathbb{Z}^+$. There exists a constant $c_3 = c_3(\Delta)$ such that, for all $(k_1, k_2) \in T^2$ and all graphs $G = (V, E)$, possibly infinite, of maximum degree at most Δ, there is an optimal labelling $f \in L(G; k_1, k_2)$ with all labels $f(v) \in D(k_1, k_2)$ in which the smallest label is 0, the largest label is $\lambda(k_1, k_2)$, and each of the labels $f(v)$ and $\lambda(k_1, k_2)$ can be expressed in the form $\sum_i a_i k_i$, where the nonnegative integer coefficients a_i are at most c_3.*

Proof. By the Scaling Property, it suffices to prove for given Δ the existence of c_3 that works for $(k, 1)$ for all $k \geq 0$. Let $G = (V, E) \in \mathcal{G}_\Delta$. Let f be an optimal labelling of G in $L(k, 1)$ as in the D-Set Theorem.

Case 1. Assume k is very small; say $0 < k \leq 1/(2\Delta^3)$.

By Lemma 2.2, f has span at most Δ^2. Thus, all labels used in f have the form $ak + b$, with nonnegative integer coefficients a, b such that $b \leq \Delta^2$. The trouble comes in trying to bound a, independent of k, no matter how small it gets. What we do is push down the labels $f(v)$ in a greedy way to produce a labelling $f' \in L(k, 1)$ such that all labels belong to a set $S \subseteq D(k, 1)$ in which the coefficients a are also bounded in terms of Δ. Since f' also has smallest label 0, and $f'(v) \leq f(v)$ for all vertices v, f' is an optimal labelling, one that satisfies the required conditions with $c_3 = \Delta^3 + \Delta^2$.

We define the set

$$S = \{ak + b : a, b \in \mathbb{Z},\ 0 \leq b \leq \Delta^2,\ 0 \leq a \leq (b + 1)\Delta\}.$$

It is important to note that our assumption that k is small means there is a gap between elements in S of the form $ak + b$ and $b + 1$. Let the set of labels used by f be given by $\{0 = l_0 < l_1 < l_2 < \cdots < l_r\}$, where $r = r(k)$ is finite, since it is contained in the finite set $D(k, 1) \cap [0, \Delta^2]$. For vertices v with label $l_0 = 0$, we set $f'(v) = f(v)$. We next take care of vertices in $f^{-1}(l_1)$, then those in $f^{-1}(l_2)$, and so on, through $f^{-1}(l_r)$. Let us suppose we are dealing with vertices v with label l_i, having already pushed down labels for vertices w with $f(w) < l_i$. Although $f^{-1}(l_i)$ can be infinite, no

two of its vertices are within distance two, so they can all be pushed simultaneously without any concern about interference. What we do have to ensure is that when we push down $f(v)$, $f'(v)$ is not too close to any $f'(w)$ for some w already pushed down. We define $f'(v)$ to be the largest element of $S \cap [0, l_i]$ that is at least k (resp., 1) away from $f'(w)$ for every w at distance one (resp., two) from v.

With this definition, f' has all of the required properties. What needs to be proved is that there is, indeed, some element of $S \cap [0, l_i]$ that is far enough away from the labels $f'(w)$ already defined. Put $B = \lfloor l_i \rfloor$ and $R = l_i - B$ so that $0 \le R < 1$.

Case 1a. Suppose that $R \le (B+1)\Delta k$. Put $A = \lfloor R/k \rfloor$. Then $f'(v)$ will be $Ak + B$. For we have $Ak + B \in S$, $0 \le l_i - (Ak + B) < k$, and $Ak + B$ is far enough from values $f'(w)$ already defined for nearby vertices w. For vertices w at distance one (resp., two) from v that were already pushed, $f(w) \le f(v) - k = R - k + B$ (resp., $f(w) \le f(v) - 1 = R + (B-1)$), so that $f'(w) \le (A-1)k + B = f'(v) - k$ (resp., $f'(w) \le Ak + (B-1) = f'(v) - 1$).

Case 1b. Suppose that $(B+1)\Delta k < R < 1$. Then $f'(v)$ will be the largest of the $\Delta + 1$ labels $ak + B$, $B\Delta \le a \le (B+1)\Delta$, that is not used as $f'(w)$ for any vertices w adjacent to v. Then $f'(v)$ is at least k away from labels $f'(w)$ for w adjacent to v. If w at distance two from v was already pushed, then $f(w) \le R + (B-1)$ means that $f'(w) \le B\Delta k + (B-1) = (B\Delta k + B) - 1 \le f'(v) - 1$.

Either way, $f'(v)$ exists as required.

Case 2. Assume k is very large; say $k \ge 2\Delta^4$.

The argument proceeds as in Case 1, though now we have to use the fact that for any vertex v, the number of vertices at distance two is at most Δ^2. This time we define our set

$$S = \{ak + b : a, b \in \mathbb{Z},\ 0 \le a \le \Delta,\ 0 \le b \le (a+1)\Delta^2\}.$$

We push down f in a similar way as before to produce an optimal labelling f' which has coefficient bound roughly Δ^4.

Case 3. Assume k is intermediate, $1/(2\Delta^3) < k < 2\Delta^2$.

By Lemma 2.2, $\lambda(k, 1) \le \Delta^2 \max\{k, 1\}$, which leads to an upper bound of $2\Delta^5$ on the coefficients a, b of the labels $ak + b$ of the optimal labelling f.

We see that $2\Delta^5$ serves as an upper bound on the coefficients for all k. □

An important note about the Delta Bound Conjecture is that it is simple to give a bound on the label coefficients (in terms of Δ and p) for which there does exist a feasible labelling f. For instance, there are the labellings described in connection with Lemma 2.2, in which every label has the form ak, where $k = \max_i\{k_i\}$, and a is at most Δ^p. But such labellings are certainly *not* optimal in general. Hence, the tough part of proving the Delta Bound Conjecture for general p is to show that for some constant c_3 there exists an *optimal* feasible labelling with coefficients bounded by c_3.

Besides being stronger than the PL Conjecture, the Delta Bound Conjecture is perhaps more natural, and easier to understand.

7. Degree bounds. Just as with chromatic numbers, it is interesting to consider how large the optimal span $\lambda(G; \mathbf{k})$ can be given the degrees of the vertices. Specifically, what if we bound the maximum degree $\Delta(G)$? Algorithms devised to achieve bounds we find are potentially useful, since they may produce reasonably efficient channel assignments.

For a connected finite graph G, an easy best-possible bound on the chromatic number is

$$\chi(G) \le \Delta(G) + 1,$$

and the well-known Brooks's Theorem implies that this bound is best-possible if and only if G is a clique K_n or an odd cycle C_{2k+1}. The $\Delta(G)+1$ bound can be achieved by arbitrarily ordering the vertices V, say, $\{v_1, v_2, \ldots\}$, and doing a greedy first-fit labelling of them one-by-one (always choose the lowest permissible color). Indeed, this works even if G is infinite.

For the basic λ-number, $\lambda(G) = \lambda(G; 2, 1)$, the analogous question was proposed in [16], cf. [29]. Of course, a connected graph G with maximum degree 0 or 1 must be a K_1 or K_2, respectively, and have λ equal to 0 or 2, respectively. After checking many examples, Griggs and Yeh made the following still-unproved conjecture.

CONJECTURE 7.1 (Delta Squared Conjecture). *If G is a connected graph with maximum degree $\Delta \geq 2$, then $\lambda(G) \leq \Delta^2$.*

This was stated for finite graphs, but would hold as well for infinite graphs by applying a compactness argument (the Rado Selection Principle, say). The quick explanation for why the bound is quadratic in Δ, instead of linear as for chromatic number, is that the interference in labels extends to distance two from a given vertex, and the number of vertices within distance two can be as large as $\Delta + \Delta(\Delta - 1) = \Delta^2$. Of course, this observation does not prove the conjecture, since there is the added restriction that labels for adjacent vertices cannot be consecutive.

The conjecture is tantalizing in part because if it fails, it is not by much. Griggs and Yeh used a simple vertex ordering and greedy first-fit labelling to show that

$$\lambda(G) \leq \Delta^2 + 2\Delta,$$

which supports the conjecture down to order $O(\Delta)$. On the other hand, they constructed graphs for infinitely many values Δ, using finite projective planes, for which

$$\lambda(G) \geq \Delta^2 - \Delta.$$

Also in support of the conjecture is that it has been shown by many researchers to hold for many classes of graphs. To mention a few, it is known to hold if G is diameter two [16], [29], and better bounds than Δ^2 have been proved for trees [16], chordal graphs [28], and planar graphs [18], [26]. Indeed, no one has found any graphs for which the Δ^2 bound is sharp, besides the short list in the original paper [16]:

- Paths and cycles, P_n and C_n, $n \geq 3$ ($\Delta = 2$),
- Petersen graph, $n = 10$ ($\Delta = 3$),
- Hoffman-Singleton graph, $n = 50$ ($\Delta = 7$),
- the 57-regular diameter-two graph on $57^2 + 1$ vertices, if it exists.

Chang and Kuo [4] managed to cut the gap in the general upper bound in half, proving that

$$\lambda(G) \leq \Delta^2 + \Delta.$$

The bound remained there for nearly ten years, before it was improved by Král' and Škrekovski.

THEOREM 7.2 (see [21]). *Let G be a graph, possibly infinite, with finite maximum degree $\Delta \geq 2$. Then $\lambda(G) \leq \Delta^2 + \Delta - 1$.*

For $\Delta = 2$, G must be a path or cycle, for which the conjecture is already verified. The next case up is $\Delta = 3$, where the best known general bound is now 11 [21]. Georges and Mauro checked many such graphs [9]. They not only found no graphs with $\lambda(G) > 9$, they found no other connected graphs with $\lambda = 9$. In fact, they found no such graphs at all with $\lambda = 8$, so they suspect (personal communication) that $\lambda(G) \leq 7$ if G is connected, has maximum degree 3, and is not the Peterson graph.

Now we consider Δ-bounds on $\lambda(G;\mathbf{k})$ for general separations $\mathbf{k}$. Earlier we gave such a bound, again by ordering the vertices and doing a greedy first-fit labelling, in Lemma 2.2:

$$\lambda(G;\mathbf{k}) \le k\Delta^p,$$

where k is the maximum k_i. However, when some k_i's are smaller than k, it is clear that $\lambda(G;\mathbf{k})$ should be smaller. A more careful argument takes advantage of such variation in the separations.

THEOREM 7.3. *Let G be a graph, possibly infinite, with finite maximum degree $\Delta \ge 0$. Let $\mathbf{k} = (k_1, \ldots, k_p) \ge 0$. Then $\lambda(G;\mathbf{k}) \le \sum_{i=1}^{p} 2k_i\Delta(\Delta-1)^{i-1}$.*

Proof. As in the proof of Lemma 5.5, it is enough to consider a single component of G in which the vertices are arbitrarily ordered $V = v_1, v_2, \ldots$. We do a greedy first-fit labelling f of the vertices, using for each vertex v the smallest label in $[0, B]$, where B is the bound in the theorem that is not too close to any previously assigned labels. To see that there is always such an available label, consider a previously labelled vertex w at distance i from v, $1 \le i \le p$. Then $f(v)$ must avoid the interval $(f(w) - k_i, f(w) + k_i)$ in order that $f \in L(\mathbf{k})$. Bounding the number of vertices at distance i, and assuming in the worst case that all of these vertices are already labelled and that their intervals are disjoint, we have a union of open intervals of lengths adding up to the bound B, so some element of $[0, B]$ is available for $f(v)$. ∎

A variation of the argument above gives a related bound that is sometimes slightly better, depending on the k_i's.

THEOREM 7.4. *Let G be a graph, possibly infinite, with finite maximum degree $\Delta \ge 0$. Let $\mathbf{k} = (k_1, \ldots, k_p) \ge 0$. Then $\lambda(G;\mathbf{k}) \le \sum_{i=1}^{p} (2\lceil k_i \rceil - 1)\Delta(\Delta-1)^{i-1}$.*

Proof. Do a greedy first-fit labelling as before, but restrict the labels $f(v)$ to integers. For a previously labelled vertex w at distance i from v, $f(v)$ must avoid the integers in the interval $(f(w) - k_i, f(w) + k_i)$ in order that $f \in L(\mathbf{k})$. These integers are from $f(w) + 1 - \lceil k_i \rceil$ to $f(w) - 1 + \lceil k_i \rceil$, a total of $2\lceil k_i \rceil - 1$ integers. The stated bound follows as before. ∎

8. Related results. The development of our theory of real number labellings was influenced by work on the triangular lattice described in the winning student MCM papers [3], [7], [11], [24], [5] and in the preprint [30]. These papers forced us to consider values $\lambda(\Gamma_\Delta; k, 1)$ for nonintegral values of k.

There is a considerable amount of work in the literature on labellings that is related to this project. We must first mention earlier work of Georges and Mauro that we only realized, after working out the concept of real-number labellings, is very much in the spirit of this project. In 1995 [8] they proved a restricted version of the D-Set Theorem: it is shown that for finite graphs G and for integers $p \ge q \ge 0$, there is an optimal labelling in $L(G; p, q)$ in which every label and $\lambda(G; p, q)$ have the form $ap + bq$, where a, b are nonnegative integers. They prove in this restricted setting that for integers $c > 0$, $\lambda(G; cp, cq) = c\lambda(G; p, q)$, a special case of our Scaling Property 2.1. They determine $\lambda(G; p, q), p \ge q$ for G being a path, a cycle, or various other graphs. In fact, our path and cycle formulas Theorems 3.2 and 3.3 above can be deduced—for $k \ge 1$—from their formulas for integers $p \ge q$ by using our real number model, the Scaling Property 2.1, Corollary 4.2, and continuity (Theorem 5.1).

Moreover, a later paper of Georges and Mauro [10] introduces what we refer to as labellings in $L(G; k, 1)$ with rational $k \ge 1$. They prove these labellings are continuous. This paper is also maybe the first to consider infinite graphs G. Its main result is to determine $\lambda(G; p, q)$ for integers $p \ge q \ge 1$ when G is the infinite Δ-regular tree.

Early versions of our results were presented at conferences going back to 2001, and slides from a presentation in 2003 are posted on the Web [13].

Georges and Mauro (personal communication) have now extended their earlier results to obtain continuity and piecewise-linearity statements for labellings with conditions at distance two, applicable to infinite graphs of bounded degree. This work appears similar to our Theorem 6.3, though their model is more restricted.

Mohar [25] has investigated a more general model, but restricted to finite graphs, in which there is a minimum separation $k_{v,w}$ for every pair of distinct vertices v and w. He actually works with the circular span of a graph, proving that it is continuous and piecewise linear, with only finitely many linear segments, as a function of the separations. He also considers an even more general directed graph model. A variation of his argument in the setting of our labellings may give a proof of the PL Conjecture 5.3 (for finite graphs). While our model is more restricted, in that separations depend only the distance between v and w, it may have special properties due to this restriction. We also work more generally with infinite graphs.

More recently, a paper by Leese and Noble [23] considers circular real number labellings with conditions at distance two, and obtains a continuous piecewise linear result in that context.

9. Directions for further research. Of course, we are anxious to see the conjectures above settled. As we completed this paper, we learned that a group in Prague (Babilon, Jelínek, Král', Valtr) is also preparing a paper on distance-dependent labellings from a somewhat different perspective [1], motivated in large part by the paper of Leese and Noble mentioned above. (Our main ideas were already presented at the DIMACS workshop [13] in October, 2003.)

In the distance two case ($p = 2$) let us consider how soon the formula becomes linear. What John Georges (personal communication) has observed in many examples is that $\lambda(G; k, 1)$ seems to be linear for $k > \Delta$—it settles down quickly. Is this true and can it be proven in general? Also, Theorem 6.3 can be used to bound the number of linear pieces in terms of Δ—but how good a bound can be given?

The authors are planning a future paper that explores the symmetry properties of optimal labellings of the triangular lattice with conditions at distance two.

It would be interesting to expand our model to consider infinite graphs with separations k_i at all distances i, not just finitely many conditions. Even for the particular examples of the triangular lattice and the square lattice, it would be interesting to characterize infinite $\mathbf{k}$ such that λ exists.

For use in many applications, we should return to considering the original problem of labelling transmitters in a planar network, using Euclidean distance, rather than graph distance. Perhaps the results for graphs can be helpful?

In some applications we have been told that many channels must be assigned to each transmitter (e.g., if each cell phone user in a particular cell must have a separate frequency). This can be accomplished by assigning an entire arithmetic progression of labels to each vertex in a graph, with the same distance d used for every progression such that, for nearby vertices v and w, every label used for v is sufficiently separated from every label for w. Equivalently, each vertex is assigned a label in the interval $[0, d)$, with distance between labels measured on the circle, that is, modulo d. The goal is to minimize the *circular span* d. There is a sizable literature on this problem for integer labellings. We have been investigating the extension of this model to allow real-number labellings, and there are analogues of some of the results given in this paper for "linear" labellings. This work will be described in a future paper.

Acknowledgments. The first author wishes to express gratitude to his old friend, Tom Savage, whose enthusiasm for this subject was an inspiration in developing this project. The first author's visit to Trinity, to speak in a special lecture series sponsored by Savage, brought the author together with Trinity faculty John Georges and David Mauro for a stimulating exchange of ideas. Both authors are thankful for helpful discussions and correspondence with many other colleagues, including Tiziana Calamoneri, Daniel Král', Renu Laskar, Daphne D.-F. Liu, Lincoln Lu, László Székely, and Craig Tovey. We also gladly acknowledge the support of the DIMACS-DIMATIA-Rényi series of workshops on graph colorings.

REFERENCES

[1] R. Babilon, V. Jelínek, D. Král', and P. Valtr, *Graph Labelings with Adjustable Weights*, ITI Report 2004-226, Institute for Theoretical Computer Science, Prague, Czech Republic, submitted.

[2] A. A. Bertossi and M. A. Bonuccelli, *Code assignment for hidden terminal interference avoidance in multihop packet radio networks*, IEEE/ACM Trans. Networking, 3 (Aug., 1995), pp. 441–449.

[3] R. E. Broadhurst, W. J. Shanahan, and M. D. Steffen , *We're sorry, you're outside the coverage area*, UMAP J., 21 (2000), pp. 327–342.

[4] G. J. Chang and D. Kuo, *The $L(2,1)$-labeling problem on graphs*, SIAM J. Discrete Math., 9 (1996), pp. 309–316.

[5] R. Chu, B. Xiu, and R. Zong, *Utilize the limited frequency resources efficiently*, UMAP J., 21 (2000), pp. 343–356.

[6] M. B. Cozzens and F. S. Roberts, *T-colorings of graphs and the channel assignment problem*, Congr. Numer., 35 (1982), pp. 191–208.

[7] D. J. Durand, J. M. Kline, and K. M. Woods, *Groovin' with the big band(width)*, UMAP J., 21 (2000), pp. 357–367.

[8] J. P. Georges and D. W. Mauro, *Generalized vertex labelings with a condition at distance two*, Congr. Numer., 109 (1995), pp. 141–159.

[9] J. P. Georges and D. W. Mauro, *On generalized Petersen graphs labeled with a condition at distance two*, Discrete Math., 259 (2003), pp. 311-318.

[10] J. P. Georges and D. W. Mauro, *Labeling trees with a condition at distance two*, Discrete Math., 269 (2003), pp. 127–148.

[11] J. Goodwin, D. Johnston, and A. Marcus, *Radio channel assignments*, UMAP J., 21 (2000), pp. 369–378.

[12] J. R. Griggs, *Judge's commentary: The outstanding channel assignment papers*, UMAP J., 21 (2000), pp. 379–386.

[13] J. R. Griggs and X. T. Jin, *Real number channel assignments with distance conditions*, DIMACS Graph Coloring Workshop lecture, Oct., 2003, posted online at http://dimacs.rutgers.edu/Workshops/GraphColor/slides.html.

[14] J. R. Griggs and X. T. Jin, *Real Number Channel Assignments for Lattices*, preprint, Sept., 2005, posted online at http://www.math.sc.edu/∼griggs/.

[15] J. R. Griggs and X. T. Jin, *Real Number Labelings for Paths and Cycles*, preprint, Sept., 2005, posted online at http://www.math.sc.edu/∼griggs/.

[16] J. R. Griggs and R. K.-C. Yeh, *labeling graphs with a condition at distance* 2, SIAM J. Discrete Math., 5 (1992), pp. 586–595.

[17] W. K. Hale, *Frequency assignment: Theory and applications*, Proc. IEEE, 68 (1980), pp. 1497–1514.

[18] J. van den Heuvel and S. McGuinness, *Coloring the square of a planar graph*, J. Graph Theory, 42 (2003), pp. 110–124.

[19] X. T. Jin, *Real Number Graph Labeling with Distance Conditions*, Ph.D. dissertation, Math. Dept., University of South Carolina, Columbia, SC August, 2005.

[20] X. T. Jin and R. K. Yeh, *Graph distance-dependent labeling related to code assignment in computer networks*, Naval Res. Logist., 52 (2005), pp. 159–164.

[21] D. Král' and R. Škrekovski, *A theorem about the channel assignment problem*, SIAM J. Discrete Math., 16 (2003), pp. 426–437.

[22] R. A. Leese, *A unified approach to the assignment of radio channels to a regular hexagonal grid*, IEEE Trans. Vehicular Tech., 46 (1997), pp. 969–980.

[23] R. A. Leese and S. D. Noble, *Cyclic labelings with constraints at two distances*, Electron. J. Combin., 11 (2004), #R16, 16pp.
[24] J. Mintz, A. Newcomer, and J. C. Prince, *A channel assignment model: The span without a face*, UMAP J., 21 (2000), pp. 311–326.
[25] B. Mohar, *Circular Colorings of Edge-Weighted Graphs*, J. Graph Theory, 43 (2003), pp. 107–116.
[26] M. Molloy and M. R. Salavatipour, *Frequency channel assignment on planar networks*, in Proceedings of the 10th Annual Europ. Sympos. Algorithms (ESA), 2002, pp. 736–747.
[27] F. S. Roberts, *Working Group Agenda,* DIMACS/DIMATIA/Renyi Working Group on Graph Colorings and their Generalizations, 2003, posted online at http://dimacs.rutgers.edu/Workshops/GraphColor/main.html.
[28] D. Sakai, *Labelling chordal graphs: Distance two condition*, SIAM J. Discrete Math., 7 (1994), pp. 133–140.
[29] R. K. Yeh, *Labeling Graphs with a Condition at Distance* 2, Ph.D. dissertation, University of South Carolina, Columbia, SC, 1990.
[30] D. Zhu and A. Shi, *Optimal Channel Assignments*, personal communication, 2001.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 328–343

A BOUND ON THE PRECISION REQUIRED TO ESTIMATE A BOOLEAN PERCEPTRON FROM ITS AVERAGE SATISFYING ASSIGNMENT*

PAUL W. GOLDBERG†

Abstract. A Boolean perceptron is a linear threshold function over the discrete Boolean domain $\{0,1\}^n$. That is, it maps any binary vector to 0 or 1, depending on whether the vector's components satisfy some linear inequality. In 1961, Chow showed that any Boolean perceptron is determined by the average or "center of gravity" of its "true" vectors (those that are mapped to 1), together with the total number of true vectors. Moreover, these quantities distinguish the function from any other Boolean function, not just from other Boolean perceptrons.

In this paper we go further, by identifying a lower bound on the Euclidean distance between the average satisfying assignment of a Boolean perceptron and the average satisfying assignment of a Boolean function that disagrees with that Boolean perceptron on a fraction ϵ of the input vectors. The distance between the two means is shown to be at least $(\epsilon/n)^{O(\log(n/\epsilon)\log(1/\epsilon))}$. This is motivated by the statistical question of whether an empirical estimate of this average allows us to recover a good approximation to the perceptron. Our result provides a mildly superpolynomial upper bound on the growth rate of the sample size required to learn Boolean perceptrons in the "restricted focus of attention" setting. In the process we also find some interesting geometrical properties of the vertices of the unit hypercube.

Key words. Boolean functions, threshold functions, geometry, inductive learning

AMS subject classifications. 68Q15, 68Q32, 52C07, 52C35

DOI. 10.1137/S0895480103426765

1. Introduction.

A *Boolean perceptron* is a linear threshold function over the domain of 0/1-vectors. (Subsequently we usually just say "perceptron" and omit the adjective "Boolean.") Thus it is specified by a weight vector $\mathbf{w}$ of n real numbers and a real-valued threshold t, and it maps a binary vector $\mathbf{x}$ to the output value 1, provided that $\mathbf{w}.\mathbf{x} \geq t$; otherwise it maps $\mathbf{x}$ to 0.

In this paper we consider the problem of estimating a perceptron from an approximate value of the mean, or "center of gravity" of its satisfying assignments. Chow [9] originally showed that any Boolean perceptron is identified by the exact value of the average of its satisfying assignments, along with the number of satisfying assignments, in the sense that there are no other Boolean functions of any kind for which the average and number of satisfying assignments is the same. The question of the extent to which an approximation to the average determines the perceptron is equivalent to the problem of learning Boolean perceptrons in the "restricted focus of attention" setting, described below.

The *Chow parameters* of a Boolean function are the coordinates of the vector sum of the satisfying vectors, together with the number of satisfying vectors. Subject to a uniform distribution over Boolean vectors, these are essentially equivalent to the conditional probabilities that the ith component of $\mathbf{x}$ is equal to 1, conditioned

*Received by the editors April 30, 2003; accepted for publication (in revised form) March 20, 2005; published electronically April 21, 2006. This work was supported by EPSRC grant GR/R86188/01, and in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication reflects only the author's views. A preliminary version of this paper was presented at the 2001 COLT conference.

http://www.siam.org/journals/sidma/20-2/42676.html

†Department of Computer Science, University of Warwick, Coventry, CV4 7AL, UK (pwg@dcs.warwick.ac.uk, http://www.dcs.warwick.ac.uk/∼pwg/).

on $\mathbf{x}$ being a satisfying assignment. Letting y denote the output value and $\mathbf{x} = ((\mathbf{x})_1, \ldots, (\mathbf{x})_n)$, these are the probabilities $\Pr((\mathbf{x})_i = 1 \mid y = 1)$, for $i = 1, \ldots, n$, together with the value $\Pr(y = 1)$.[1] Chow's result says that these values uniquely define the function, provided that it is a Boolean perceptron. (Bruck [8] shows, more generally, that a threshold function G over a set of monomials is characterized by the spectral coefficients of G that correspond to those monomials.) Hence a weights-based parametrization $(\mathbf{w}, t)$ should in principle be derivable from the Chow parameters; there will be some amount of freedom for $(\mathbf{w}, t)$ to vary while preserving the functional behavior on binary inputs.

In this paper we show that additive approximations of the Chow parameters determine the approximate behavior of the function, to within a mildly superpolynomial factor. That is in contrast to the situation for the weights-based parametrization of a perceptron, for which a tiny perturbation of some parameter may result in a large change to the set of points that are mapped to output value 1. In this sense the Chow parameters, as a description of a Boolean perceptron, are a more robust parametrization.

1.1. Background and previous results. Chow's paper gave rise to subsequent work that addressed the algorithmic problem of recovering a weights-based parametrization of a perceptron from its Chow parameters. This problem and related ones were later reconsidered in the computational learning theory literature, notably work on probably approximately correct (PAC)-learning in the so-called "restricted focus of attention" setting.

Earlier work that followed from [9] includes an algorithm by Kaszerman [16] for recovering a linear threshold function from its Chow parameters. The algorithm is iterative and somewhat related to the perceptron algorithm [19]; it does not have a good bound on the number of iterations and assumes that exact values of the parameters are given. A paper of Winder [20] compares seven functions (four of which were proposed in previous papers) for rescaling Chow parameters to obtain weights for a linear-threshold function. None of these functions has perfect performance, and it is uncertain that any function exists from individual Chow parameters to good weights—it may be necessary to deal with them collectively rather than individually. A further paper by Winder [21] investigates the class of Boolean functions that are uniquely defined by their Chow parameters, and shows among other things that it lies properly between the class of linear threshold functions and the class of monotonic functions.

The problem of learning a function f means reconstructing it (exactly or approximately) from a limited collection of observations of its input vectors $\mathbf{x}$ and associated values $f(\mathbf{x})$. There is much known about learning Boolean perceptrons in various settings, for example irrelevant attributes [17], classification noise [6], and learning from a source of "helpful" examples [2]. Special cases include monomials, decision lists [18, 12], and Boolean threshold functions. Further work on this topic occurs in the more general context of perceptrons over the real as opposed to the Boolean domain. An example is that they may be PAC-learned in a time polynomial in the dimension n and the PAC parameters ϵ and δ, using the Vapnik–Chervonenkis (VC) dimension theory [7]. Chapter 24 of [1] and references therein are a good introduction to results

[1] If the coordinates of the sum of all satisfying vectors are rescaled down by the number of satisfying vectors, one obtains the average satisfying assignment, whose coordinates are the probabilities $\Pr((\mathbf{x})_i = 1 \mid y = 1)$. The Chow parameters are recovered by multiplying this average by $2^n \cdot \Pr(y = 1)$.

on learning Boolean perceptrons.

Restricted focus of attention (RFA) learning was introduced and developed in the papers [3, 4, 5]. The k-RFA setting (where k is a positive integer) allows an algorithm to see only a subset of size k of the input attributes of any training example. The usual assumption has been that the distribution of input vectors $\mathbf{x}$ is known to be a product distribution (with no other information given about it). Clearly, 1-RFA learning (in which only one input attribute of each example is visible) is a very restrictive setting, making positive results of particular interest. In [13] we studied in detail the problem of learning linear-threshold functions over the real domain in the 1-RFA setting, so that each example of input/output behavior of the target function has only a single input component value, together with the binary value of the output, revealed to the learning algorithm. We showed that the input distribution (in [13], not necessarily a product distribution) needs to be at least partly known, and that the sample size required for learning depends sensitively on the input distribution. We identified measures of "well-behavedness" of the input distribution and gave sample size bounds in terms of these measures.

This paper addresses the topic of 1-RFA learning of perceptrons where the input distribution is uniform over V, the vertices of the unit hypercube. From [5] we have that a random sample of 1-RFA data is equivalent, in terms of the information it conveys, to approximations of the conditional probabilities $\Pr(y = 1 \mid (\mathbf{x})_i = b)$, for $b \in \{0, 1\}$ (where $(\mathbf{x})_i$ denotes the ith component of $\mathbf{x}$), together with the probability $\Pr(y = 1)$, and these approximations have additive error inversely proportional to the sample size. The coordinates of the average satisfying assignment are related as follows:

$$\begin{aligned}\Pr((\mathbf{x})_i = 1 \mid y = 1) &= \frac{\Pr((\mathbf{x})_i = 1)}{\Pr(y = 1)} \Pr(y = 1 \mid (\mathbf{x})_i = 1) \\ &= \frac{1}{2\Pr(y = 1)} \Pr(y = 1 \mid (\mathbf{x})_i = 1).\end{aligned}$$

Provided that $\Pr(y = 1)$ is not too small, we obtain good estimates of the coordinates of the average satisfying assignment from estimates of probabilities $\Pr(y = 1 \mid (\mathbf{x})_i = 1)$ (and vice versa). Our analysis handles low values of $\Pr(y = 1)$ as a special case.

The reason why the uniform distribution on V (for which bounds of [13] are inapplicable) is of particular interest is that it is the most natural and widely studied input distribution from the perspective of computational learning theory. The question of whether this learning problem is solvable with polynomial time or sample size was previously discussed in [10] and [13] and is currently known to be solvable under the restriction that weights are polynomially bounded. Birkendorf et al. [5] suggest the following rule: for $1 \leq i \leq n$ and $b \in \{0, 1\}$, let p_b^i be the observed conditional probability $\Pr(y = 1 \mid (\mathbf{x})_i = b)$ and let $p = \Pr(y = 1)$. Then take $\mathbf{x}$ to be a positive instance if $\frac{1}{n}\sum_{i=1}^{n} p^i_{(\mathbf{x})_i} > p$; otherwise label $\mathbf{x}$ as negative. It is left as an open problem whether the rule is valid.

We show here that, given a perceptron F and any Boolean function that disagrees with F on at least a fraction ϵ of input vectors, their average satisfying assignments must differ by $(\epsilon/n)^{O(\log(n/\epsilon)\log(1/\epsilon)}$ in the L_2 metric. The computational learning-theoretic result that follows is a mildly superpolynomial bound (of the order of $\log(\delta^{-1})(n/\epsilon)^{O(\log(n/\epsilon)\log(1/\epsilon))}$) on the asymptotic growth rate of sample size requirement for PAC-learning a perceptron from 1-RFA data. This is a purely "information-theoretic" result; we do not have any algorithm whose runtime has an asymptotic growth rate that improves substantially on a brute-force approach.

1.2. Notation and terminology. Let V be the input domain, i.e., the vertices of the unit hypercube, or 0/1-vectors. By a *vertex* we mean a member of V, i.e., a 0/1-vector of length n.

F will denote a Boolean perceptron, typically the "target function," and G will denote a Boolean function (not necessarily a Boolean perceptron), for example an estimate of F returned by an algorithm. The *positive* (respectively, *negative*) examples of a function are those that are mapped to 1 (respectively, 0). Let $pos(F)$, $neg(F)$, $pos(G)$, $neg(G)$ denote the positive and negative examples of F and G. (So $pos(F) = \{F^{-1}(1)\}$, etc.) F and G divide V into four subsets defined as follows:

$$V_{00} = neg(F) \cap neg(G), \qquad V_{01} = neg(F) \cap pos(G),$$

$$V_{10} = pos(F) \cap neg(G), \qquad V_{11} = pos(F) \cap pos(G).$$

For $R \subseteq \mathbb{R}^n$, let $m(R)$ be the number of elements of V that lie in R. Let $a(R)$ be the vector sum of elements of $V \cap R$. Let $\mu(R)$ denote the (unweighted) average of members of V that lie in the region R, so that $\mu(R) = a(R)/m(R)$, well-defined provided that $m(R) > 0$. The region of disagreement of F and G is $V_{01} \cup V_{10}$; thus the disagreement rate between F and G, over the uniform distribution on V, is $(m(V_{01}) + m(V_{10}))/2^n$.

Throughout, logarithms are to the base 2.

When we refer to subspaces, or spanning, or dimension, we mean in the affine sense, so that a "subspace" does not necessarily contain the origin, and the spanning set of $S \subseteq \mathbb{R}$, denoted Span(S), is the set of points that are expressible as the sum of one member of the spanning set plus a weighted sum of differences between pairs of points in S. A *line* means a 1-dimensional affine subspace.

We adopt the following usage of alphabetic symbols throughout the paper, which extends to variants embellished with primes or subscripts:

1. H denotes a hyperplane in $\mathbb{R}^n$ (an affine subspace with dimension $n-1$).
2. A denotes an affine subspace with possibly lower dimension.
3. S denotes a finite set of points in $\mathbb{R}^n$.
4. A point in $\mathbb{R}^n$ or an n-dimensional vector will be denoted by a lowercase boldface letter such as $\mathbf{x}$, and $(\mathbf{x})_i$ denotes the ith entry or component of $\mathbf{x}$. $\mathbf{v}$ is used to denote an element of V.

For $\mathbf{x} = ((\mathbf{x})_1, \ldots, (\mathbf{x})_n)$ let $\|\mathbf{x}\|$ denote the Euclidean norm of $\mathbf{x}$, i.e., $(\sum_{i=1}^n ((\mathbf{x})_i)^2)^{1/2}$. Let $d_E(\mathbf{x}, Z)$ denote the Euclidean distance between $\mathbf{x} \in \mathbb{R}^n$ and the closest point to $\mathbf{x}$ in $Z \subseteq \mathbb{R}^n$.

2. Geometric results. In this section we give various geometric results about the vertices of the unit hypercube, which we use in section 3 to deduce the bound on sample size requirement in the inductive learning context described in the last section. We start with an informal summary of the results of this section:

1. Lemma 1 gives a simple upper bound on the number of elements of V contained in a linear subspace, in terms of the dimension of that subspace.
2. Theorem 2 shows that if a hyperplane contains a large number of elements of V, then the coefficients of that hyperplane have a large common denominator. (A lower bound on the common denominator is given in terms of the number of elements of V contained by the hyperplane.)
3. Theorem 3 uses Theorem 2 to show that any hyperplane that "narrowly misses" a large fraction of V can be perturbed slightly so that it actually contains all those vertices. The resulting hyperplane no longer "narrowly misses" any other vertices. More precisely, if a hyperplane comes within

distance $O((1/\alpha)(n\log(n/\alpha))^{\log(n/\alpha)})$ of a fraction α of the 2^n vertices, then all those $\alpha \cdot 2^n$ vertices lie on the perturbed hyperplane.

4. Theorem 4 uses Theorem 3 to derive a lower bound on the distance between $\mu(V_{01})$ and $\mu(V_{10})$ (the means of the two regions of disagreement between two Boolean functions, one of which is a perceptron) in terms of their disagreement rate $m(V_{01} \cup V_{10})/2^n$.

LEMMA 1. *Any affine subspace A of $\mathbb{R}^n$ of dimension d contains at most 2^d elements of the vertices of the unit hypercube.*

Proof. The proof proceeds by induction on d. The lemma clearly holds for $d = 0$, when A consists of a single point.

Suppose $d > 0$. Assume that A contains at least two elements of V (if not, we are done). For $\mathbf{v}_1, \mathbf{v}_2 \in V \cap A$, suppose that $\mathbf{v}_1$ and $\mathbf{v}_2$ differ in the ith component, so that $(\mathbf{v}_1)_i \neq (\mathbf{v}_2)_i$.

Divide V into two subcubes V' and V'', where V' is elements $\mathbf{v} \in V$ such that $(\mathbf{v})_i = 0$, and V'' is elements $\mathbf{v} \in V$ with $(\mathbf{v})_i = 1$. By construction, $A \cap V' \neq \emptyset$ and $A \cap V'' \neq \emptyset$.

Since A intersects V', we have that $A \cap \mathrm{Span}(V'')$ is a proper subspace of A, and similarly, $A \cap \mathrm{Span}(V')$ is a proper subspace of A. The inductive hypothesis tells us that each of these subspaces contains at most 2^{d-1} elements of V, for a total of at most 2^d elements of V, as required. □

Observation 1. Let $S \subseteq V$, $|S| = \alpha \cdot 2^n$ (where $0 \leq \alpha \leq 1$). Let $d = n - \lfloor \log(1/\alpha) \rfloor - 1$. Then, given any subset of size d of the n components, there exist two distinct elements of S that agree on all those d components.

Proof. At most 2^d elements of V can be distinguished from each other via their values on a set of d coordinates. We assumed that $|S| = \alpha \cdot 2^n$. Since $d = n - \lfloor \log(1/\alpha) \rfloor - 1$, we can deduce that $\alpha > 2^{d-n}$, and hence $|S| > 2^d$. By the pigeonhole principle, two distinct elements of S agree on the d coordinates. □

THEOREM 2. *Let H be a hyperplane in $\mathbb{R}^n$, and suppose that H contains a fraction α of the vertices of the unit hypercube and that H is spanned by the vertices that it contains. Suppose that H is described as the set of points $\{\mathbf{x} : \mathbf{w}.\mathbf{x} = t\}$, with parameters $(\mathbf{w}, t)$ rescaled so that $\|\mathbf{w}\| = 1$. Then all the components of $\mathbf{w}$ are integer multiples of some quantity at least as large as*

$$\left(\sqrt{n}(1 + \lfloor \log(1/\alpha) \rfloor)! n^{(1+\lfloor \log(1/\alpha) \rfloor)}\right)^{-1}.$$

Proof. We construct a linear system that must be satisfied by the weights $\{(\mathbf{w})_i : 1 \leq i \leq n\}$ such that when we solve it (invert a matrix), elements of the inverted matrix have a large common denominator. Initially the system will be satisfied by the $(\mathbf{w})_i$ values when they are rescaled so that their maximum (in absolute value) is equal to 1. Afterwards we will rescale so that $\|\mathbf{w}\| = 1$.

Let $x_1 \in \arg\max_i(|(\mathbf{w})_i|)$. The first linear equality is $(\mathbf{w})_{x_1} = 1$. This does the job of rescaling the $(\mathbf{w})_i$ values such that their maximum (in absolute value) is 1.

Let $d = n - \lfloor \log(1/\alpha) \rfloor - 1$, as in Observation 1. For $\mathbf{v} \in V$, $(\mathbf{v})_i$, the ith component of $\mathbf{v}$, is equal to 0 or 1. We identify a subset of the component indices $\{x_2, \ldots, x_d\} \subseteq \{1, \ldots, n\}$ together with $2(d-1)$ vertices $\{\mathbf{v}_2, \mathbf{v}'_2, \ldots, \mathbf{v}_d, \mathbf{v}'_d\} \subseteq H \cap V$ such that

$$\begin{aligned} (\mathbf{v}_j)_{x_j} - (\mathbf{v}'_j)_{x_j} = 1 \quad &\text{for} \quad 2 \leq j \leq d, \\ (\mathbf{v}_j)_{x_i} = (\mathbf{v}'_j)_{x_i} \quad &\text{for} \quad 2 \leq j \leq d,\ 1 \leq i \leq d,\ j \neq i. \end{aligned}$$

For $\mathbf{v}$, $\mathbf{v}' \in H \cap V$, $\mathbf{w}$ satisfies $(\mathbf{v} - \mathbf{v}').\mathbf{w} = 0$. The next $d-1$ linear equalities are $(\mathbf{v}_j - \mathbf{v}'_j).\mathbf{w} = 0$ for $2 \le j \le d$. These linear constraints on $\mathbf{w}$ are independent of each other, since for the subset $\{x_2, \ldots, x_d\} \subset \{1, \ldots, n\}$, the linear constraint $(\mathbf{v}_j - \mathbf{v}'_j).\mathbf{w} = 0$ has coefficient 1 on the x_jth component of $\mathbf{w}$ and 0 on the other components in L_d. We continue by demonstrating how to find a suitable set $\{\mathbf{v}_2, \mathbf{v}'_2, \ldots, \mathbf{v}_d, \mathbf{v}'_d\}$. Let

$$R_1 = \{1, \ldots, n\} \setminus \{x_1\},$$
$$L_1 = \{x_1\}.$$

Choose $\mathbf{v}_2, \mathbf{v}'_2 \in H \cap V$ such that

$$\{\mathbf{v}_2, \mathbf{v}'_2\} \in \arg \max_{\{\mathbf{v},\mathbf{v}'\} \subseteq H \cap V; \mathbf{v} \neq \mathbf{v}'; (\mathbf{v})_\ell = (\mathbf{v}')_\ell \text{ for } \ell \in L_1} \Big(|\{i \in R_1 \ : \ (\mathbf{v})_i = (\mathbf{v}')_i\}| \Big).$$

Thus $\mathbf{v}_2$ and $\mathbf{v}'_2$ are chosen to be two distinct vertices in $H \cap V$, which have minimum Hamming distance from each other, subject to the requirement that they agree on component x_1.

Since $\mathbf{v}_2 \neq \mathbf{v}'_2$, there exists $x_2 \in R_1$ such that $(\mathbf{v}_2)_{x_2} \neq (\mathbf{v}'_2)_{x_2}$. We may assume that $(\mathbf{v}_2)_{x_2} = 1$ and $(\mathbf{v}'_2)_{x_2} = 0$. Let

$$R_2 = \{i \in R_1 \ : \ (\mathbf{v}_2)_i = (\mathbf{v}'_2)_i\},$$
$$L_2 = \{x_1, x_2\}.$$

R_2 is a maximal subset of R_1 such that two distinct vertices agree on coordinates indexed by R_2 and L_1. By Observation 1, $|R_2| \ge n - \lfloor \log(1/\alpha) \rfloor - 2$.

Generally, for $j > 2$, construct $x_j \in R_{j-1}$, $R_j \subseteq R_{j-1} \setminus \{x_j\}$, and $L_j = L_{j-1} \cup \{x_j\}$ as follows. Choose $\mathbf{v}_j, \mathbf{v}'_j \in H \cap V$ such that

$$\{\mathbf{v}_j, \mathbf{v}'_j\} \in \arg \max_{\{\mathbf{v},\mathbf{v}'\} \subseteq H \cap V; \mathbf{v} \neq \mathbf{v}'; (\mathbf{v})_\ell = (\mathbf{v}')_\ell \text{ for } \ell \in L_{j-1}} \Big(|\{i \in R_{j-1} \ : \ (\mathbf{v})_i = (\mathbf{v}')_i\}| \Big).$$

Thus $\mathbf{v}_j$ and $\mathbf{v}'_j$ are chosen to be two distinct vertices in $H \cap V$ that have minimum Hamming distance over coordinates indexed by R_{j-1}, subject to the constraint that they agree on coordinates indexed by L_{j-1}.

We claim that there exists $x_j \in R_{j-1}$ such that $(\mathbf{v}_j)_{x_j} \neq (\mathbf{v}'_j)_{x_j}$.

Suppose that the claim is false. Then $(\mathbf{v}_j)_i = (\mathbf{v}'_j)_i$ for all $i \in R_{j-1}$, and $(\mathbf{v}_j)_\ell = (\mathbf{v}'_j)_\ell$ for all $\ell \in L_{j-1}$ (and note that for $\ell \in L_{j-1}$, $\ell \notin R_{j-1}$). This contradicts the choice of $\{\mathbf{v}_{j-1}, \mathbf{v}'_{j-1}\}$ as a pair of vertices that have minimum Hamming distance on coordinates indexed by R_{j-2} (which contains R_{j-1}) while also agreeing on coordinates indexed by L_{j-2}. Note that

1. $\mathbf{v}_{j-1}$ and $\mathbf{v}'_{j-1}$ agree on coordinates indexed by L_{j-2}. They agree on $|R_{j-1}|$ elements of R_{j-2}.
2. $\mathbf{v}_j$ and $\mathbf{v}'_j$ agree on coordinates indexed by $L_{j-1} = L_{j-2} \cup \{x_{j-1}\}$, where $x_{j-1} \in R_{j-2}$. They also agree on all elements of $R_{j-1} \subseteq R_{j-2}$.
3. From the above two points, amongst pairs of vertices $\mathbf{v}$ and $\mathbf{v}'$ that agree on L_{j-2}, $\mathbf{v}_j$ and $\mathbf{v}'_j$ agree on more elements of R_{j-2} than do $\mathbf{v}_{j-1}$ and $\mathbf{v}'_{j-1}$.

Hence there exists $x_j \in R_{j-1}$ such that $(\mathbf{v}_j)_{x_j} \neq (\mathbf{v}'_j)_{x_j}$, and we can assume $(\mathbf{v}_j)_{x_j} = 1$ and $(\mathbf{v}'_j)_{x_j} = 0$. Let

$$R_j = \{i \in R_{j-1} \ : \ (\mathbf{v}_j)_i = (\mathbf{v}'_j)_i\},$$
$$L_j = L_{j-1} \cup \{x_j\}.$$

R_j is a maximal subset of R_{j-1} (where $|R_{j-1}| \geq n - \lfloor \log(1/\alpha) \rfloor - (j-1)$) such that $\mathbf{v}_j$ agrees with $\mathbf{v}'_j$ on coordinates indexed by R_j (and the $j-1$ coordinates indexed by L_{j-1}). By Observation 1, $|R_j| \geq n - \lfloor \log(1/\alpha) \rfloor - j$.

Recall that $d = n - \lfloor \log(1/\alpha) \rfloor - 1$, as in Observation 1. Since $|R_j| \geq n - \lfloor \log(1/\alpha) \rfloor - j$, the above construction can be carried out for $2 \leq j \leq d$.

By our assumption that $\mathrm{Span}(H \cap V) = H$, there exists a set $\{\mathbf{v}_{d+1}, \mathbf{v}'_{d+1} \ldots, \mathbf{v}_n, \mathbf{v}'_n\} \subset H \cap V$ such that each pair of vertices $\{\mathbf{v}_j, \mathbf{v}'_j\}$ for $d+1 \leq j \leq n$ imposes on $\mathbf{w}$ a new linear constraint $(\mathbf{v}_j - \mathbf{v}'_j).\mathbf{w} = 0$ that is linearly independent of the others.

Let M be a matrix whose first row is all zero apart from the x_1th entry, which contains the value 1. The jth row (for $2 \leq j \leq n$) is the components of $(\mathbf{v}_j - \mathbf{v}'_j)$. We have $M.\mathbf{w} = \mathbf{r}$, where $\mathbf{r}$ is all zero apart from $(\mathbf{r})_1 = 1$. Now rearrange the columns of M in the order $x_1, \ldots, x_n$ (where $\{x_{d+1}, \ldots, x_n\} = \{1, \ldots, n\} \setminus \{x_1, \ldots, x_d\}$), and let $\mathbf{r} = (1, 0, \ldots, 0)^T$. We have constructed a linear system $M.\mathbf{w}^P = \mathbf{r}$, where $\mathbf{w}^P$ is a permutation of $\mathbf{w}$ and

1. M is an invertible $n \times n$ matrix with entries in $\{0, 1, -1\}$;
2. the $d \times d$ submatrix of M comprising the first d rows and columns is the identity matrix;
3. $\mathbf{r} = (1, 0, \ldots, 0)^T$.

Hence $\mathbf{w}^P = M^{-1}\mathbf{r}$. The (i,j)th entry of M^{-1} is given by $\det(M_{i,j})/\det(M)$, where $\det(M)$ denotes the determinant of matrix M, and $M_{i,j}$ is the submatrix of M obtained by removing column i and row j. We will upper-bound the determinant of M.

Construct M' by adding (respectively, subtracting) row j (for $1 \leq j \leq d$) to row j' (for $d+1 \leq j' \leq n$) whenever the jth entry of row j' is equal to -1 (respectively, 1). $M' = (m)_{ij}$ satisfies

$$\begin{aligned} m_{ij} &= 0 \quad \text{for} \quad d+1 \leq i \leq n,\ 1 \leq j \leq d, \\ -n \leq m_{ij} &\leq n \quad \text{for} \quad d+1 \leq i \leq n,\ d+1 \leq j \leq n. \end{aligned}$$

Here $\det(M') = \det(M)$, the first d rows and columns of M' is still the identity matrix, and so from the features of M' noted above, $\det(M')$ is equal to $\det(M'')$, where M'' is the $(n-d) \times (n-d)$ submatrix of M' in the bottom right-hand corner of M'.

Now observe that the determinant of any $i \times i$ matrix with entries in $\{-n, -(n-1), \ldots, n-1, n\}$ is upper bounded[2] by $i!n^i$, so that $|\det(M)| \leq (n-d)!n^{n-d}$. Accordingly, entries of M^{-1} (and consequently, components of $\mathbf{w}$) must be integer multiples of a quantity greater than or equal to

$$\left((n-d)!n^{n-d}\right)^{-1} = \left((1 + \lfloor \log(1/\alpha) \rfloor)!n^{(1+\lfloor \log(1/\alpha) \rfloor)}\right)^{-1},$$

and so components of $\mathbf{w}$ are also integer multiples of this quantity.

The maximum absolute value of a component of $\mathbf{w}$ (or $\mathbf{w}^P$) is 1, so $1 \leq \|\mathbf{w}\| \leq \sqrt{n}$. Rescaling $\mathbf{w}$ to get $\|\mathbf{w}\| = 1$, we find that the components of $\mathbf{w}$ are integer multiples of a quantity at least as large as the above, divided by $\sqrt{n}$. That is,

$$\left(\sqrt{n}(1 + \lfloor \log(1/\alpha) \rfloor)!n^{(1+\lfloor \log(1/\alpha) \rfloor)}\right)^{-1},$$

as in the statement of the theorem. □

[2]There is not a substantially better upper bound on the determinant of this matrix that uses the fact that the matrix is over integers with absolute value at most n; from Hadamard [14], the determinant of a $i \times i$ matrix over $\{1, -1\}$ may be as high as $i^{i/2}$. This becomes $n^i.i^{i/2}$ when the entries 1 and -1 are replaced with n and $-n$, respectively.

We use Theorem 2 to prove the following.

THEOREM 3. *Given any hyperplane in $\mathbb{R}^n$ whose β-neighborhood contains a subset S of vertices of the unit hypercube, where $|S| = \alpha \cdot 2^n$, there exists a hyperplane which contains all elements of S, provided that*

$$0 \le \beta \le \left((2/\alpha) \cdot n^{(5+\lfloor \log(n/\alpha) \rfloor)} \cdot (2 + \lfloor \log(n/\alpha) \rfloor)!\right)^{-1}.$$

Proof. Let $H = \{\mathbf{x} \ : \ \mathbf{w}.\mathbf{x} = t\}$, where by rescaling we can assume $\|\mathbf{w}\| = 1$. Assume that the β-neighborhood of H contains S. Then for $\mathbf{v} \in S$, we have $\mathbf{w}.\mathbf{v} \in [t - \beta, t + \beta]$.

Define a new weight vector $\mathbf{w}'$ derived from $\mathbf{w}$ by taking each weight in $\mathbf{w}$ and rounding it off to the nearest integer multiple of β (rounding down in the event of a tie). Then we claim that scalar products $\mathbf{w}'.\mathbf{v}$ can take at most $n + 2$ distinct values for $\mathbf{v} \in S$. To see this, note that for $\mathbf{v} \in S$,

1. $\mathbf{w}'.\mathbf{v} < \mathbf{w}.\mathbf{v} + n\beta/2 \le t + \beta + n\beta/2$,
2. $\mathbf{w}'.\mathbf{v} \ge \mathbf{w}.\mathbf{v} - n\beta/2 \ge t - \beta - n\beta/2$,
3. $\mathbf{w}'.\mathbf{v}$ is an integer multiple of β for $\mathbf{v} \in V$.

Items 1 and 2 show that $\mathbf{w}'.\mathbf{v}$ lies in a semiopen interval of length $\beta(n+2)$, and with 3 there are only at most $(n + 2)$ possible values in the interval. Let T be the set of these $n + 2$ values.

Let t' be the member of T which maximizes the number of vertices $\mathbf{v} \in S$ satisfying $\mathbf{w}'.\mathbf{v} = t'$. Then there are at least $\alpha \cdot 2^n/(n + 2)$ vertices $\mathbf{v} \in S$ that satisfy $\mathbf{w}'.\mathbf{v} = t'$. Let

$$\begin{aligned} A_1 &= \mathrm{Span}(\{\mathbf{v} \in S \ : \ \mathbf{w}'.\mathbf{v} = t'\}), \\ H_1 &= \{\mathbf{x} \in \mathbb{R}^n \ : \ \mathbf{w}'.\mathbf{x} = t'\}. \end{aligned}$$

Note that $|A_1 \cap V| \ge \alpha \cdot 2^n/(n + 2)$, and hence by Lemma 1,

$$\dim(A_1) \ge n - \log(1/\alpha) - \log(n + 2). \tag{1}$$

We next show that for all $\mathbf{v} \in S$,

$$d_E(\mathbf{v}, H_1) \le 2n\beta. \tag{2}$$

Note that $\|\mathbf{w}' - \mathbf{w}\| \le \sqrt{n}\beta/2$. $\|\mathbf{w}\| = 1$, and since the Euclidean norm is a metric,

$$\|\mathbf{w}'\| \in [1 - \sqrt{n}\beta/2, 1 + \sqrt{n}\beta/2].$$

For $\mathbf{v} \in S$, $\mathbf{w}'.\mathbf{v} - t' \in [-(n + 2)\beta, (n + 2)\beta]$. Let $(\mathbf{w}'', t'')$ be $(\mathbf{w}', t')$ rescaled so that $\|w''\| = 1$. Then

$$\begin{aligned} & \mathbf{w}''.\mathbf{v} - t'' \in [-(n+2)\beta/(1 - \sqrt{n}\beta/2), (n+2)\beta/(1 - \sqrt{n}\beta/2)] \\ \Rightarrow \quad & \mathbf{w}''.\mathbf{v} - t'' \in [-2n\beta, 2n\beta] \qquad (\text{since } \sqrt{n}\beta \ll 1) \\ \Rightarrow \quad & \mathbf{w}''.\mathbf{v} \in [t'' - 2n\beta, t'' + 2n\beta]. \end{aligned}$$

Since $\|\mathbf{w}''\| = 1$, $\mathbf{v}$ is within Euclidean distance $2n\beta$ of H_1. This establishes (2).

We want to show that $\dim(\mathrm{Span}(S)) \le n - 1$. We next find a hyperplane H_k that contains A_1 and other elements of S such that $\mathrm{Span}(H_k \cap S) = H_k$ (allowing Theorem 2 to apply to H_k) and such that we also obtain a bound on $d_E(\mathbf{v}, H_k)$ for $\mathbf{v} \in S$.

We know that $\dim(A_1) < n$. If $\dim(A_1) = n-1$, then set $k = 1$ and use $H_k = H_1 = A_1$. Suppose that $\dim(A_1) < n-1$. Then let A'_1 be a subspace of H_1 such that $\dim(A'_1) = n-2$ and $A_1 \subseteq A'_1$. Let $\mathbf{v}_1 \in \arg\max_{\mathbf{v}\in S}(d_E(\mathbf{v}, A'_1))$.

Let H_2 be the hyperplane $\mathrm{Span}(A'_1 \cup \{\mathbf{v}_1\})$. Then for all $\mathbf{v} \in S$, using (2),

$$d_E(\mathbf{v}, H_2) \leq d_E(\mathbf{v}, H_1) + d_E(\mathbf{v}_1, H_1) \leq 4n\beta.$$

Let $A_2 = \mathrm{Span}(A_1 \cup \{\mathbf{v}_1\})$. Since $\mathbf{v}_1 \notin A_1$ we have $\dim(A_2) = \dim(A_1) + 1$.

Generally, for $j \geq 1$, if $A_j \subset H_j$, $A_j \neq H_j$, construct A_{j+1} and H_{j+1} from A_j and H_j as follows. Choose A'_j of dimension $n-2$ such that

$$A_j \subseteq A'_j \subset H_j.$$

Then choose

$$\mathbf{v}_j \in \arg\max_{\mathbf{v}\in S}(d_E(\mathbf{v}, A'_j)).$$

Then let $H_{j+1} = \mathrm{Span}(A'_j \cup \{\mathbf{v}_j\})$ and $A_{j+1} = \mathrm{Span}(A_j \cup \{x_j\})$. Then for all $\mathbf{v} \in S$,

$$d_E(\mathbf{v}, H_{j+1}) \leq d_E(\mathbf{v}, H_j) + d_E(\mathbf{v}_j, H_j) \leq 2^{j+1}n\beta.$$

$A_{j+1} \subseteq H_{j+1}$ and $\dim(A_{j+1}) = 1 + \dim(A_j)$. The maximum value that j can take is

$$k = n - \dim(A_1) \leq \log(1/\alpha) + \log(n+2) \tag{3}$$

(the inequality follows from (1)), at which point we obtain $A_k = H_k$ with $\dim(H_k) = n-1$. H_k satisfies

1. $H_k = \mathrm{Span}(H_k \cap S)$,
2. $\dim(H_k) = n-1$,
3. $|H_k \cap S| \geq \alpha \cdot 2^n/(n+2)$,
4. for all $\mathbf{v} \in S$, $d_E(\mathbf{v}, H_k) \leq 2^k n\beta \leq (1/\alpha)(n+2)n\beta$, using (3).

Hence by properties 1–3 above and Theorem 2, H_k takes the form

$$H_k = \{\mathbf{x} \ : \ \mathbf{w}_k.\mathbf{x} = t_k\},$$

where $\|\mathbf{w}_k\| = 1$ and entries of $\mathbf{w}_k$ and t_k are multiples of

$$E = \left(\sqrt{n}\left(1 + \left\lfloor \log\left(\frac{n+2}{\alpha}\right)\right\rfloor\right)! n^{(1+\lfloor \log((n+2)/\alpha)\rfloor)}\right)^{-1}$$

(the expression from Theorem 2 with $\alpha/(n+2)$ plugged in for α).

$\mathbf{w}_k.\mathbf{v}$ is an integer multiple of E for all $\mathbf{v} \in V$. Hence if $t_k - E < \mathbf{w}_k.\mathbf{v} < t_k + E$, then $\mathbf{w}_k.\mathbf{v} = t_k$.

From property 4 of H_k, for all $\mathbf{v} \in S$, $\mathbf{w}_k.\mathbf{v} = t_k$, provided that we have

$$(1/\alpha)(n+2)n\beta < E.$$

Equivalently,

$$\beta < \left((1/\alpha)(n+2)n\sqrt{n}\left(1 + \left\lfloor \log\left(\frac{n+2}{\alpha}\right)\right\rfloor\right)! n^{(1+\lfloor \log((n+2)/\alpha)\rfloor)}\right)^{-1}.$$

The expression for β given in the statement of this theorem satisfies the inequality. □

THEOREM 4. *Let F be a Boolean perceptron and let G be a Boolean function that disagrees with F on a fraction ϵ of the 2^n elements of V. Assume also that $|V_{01}| \geq \frac{1}{4}\epsilon \cdot 2^n$ and $|V_{10}| \geq \frac{1}{4}\epsilon \cdot 2^n$. Then the Euclidean distance between $\mu(V_{01})$ and $\mu(V_{10})$ is lower bounded by*

$$\left((4/\epsilon) \cdot n^{(5+\lfloor \log(2n/\epsilon) \rfloor)} \cdot (2 + \lfloor \log(2n/\epsilon) \rfloor)!\right)^{-4\log(1/\epsilon)},$$

which is $(\epsilon/n)^{O(\log(n/\epsilon)\log(1/\epsilon))}$.

Proof. If l is a line and S is a set of points, let $l(S)$ denote the set of points obtained by projecting elements of S onto their closest points on l.

Let H_F denote a hyperplane defining F, and let l_1 be a line normal to H_F. We may assume that H_F does not contain any elements of V. Observe that members of $l_1(V_{01})$ are separated from members of $l_1(V_{10})$ by the point of intersection of l_1 and H_F (which itself is $l_1(H_F)$). Let

$$\beta = \left((4/\epsilon) \cdot n^{(5+\lfloor \log(2n/\epsilon) \rfloor)} \cdot (2 + \lfloor \log(2n/\epsilon) \rfloor)!\right)^{-1} \tag{4}$$

(where we have plugged $\epsilon/2$ for α into the expression for β in the statement of Theorem 3). Our analysis uses a sequence of $\lfloor \log(1/\epsilon) \rfloor$ cases.

Case 1. Suppose that at least a fraction $\beta^{4\log(1/\epsilon)-2}$ of elements of $V_{01} \cup V_{10}$ (i.e., at least $(\epsilon \cdot 2^n)\beta^{4\log(1/\epsilon)-2}$ vertices altogether) have projections onto l_1 that are more than β distant from $l_1(H_F)$. In this case we have

$$\|\mu(V_{01}) - \mu(V_{10})\| \geq \beta \cdot \beta^{4\log(1/\epsilon)-2}.$$

The alternative is that at least a fraction $(1-\beta^{4\log(1/\epsilon)-2})$ of elements of $V_{01} \cup V_{10}$ (thus, at least $(\epsilon \cdot 2^n)(1-\beta^{4\log(1/\epsilon)-2})$ points altogether) have projections onto l_1 that are less than β distant from $l_1(H_F)$.

In this case we apply Theorem 3 to obtain a hyperplane A_1 that contains all these points, that is, at least a fraction $1-\beta^{4\log(1/\epsilon)-2}$ of elements of $V_{01} \cup V_{10}$. (Theorem 3 applies since $\epsilon(1-\beta^{4\log(1/\epsilon)-2})$ plays the role of α, and $\epsilon(1-\beta^{4\log(1/\epsilon)-2}) > \frac{1}{2}\epsilon$ (thus, with (4), the corresponding β-value is sufficiently small).)

Case 2. Let $A_2' = H_F \cap A_1$; since H_F does not contain any elements of V, H_F does not contain A_1. $A_2' \subset A_1$ separates $V_{01} \cap A_1$ from $V_{10} \cap A_1$. Let $l_2 \subseteq A_1$ be a line normal to A_2'.

Now suppose that at least a fraction $\beta^{4\log(1/\epsilon)-4}$ of elements of $V_{01} \cup V_{10}$ lie in A_1 and have projections onto l_2 that are more than β distant from $l_2(A_2')$. Then

$$\|\mu(A_1 \cap V_{01}) - \mu(A_1 \cap V_{10})\| \geq \beta \cdot \beta^{4\log(1/\epsilon)-4}.$$

$|(V_{01} \setminus A_1)|/|V_{01}| \leq \epsilon\beta^{4\log(1/\epsilon)-2}/(\epsilon/4)$, and since all vertices lie within $\sqrt{n}$ of each other, the distance $\|\mu(V_{01}) - \mu(V_{01} \setminus A_1)\|$ is at most $(4\sqrt{n})\beta^{4\log(1/\epsilon)-2}$. A similar argument applies to V_{10}. Hence we have

$$\begin{aligned}\|\mu(V_{01}) - \mu(V_{10})\| &\geq \beta \cdot \beta^{4\log(1/\epsilon)-4} - 2(4\sqrt{n})\beta^{4\log(1/\epsilon)-2} \\ &= \beta^{4\log(1/\epsilon)-4}(\beta - \beta^2 8\sqrt{n}) \geq \beta^{4\log(1/\epsilon)}.\end{aligned}$$

It remains to cover the cases where a fraction less than $\beta^{4\log(1/\epsilon)-4}$ of the members of $V_{01} \cup V_{10}$ have projections onto l_2 that are more than β distant from $l_2(A_2')$. Generally case j arises when a subspace A_{j-1} of dimension $n-(j-1)$ has been

identified that contains at least a fraction $1-\sum_{\ell=1}^{j-1}\beta^{(4\log(1/\epsilon)-2\ell)}$ of the elements of $V_{01}\cup V_{10}$ (and we have not yet found a hyperplane separating enough of V_{01} from V_{10} with a sufficiently large margin).

Case j. Subspace A_{j-1} with $\dim(A_{j-1})=n-(j-1)$ satisfies

$$\frac{|A_{j-1}\cap(V_{01}\cup V_{10})|}{|V_{01}\cup V_{10}|}\geq 1-\sum_{\ell=1}^{j-1}\beta^{(4\log(1/\epsilon)-2\ell)}.$$

Let $A'_j=A_{j-1}\cap H_F$ and $\dim(A'_j)=n-j$. Let $l_j\subseteq A_{j-1}$ be a line normal to A'_j.

Suppose that at least a fraction $\beta^{(4\log(1/\epsilon)-2j)}$ of elements of $V_{01}\cup V_{10}$ lie in A_{j-1} and have projections onto l_j that are more than β distant from $l_j(A'_j)$. Then

$$\|\mu(A_{j-1}\cap V_{01})-\mu(A_{j-1}\cap V_{10})\|\geq\beta\cdot\beta^{(4\log(1/\epsilon)-2j)}.$$

Note that

$$\frac{|(V_{01}\cup V_{10})\setminus A_{j-1}|}{|V_{01}\cup V_{10}|}\leq\sum_{\ell=1}^{j-1}\beta^{(4\log(1/\epsilon)-2\ell)}.$$

Since $\beta<\frac{1}{2}$, this fraction is less than $2\beta^{(4\log(1/\epsilon)-2(j-1))}$. Hence

$$\begin{aligned}\|\mu(V_{01})-\mu(V_{10})\|&\geq\beta\cdot\beta^{(4\log(1/\epsilon)-2j)}-(4\sqrt{n})2\beta^{(4\log(1/\epsilon)-2(j-1))}\\&=\beta^{(4\log(1/\epsilon)-2j)}(\beta-2\beta^2 4\sqrt{n})\\&\geq\beta^{4\log(1/\epsilon)}.\end{aligned}$$

If, alternatively, a fraction at least $1-\beta^{(4\log(1/\epsilon)-2j)}$ of elements of $V_{01}\cup V_{10}$ have projections onto l_j at most β from $l_j(A'_j)$, then we construct A_j of dimension $n-j$ that contains all these points.

Let $V_j\subseteq(V_{01}\cup V_{10})$ denote this set of points. Let S_j be a set of $j-1$ vertices such that $\dim(\mathrm{Span}(A_{j-1}\cup S_j))=n$. The hyperplane $\mathrm{Span}(A'_j\cup S_j)$ lies within Euclidean distance β of elements of V_j, where $|V_j|\geq\frac{1}{2}\epsilon\cdot 2^n$. (For $j\leq\lfloor\log(\epsilon^{-1})\rfloor$, the fraction of elements of $V_{01}\cup V_{10}$ that are in V_j is at least $1-\beta^{(4\log(1/\epsilon)-2j)}$, so that $|V_j|\geq\frac{1}{2}\epsilon$.) Use Theorem 3 (and (4)) to obtain hyperplane H_j, which contains $V_j\cup S_j$. Let $A_j=H_j\cap A_{j-1}$. H_j cannot contain A_{j-1} since H_j also contains S_j and we have $\mathrm{Span}(A_{j-1}\cup S_j)=n$. Hence $\dim(A_j)=n-j$.

For $j<\lfloor\log(\epsilon^{-1})\rfloor$,

$$\frac{|A_j\cap(V_{01}\cup V_{10})|}{|V_{01}\cup V_{10}|}=\frac{|V_j|}{|V_{01}\cup V_{10}|}\geq 1-\beta^{4\log(1/\epsilon)-2j}>1-\sum_{\ell=1}^{j}\beta^{4\log(1/\epsilon)-2\ell}>\frac{1}{2}\epsilon,$$

and thus for $j<\lfloor\log(\epsilon^{-1})\rfloor$ we are ready for case $j+1$.

By Lemma 1 the number of cases (and hence j) is indeed upper bounded by $\lfloor\log(\epsilon^{-1})\rfloor$, since otherwise the subspace A_j does not have sufficient dimension to hold a fraction $\frac{1}{2}\epsilon$ of elements of V. Each of these cases provides a lower bound on $\|\mu(V_{01})-\mu(V_{10})\|$ of $\beta^{4\log(1/\epsilon)}$, which is

$$\Big((4/\epsilon)\cdot n^{(5+\lfloor\log(2n/\epsilon)\rfloor)}\cdot(2+\lfloor\log(2n/\epsilon)\rfloor)!\Big)^{-4\log(1/\epsilon)},$$

as in the statement of the theorem. □

3. Statistical learning-theoretic consequences. For domain $V = \{0,1\}^n$ let $U(V)$ denote the uniform distribution on V. For a Boolean function G having at least one satisfying assignment, let $Y_{G,0}$ be the following Bernoulli random variable: $Y_{G,0} = 1$ if for $\mathbf{v} \sim U(V)$ we have $G(\mathbf{v}) = 1$. Recall that $(\mathbf{v})_i$ denotes the 0/1 value of the ith component of $\mathbf{v}$. For $1 \le i \le n$ let $Y_{G,i}$ be the following Bernoulli random variable: $Y_{G,i} = 1$ if for $\mathbf{v} \sim U(\{\mathbf{u} \in V \ : \ (\mathbf{u})_i = 1\})$ we have $G(\mathbf{v}) = 1$.

To learn a Boolean perceptron in the 1-RFA regime (over the uniform distribution on $V = \{0,1\}^n$), a "target perceptron" F is selected by an adversary. A learning algorithm may (in unit time) generate an observation $(\mathbf{v}, \ell)$, where $\mathbf{v} \sim U(V)$ and $\ell = F(\mathbf{v})$. The algorithm has access to the value ℓ and may select $i \in \{1, \ldots, n\}$, so as to observe the value $(\mathbf{v})_i$. The remainder of $\mathbf{v}$ is not available to the algorithm. This is equivalent to being given access to repeated observations of the random variables $Y_{F,i}$ above, for $0 \le i \le n$. The objective is to output, with probability $1-\delta$, a function G (the "hypothesis," an estimate of F) such that G disagrees with F on a fraction at most ϵ of elements of V. (An alternative formulation of RFA learning assumes that the indices of the observed components of an input vector $\mathbf{v}$ are selected uniformly at random. We noted in [13] that for 1-RFA learning this is equivalent, for the purpose of obtaining polynomial bounds, to the assumption that the index is chosen by the algorithm.)

We continue by using the results of section 2 to obtain a bound on the sample size required to learn a Boolean perceptron in the 1-RFA setting. Thus we show how a computationally unbounded (but with limited sample size) algorithm can select a good hypothesis from the entire set of Boolean perceptrons, using sample size $\log(\delta^{-1}) \cdot (n/\epsilon)^{\log(n/\epsilon)\log(1/\epsilon)}$, where δ is the probability that the hypothesis has error greater than ϵ. For any Boolean function G let

$$p_{G,0} = \Pr_{\mathbf{v} \sim U(V)}(G(\mathbf{v}) = 1),$$
$$p_{G,i} = \Pr_{\mathbf{v} \sim U(V)}(G(\mathbf{v}) = 1 \mid (\mathbf{v})_i = 1).$$

For a Boolean function G define cost function $c_F(G)$ and empirical cost function $\hat{c}_F(G)$ as

$$c_F(G) = \max_{0 \le i \le n}(|p_{G,i} - p_{F,i}|),$$
$$\hat{c}_F(G) = \max_{0 \le i \le n}(|p_{G,i} - \hat{p}_{F,i}|),$$

where $\hat{p}_{F,i}$ is defined in Figure 1. Note that $c_F(F) = 0$.

LEMMA 5. *Let F be a Boolean perceptron that is satisfied by at least $(\epsilon/2) \cdot 2^n$ input vectors. Let Boolean function G disagree with F on at least a fraction ϵ of inputs. Then*

$$c_F(G) \ge \left(\frac{\epsilon^2}{32\sqrt{n}}\right)\left((4/\epsilon) \cdot n^{(5+\lfloor \log(2n/\epsilon) \rfloor)} \cdot (2 + \lfloor \log(2n/\epsilon) \rfloor)!\right)^{-4\log(1/\epsilon)}.$$

Proof. We consider two cases. As in the proof of Theorem 4, let $\beta = ((4/\epsilon) \cdot n^{(5+\lfloor \log(2n/\epsilon) \rfloor)} \cdot (2 + \lfloor \log(2n/\epsilon) \rfloor)!)^{-1}$.

Case 1. $|p_{F,0} - p_{G,0}| \ge \frac{\epsilon^2}{32\sqrt{n}} \cdot \beta^{4\log(1/\epsilon)}$ (that is, there is a difference of at least $\frac{\epsilon^2}{32\sqrt{n}} \cdot \beta^{4\log(1/\epsilon)}$ between the probability that $F(\mathbf{v}) = 1$ and the probability that $G(\mathbf{v}) = 1$). Then $c_F(G) \ge \frac{\epsilon^2}{32\sqrt{n}} \cdot \beta^{4\log(1/\epsilon)}$, which implies the statement of the lemma.

1. *Draw a sample S_0 of observations, where $|S_0| = \Theta((1/\epsilon)\log(1/\delta))$.*
2. *Let $\hat{p}_{F,0}$ be the fraction of examples in S_0 which satisfy F (we do not look at any component of the input vectors).*
3. *If $\hat{p}_{F,0} < \frac{3}{4}\epsilon$, then output G, where $G(\mathbf{v}) = 0$ for all $\mathbf{v} \in \{0,1\}^n$.*
4. *Else*
   (a) *For $1 \le i \le n$, draw a sample S_i of observations, where $|S_i| = (\log(1/\delta)(n/\epsilon)^{O(\log(n/\epsilon)\log(1/\epsilon))})$. Look at the ith component of each input $\mathbf{v}$ in S_i.*
   (b) *For $0 \le i \le n$, let $\hat{p}_{F,i}$ be the fraction of all examples with $(\mathbf{v})_i = 1$ in S_i which are positive (satisfy F).*
   (c) *For every satisfiable Boolean function G let $p_{G,i} = \Pr(Y_{G,i} = 1)$ (for $0 \le i \le n$).*
   (d) *Let $\hat{c}(G) = \max_{0\le i\le n}(|\hat{p}_{F,i} - p_{G,i}|)$.*
   (e) *Output a Boolean function from $\arg\min_G(\hat{c}(G))$.*

FIG. 1. *Rule for selecting low-error perceptron.*

Case 2. If $|p_{F,0} - p_{G,0}| < \frac{\epsilon^2}{32\sqrt{n}} \cdot \beta^{4\log(1/\epsilon)}$, then $|V_{01}| \ge (\epsilon/4)\cdot 2^n$ and $|V_{10}| \ge (\epsilon/4)\cdot 2^n$. So Theorem 4 applies to F and G, and we have

$$\|\mu(V_{01}) - \mu(V_{10})\| \ge \beta^{4\log(1/\epsilon)}.$$

Let $\lambda = |V_{10}|/(|V_{10}|+|V_{11}|)$, $\lambda' = |V_{01}|/(|V_{01}|+|V_{11}|)$. If $|V_{10}| \ge |V_{01}|$, then $\lambda \ge \lambda'$ and

$$\lambda - \lambda' \le \frac{|V_{10}| - |V_{01}|}{|V_{01}| + |V_{11}|} \le \frac{|V_{10}| - |V_{01}|}{|V_{01}|} \le \frac{(\epsilon^2/32\sqrt{n})\beta^{4\log(1/\epsilon)}}{\epsilon/4} = \frac{\epsilon}{8\sqrt{n}}\beta^{4\log(1/\epsilon)}.$$

If $|V_{01}| \ge |V_{10}|$, we have the same upper bound on $\lambda' - \lambda \ge 0$.

$$\begin{aligned}\mu(pos(F)) &= (1-\lambda)\cdot\mu(V_{11}) + \lambda\mu(V_{10}),\\ \mu(pos(G)) &= (1-\lambda')\cdot\mu(V_{11}) + \lambda'\mu(V_{01})\\ &= (1-\lambda)\cdot\mu(V_{11}) + \lambda\mu(V_{01}) + (\lambda-\lambda')(\mu(V_{11}) - \mu(V_{01})).\end{aligned}$$

Hence (note that $\lambda \ge \frac{\epsilon}{4}$):

$$\begin{aligned}\|\mu(pos(F)) - \mu(pos(G))\| &\ge \lambda\|(\mu(V_{10}) - \mu(V_{01}))\| - (\lambda-\lambda')\|\mu(V_{11}) - \mu(V_{01})\|\\ &\ge \tfrac{\epsilon}{4}\|\mu(V_{10}) - \mu(V_{01})\| - (\lambda-\lambda')\sqrt{n}\\ &\ge \tfrac{\epsilon}{4}\beta^{4\log(1/\epsilon)} - \tfrac{\epsilon}{8}\beta^{4\log(1/\epsilon)}.\end{aligned}$$

The statement of the lemma follows—there exists $i \in \{1,\dots,n\}$ such that the ith component of $\mu(pos(F))$ differs from the ith component of $\mu(pos(G))$ by at least the above quantity divided by $\sqrt{n}$. □

THEOREM 6. *Let F be an arbitrary Boolean perceptron, and suppose that we have access to a source of observations of the form $((\mathbf{v})_i, F(\mathbf{v}))$, where $\mathbf{v} \sim U(V)$ and where we may select the value of $i \in \{1,\dots,n\}$ for each observation. Then (ignoring issues of computational efficiency) it is possible to find, with probability $1-\delta$, a Boolean function G such that $\Pr_{\mathbf{v}\sim U(V)}(F(\mathbf{v}) \ne G(\mathbf{v})) \le \epsilon$, and the number of observations required is*

$$\log(1/\delta)\cdot(n/\epsilon)^{O(\log(n/\epsilon)\log(1/\epsilon))}.$$

Proof. We use the procedure illustrated in Figure 1. Note that symbols denoting various quantities are introduced in Figure 1.

Choose $N = |S_0|$ to ensure that with probability $1-\frac{1}{2}\delta$, if $\hat{p}_{F,0} < \frac{3}{4}\epsilon$, then $p_{F,0} \le \epsilon$. As a result, the function G output in line 3, which has no satisfying assignments, has error at most ϵ. We show as follows that $N = O((1/\epsilon)\log(1/\delta))$ is large enough.

Recall Hoeffding's inequality: Let $Y_1, \ldots, Y_N$ be Bernoulli trials with probability p of success. Let $T = Y_1 + \cdots + Y_N$ denote the total number of successes. Then for $\gamma \in [0,1]$,

$$\Pr(|T - pN| > \gamma N) \le 2e^{-2N\gamma^2}.$$

Set $\gamma = \frac{1}{4}\epsilon$ to ensure that with high probability

$$|\hat{p}_{F,0} - p_{F,0}| < \frac{1}{4}\epsilon. \tag{5}$$

$N = |S_0|$ must then satisfy $2e^{-2N(\epsilon/4)^2} \le \frac{1}{2}\delta$, which is satisfied by $N = O(\epsilon^{-1}\log(\delta^{-1}))$.

Equation (5) ensures that if $\hat{p}_{F,0} \ge \frac{3}{4}\epsilon$, then $p_{F,0} \ge \frac{1}{2}\epsilon$. Thus line 3 of Figure 1 is (with probability $1 - \frac{1}{2}\delta$) used only when $p_{F,0} \ge \frac{1}{2}\epsilon$ (and Lemma 5 is applicable). As in the proofs of Theorem 4 and Lemma 5, let $\beta = ((4/\epsilon) \cdot n^{(5+\lfloor \log(2n/\epsilon)\rfloor)} \cdot (2 + \lfloor \log(2n/\epsilon)\rfloor)!)^{-1}$.

We choose the size of each S_i large enough to ensure that with probability $1-\delta/4$ each S_i contains at least N' examples $(\mathbf{v}, F(\mathbf{v}))$ with $(\mathbf{v})_i = 1$, where N' is large enough to ensure that

$$\text{with probability } 1-\delta/4, \text{ for } 1 \le i \le n, \qquad |\hat{p}_{F,i} - p_{F,i}| < \left(\frac{\epsilon^2}{64\sqrt{n}}\right)\beta^{4\log(1/\epsilon)}. \tag{6}$$

The above can be ensured by taking a union bound if we have

$$\text{for } 1 \le i \le n, \text{ with probability } 1-\delta/4n, \qquad |\hat{p}_{F,i} - p_{F,i}| < \left(\frac{\epsilon^2}{64\sqrt{n}}\right)\beta^{4\log(1/\epsilon)}.$$

By Hoeffding's inequality it is sufficient for N' to satisfy $2\exp(-2N'(\epsilon^2/64\sqrt{n})\beta^{4\log(1/\epsilon)}) < \delta/4n$, which is satisfied by $N' = O((n/\epsilon^2)\log(n/\delta)/\beta^{4\log(1/\epsilon)})$.

Set $|S_i| = 4N'$. A standard Chernoff bound (see, for example, [1, p. 361]) tells us that if T is the number of successes in N Bernoulli trials with probability p of success,

$$\Pr\left(T < \frac{1}{2}Np\right) \le \exp\left(-\frac{Np}{8}\right).$$

Here $|S_i| = 4N'$, and so the expected number of examples with $(\mathbf{v})_i = 1$ is $2N'$ (since $\Pr((\mathbf{v})_i = 1) = \frac{1}{2}$), and the probability that we fail to obtain N' of these examples is $O(\exp(-N'(\epsilon/2)/8)) = O(\delta/n)$. For $N' = O((n/\epsilon^2)\log(n/\delta)/\beta^{4\log(1/\epsilon)})$ this failure probability can be made as low as $\delta/4n$, so that with probability at least $1 - \frac{1}{4}\delta$, for $1 \le i \le n$, S_i contains at least N' examples with $(\mathbf{v})_i = 1$.

Equation (6) implies

$$\text{with probability } 1-\delta/4, \text{ for all } G, \qquad |\hat{c}_F(G) - c_F(G)| < \left(\frac{\epsilon^2}{64\sqrt{n}}\right)\beta^{4\log(1/\epsilon)}.$$

Then by Lemma 5 (and noting that $c_F(F) = 0$), $\hat{c}_F(F) < \hat{c}_F(G)$ for all Boolean functions G that disagree with F on a fraction at least ϵ of inputs.

The total sample size is $O(n \cdot N')$, which is $O((n^2/\epsilon^2)\log(n/\delta)/\beta)$, which is $\log(1/\delta) \cdot (n/\epsilon)^{O(\log(n/\epsilon)\log(1/\epsilon))}$. ∎

3.1. Conclusions and open problems. The problem of PAC-learning a Boolean perceptron from empirical estimates of its Chow parameters has been raised in various papers in computational learning theory. We have so far just shown a bound on the asymptotic growth rate of sample-size required (the problem of how to best select the right hypothesis, given sufficient data, having not been addressed), and that bound is still superpolynomial. We suspect that the true growth rate is polynomially bounded as a function of n/ϵ.

Our results show that an algorithm can minimize over the set of all Boolean functions; we do not have to restrict ourselves to Boolean perceptrons. This demonstrates how the usage of a set of statistics, as opposed to empirical risk minimization, can automatically avoid over-fitting. However, there is the possibility that there should exist a better bound on the distance between the average satisfying assignment of two functions if both, and not just one, of them are perceptrons.

There may be a practical advantage to minimizing over all Boolean functions, in that if the minimization is being done by local search, it may reduce problems with local optima. However, in principle one can just minimize over the set of all Boolean perceptrons. The algorithm uses the values $p_{G,i}$ for Boolean functions G, and for Boolean perceptrons computing these quantities exactly is $\sharp P$-hard since it is the 0/1 knapsack problem [11]. However, sufficiently good approximations to these quantities could be found by generating a polynomial-size collection of inputs from $U(V)$ and using the empirical values.

Håstad [15] has shown that some Boolean perceptrons need weights of size around $2^{(n \log n)/2-n}$ to be represented exactly. For $n = \lfloor \log(\epsilon^{-1}) \rfloor$ (n being the dimension of the domain), an approximation with error less than ϵ must be exact. This implies that we may need to learn a weight of size more than polynomial in ϵ, in order to recover a weights-based parametrization—weights may be as high as $(1/\epsilon)^{\log \log(1/\epsilon)}$. This eliminates one natural-looking way of obtaining the desired polynomial growth rate in ϵ^{-1} (namely, looking for a perceptron whose coefficients are polynomially bounded as a function of the dimension and the quality of the approximation).

Acknowledgment. I would like to thank the referees for their corrections and comments.

REFERENCES

[1] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, Cambridge, UK, 1999.

[2] M. Anthony, G. Brightwell, and J. Shawe-Taylor, *On specifying Boolean functions by labelled examples*, Discrete Appl. Math., 61 (1995), pp. 1–25.

[3] S. Ben-David and E. Dichterman, *Learning with restricted focus of attention*, J. Comput. System Sci., 56 (1998), pp. 277–298.

[4] S. Ben-David and E. Dichterman, *Learnability with restricted focus of attention guarantees noise-tolerance*, in Proceedings of the 5th International Workshop on Algorithmic Learning Theory, Lecture Notes in Comput. Sci. 872, Springer, New York, 1994, pp. 248–259.

[5] A. Birkendorf, E. Dichterman, J. Jackson, N. Klasner, and H. U. Simon, *On restricted-focus-of-attention learnability of Boolean functions*, Machine Learning, 30 (1998), pp. 89–123.

[6] A. Blum, A. Frieze, R. Kannan, and S. Vempala, *A polynomial-time algorithm for learning noisy linear threshold functions*, Algorithmica, 22 (1998), pp. 35–52.

[7] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, *Learnability and the Vapnik–Chervonenkis dimension*, J. ACM, 36 (1989), pp. 929–965.

[8] J. Bruck, *Harmonic analysis of polynomial threshold functions*, SIAM J. Discrete Math., 3 (1990), pp. 168–177.

[9] C. K. Chow, *On the characterization of threshold functions*, in Proceedings of the Sympo-

sium on Switching Circuit Theory and Logical Design, American Institute of Electrical Engineers, 1961, pp. 34–38.

[10] E. Dichterman, *Learning with Limited Visibility*, CDAM Research Reports Series, LSE-CDAM-98-01, London School of Economics, London, 1998.

[11] M. E. Dyer, A. M. Frieze, R. Kannan, A. Kapoor, L. Perkovic, and U. Vazirani, *A mildly exponential time algorithm for approximating the number of solutions to a multidimensional knapsack problem*, Combin. Probab. Comput., 2 (1993), pp. 271–284.

[12] T. Eiter, T. Ibaraki, and K. Makino, *Decision Lists and Related Boolean Functions*, Institut Für Informatik JLU Giessen (IFIG) Research Reports 9804, Justus-Liebig Universitat, Giessen, Germany, 1998.

[13] P. W. Goldberg, *Learning fixed-dimension linear thresholds from fragmented data*, Inform. and Comput., 171 (2001), pp. 98–122.

[14] J. Hadamard, *Résolution d'une question relative aux déterminants*, Bull. Sci. Math., 2 (1893), pp. 240–246.

[15] J. Håstad, *On the size of weights for threshold gates*, SIAM J. Discrete Math., 7 (1994), pp. 484–492.

[16] P. Kaszerman, *A geometric test-synthesis procedure for a threshold device*, Inform. and Control, 6 (1963), pp. 381–398.

[17] N. Littlestone, *Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm*, Machine Learning, 2 (1988), pp. 285–318.

[18] R. L. Rivest, *Learning decision lists*, Machine Learning, 2 (1996), pp. 229–246.

[19] F. Rosenblatt, *Principles of Neurodynamics*, Spartan Books, New York, 1962.

[20] R. O. Winder, *Threshold gate approximations based on Chow parameters*, IEEE Trans. Comput., 18 (1969), pp. 372–375.

[21] R. O. Winder, *Chow parameters in threshold logic*, J. ACM, 18 (1971), pp. 265–289.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 344–356

THE MINOR CROSSING NUMBER*

DRAGO BOKAL†, GAŠPER FIJAVŽ‡, AND BOJAN MOHAR§

Abstract. The minor crossing number of a graph G is defined as the minimum crossing number of all graphs that contain G as a minor. Basic properties of this new invariant are presented. We study topological structure of graphs with bounded minor crossing number and obtain a new strong version of a lower bound based on the genus. We also give a generalization of an inequality of Moreno and Salazar crossing numbers of a graph and its minors.

Key words. crossing number, graph minor

AMS subject classifications. 05C10, 05C83

DOI. 10.1137/05062706X

1. Preliminaries.

Crossing numbers of graphs have been thoroughly studied [20], yet only a few exact results are known, and new ideas seem to be needed. Crossing numbers in general give a measure of nonplanarity of graphs. Unfortunately, they are not monotone with respect to graph minors. Seymour (see Archdeacon [1]) asked "How to define a crossing number that would work well with minors?" In this paper we propose two possible answers to this question and study one of them in greater detail. Our approach is based on general principles of how a graph invariant can be transformed into a minor-monotone graph invariant [4].

Crossing numbers of graphs are believed to have applications in VLSI design where one wants a design of a (huge) electrical network such that the number of crossing edges (wires) is minimized [3, 10, 11]. However, today's chip manufacturers replace vertices of high degree by binary trees. The minor crossing number treated in this paper does precisely this—each vertex is expanded into a cubic tree in such a way that the resulting graph can be realized with as few crossings as possible. It turns out that this interpretation of crossing numbers has rich mathematical structure, whose basics are uncovered in this work.

Let $G = (V_G, E_G)$ be a graph and Σ a closed surface. If Σ has Euler characteristics χ, then the number $g = 2-\chi$ is called the *Euler genus* of Σ. The nonorientable surface of Euler genus $g \geq 1$ is denoted by $\mathbb{N}_g$, and the orientable surface of Euler genus $2g$ ($g \geq 0$) is denoted by $\mathbb{S}_g$.

A *drawing* $D = (\varphi, \varepsilon)$ of G in (PL) surface Σ consists of a one-to-one mapping $\varphi : V_G \to \Sigma$ and a mapping $\varepsilon : E_G \to \Omega(\Sigma)$ that maps edges of G to simple (polygonal) curves in Σ such that endpoints of $\varepsilon(uv)$ are $\varphi(u)$ and $\varphi(v)$, $\varphi(V_G)$ does not intersect interiors of images of edges, and the intersection of interiors of ε-images of any two distinct edges contains at most one point.

*Received by the editors March 17, 2005; accepted for publication (in revised form) November 28, 2005; published electronically April 21, 2006. This work was supported in part by the Ministry of Higher Education, Science and Technology of Slovenia, Research Project L1-5014 and Research Program P1-0297.

http://www.siam.org/journals/sidma/20-2/62706.html

†Department of Mathematics, IMFM, SI-1000 Ljubljana, Slovenia (drago.bokal@imfm.uni-lj.si).

‡Faculty of Computer and Information Science, University of Ljubljana, SI-1000 Ljubljana, Slovenia (gasper.fijavz@fri.uni-lj.si).

§Department of Mathematics, University of Ljubljana, SI-1000 Ljubljana, Slovenia (bojan.mohar@uni-lj.si). Current address: Department of Mathematics, Simon Fraser University, Burnaby, B.C., Canada.

Let e and f be distinct edges of G, let r and s be their images in Σ, and suppose that $x \in r \cap s$. Let U be a neighborhood of x so that for each disk neighborhood $B \subseteq U$ of x both $B \cap r \cap s = \{x\}$ and $|\partial B \cap (r \cup s)| = 4$. We say that e and f or that r and s *cross* at x (and call x a *crossing*) if points of r and s interlace along ∂B for every such B, and say that r and s *touch* otherwise. In the latter case we call x a *touching* of r and s (or of e and f).

A drawing D is *normal* if it has no touchings and for each crossing x there are precisely two edges of G whose crossing is x.

Crossing number of a graph G in Σ, $\mathrm{cr}(G, \Sigma)$, is defined as the minimum number of crossings in any normal drawing of G in Σ, and with $\mathrm{cr}(G)$ we denote the crossing number of G in the sphere. For a drawing $D = (\varphi, \varepsilon)$ of G in Σ, connected regions of $\Sigma \setminus \varepsilon(E_G)$ are called *faces of* D. By our standards, a drawing of G in the plane $\mathbb{R}^2$ is a drawing of G in the sphere $\mathbb{S}_0$, equipped with an *infinite point* ∞ avoiding the image of G. The *infinite face* of a drawing of G in the plane is the face containing ∞. Further, an *embedding* is a drawing without crossings. Besides this terminology, the reader is referred to [15] for other notions related to graph embeddings.

For a given graph G, the *minor crossing number* is defined as the minimum crossing number of all graphs that contain G as a minor:

$$\mathrm{mcr}(G, \Sigma) := \min\{\mathrm{cr}(H, \Sigma) \mid G \leq_m H\}. \tag{1.1}$$

By $\mathrm{mcr}(G)$ we denote $\mathrm{mcr}(G, \mathbb{S}_0)$.

Similarly, the *major crossing number* of G is the maximum crossing number taken over all minors of G:

$$\mathrm{Mcr}(G, \Sigma) := \max\{\mathrm{cr}(H, \Sigma) \mid H \leq_m G\}. \tag{1.2}$$

The following two lemmas follow directly from the definitions.

LEMMA 1.1. *For every graph G and every surface Σ,*

$$\mathrm{mcr}(G, \Sigma) \leq \mathrm{cr}(G, \Sigma) \leq \mathrm{Mcr}(G, \Sigma).$$

LEMMA 1.2. *If G is a minor of H, then for every surface Σ,*

$$\mathrm{mcr}(G, \Sigma) \leq \mathrm{mcr}(H, \Sigma) \quad \textit{and} \quad \mathrm{Mcr}(G, \Sigma) \leq \mathrm{Mcr}(H, \Sigma).$$

Lemma 1.2 immediately yields the following.

COROLLARY 1.3. *Let $k \geq 0$ be an integer and Σ a surface. The families of graphs $\omega(k, \Sigma) := \{G \mid \mathrm{mcr}(G, \Sigma) \leq k\}$ and $\Omega(k, \Sigma) := \{G \mid \mathrm{Mcr}(G, \Sigma) \leq k\}$ are minor-closed.*

For each graph G there exists a graph $\bar{G}$ such that $G \leq_m \bar{G}$ and $\mathrm{mcr}(G, \Sigma) = \mathrm{cr}(\bar{G}, \Sigma)$. We call such a graph $\bar{G}$ a *realizing graph of* G, and an optimal drawing of $\bar{G}$ in Σ is called a *realizing drawing of* G (with respect to Σ). By no means are a realizing graph or drawing uniquely determined, but we shall always assume that G and $\bar{G}$ have the same number of connected components.

As G is a minor of its realizing graph $\bar{G}$, G can be obtained as a contraction of a subgraph of $\bar{G}$. In other words, $G = (\bar{G} - R)/C$ for suitable edge sets $R, C \subseteq E_{\bar{G}}$. The edges of R are called *removed edges*, and those in C are *contracted edges*. Note that the edge-set C is acyclic and that $E_G = E_{\bar{G}} \setminus (R \cup C)$ are the original edges of G. It is clear that every graph G has a realizing graph $\bar{G}$ such that $R = \emptyset$.

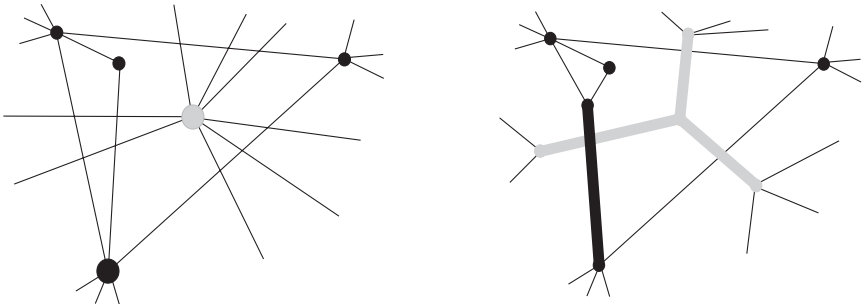


FIG. 1. mcr *as an extension of* cr.

For each vertex $v \in V_G$ there is a unique maximal tree $T_v \subseteq \bar{G}[C]$ which is contracted to v. In the figures, the original edges will be drawn as thin lines and the contracted edges as thick lines.

The minor crossing number can be considered a natural extension of the usual crossing number. Clearly, if $e, f \in E_{\bar{G}}$ cross in a realizing drawing of G, then $e, f \in C \cup E_G$. If both belong to C, then their crossing is a *vertex-vertex* crossing; if both belong to E_G, then they cross in an *edge-edge* crossing; and otherwise they cross in an *edge-vertex* crossing. This point of view is illustrated in Figure 1. Note that by subdividing the original edges appropriately, all the crossings in the realizing drawing can be forced to be vertex-vertex crossings.

If G is a cubic graph, then clearly $\mathrm{mcr}(G, \Sigma) = \mathrm{cr}(G, \Sigma)$. Hliněný proved in [6] that computing the planar crossing number of cubic graphs is NP-hard and has remarked that this implies that the same holds for computing $\mathrm{mcr}(G)$ for any graph G. Crossing numbers of cubic graphs were also studied by McQuillan and Richter [13] and Richter [17].

PROPOSITION 1.4. *For every graph G and every surface Σ there exists a cubic realizing graph H. Moreover, if $\delta(G) \geq 3$, then G can be obtained from H by contracting edges only.*

Proof. Let H_0 be a realizing graph of G without removed edges, and let $D_0 = (\varphi, \varepsilon)$ be an optimal drawing of H_0. We shall describe H in terms of its drawing D obtained from D_0. For each vertex v of H_0 of degree $d := d_{H_0}(v) \neq 3$ let U_v be a closed disk containing $\varphi(v)$ in its interior, so that a small neighborhood of U_v contains no crossings, U_v is disjoint from U_u for $u \in V_{H_0} \setminus \{v\}$, and $U_v \cap \varphi(E_{H_0})$ is connected.

For each of the cases $d > 3$, $d = 2$, and $d = 1$, we modify D_0 in U_v as indicated in Figure 2. Let D be this new drawing and H the graph defined by D.

Clearly $G \leq_m H$, and so $\mathrm{cr}(H, \Sigma) \geq \mathrm{mcr}(G, \Sigma)$. As D contains no new crossings, we have $\mathrm{mcr}(G, \Sigma) = \mathrm{cr}(H_0, \Sigma) = \mathrm{cr}(D, \Sigma) \geq \mathrm{cr}(H, \Sigma)$. A combination of these two inequalities proves that $\mathrm{cr}(H, \Sigma) = \mathrm{mcr}(G, \Sigma)$.

If $\delta(G) \geq 3$, then we can assume $\delta(H_0) \geq 3$, which implies $|E_H| - |V_H| = |E_{H_0}| - |V_{H_0}|$. As $H_0 \leq_m H$, we can obtain G from H by contracting edges only. ∎

2. Minor crossing number and maximum degree. In this section we present a generalization of the following result (cf. also section 6).

THEOREM 2.1 (see Moreno and Salazar [16]). *Let G be a minor of a graph H with $\Delta(G) \leq 4$. Then $\frac{1}{4}\,\mathrm{cr}(G, \Sigma) \leq \mathrm{cr}(H, \Sigma)$ for every surface Σ.*

Suppose that $G = H/e$ for $e = v_1v_2 \in E_H$. For $i = 1, 2$, let $d_i = \deg_H(v_i) - 1$ be the number of edges incident with v_i and distinct from e. We may assume that

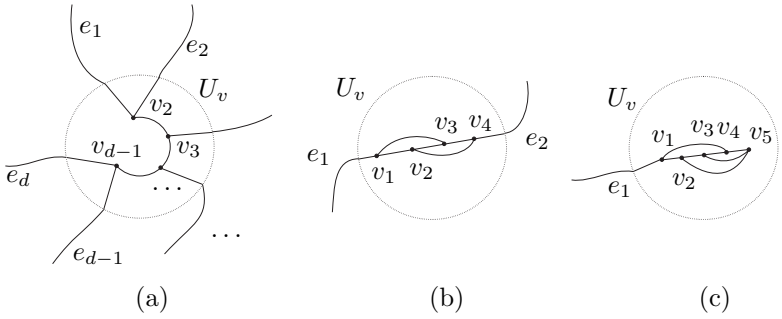


FIG. 2. *Drawing a cubic realizing graph; cf. Proposition* 1.4.

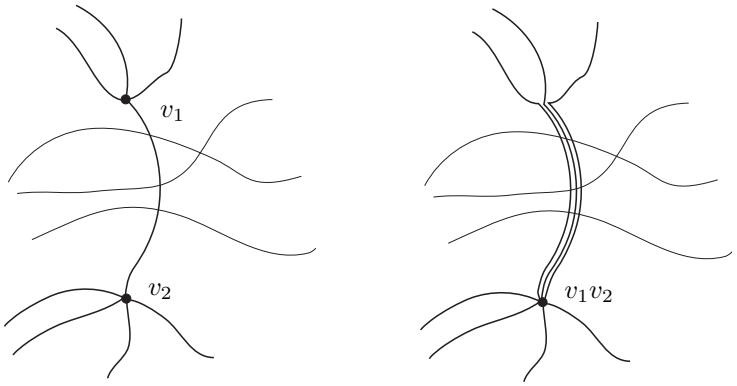


FIG. 3. *Contracting edges on a drawing.*

$d_1 \le d_2$. As shown in Figure 3, any given drawing of H can be changed into a drawing of G such that every crossing involving e is replaced by d_1 new crossings.

More generally, let G be a minor of H. We assume that $G = (H - R)/C$. Then $E_G = E_H \setminus (R \cup C)$. Let $D_H = (\varphi_H, \varepsilon_H)$ be a normal drawing of H. Then D_H determines a normal drawing of $H - R$ in Σ in which no new crossings arise. On the other hand, by contracting the edges in C, the number of crossings can increase. If we perform edge-contractions one by one, and every time apply the redrawing procedure as described above, then we can control the number of new crossings. To do the counting properly, we need some additional notation.

Let us define $w(G, H) : E_H \to \mathbb{N}$ by setting $w(G, H, e) = 0$ if $e \in R$ and $w(G, H, e) = 1$ if $e \in E_G$. If $e \in C$, let T_v be the maximal tree induced by C containing e (which contracts to the vertex v in G). Let T_1, T_2 be the components of $T_v - e$, and let d_i $(i = 1, 2)$ denote the number of edges in E_G that are incident with T_i. Then we set $w(G, H, e) = \min\{d_1, d_2\}$. For $e \in E_H$ we call $w(G, H, e)$ the *weight* of the edge e.

Let $G \le_m H_1 \le_m H$, so that $G = (H_1 - R_1)/C_1$, $H_1 = (H - R')/C'$, and $G = (H - R)/C$, where $R = R_1 \cup R'$ and $C = C_1 \cup C'$. Let D_H be a normal drawing of H. Further, let D_1 be a drawing of H_1 obtained from D_H by removing the edges of R' and applying the described contractions of the edges in C' one after another. When doing these contractions, we proceed much as shown in Figure 3 except that the criterion for whether to contract towards v_1 or v_2 is not the degree of v_1 or v_2 but the quantities d_1 or d_2 introduced in the previous paragraph. Similarly, let D_G

be obtained from D_1 by using R_1 and C_1. If D is a drawing, let $X(D)$ be the set of crossings of D, and for $x \in X(D)$ let e_x and f_x be the edges that cross at x.

LEMMA 2.2. *Let G, H, H_1 and their drawings D_G, D_H, D_1 be as defined in the previous paragraph. Then*

$$\sum_{x \in X(D_1)} w(G, H_1, e_x)\, w(G, H_1, f_x) \leq \sum_{x \in X(D_H)} w(G, H, e_x)\, w(G, H, f_x). \tag{2.1}$$

Proof. It is enough to prove this for the case when H_1 and H differ only in a single minor operation with respect to G, i.e., $R' \cup C' = \{e\}$. If $H_1 = H - e$, then $w(G, H, e) = 0$ and the sums are equal.

Suppose now that $H_1 = H/e$. As simplifying the image of e decreases the right-hand sum, we may assume that $\varepsilon_H(e)$ is a simple arc. We adopt the notation introduced above. The edge e is contracted, and thus $e \in C$. After the contraction of e, all weights remain the same; i.e., $w(G, H_1, f) = w(G, H, f)$ for every $f \in E_H - e$. Hence, the difference between the left- and the right-hand sides in (2.1) is that the crossings of e in D_H are replaced by newly introduced crossings in D_1 (as shown in Figure 3). Let $x \in X(D_H)$ with $e_x = e = v_1v_2$, and let E_1 be the set of edges incident with v_1. Since $\sum_{f \in E_1 - e} w(G, H_1, f) = \sum_{f \in E_1 - e} w(G, H, f) = w(G, H, e)$ and to each crossing x of e with some e' in D_1 there correspond exactly the crossings of $E_1 - e$ with the edge e', the inequality (2.1) follows. □

THEOREM 2.3. *Let G be a minor of a graph H, Σ be a surface, and $\tau := \left\lfloor \frac{1}{2}\Delta(G) \right\rfloor$. Then*

$$\operatorname{cr}(G, \Sigma) \leq \tau^2 \operatorname{cr}(H, \Sigma).$$

Proof. Let D_H be an optimal drawing of H, and let D_G be the drawing of G, obtained from D_H as described before Lemma 2.2. We apply Lemma 2.2 with $H_1 = G$. Obviously, $\operatorname{cr}(G, \Sigma) \leq \operatorname{cr}(D_G, \Sigma)$. As all edges in G have weight $w(G, G, e) = 1$, the left-hand side of inequality (2.1) equals the number of crossings in D_G. Since the weights $w(G, H, e)$ of edges in H are bounded from above by τ, the theorem follows. □

By using Theorem 2.3 together with definition (1.1) and Lemma 1.2, we obtain the following corollary.

COROLLARY 2.4. *Let G be a graph, Σ a surface, and $\tau := \left\lfloor \frac{1}{2}\Delta(G) \right\rfloor$. Then*

$$\operatorname{mcr}(G, \Sigma) \leq \operatorname{cr}(G, \Sigma) \leq \tau^2 \operatorname{mcr}(G, \Sigma).$$

3. Minor crossing number and genus. In this section we derive some genus-related lower bounds for minor crossing number of graphs. For additional terminology, we refer the reader to [15].

THEOREM 3.1. *Let G be a graph with genus $g(G)$ and nonorientable genus $\tilde{g}(G)$. If Σ is an orientable surface of genus $g(\Sigma)$, then $\operatorname{mcr}(G, \Sigma) \geq g(G) - g(\Sigma)$ and $\operatorname{mcr}(G, \Sigma) \geq \tilde{g}(G) - 2g(\Sigma)$.*

If Σ is a nonorientable surface with nonorientable genus $g(\Sigma)$, then $\operatorname{mcr}(G, \Sigma) \geq \tilde{g}(G) - g(\Sigma)$.

Proof. Let D be an optimal drawing of a realizing graph $\bar{G}$ in an orientable surface Σ. For each crossing in D we add a handle to Σ and obtain an embedding of $\bar{G}$ in a surface Σ' of genus $g(\Sigma') = g(\Sigma) + \operatorname{mcr}(G, \Sigma)$. Using minor operations on D, we can obtain an embedding of G in Σ', which yields $g(\Sigma') \geq g(G)$. Thus, we have $\operatorname{mcr}(G, \Sigma) \geq g(G) - g(\Sigma)$.

The other two claims can be proved in a similar way by adding crosscaps at crossings of D. Note also that adding a crosscap to an orientable surface of genus g results in a surface of nonorientable genus $2g+1$. □

When the genus of a graph is not known, one can derive the following lower bound using the Euler formula and the same technique as in the preceding proof.

PROPOSITION 3.2. *Let G be a graph with $n = |V_G|$, $m = |E_G|$, and girth r, and let Σ be a surface of Euler genus g. Then* $\operatorname{mcr}(G,\Sigma) \geq \frac{r-2}{r}m - n - g + 2$.

Proof. As in the proof of Theorem 3.1, we obtain an embedding D of G in $\mathbb{N}_{g+k}$, where $k = \operatorname{mcr}(G,\Sigma)$. Let f be the number of faces in D. All faces have length at least r, and thus $f \leq \frac{2m}{r}$. The Euler formula results in $2-(g+k) = n-m+f \leq n-\frac{r-2}{r}m$, which yields the claimed bound. □

In section 5 we derive an improvement over Proposition 3.2; see Theorem 5.6.

The following proposition relates minor crossing numbers in different surfaces with the one in the plane.

PROPOSITION 3.3. *The inequality* $\operatorname{mcr}(G,\Sigma) \leq \max(0, \operatorname{mcr}(G) - g(\Sigma))$ *holds for every surface Σ and every graph G, where $g(\Sigma)$ denotes the (non)orientable genus of Σ.*

Proof. Let us start with a realizing drawing of G in the sphere. We can remove at least one existing crossing by adding either a crosscap (if the surface is nonorientable) or a handle. This increases the genus of the surface by 1, and the result follows. □

4. Minor crossing number and connectivity. Let $G_1, \ldots, G_k$ be the components of a graph G. It is easy to see that $\operatorname{mcr}(G) = \sum_{i=1}^{k} \operatorname{mcr}(G_i)$. We shall extend this fact to the blocks (2-connected components) of G, even in the setting of the minor crossing number in a surface.

Let Σ be a surface and k a positive integer. We say that a collection $\Sigma_1, \ldots, \Sigma_k$ of surfaces is a *decomposition* of Σ and write $\Sigma = \Sigma_1 \# \cdots \# \Sigma_k$ if Σ is homeomorphic to the connected sum of $\Sigma_1, \ldots, \Sigma_k$.

THEOREM 4.1. *Let Σ be a surface and let G be a graph with blocks $G_1, \ldots, G_k$. Then*

$$\sum_{i=1}^{k} \operatorname{mcr}(G_i,\Sigma) \leq \operatorname{mcr}(G,\Sigma) \leq \min\left\{ \sum_{i=1}^{k} \operatorname{mcr}(G_i,\Sigma_i) \;\middle|\; \Sigma = \Sigma_1 \# \cdots \# \Sigma_k \right\}.$$

Proof. Let D be an optimal drawing of a realizing graph $\bar{G}$ in Σ. For each G_i it contains an induced subdrawing D_i of some graph $\tilde{G}_i$ with G_i as a minor. G_i and G_j are either disjoint (implying that $\tilde{G}_i$ and $\tilde{G}_j$ are disjoint), or they have a cutvertex v in common (implying that $\tilde{G}_i$ and $\tilde{G}_j$ intersect in a part of the tree T_v). As there are at least $\operatorname{mcr}(G_i,\Sigma)$ crossings in D_i and there are no crossings in the subdrawing induced by T_v for any $v \in V_G$, the lower bound follows.

Let us reorder the blocks of G in such a way that for $i = 2, \ldots, k$ the block G_i shares at most one vertex with the graph $H_i := \bigcup_{j=1}^{i-1} G_j$. This can be done using the block-cutvertex forest of G.

Let $\Sigma_1, \ldots, \Sigma_k$ be a decomposition of Σ where the minimum is attained. For $i = 1, \ldots, k$ let the D_i be some optimal drawing of $\bar{G}_i$ in Σ_i. Set $\tilde{D}_1 = D_1$, $\tilde{H}_1 = \bar{G}_1$, and $\Pi_1 = \Sigma_1$. For $i = 2, \ldots, k$ we choose a face f_i of $\tilde{D}_{i-1}$ in Π_{i-1} and f_i' of D_i in Σ_i. If H_{i-1} and G_i share a vertex v, then we choose f_i incident with some vertex x_i of $T_v \subseteq \tilde{H}_{i-1}$ and f_i' incident with some vertex y_i of $T_v \subseteq \bar{G}_i$; otherwise the choice can be arbitrary. By constructing a connected sum of faces f_i, f_i' and, if necessary, connecting x_i with y_i in the new face $f_i \# f_i'$, we obtain a drawing $\tilde{D}_i$ of $\tilde{H}_i$ in $\Pi_i := \Pi_{i-1} \# \Sigma_i$.

It is clear that $G \leq_m \tilde{H}_k$ and that $\tilde{D}_k$ is a drawing of $\tilde{H}_k$ in Σ with at most $\sum_{i=1}^{k} \mathrm{mcr}_{\Sigma_i}(G_i)$ crossings. This proves the upper bound inequality. □

COROLLARY 4.2. *Let G be a graph with blocks $G_1, \ldots, G_k$. Then*

$$\mathrm{mcr}(G) = \sum_{i=1}^{k} \mathrm{mcr}(G_i).$$

Proof. To prove this, one just has to observe that, for $\Sigma = \mathbb{S}_0$, the left-hand side and the right-hand side in the inequalities in Theorem 4.1 are equal. □

5. Structure of graphs with bounded $\mathrm{mcr}(G, \Sigma)$. As mentioned in section 1, the family $\omega(k, \Sigma)$ of all graphs, whose $\mathrm{mcr}(G, \Sigma)$ is at most k, is minor-closed. Let us denote by $F(k, \Sigma)$ the set of minimal forbidden minors for $\omega(k, \Sigma)$. $F(k)$ and $\omega(k)$ stand for $F(k, \mathbb{S}_0)$ and $\omega(k, \mathbb{S}_0)$, respectively.

Graphs in $\omega(0, \Sigma)$ have a simple topological characterization—they are precisely the graphs that can be embedded in Σ. A similar topological characterization holds for graphs in $\omega(1)$. They are precisely the graphs that can be embedded in the projective plane with face-width at most 2. This was observed by Robertson and Seymour [18], who determined the set $F(1)$ of minimal forbidden minors for $\omega(1)$ as follows.

THEOREM 5.1 (see Robertson and Seymour [18]). *The set $F(1)$ contains precisely the 41 graphs $G_1, \ldots, G_{35}$ and $Q_1, \ldots, Q_6$, where $G_1, \ldots, G_{35}$ are the minimal forbidden minors for embeddability in the projective plane and $Q_1, \ldots, Q_6$ are* projective planar *graphs that can be obtained from the Petersen graph by successively applying the $Y\Delta$ and ΔY operations.*

This theorem establishes the following linear time algorithm for testing whether $\mathrm{mcr}(G)$ is at most 1: first embed G in the projective plane [14] and then check whether the face-width of the embedding is less than or equal to 2 (see [8]).

Let us remark that the forbidden minors for the projective plane have been determined by Glover, Huneke, and Wang [7] and Archdeacon [2]. There are seven graphs that can be obtained from the Petersen graph by $Y\Delta$ and ΔY operations (known as the Petersen family), but one of them cannot be embedded in the projective plane and is one of the forbidden minors for the projective plane.

We will prove that every family $\omega(k, \Sigma)$ has a similar topological representation, for which we need some further definitions.

Let γ be a one-sided simple closed curve in a nonorientable surface Π of Euler genus g. Cutting Π along γ and pasting a disk to the resulting boundary yields a surface denoted by Π/γ of Euler genus $g-1$. We say that Π/γ is obtained from Π by *annihilating* a crosscap at γ.

Let us call a set of pairwise noncrossing, onesided, simple closed curves $\Gamma = \{\gamma_1, \ldots, \gamma_k\}$ in a nonorientable surface Π a *k-system* in Π. It is easy to see that for distinct $\gamma_i, \gamma_j \in \Gamma$ the surface $(\Pi/\gamma_i)/\gamma_j$ is homeomorphic to $(\Pi/\gamma_j)/\gamma_i$. Therefore the order in which we annihilate the crosscaps at prescribed curves is irrelevant, and we define $\Pi/\Gamma := \Pi/\gamma_1/\ldots/\gamma_k$. We say that the k-system Γ in Π is an *orienting k-system* if the surface Π/Γ is orientable.

Suppose that D is a drawing of G in a nonorientable surface Π with at most c crossings. If there exists an (orienting) k-system Γ in Π with each $\gamma \in \Gamma$ intersecting D in at most two points, then we say that D is (*orientably*) *(c, k)-degenerate*, and we call Γ an (*orienting*) *k-system* of D. If $c = 0$, then D is an embedding and we also say that it is *k-degenerate*. Let us observe that an embedding of a graph in the projective plane is 1-degenerate precisely when the face-width of the embedding is at most 2.

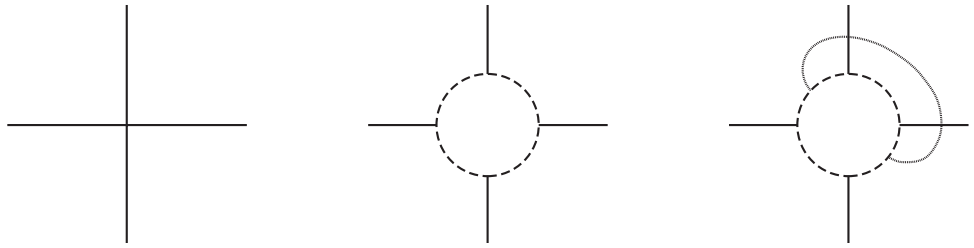


FIG. 4. *Replacing a crossing by a crosscap and a respective annihilating curve.*

LEMMA 5.2. *Let Σ be an (orientable) surface of Euler genus g, and let $k \geq 1$ be an integer. Then, for any $l \in \{1, \ldots, k\}$, the family $\omega(k, \Sigma)$ consists precisely of all those graphs $G \in \omega(k-l, \mathbb{N}_{g+l})$, for which there exists a graph $\tilde{G}$ that contains G as a minor and that can be drawn in the nonorientable surface $\mathbb{N}_{g+l}$ of Euler genus $g+l$ with (orienting) degeneracy $(k-l, l)$.*

Proof. Let $G \in \omega(k, \Sigma)$ and let $\bar{G}$ be its realizing graph, drawn in Σ with at most k crossings. Choose a subset of l crossings of $\bar{G}$. By replacing a small disk around each of the chosen crossings with a Möbius band, we obtain a drawing of $\bar{G}$ in $\mathbb{N}_{g+l}$ with (orienting) degeneracy $(k-l, l)$. The replacement at one such crossing and the corresponding curve annihilating the crosscap are illustrated in Figure 4.

For the converse we first prove the induction basis $l = 1$.

Let $\tilde{G}$ be the graph that contains G as a minor and is drawn in $\mathbb{N}_{g+1}$ with at most $k-1$ crossings, and let us assume that a one-sided curve γ intersects the drawing of $\tilde{G}$ in at most two points, x and y. After cutting the surface along γ and pasting a disc Δ on the resulting boundary, we get a surface of Euler genus g. On the boundary of Δ, two copies of x and y interlace. By adding paths P_x and P_y joining the copies of x and y (respectively), we obtain a drawing D' of a graph G', which contains $\tilde{G}$ (and hence also G) as a minor. Clearly, D' has one crossing more (the one between P_x and P_y) than the drawing of $\tilde{G}$. So, D' is $(k-1, 1)$-degenerate.

If $l \geq 2$, we may annihilate the crosscaps consecutively, as the curves in the corresponding l-system are noncrossing. Note that if the l-system is orienting, we obtain an orientable surface Σ. □

LEMMA 5.3. *Let $\tilde{G}$ be a graph with an (orientably) k-degenerate embedding in a surface Σ. If G is a surface minor of $\tilde{G}$, then G is also (orientably) k-degenerate.*

Proof. It suffices to verify the claim for edge-deletions and edge-contractions. For edge-deletions, there is nothing to be proved, and for edge contractions, one has to show only that a k-system for $\tilde{G}$ can be transformed into a k-system for $\tilde{G}/e$. We leave the details to the reader. □

Lemma 5.3 can be extended to drawings with crossings if we restrict edge-contraction to edges that are not involved in crossings.

As a direct consequence of Lemmas 5.2 and 5.3 we have the following result.

THEOREM 5.4. *Let Σ be an (orientable) surface of Euler genus g, and let $k \geq 1$ be an integer. Then $\omega(k, \Sigma)$ consists of precisely all the graphs that can be embedded in the nonorientable surface $\mathbb{N}_{g+k}$ of Euler genus $g+k$ with (orienting) degeneracy k.*

Figure 5(a) exhibits the geometric structure of a realizing drawing in the Klein bottle, (b) shows the general structure of its minors G with $\mathrm{mcr}(G) \leq 2$, and (c) is a degenerate example of this structure in which the curves of the corresponding

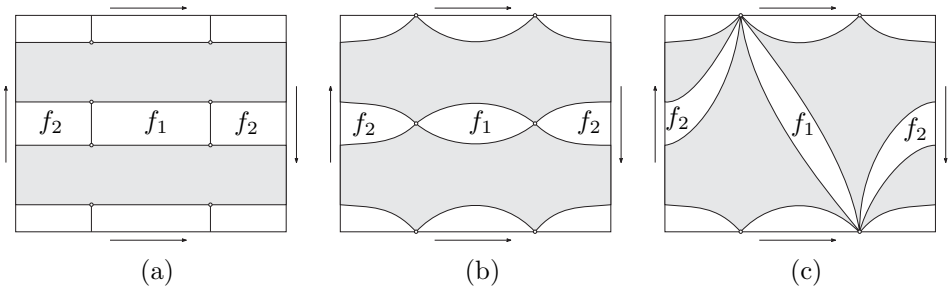


FIG. 5. *Embeddings in the Klein bottle with orienting degeneracy* 2.

2-system $\{\gamma_1, \gamma_2\}$ touch twice.

Theorem 5.4 can be used to express a more intimate relationship between the graphs in $\omega(k, \Sigma)$ and $\omega(0, \Sigma)$, as follows.

COROLLARY 5.5. *Let Σ be a surface of Euler genus g, $k \geq 0$ be an integer, and $G \in \omega(k, \Sigma)$. Then there exists a graph H, which embeds in Σ, such that G can be obtained from H by identifying at most k pairs of vertices.*

Theorem 5.4 can be used to improve the lower bound of Proposition 3.2.

THEOREM 5.6. *Let G be a simple graph with $n = |V_G|$, $m = |E_G|$ and let Σ be a surface of Euler genus g. Then*

$$\operatorname{mcr}(G, \Sigma) \geq \tfrac{1}{2}(m - 3(n + g) + 6).$$

Two technical lemmas are needed for the proof of this result. Let Σ be a closed surface and $x, y \in \Sigma$. Let $\Gamma = \{\gamma_1, \ldots, \gamma_k\}$ be a k-system of one-sided noncrossing simple closed curves in Σ such that $\gamma_i \cap \gamma_j = \{x, y\}$ for all $1 \leq i < j \leq k$. Let $\gamma_i = \gamma_i^1 \cup \gamma_i^2$, where γ_i^l is an arc from x to y. If a curve $\gamma_i^l \cup \gamma_j^m$ $(i \neq j)$ bounds a disk in Σ whose interior contains no segment of curves in Γ, then we say that $\gamma_i^l \cup \gamma_j^m$ is a Γ-*digon*.

LEMMA 5.7. *Every k-system Γ has at most $k - 1$ Γ-digons.*

Proof. We assume the notation introduced above. Let us contract one of the segments, say γ_1^1. Then each other γ_i^l becomes a loop in Σ. Since Γ is a k-system of one-sided noncrossing loops, the loops in Γ generate a k-dimensional subspace of the first homology group $H_1(\Sigma; \mathbb{Z}_2)$. This implies that the $2k - 1$ loops $L = \{\gamma_i^l \mid 1 \leq i \leq k, l = 1, 2\} \setminus \{\gamma_1^1\}$ also generate at least k-dimensional subspace. If there are k Γ-digons, then k of the loops could be removed from L, and the remaining $k - 1$ loops would still generate the same k-dimensional subspace. This contradiction completes the proof. □

Let G be a graph and D its k-degenerate embedding in a surface Σ. Let $\Gamma = \{\gamma_1, \ldots, \gamma_k\}$ be the corresponding k-system of D. The curves γ_i are pairwise noncrossing, so we may assume that γ_i and γ_j $(i \neq j)$ intersect (touch) only in points where they intersect the graph. We subdivide edges of D in such a way that every γ_i intersects D only at vertices. If γ_i intersects D at vertices u_i and v_i, we add to D two new edges e_i, f_i with ends u_i, v_i whose embedding in Σ coincides with γ_i. (If $u_i = v_i$, we add one loop e_i at v_i.) We call the resulting embedding D' a k-*augmented embedding* of D and the corresponding graph G' a k-*augmented graph* of G (with respect to Γ). Let us observe that we may assume that curves in Γ intersect D only at vertices. In that case, subdivision of edges is not necessary, and then G is a subgraph of G'.

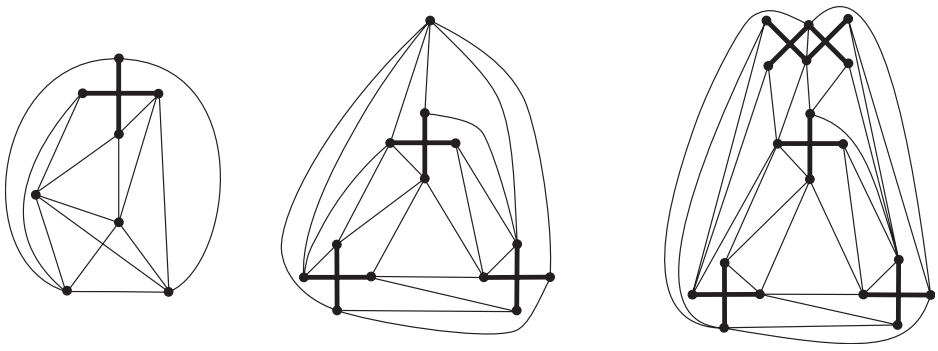


FIG. 6. *Realizing drawings of K_6, K_7, and K_8, respectively.*

LEMMA 5.8. *Let D be a k-degenerate embedding of a simple graph G in a nonorientable surface Σ, and let D' be a k-augmented embedding of D. Then D' has at most k faces of length two and has no faces of length one.*

Proof. We shall use the notation introduced before the lemma. Since G is a simple graph, any face of length 1 or 2 involves some edge e_i, f_i ($i \in \{1, \dots, k\}$). If e_i is a loop, it cannot bound a face since γ_i is a onesided curve in Σ. Two loops cannot form a facial boundary since then they would be homotopic, and homotopic onesided curves always cross each other. So, an edge e_i or f_i can be part of a face of length two only when $u_i \neq v_i$.

For simplicity of notation, suppose that $\gamma_1, \dots, \gamma_t$ all contain the same pair of vertices u_1 and v_1. It suffices to see that the edges e_i, f_i ($i = 1, \dots, t$) and possible edge $e_0 = u_1v_1$ of G together form at most t faces of length 2. By Lemma 5.7, $\{e_i, f_i \mid 1 \leq i \leq t\}$ form at most $t-1$ faces of length 2, and e_0 can give rise to one additional such face. This proves the claim, and the application of this claim to all pairs u_i, v_i completes the proof of the lemma. □

Proof of Theorem 5.6. Let $\mathrm{mcr}(G, \Sigma) = k$. By Theorem 5.4, there exists an embedding D of G in $\mathbb{N}_{g+k}$ with crossing degeneracy k. Let D' be a k-augmented embedding of D, and let G' be its graph. By Lemma 5.8, removing at most k edges from G' yields an embedding D'' without faces of length two, implying $|F_{D''}| \leq \frac{2}{3}|E_{D''}|$. Euler formula implies $n - |E_{D''}| + |F_{D''}| = 2 - (g+k)$. The stated inequality follows. □

If one would like to extend the bound of Proposition 3.2 for graphs of girth $r \geq 4$, additional arguments would be needed.

6. Examples. We have so far developed some tools to find lower bounds of the minor crossing number. In this section, they are applied to several families of graphs. In general, Theorem 2.3 yields better bounds for graphs of small maximum degree (cubes, $C_n \square C_m$), while Theorem 3.1 suits graphs with large maximum degree better, e.g., complete bipartite graphs. Theorem 5.6 performs best on dense graphs of girth three, for instance complete graphs.

6.1. Complete graphs. Theorem 5.6 implies the following inequality, which is sharp for $n \in \{3, \dots, 8\}$, as demonstrated in Figure 6.

PROPOSITION 6.1. *Let $n \geq 3$. Then* $\mathrm{mcr}(K_n) \geq \left\lceil \frac{1}{4}(n-3)(n-4) \right\rceil$.

The following proposition establishes an upper bound.

PROPOSITION 6.2. *For $n \geq 9$,* $\mathrm{mcr}(K_n) \leq \left\lfloor \frac{1}{2}(n-5)^2 \right\rfloor + 4$.

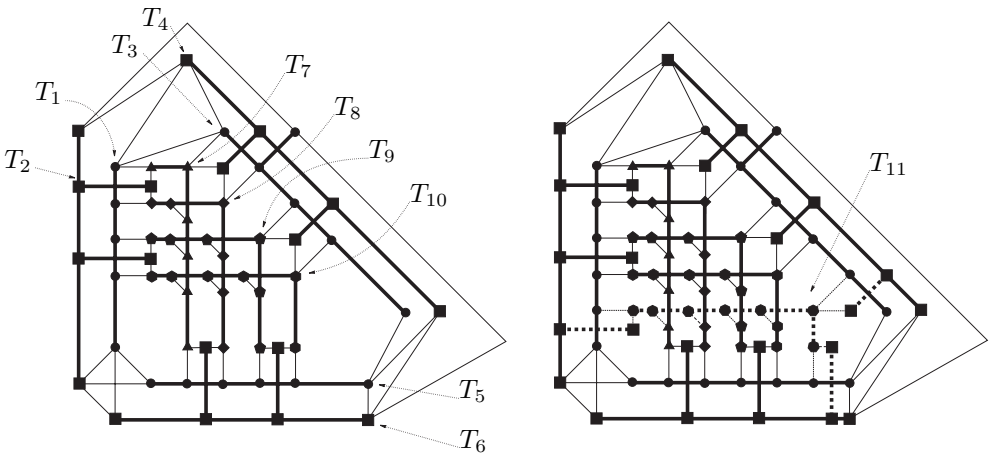


FIG. 7. *Drawings of graphs $\tilde{K}_{10}$ and $\tilde{K}_{11}$.*

Proof (sketch). We shall exhibit graphs $\tilde{K}_n$ ($n \geq 9$) together with their drawings D_n so that $\tilde{K}_n$ contains K_n as a minor and that $\mathrm{cr}(D_n) = \left\lfloor \frac{1}{2}(n-5)^2 \right\rfloor + 4$. Figure 7 presents drawings of $\tilde{K}_{10}$ and $\tilde{K}_{11}$. Different vertex symbols (diamond, circle, triangle, etc.) represent vertices in the same tree T_v, $v \in V_{K_n}$, which contracts to the vertex v in the K_n minor. By contracting the thick edges of the graphs in Figure 7, we obtain K_{10} and K_{11}, respectively.

The reader should have no difficulty placing the tree T_{n+1} into D_n in order to obtain D_{n+1}. The tree T_{n+1} crosses precisely each T_v with $7 \leq v \leq n$. To connect T_{n+1} with the trees $T_1, \ldots, T_6$, we need three new crossings if n is even (T_1 with T_2, T_3 with T_4, and T_5 with T_6) and no new crossing if n is odd.

Let c_n denote the number of crossings in the drawing of $\tilde{K}_n$ described above, and let $a_k = c_{2k}$. We have $a_4 = 6$, $a_5 = 14$, $a_6 = 26$, and a recurrence equation

$$\begin{aligned} a_{k+1} = c_{2k+2} &= c_{2k+1} + (2k-1-6) \\ &= c_{2k} + (2k-6) + 3 + (2k-1-6) \\ &= c_{2k} + 4k - 8 \\ &= a_k + 4k - 8, \end{aligned}$$

whose solution is $a_k = 2k^2 - 10k + 14$. For even n this yields

$$c_n = \tfrac{1}{2}((n-5)^2 + 3),$$

and for odd n

$$c_n = \tfrac{1}{2}(n-5)^2 + 4. \qquad \square$$

COROLLARY 6.3. *Let Σ be a fixed surface. For $n \in \mathbb{N}$, let $c_n = \frac{\mathrm{mcr}(K_n,\Sigma)}{n(n-1)}$. The sequence $\{c_n\}_{n=1}^{\infty}$ is nondecreasing and*

$$c_\infty := \lim_{n\to\infty} c_n \in \left[\tfrac{1}{4}, \tfrac{1}{2}\right].$$

Proof. First we prove the following claim: Let $\mathrm{mcr}(K_n, \Sigma) \geq c\,n(n-1)$. Then $\mathrm{mcr}(K_m, \Sigma) \geq c\,m(m-1)$ for every $m \geq n$.

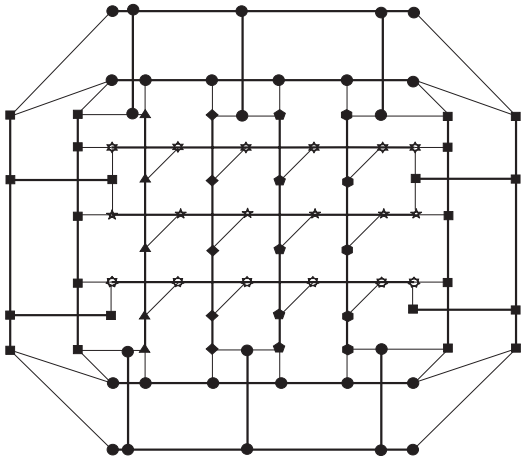


FIG. 8. *A drawing of the graph* $\tilde{K}_{8,7}$ *with* 22 *crossings.*

Clearly it suffices to prove this for $m = n + 1$. Let $\bar{D}$ be a realizing drawing of K_{n+1} in Σ. Let T_i be the tree in $\bar{D}$ which contracts to the vertex i of K_{n+1}. If we remove T_i and all incident edges from $\bar{D}$, we obtain a drawing of a graph with K_n minor. This can be done in $n+1$ different ways. These $n+1$ drawings contain at least $(n+1)\operatorname{mcr}(K_n, \Sigma)$ crossings altogether. We may assume that there are no removed edges in $\bar{D}$, as their number can only increase the number of crossings. Then each crossing from $\bar{D}$ appears in at most $n-1$ of these drawings. Therefore, $(n-1)\operatorname{mcr}(K_{n+1}, \Sigma) \geq (n+1)\operatorname{mcr}(K_n, \Sigma) \geq c\,(n+1)n(n-1)$.

The stated bounds on c_∞ follow easily from Propositions 6.1 and 6.2. □

We believe that the minor crossing number of complete graphs lies close to the upper bound from Proposition 6.2, so that the following asymptotic holds: $\operatorname{mcr}(K_n) = \frac{1}{2}n^2 + O(n)$.

6.2. Complete bipartite graphs. The genus of complete bipartite graphs [15, Theorem 4.4.7] in combination with Theorem 3.1 establishes the following proposition.

PROPOSITION 6.4. *Let* $3 \leq m \leq n$. *Then*

$$\operatorname{mcr}(K_{m,n}) \geq \left\lceil \tfrac{1}{2}(m-2)(n-2) \right\rceil .$$

For the upper bound, consider a set of graphs $\tilde{K}_{m,n}$. They are constructed in a way similar to that of their complete analogues $\tilde{K}_n$, and an example is presented in Figure 8.

PROPOSITION 6.5. *Let* $4 \leq m \leq n$. *Then*

$$\operatorname{mcr}(K_{m,n}) \leq (m-3)(n-3) + 5.$$

Also in the case of complete bipartite graphs we think that the upper bound from Proposition 6.5 lies close to the actual minor crossing number: $\operatorname{mcr}(K_{m,n}) = m \cdot n + O(m+n)$.

6.3. Hypercubes. Applying Proposition 3.2 to hypercubes yields the following result.

PROPOSITION 6.6. *Let* $n \geq 4$. *Then* $\operatorname{mcr}(Q_n) \geq (n-4)2^{n-2} + 2$.

Using the best known lower bound for crossing number of hypercubes, $\mathrm{cr}(Q_n) > 4^n/20 - (n^2+1)2^{n-1}$ by Sýkora and Vrťo [19] in combination with Theorem 2.3, we can deduce the following alternative lower bound, which is stronger for large values of n.

PROPOSITION 6.7. *Let $n \geq 4$. Then* $\mathrm{mcr}(Q_n) > \frac{1}{n^2}\left(\frac{1}{5}\,4^n - 2^{n+1}\right) - 2^{n+1}$.

Following the same idea as in [12, Figures 2 and 3], one can obtain a family of graphs $\tilde{Q}_n$ and their drawings D_n with $\Delta(\tilde{Q}_n) = 4$ and $\tilde{Q}_n$ having Q_n as a minor. They establish the following upper bound.

PROPOSITION 6.8. *Let $n \geq 2$. Then* $\mathrm{mcr}(Q_n) \leq 2 \cdot 4^{n-2} - (n-1)2^{n-1}$.

6.4. Cartesian products of cycles $C_m \Box C_n$. Combining the results presented in [5] and Theorem 2.3 implies the following fact.

PROPOSITION 6.9. *Suppose that $n \geq m$ and either $m \leq 7$ or $m \geq 3$ and $n \geq m(m+1)$. Then $\frac{1}{4}(m-2)n \leq \mathrm{mcr}(C_m \Box C_n) \leq (m-2)n$.*

REFERENCES

[1] D. ARCHDEACON, *Problems in Topological Graph Theory*, Department of Mathematics and Statistics, University of Vermont, Burlington, VT; online at http://www.emba.uvm.edu/~archdeac/problems/minorcr.htm, 1995.

[2] D. ARCHDEACON, *A Kuratowski theorem for the projective plane*, J. Graph Theory, 5 (1981), pp. 243–246.

[3] S. N. BHATT AND F. T. LEIGHTON, *A framework for solving VLSI graph layout problems*, J. Comput. System Sci., 28 (1984), pp. 300–343.

[4] G. FIJAVŽ, *Graph Minors and Connectivity*, Ph.D. thesis, Department of Mathematics, University of Ljubljana, Slovenia, 2001 (in Slovene).

[5] L. Y. GLEBSKY AND G. SALAZAR, *The crossing number of $C_m \Box C_n$ is as conjectured for $n \geq m(m+1)$*, J. Graph Theory, 47 (2004), pp. 53–72.

[6] P. HLINĚNÝ, *Crossing number is hard for cubic graphs (extended abstract)*, in Math Foundations of Computer Science MFCS 2004, Lecture Notes in Comput. Sci. 3153, Springer-Verlag, New York, 2004, pp. 772–782.

[7] H. H. GLOVER, J. P. HUNEKE, AND C.-S. WANG, 103 *graphs that are irreducible for the projective plane*, J. Combin. Theory Ser. B, 27 (1979), pp. 332–370.

[8] M. JUVAN AND B. MOHAR, *An Algorithm for Embedding Graphs in the Torus*, manuscript.

[9] K. KURATOWSKI, *Sur le problème des courbes gauches en topologie*, Fund. Math., 15 (1930), pp. 271–283.

[10] F. T. LEIGHTON, *Complexity Issues in VLSI*, MIT Press, Cambridge, MA, 1983.

[11] F. T. LEIGHTON, *New lower bound techniques for VLSI*, Math. Systems Theory, 17 (1984), pp. 47–70.

[12] T. MADEJ, *Bounds for the crossing number of the N-cube*, J. Graph Theory, 15 (1991), pp. 81–97.

[13] D. MCQUILLAN AND R. B. RICHTER, *On 3-regular graphs having crossing number at least 2*, J. Graph Theory, 18 (1994), pp. 831–893.

[14] B. MOHAR, *Projective planarity in linear time*, J. Algorithms, 15 (1993), pp. 482–502.

[15] B. MOHAR AND C. THOMASSEN, *Graphs on Surfaces*, Johns Hopkins University Press, Baltimore, MD, 2001.

[16] E. G. MORENO AND G. SALAZAR, *Bounding the crossing number of a graph in terms of the crossing number of a minor with small maximum degree*, J. Graph Theory, 36 (2001), pp. 168–173.

[17] R. B. RICHTER, *Cubic graphs with crossing number two*, J. Graph Theory, 12 (1988), pp. 363–374.

[18] N. ROBERTSON AND P. SEYMOUR, *Excluding a graph with one crossing*, Contemp. Math., 147 (1993), pp. 669–675.

[19] O. SÝKORA AND I. VRŤO, *On crossing numbers of hypercubes and cube connected cycles*, BIT, 33 (1993), pp. 232–237.

[20] I. VRŤO, *Crossing number of graphs: A bibliography*, Institute of Mathematics, Slovak Academy of Sciences, Bratislava, Slovak Republic; available online from ftp://ftp.ifi.savba.sk/pub/imrich/crobib.pdf.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 357–371

THE BIDIMENSIONAL THEORY OF BOUNDED-GENUS GRAPHS*

ERIK D. DEMAINE†, MOHAMMADTAGHI HAJIAGHAYI†, AND DIMITRIOS M. THILIKOS‡

Abstract. Bidimensionality provides a tool for developing subexponential fixed-parameter algorithms for combinatorial optimization problems on graph families that exclude a minor. This paper extends the theory of bidimensionality for graphs of bounded genus (which is a minor-excluding family). Specifically we show that, for any problem whose solution value does not increase under contractions and whose solution value is large on a grid graph augmented by a bounded number of handles, the treewidth of any bounded-genus graph is at most a constant factor larger than the square root of the problem's solution value on that graph. Such bidimensional problems include vertex cover, feedback vertex set, minimum maximal matching, dominating set, edge dominating set, r-dominating set, connected dominating set, planar set cover, and diameter. On the algorithmic side, by showing that an augmented grid is the prototype bounded-genus graph, we generalize and simplify many existing algorithms for such problems in graph classes excluding a minor. On the combinatorial side, our result is a step toward a theory of graph contractions analogous to the seminal theory of graph minors by Robertson and Seymour.

Key words. treewidth, grids, graphs on surfaces, graph minors, graph contractions

AMS subject classifications. 05C83, 05C85

DOI. 10.1137/040616929

1. Introduction.

The recent theory of fixed-parameter algorithms and parameterized complexity [DF99] has attracted much attention in its less than 10 years of existence. In general the goal is to understand when NP-hard problems have algorithms that are exponential only in a parameter k of the problem instead of the problem size n. Fixed-parameter algorithms whose running time is polynomial for fixed parameter values—or more precisely $f(k) \cdot n^{O(1)}$ for some (superpolynomial) function $f(k)$—make these problems efficiently solvable whenever the parameter k is reasonably small.

In the last five years, several researchers have obtained exponential speedups in fixed-parameter algorithms for various problems on several classes of graphs. While most previous fixed-parameter algorithms have a running time of $2^{O(k)}n^{O(1)}$ or worse, the exponential speedups result in subexponential algorithms with typical running times of $2^{O(\sqrt{k})}n^{O(1)}$. For example, the first fixed-parameter algorithm for finding a dominating set of size k in planar graphs [AFF+01] had running time $O(8^k n)$; subsequently, a sequence of subexponential algorithms and improvements have been obtained, starting with running time $O(4^{6\sqrt{34k}}n)$ [ABF+02], then $O(2^{27\sqrt{k}}n)$ [KP02], and finally $O(2^{15.13\sqrt{k}}k + n^3 + k^4)$ [FT03]. Other subexponential algorithms for

*Received by the editors October 13, 2004; accepted for publication (in revised form) August 29, 2005; published electronically May 3, 2006. A preliminary version of this paper appeared in *Proceedings of the 29th International Symposium on Mathematical Foundations of Computer Science*, Prague, 2004, pp. 191–203.

http://www.siam.org/journals/sidma/20-2/61692.html

†MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar St., Cambridge, MA 02139 (edemaine@mit.edu, hajiagha@mit.edu).

‡Department of Mathematics, National and Capodistrian University of Athens, Panepistimioupolis, GR-15784, Athens, Greece (sedthilk@math.uoa.gr). This author's work was supported by the EU within the 6th Framework Programme under contract 001907 (DELIS) and by the Spanish CICYT project TIC-2002-04498-C05-03 (TRACER).

other domination and covering problems on planar graphs have also been obtained [ABF+02, AFN04, CKL01, KLL02, GKLY05].

All subexponential fixed-parameter algorithms developed so far are based on showing a "treewidth-parameter bound": Any graph whose optimal solution has value k has treewidth at most some function $f(k)$. In many cases, $f(k)$ is sublinear in k, often $O(\sqrt{k})$. Combined with algorithms that are singly exponential in treewidth and polynomial in problem size, such a bound immediately leads to subexponential fixed-parameter algorithms.

A series of papers [DFHT05, DFHT04b, DFHT04a] introduce the notion of *bidimensionality* as a general approach for obtaining treewidth-parameter bounds and therefore subexponential algorithms. This theory captures essentially all subexponential algorithms obtained so far. Roughly speaking, a parameterized problem is *bidimensional* if the parameter is large in a "grid-like graph" (linear in the number of vertices) and either closed under contractions (*contraction-bidimensional*) or closed under minors (*minor-bidimensional*). Examples of bidimensional problems include vertex cover, feedback vertex set, minimum maximal matching, dominating set, edge dominating set, r-dominating set,[1] connected dominating set, planar set cover, and diameter. Diameter is a simple computational problem, but its bidimensionality has important consequences as it forms the basis of locally bounded treewidth for minor-closed graph families [DH04a].

Treewidth-parameter bounds have been established for all minor-bidimensional problems in H-minor-free graphs for any fixed graph H [DFHT04b, DFHT04a]. In this case, the notion of "grid-like graph" is precisely the regular $(r \times r)$-square grid. However, contraction-bidimensional problems (such as dominating set) have proved substantially harder. In particular, the largest class of graphs for which a treewidth-parameter bound can be obtained is apex-minor-free graphs instead of general H-minor-free graphs [DFHT04a]. ("Apex-minor-free" means "H-minor-free" where H is a graph in which the removal of one vertex leaves a planar graph.) Such a treewidth-parameter bound has been obtained for all contraction-bidimensional problems in apex-minor-free graphs [DFHT04a]. In this case, the notion of a "grid-like graph" is an $r \times r$ grid augmented with additional edges such that each vertex is incident to $O(1)$ edges to nonboundary vertices of the grid. (Here $O(1)$ depends on H.) Unfortunately, this treewidth-parameter bound is large: $f(k) = (\sqrt{k})^{O(\sqrt{k})}$. For a subexponential algorithm, we essentially need $f(k) = o(k)$. For apex-minor-free graphs, such a bound is known only for the special cases of dominating set and vertex cover [DH04b, DFHT04b].

The biggest graph classes for which we know a sublinear (indeed, $O(\sqrt{k})$) treewidth-parameter bound for many contraction-bidimensional problems are single-crossing-minor-free graphs and bounded-genus graphs. ("Single-crossing-minor-free" means "H-minor-free" where H is a minor of a graph that can be drawn in the plane with one crossing.) For single-crossing-minor-free graphs [DHT05, DHN+04] (in particular, planar graphs [DFHT05]), all contraction-bidimensional problems have a bound of $f(k) = O(\sqrt{k})$. In this case, the notion of "grid-like graph" is an $r \times r$ grid partially triangulated by additional edges that preserve planarity. For bounded-genus graphs [DFHT04b], a bound of $f(k) = O(\sqrt{k})$ has been shown, for the same notion of "grid-like graphs" but only for contraction-bidimensional problems with an additional

[1] A set S of vertices is an *r-dominating set* of graph G if any vertex of G has distance at most r from some vertex in S.

property called α-*splittability*: Upon splitting a vertex, the parameter should increase by at most $\alpha = O(1)$ (or decrease).

In this paper we extend the theory of bidimensionality for bounded-genus graphs by establishing a sublinear ($f(k) = O(\sqrt{k})$) treewidth-parameter bound for general contraction-bidimensional problems in bounded-genus graphs. Our notion of "grid-like graph" is somewhat broader: a partially triangulated $r \times r$ grid (as above) with up to g additional edges ("handles"), where g is the genus of the original graph. This form of contraction-bidimensionality is more general than α-splittability,[2] and thus we generalize the results for α-splittable contraction-bidimensional problems from [DFHT04b]. It is easy to construct a parameter that is contraction-bidimensional but not α-splittable, although these parameters are not "natural." So far all "natural" contraction-bidimensional parameters we have encountered are α-splittable, though we expect other interesting problems to arise that violate α-splittability.

Our results show that a partially triangulated grid with g additional edges is the prototype graph of genus g, as observed by Lovász [Lov03]. At a high level, this property means that, to solve an (algorithmic or combinatorial) problem on a general graph of genus g, the "hardest" instance on which we should focus is the prototype graph. This property generalizes the well-known result in graph theory that the grid is the prototype planar graph. This also extends our theory of constructing such prototypes for bidimensional problems.

Further algorithmic applications of our results follow from the graph-minor theory of Robertson and Seymour (e.g., [RS85]) and its extensions [DFHT04b, DH04b]. In particular, [RS03, DFHT04b] show how to reduce many problems on general H-minor-free graphs to subproblems on bounded-genus graphs. Essentially, the difference between bounded-genus graphs and H-minor-free graphs are "apices" and "vortices," which are usually not an algorithmic barrier. Applying our new theory for bounded-genus graphs, we generalize the algorithmic extensions of [DFHT04b, DH04b]. Indeed, we simplify the approaches of both [DFHT04b] and [DH04b], where it was necessary to "split" bounded-genus graphs into essentially planar graphs because of a lack of general understanding of bounded-genus graphs. Specifically, we remove the necessity of Lemmas 7.4–7.8 in [DH04b].

Last but not least are the combinatorial aspects of our results. In a series of 20 papers (so far), Robertson and Seymour (e.g., [RS85]) developed the seminal theory of graphs excluding a minor, which has had many algorithmic and combinatorial applications. Our understanding of contraction-bidimensional parameters can be viewed as a step toward generalizing the theory of graph minors to a theory of graph contractions. Specifically, we show that any graph of genus g can be contracted to its core of a partially triangulated grid with at most g additional edges; this result generalizes an analogous result from [RS03] when permitting arbitrary minor operations (contractions and edge deletions). Avoiding edge deletions in this sense is particularly important for algorithmic applications because many parameters are not closed under edge deletions, while many parameters are closed under contraction.

This paper is part of a series of papers on bidimensionality [DHT05, DHN$^+$04, DFHT05, DH04a, DFHT04b, DH04b, DFHT04a, DH05b, DH05a]. The theory of bidimensionality has become a comprehensive body of algorithmic and combinatorial results, with consequences including tight parameter-treewidth bounds, direct separator

[2] This statement is the contrapositive of the following property: If the parameter is k for the partially triangulated grid with g additional edges, then by α-splitting the additional edges, the parameter is at most $k + \alpha g$ on the partially triangulated grid.

theorems, linearity of local treewidth, subexponential fixed-parameter algorithms, and polynomial-time approximation schemes for a broad class of problems on graphs that exclude a fixed minor. See [DH04c] for a survey of this work and the role of this paper. In particular, the results of this paper are used in the subsequent papers [DH05b, DH05a].

2. Preliminaries. All the graphs in this paper are undirected without loops or multiple edges. Given a graph G, we denote by $V(G)$ the set of its vertices and by $E(G)$ the set of its edges. For any vertex $v \in V(G)$ we denote by E_v the set of edges incident to v. Moreover, we use the notation $N_G(v)$ (or simply $N(v)$) for the set of neighbors of v in G (i.e., vertices adjacent to v).

Given an edge $e = \{x, y\}$ of a graph G, the graph obtained from G by contracting the edge e is the graph we get if we identify the vertices x and y and remove all loops and duplicate edges. A graph H obtained by a sequence of edge-contractions is said to be a *contraction* of G. A graph class $\mathcal{C}$ is a *contraction-closed* class if any contraction of any graph in $\mathcal{C}$ is also a member of $\mathcal{C}$. A contraction-closed graph class $\mathcal{C}$ is *H-contraction-free* if $H \notin \mathcal{C}$. Given any graph class $\mathcal{H}$, we say that a contraction-closed graph class $\mathcal{C}$ is *$\mathcal{H}$-contraction-free* if $\mathcal{C}$ is H-contraction-free for *any* $H \in \mathcal{H}$.

2.1. Treewidth and branchwidth. The notion of treewidth was introduced by Robertson and Seymour [RS86] and plays an important role in their fundamental work on graph minors. To define this notion, first we consider the representation of a graph as a tree, which is the basis of our algorithms in this paper. A *tree decomposition* of a graph G, denoted by $TD(G)$, is a pair (χ, T) in which T is a tree and $\chi = \{\chi_i \mid i \in V(T)\}$ is a family of subsets of $V(G)$ such that (1) $\bigcup_{i \in V(T)} \chi_i = V(G)$; (2) for each edge $e = \{u, v\} \in E(G)$ there exists an $i \in V(G)$ such that both u and v belong to χ_i; and (3) for all $v \in V(G)$, the set of nodes $\{i \in V(T) \mid v \in \chi_i\}$ forms a connected subtree of T. To distinguish between vertices of the original graph G and vertices of T in $TD(G)$, we call vertices of T *nodes* and their corresponding χ_i's *bags*. The maximum size of a bag in $TD(G)$ minus one is called the *width* of the tree decomposition. The *treewidth* of a graph G ($\mathbf{tw}(G)$) is the minimum width over all possible tree decompositions of G.

A *branch decomposition* of a graph (or a hypergraph) G is a pair (T, τ), where T is a tree with vertices of degree 1 or 3 and τ is a bijection from the set of leaves of T to $E(G)$. The *order* of an edge e in T is the number of vertices $v \in V(G)$ such that there are leaves t_1, t_2 in T in different components of $T(V(T), E(T) - e)$ with $\tau(t_1)$ and $\tau(t_2)$ both containing v as an endpoint.

The *width* of (T, τ) is the maximum order over all edges of T, and the *branchwidth* of G, $\mathbf{bw}(G)$, is the minimum width over all branch decompositions of G. (In the case where $|E(G)| \leq 1$, we define the branchwidth to be 0; if $|E(G)| = 0$, then G has no branch decomposition; if $|E(G)| = 1$, then G has a branch decomposition consisting of a tree with one vertex—the width of this branch decomposition is considered to be 0.)

It is easy to see that, if H is a minor of G, then $\mathbf{bw}(H) \leq \mathbf{bw}(G)$. The following result is due to Robertson and Seymour [RS91, Theorem 5.1].

LEMMA 2.1 (see [RS91]). *For any connected graph G where $|E(G)| \geq 3$, $\mathbf{bw}(G) \leq \mathbf{tw}(G) + 1 \leq \frac{3}{2}\mathbf{bw}(G)$.*

The main combinatorial result of this paper is Theorem 4.8 (see the end of section 4.2), which determines, for any k and g, a family of graphs $\mathcal{H}_{k,g}$ such that any $\mathcal{H}_{k,g}$-contraction-free graph G with genus g will have branchwidth $O(gk)$. To describe such a family, we will need some definitions on graph embeddings.

2.2. Graph embeddings. Most of the notions defined in this subsection can be found in [MT01].

A *surface* Σ is a compact 2-manifold without boundary. We will always consider connected surfaces. We denote by $\mathbb{S}_0$ the sphere $\{(x,y,z) \mid x^2+y^2+z^2=1\}$. A *line* in Σ is a subset homeomorphic to $[0,1]$. An *O-arc* is a subset of Σ homeomorphic to a circle. A subset of Σ is an *open disk* if it is homeomorphic to $\{(x,y) \mid x^2+y^2<1\}$, and it is a *closed disk* if it is homeomorphic to $\{(x,y) \mid x^2+y^2\leq 1\}$.

A 2-*cell embedding* of a graph G in a surface Σ is a drawing of the vertices as points in Σ and the edges as lines in Σ such that every face (connected component of $\Sigma - E(G) - V(G)$) is an open disk. To simplify notations we do not distinguish between a vertex of G and the point of Σ used in the drawing to represent the vertex or between an edge and the line representing it. We also consider G as the union of the points corresponding to its vertices and edges. That way, a subgraph H of G can be seen as a graph H where $H \subseteq G$. We use the notation $V(G)$, $E(G)$, and $F(G)$ for the set of the vertices, edges, and faces of the embedded graph G. For $\Delta \subseteq \Sigma$, $\overline{\Delta}$ is the *closure* of Δ. The boundary of Δ is $\mathbf{bd}(\Delta) = \overline{\Delta} \cap \overline{\Sigma - \Delta}$, and the interior is $\mathbf{int}(\Delta) = \overline{\Delta} - \mathbf{bd}(\Delta)$.

A subset of Σ meeting the drawing only in vertices of G is called *G-normal*. If an O-arc is G-normal, then we call it a *noose*. The length of a noose is the number of vertices it meets.

Representativity [RS88] is the measure of the "density" of the embedding of a graph in a surface. The *representativity* (or *facewidth*) $\mathbf{rep}(G)$ of a graph G embedded in surface $\Sigma \neq \mathbb{S}_0$ is the smallest length of a noncontractible noose in Σ. In other words, $\mathbf{rep}(G)$ is the smallest number k such that Σ contains a noncontractible (non–null-homotopic in Σ) closed curve that intersects G in k points.

It is more convenient to work with Euler genus. The *Euler genus* $\mathbf{eg}(\Sigma)$ of a surface Σ is equal to the nonorientable genus $\tilde{g}(\Sigma)$ (or the crosscap number) if Σ is a nonorientable surface. If Σ is an orientable surface, $\mathbf{eg}(\Sigma)$ is $2g(\Sigma)$, where $g(\Sigma)$ is the orientable genus of Σ. Given a graph G, its Euler genus $\mathbf{eg}(G)$ is the minimum $\mathbf{eg}(\Sigma)$ where Σ is a surface in which G can be embedded.

2.3. Splitting graphs and surfaces. In this section we describe precisely how to cut along a noncontractible noose in order to decrease the genus of the graph until we obtain a planar graph.

Let G be a graph and let $v \in V(G)$. Also suppose we have a partition $\mathcal{P}_v = (N_1, N_2)$ of the set of the neighbors of v. Define the *splitting* of G with respect to v and P_v to be the graph obtained from G by (i) removing v and its incident edges; (ii) introducing two new vertices v^1, v^2; and (iii) connecting v^i with the vertices in $N_i, i = 1, 2$. If H is the result of the consecutive application of the above operation on some graph G, then we say that H is a *splitting* of G. If additionally in such a splitting process we do not split vertices that are results of previous splittings, then we say that H is a *fair splitting* of G.

The following lemma defines how to find a fair splitting for a given noncontractible noose. It will serve as a link between Lemmas 4.4 and 4.7 in the proof of the main result of this paper. Its proof is straightforward, following lines similar to those of [DFHT04b].

LEMMA 2.2. *Let G be a connected graph 2-cell embedded in a nonplanar surface Σ, and let N be a noncontractible noose of Σ. Then there is a fair splitting G' of G affecting the set $S = (v_1, \ldots, v_\rho)$ of the vertices of G met by N, such that* (i) *G' has at most two connected components;* (ii) *each connected component of G'*

can be 2*-cell embedded in a surface with Euler genus strictly smaller than the Euler genus of* Σ*; and* (iii) *there are two faces* f_1 *and* f_2*, each in the* 2*-cell embedding of a connected component of* G' *(and the connected components are different for the two faces if* G' *is disconnected), such that the boundary of* f_i*, for* $i \in \{1, 2\}$*, contains* $S_i = (v_1^i, \ldots, v_\rho^i)$*, where* v_j^1 *and* v_j^2 *are the vertices created after the splitting of the vertex* v_j*, for* $j = 1, \ldots, \rho$*.*

3. Incomplete embeddings and their properties. In this section we give a series of definitions and results that support the proof of the main theorem of the next section. In particular, we will need special embeddings of graphs that are incomplete; i.e., only *some* of the edges and vertices of the graph are embedded in a surface. Moreover, we will extend the definition of a contraction so that it will also consider contractions of faces for the part of the graph that is embedded.

Let Σ be a surface (orientable or not). Given a graph G, a vertex set $V \subseteq V(G)$, and an edge set $E \subseteq E(G)$ such that $\cup_{v \in V} E_v \subseteq E$, we denote by G^- the graph obtained by G by removing all vertices in V and all edges in E, i.e., the graph $G^- = (V(G) - V, E(G) - E)$.[3] We also say that G is (V, E)*-embeddable in* Σ if G^- has a 2-cell embedding in Σ. We call the graph G^- the *ground* of G and we call the edges and vertices of G^- *landed*. On the other hand, we call the vertices in V and E *flying*. Notice that the flying edges are partitioned into three categories: those that have both endpoints in $V(G) - V$ (we call them *bridges*), those with one endpoint in $V(G) - V$ and one endpoint in V (we call them *pillars*), and those with both endpoints in V (we call them *clouds*). From now on, whenever we refer to a graph (V, E)-embeddable in Σ we will accompany it with the corresponding 2-cell embedding of G^- in Σ.

The set of *atoms* of G with respect to some (V, E)-embedding of G in Σ is the set $A(G) = V(G) \cup E(G) \cup F(G)$, where $F(G)$ is the set of faces of the 2-cell embedding of G^- in Σ. Notice that a flying atom can only be a vertex or an edge. In this paper, we will consider the faces as open sets whose boundaries are cyclic sequences of edges and vertices.

3.1. Contraction mappings. A strengthening of a graph being a contraction of another graph is for there to be a "contraction mapping" which preserves some aspects of the embedding in a surface during the contractions. See Figure 3.1 for an example. Given two graphs G and H that are $(V^{(G)}, E^{(G)})$- and $(V^{(H)}, E^{(H)})$-embeddable in Σ and Σ', respectively, we say that $\phi : A(G) \to A(H)$ is a *contraction mapping* from G to H with respect to their corresponding embeddings if the following conditions are satisfied:

1. For any $v \in V(G)$, $\phi(v) \in V(H)$.
2. For any $e \in E(G)$, $\phi(e) \in E(H) \cup V(H)$.
3. For any $f \in F(G)$, $\phi(f) \in F(H) \cup E(H) \cup V(H)$.
4. For any $v \in V(H)$, $G[\phi^{-1}(v)]$ is a connected subgraph of G.
5. $\{\phi^{-1}(v) \mid v \in V(H)\}$ is a partition of $V(G)$.
6. If $\phi(\{x, y\}) = v \in V(H)$, then $\phi(x) = \phi(y) = v$.
7. If $\phi(\{x, y\}) = e \in E(H)$, then $\{\phi(x), \phi(y)\} \in E(H)$.
8. If $f \in F(G)$ and $\phi(f) = v \in V(G)$ and $f = (x_0, \ldots, x_{r-1})$, then $\phi(\{x_i, x_{i+1}\}) = \phi(x_i) = v$ for any $i = 0, \ldots, r-1$ (where indices are taken modulo r).

[3] In this paper, the vertices and edges of a graph G are referred to as $V(G)$ and $E(G)$, respectively, while V and E are subsets.

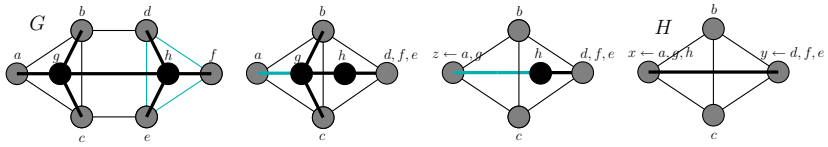


FIG. 3.1. *An example of a contraction of a graph (V, E)-embeddable in $\mathbb{S}_0$, where $V = \{g, h\}$ and $E = \{\{g, a\}, \{g, b\}, \{g, c\}, \{h, g\}, \{h, d\}, \{h, f\}, \{h, e\}\}$. The contraction is shown in a three-step sequence: contracting the edges of the face $\{d, e, f\}$, then the edge $\{a, g\}$, and then edge $\{z, h\}$. A contraction mapping from G to H is defined as follows: $\phi(a) = \phi(g) = \phi(h) = \phi(\{a, g\}) = \phi(\{g, h\}) = x$, $\phi(b) = b$, $\phi(c) = c$, $\phi(d) = \phi(f) = \phi(e) = \phi(\{f, d\}) = \phi(\{d, e\}) = \phi(\{e, f\}) = \phi(\{d, e, f\}) = y$, $\phi(\{a, b\}) = \phi(\{g, b\}) = \{x, b\}$, $\phi(\{a, c\}) = \phi(\{g, c\}) = \{x, c\}$, $\phi(\{b, c\}) = \{b, c\}$, $\phi(\{b, d\}) = \{b, y\}$, $\phi(c, e) = \{c, y\}$, $\phi(\{a, b, c\}) = \{x, b, c\}$, $\phi(\{b, d, e, c\}) = \{b, c, y\}$, $\phi(\{h, d\}) = \phi(\{h, e\}) = \phi(\{h, f\}) = \{x, y\}$, $\phi(\{a, b, d, f, e, c\}) = \{x, b, y, c\}$.*

9. If $f \in F(G)$ and if $\phi(f) = e$ (an edge of H), then there are two edges of f contained in $\phi^{-1}(e)$.
10. If $f \in F(G)$ and if $\phi(f) = g$ (a face of H), then each edge of g is landed and is the image of some edge in f.

Notice that, from conditions 1, 2, and 3, the preimages of the faces of H are faces of G.

The following lemma is easy.

LEMMA 3.1. *If there exists some contraction mapping from a graph G to a graph H with respect to some embedding of G and H, then H is a contraction of G.*

Proof. Observe that H can be obtained from G if we contract all the edges of $\bigcup_{v \in V(H)} G[\phi^{-1}(v)]$. ∎

3.2. Properties of contraction mappings. It is important that the two notions (contraction and existence of a contraction mapping) are identical in the case where G and H have no flying atoms, i.e., $V^{(G)} = V^{(H)} = E^{(G)} = E^{(H)} = \emptyset$. We choose to work with contraction mappings instead of simple contractions because they include stronger information that is sufficient to build the induction argument of Lemma 4.7.

LEMMA 3.2. *If G and H are graphs and H is a contraction of G, then for any $(\emptyset, \emptyset)$-embedding of G and H on the same surface Σ there exists a contraction mapping from G to H with respect to their corresponding embeddings.*

Proof. We partition the contracted edges of H into connected subsets such that no edges belonging to different subsets are connected by a path of contracted edges. We map all edges of each such subset to the vertex of H that remains after their contraction. We also observe that an edge that does not belong to such a subset survives after the contraction and we map it to its copy in H. Notice that no edges incident to the same vertex belong to different subsets. Using this fact, we map each vertex of G to a vertex of H as follows: If v is incident to some contracted edge, then we map it to the same vertex to which this contracted edge is mapped. If not, then this vertex survives after the contraction procedure and therefore it is mapped to its copy in H. It remains now to map any face f of G to atoms of H. Notice that, if some face of G is incident to noncontractible edges, then these edges should be at least two. Using this fact, we distinguish three cases: In the first, all the edges in $\mathbf{bd}(f)$ belong to the same subset of the partition. Then we map f to the vertex occurring by the construction of the edges in this subset. In the second case, there are exactly two noncontractible edges of G in $\mathbf{bd}(f)$. Then these two edges should be mapped to the same edge e of H, and we map f to e. In the third case, we have that $\mathbf{bd}(f)$ contains

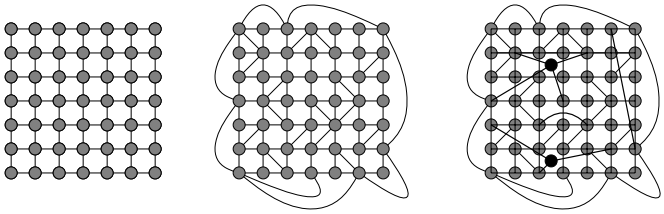


FIG. 3.2. *A* (7×7)*-grid, a partially triangulated* (7×7)*-grid, and a* $(7, 9)$*-gridoid (the flying edges and vertices are the distinguished ones).*

more than 2 noncontractible edges of G. We observe that the noncontractible edges in $\mathbf{bd}(f)$ define a face g in H and we map f to g. It is now easy to verify that the mapping we just defined satisfies conditions 1–10. □

The following lemma is a useful generalization of Lemma 3.2.

LEMMA 3.3. *Let G be a graph (V, E)-embeddable on some surface Σ and let H be the graph occurring from G after contracting edges in $E(G^-)$. Then $G[V] = H[V]$, H is also (V, E)-embeddable in Σ, and there exists a contraction mapping ϕ from G to H with respect to their corresponding embeddings.*

Proof. Let H^- be the result of the application of the same contractions on G^- embeddable on the surface Σ. From Lemma 3.2, there exists a mapping ϕ' from G^- to H^-. Add in H^- all the flying vertices and all the clouds of G. This implies that $G[V] = H[V]$. Then, for any flying vertex v, add in H^- all pillar edges connecting it to the vertices in $\{\phi^{-1}(u) \mid u \in V(G^-) \text{ and } u \in N_G(v)\}$. Finally, for any bridge $\{v, u\}$ of G where $\phi'(v) \neq \phi'(u)$, we add in H^- the bridge $\{\phi'(v), \phi'(u)\}$. Notice that after the aforementioned edge additions transform H^- to H, it is embeddable in Σ.

We now construct the required map ϕ as follows: For any $a \in A(G^-), \phi(a) = \phi'(a)$; for any $v \in V, \phi(v) = v$. Finally, for any $\{x, y\} \in E$, we define $\phi(\{x, y\})$ as follows. If $\phi(x) \neq \phi(y)$, we set $\phi(\{x, y\}) = \{\phi(x), \phi(y)\}$, and if $\phi(x) = \phi(y)$, we set $\phi(\{x, y\}) = \phi(x)$. □

3.3. Gridoids. A *partially triangulated* $(r \times r)$*-grid* is any graph that contains an $(r \times r)$-grid as a subgraph and is a subgraph of some triangulation of the same $(r \times r)$-grid.

We call a graph G an (r, k)*-gridoid* if it is (V, E)-embeddable in $\mathbb{S}_0$ for some pair V, E, where $|E| \leq k$, $E(G[V]) = \emptyset$ (i.e., G does not have clouds), and G^- is a partially triangulated $(r' \times r')$-grid embedded in $\mathbb{S}_0$ for some $r' \geq r$. For an example of a $(7, 9)$-gridoid and its construction, see Figure 3.2.

4. Main result. In this section we will prove that, if a graph G has branchwidth more than $4k(\mathbf{eg}(G)+1)$, then G contains as a contraction some $(k-12\mathbf{eg}(G), \mathbf{eg}(G))$-gridoid, where $k \geq 12\mathbf{eg}(G)$.

4.1. Transformations of gridoids.

LEMMA 4.1. *Let G be an (r, k)-gridoid $(\emptyset, E)$-embeddable in $\mathbb{S}_0$ and let $v \in V(G^-)$. Then there exists some contraction mapping ϕ from G to some $(r-4, k+1)$-gridoid $(\{v\}, E \cup \{\{v, y\}\})$-embeddable in $\mathbb{S}_0$ such that $\phi(v) = v$.*

Proof. Let G^* be the grid from which G is constructed. Let (x, y) denote the coordinates of the vertex v in G^*. We define the required map ϕ by distinguishing two cases.

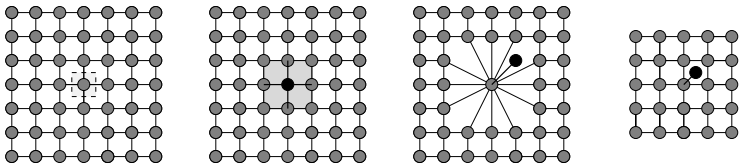


FIG. 4.1. *An example of the first case in the proof of Lemma* 4.1.

Case 1. (x, y) is a vertex of degree 4 in G^*, i.e., $x, y \notin \{1, r\}$. Refer to Figure 4.1 for an example. Let $f_1, \ldots, f_\rho$ be the faces of G^- containing v, cyclically ordered in the way they appear in the embedding of G^- in Σ, and set $f = \cup_{i=1,\ldots,\rho} \overline{f}_i$. We first consider a modified embedding of G where now v is a flying vertex (we add it in V) and the remaining ground graph has the same embedding as before with the difference that now $f - \mathbf{bd}(f)$ is a face replacing the faces $f_1, \ldots, f_r$ that disappear. We construct a graph J that is $(\{v\}, E \cup \{\{v, y\}\})$-embeddable in Σ by contracting all the edges in $\mathbf{bd}(f)$ to a single vertex y. This makes the face $f - \mathbf{bd}(f)$ "disappear" toward creating y and the pillars adjacent to v shrink to a single edge connecting v with y. We construct a mapping $\phi' : A(G) \to A(J)$ as follows. Notice that any atom a of G that is not contained in f is also an atom of J. If a is such an atom, then set $\phi'(a) = a$. If $a \in \mathbf{bd}(f)$, then $\phi'(a) = y$. If $a \in f - \mathbf{bd}(f) - \{v\}$, then set $\phi'(a) = \{y, v\}$ and, finally, set $\phi'(v) = v$. It is easy to verify that ϕ' is a contraction mapping G to J such that $\phi'(v) = v$.

We now further contract in J^- all the edges in $\{\{(x-1, i), (x, i)\}, \{(x, i), (x+1, i)\} \mid i = 1, \ldots, y-2, y+2, \ldots, r\}$ and in $\{\{(i, y-1), (i, y)\}, \{(i, y), (i, y+1)\} \mid i = 1, \ldots, x-2, x+2, \ldots, r\}$, and we call H the resulting graph (these contractions are well defined because these edges are not contracted during the previous transformation of G to J). Observe that H is an $(r-2, k+1)$-gridoid and that applying Lemma 3.3 we construct a contraction mapping ϕ'' from J to H with respect to their $(\{v\}, E \cup \{\{v, y\}\})$-embeddings in $\mathbb{S}_0$, where $\phi''(v) = v$. It remains to observe that $\phi = \phi' \circ \phi''$ is the required map and $\phi(v) = v$.

Case 2. We now examine the case where $v = (x, y)$ is a vertex of G^* with degree 2 or 3. Refer to Figure 4.2 for an example. Let q be the union of all the squares of G^* that have common edges with the unique face of G^* that is not a square (we call the cycle defined by the boundary of this face the *exterior cycle*). We construct a minor of G^- by contracting all the edges in $\mathbf{bd}(\bar{q})$. $\mathbf{bd}(\bar{q})$ contains two connected components that are disjoint cycles, and one of them is the exterior cycle of G^*. These components are shrunk to two distinct adjacent vertices v and u, and we can assume that v is the one of degree 1. We further contract some edge incident to u that is different from $\{v, u\}$. The remaining graph is a partially triangulated $(r-4, r-4)$-grid with some additional pending edge adjacent to its exterior cycle. Let G' be the graph occurring from G after applying to G the same sequence of contractions as in G^-. From Lemma 3.3 we have that G' is also $(\emptyset, E)$-embeddable in Σ and there exists a contraction mapping ϕ' from G to G' with respect to their corresponding embeddings. Moreover, as v is an endpoint of the edges contracted toward forming the vertex v of G', we have $\phi'(v) = v$. Now we update the embedding of G' so that v becomes a flying vertex (we move it in V) and the remaining ground graph has the same embedding as before with the difference that now $\{v, u\}$ is not drawn anymore in the surface (it becomes a pillar). We will use the notation H in order to denote G' $(\{v\}, E \cup \{\{v, u\}\})$-embeddable in Σ in the updated way. We also define a mapping

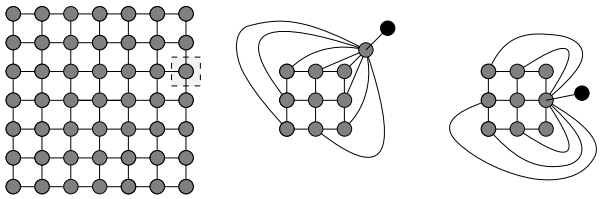


FIG. 4.2. *An example of the second case in the proof of Lemma* 4.1.

ϕ'' from G' to H with respect to their corresponding embeddings so that $\phi''(a) = a$ for any $a \in A(G')$ that is not the face of G' containing the edge $\{v, u\}$. For this face f we set $\phi''(f) = f'$, where f' is the face created in H^- after the removal of $\{v, u\}$ from the interior of f in G'^-.

Observe that H is an $(r-4, k+1)$-gridoid and that $\phi = \phi' \circ \phi''$ is a contraction mapping from G to H with respect to the $(\emptyset, E)$-embedding of G and the $(\{v\}, E \cup \{\{v, y\}\})$-embedding of H in $\mathbb{S}_0$ where $\phi(v) = v$. This completes the proof as $\phi(v) = v$. □

LEMMA 4.2. *Let G be an (r, k)-gridoid $(\emptyset, E)$-embeddable in $\mathbb{S}_0$, and let e be some of its flying edges. Then there exists some $(r-4, k)$-gridoid H $(\emptyset, E')$-embeddable in $\mathbb{S}_0$ for some E' and a contraction mapping ϕ of G to H such that $\phi(e) \in V(H)$.*

Proof. Let $e = \{v, u\}$. Refer to Figure 4.3 for an example. According to Lemma 4.1, there exists some contraction mapping ϕ from G to some $(r-4, k+1)$-gridoid G' $(\{v\}, E \cup \{\{v, y\}\})$-embeddable in $\mathbb{S}_0$ such that $\phi(v) = v$. We construct a new graph H $(\emptyset, E - \{v, u\} \cup \{v, y\})$-embeddable in $\mathbb{S}_0$ by simply contracting the edge $\{v, u\}$ to the vertex v. We define a contraction mapping ϕ' from G' to H as follows: If $a \in A(G') - \{v, u, \{v, u\}\}$, then $\phi'(a) = a$; otherwise $\phi'(a) = v$. Finally, we observe that $\phi \circ \phi'$ is a contraction mapping ϕ from G to H such that $\phi(e) \in V(H)$. □

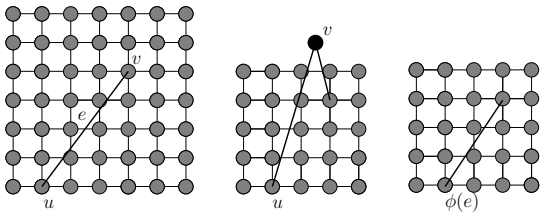


FIG. 4.3. *An example of the proof of Lemma* 4.2.

LEMMA 4.3. *Let G be an (r, k)-gridoid $(\emptyset, E)$-embeddable in $\mathbb{S}_0$, and let a be some of its atoms. Then there exists some $(r-4, k)$-gridoid $(\emptyset, E)$-embeddable in $\mathbb{S}_0$ and a contraction mapping ϕ from G to H with respect to their corresponding embeddings such that $\phi(a) \in V(H)$.*

Proof. We will denote as G^* the $(r \times r)$-grid that should be triangulated in order to construct G^-.

The lemma follows directly from Lemma 4.2 in the case where a is a flying edge of G. If this is not the case, then a is an atom of G^- that is either a vertex or an edge or a face. If a is a face, then either it is a square or triangular face included in some square $C = ((x, y), (x, y+1), (x+1, y), (x+1, y+1))$ of G^* or it is a face with all vertices in the exterior face of G^*.

We will first examine all the aforementioned cases except for the last one. We take the $(r-1, k)$-gridoid H^- that is constructed if we contract in G^- all the edges

in $\{\{(x,i),(x+1,i)\} \mid i = 1,\dots,r\}$ and in $\{\{(i,y),(i,y+1)\} \mid i = 1,\dots,r\}$.

We will now examine the case where a is a face with all vertices in the exterior face of G^*. Then we take the $(r-2,k)$-gridoid H^- that is constructed if we contract in G^- all the edges included in the exterior face of G^* to a single vertex q and then contract some edge incident to q.

Because in both cases H^- is a contraction of H, we can use Lemma 3.3 to construct a contraction $(\emptyset,E)$-mapping ϕ from G to H with respect to their $(\emptyset,E)$-embeddings in Σ. Notice also that $\phi(a) \in V(H^-)$ because, in both cases, all the edges of the cycle $(x,y),(x,y+1),(x+1,y),(x+1,y+1)$ are contracted (and therefore mapped) to a single vertex of H^-. □

4.2. Excluding gridoids as contractions.

LEMMA 4.4. *Let G be a graph $(\emptyset,\emptyset)$-embeddable on some surface Σ. Let H be an (r,k)-gridoid $(\emptyset,E)$-embeddable on the sphere, and assume that ϕ is a contraction mapping from G to H with respect to their corresponding embeddings.*

Let $\{v_1^i,\dots,v_\rho^i\}$, $i = 1,2$, be subsets of the vertices of two faces f_i, $i = 1,2$, of the embedding of G where $f_1 \cap f_2 = \emptyset$ (we assume that the orderings of the indices in each subset respect the cyclic orderings of the vertices in $f_i, i = 1,2$). Let G' be the graph obtained if we identify in G the vertex v_1^1 with the vertex v_i^2. Then, the following hold:

(a) *G' has some 2-cell embedding on a surface of bigger Euler genus.*

(b) *There exists some $(r-12,k+1)$-gridoid H, $(\emptyset,E\cup\{\{e\}\})$-embeddable on the sphere such that there exists some contraction mapping from G' to H with respect to their corresponding embeddings.*

Proof. (a) Let Σ be the surface where G is embedded. We define a surface Σ^- from Σ by removing the two "patches" defined by the (internal) points of the faces f_1 and f_2. Notice that G is still embeddable on Σ^- and that Σ^- is a surface with a boundary whose connected components are the boundaries B_1, B_2 of the faces f_1 and f_2. We now construct a new surface from Σ^- by identifying the boundaries B_1 and B_2 in a way that v_i^1 is identified with v_i^2. Notice that the embedding that follows is a 2-cell embedding and that the new surface has bigger Euler genus.

(b) From conditions 1, 2, and 3 in subsection 3.1, $\phi(f_1)$ is either a vertex, an edge, or a face of H. We apply Lemma 4.3 to construct a contraction mapping σ_1 from H to some $(r-4,k)$-gridoid H_1, where $\sigma_1(\phi(f_1)) \in V(H_1)$. Notice again that $\sigma_1(\phi(f_2))$ is either a vertex, an edge, or a face of H_1. We again use Lemma 4.3 to construct a contraction mapping σ_2 from H_1 to some $(r-8,k)$-gridoid H_2, where $\sigma_2(\sigma_1(\phi(f_i))) = v_i \in V(H_2), i = 1,2$. We now apply Lemma 4.1 for v_1 and construct some contraction mapping σ_3 from H_2 to some $(r-12,k+1)$-gridoid H_3, $(\{v_1\}, E \cup \{\{v_1,y\}\})$-embeddable in $\mathbb{S}_0$ such that $\sigma_3(v_1) = v_1$. Summing up, we have that $\phi' = \phi \circ \sigma_1 \circ \sigma_2 \circ \sigma_3$ is a map from G to H_3 with respect to the $(\emptyset,\emptyset)$-embedding of G on Σ and the $(\{v_1\}, E \cup \{\{v_1,y\}\})$-embedding of H_3 in $\mathbb{S}_0$. Moreover, we have that $\phi'(f_1) = v_1$ and $\phi'(f_2) = v_2 \in V(H_3)$ (to facilitate the notation we assume that $\sigma_3(v_2) = v_2$).

Notice now that if v is the result of the identification in H_3 of the vertex v_1 with the vertex v_2, we take a new graph H $(\emptyset, E \cup \{\{v,y\}\})$-embeddable in $\mathbb{S}_0$. Let A' be all the atoms of G that are not included in the faces f_1 and f_2. Notice that these atoms are not harmed while constructing G' from G, and we set $\mu(a) = \phi'(a)$ for each $a \in A'$. Finally, for each atom $a \in A(G') - A$, we set $\mu(a) = v$. It now is easy to check that μ is a contraction mapping from G' to H with respect to their corresponding embeddings. Because H is an $(r-12,k+1)$-gridoid, we are done. □

The following is one of the main results in [DFHT04b].

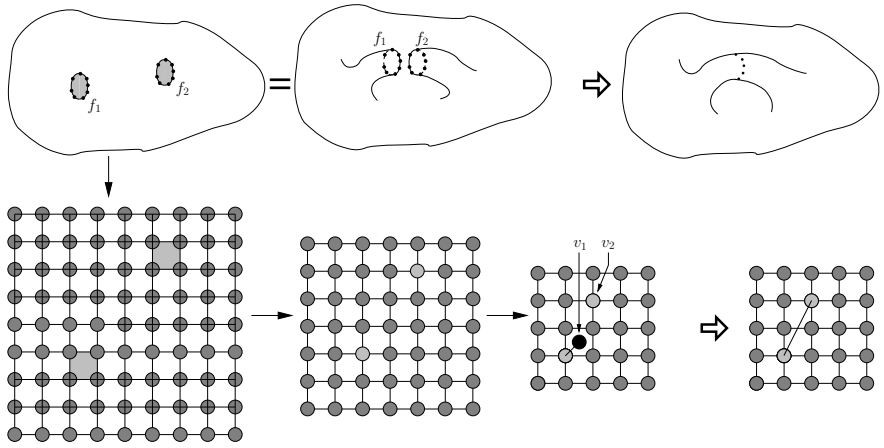


FIG. 4.4. *An example of the transformations in the proof of Lemma* 4.7.

THEOREM 4.5. *Let G be a graph 2-cell embedded in a nonplanar surface Σ of representativity at least θ. Then one can contract edges in G to obtain a partially triangulated $(\theta/4 \times \theta/4)$-grid.*

We also need the following easy lemma.

LEMMA 4.6 (see [DFHT04b]). *Let G be a graph and let H be the graph occurring from G after splitting some vertex $v \in V(G)$. Then $\mathbf{bw}(G) \leq \mathbf{bw}(H) + 1$.*

We are now ready to prove the central result of this section.

LEMMA 4.7. *Let G be a graph $(\emptyset, \emptyset)$-embeddable on a surface Σ of Euler genus g and assume that $\mathbf{bw}(G) \geq 4(r - 12g)(g + 1)$. Then there exists some $(r - 12g, g)$-gridoid H, $(\emptyset, E)$-embeddable in $\mathbb{S}_0$ such that there exists some contraction mapping from G to H with respect to their corresponding embeddings.*

Proof. First, if the graph G is disconnected, we discard all but one connected component C such that $\mathbf{bw}(C) = \mathbf{bw}(G)$.

We use induction on g. Clearly, if $g = 0$, G is a planar graph and after applying Lemma 3.1, the result follows from the planar exclusion theorem of [RST94]. (The induction base relies heavily on the fact that for conventional embeddings the contraction relation is identical to our mapping.)

Suppose now that $g \geq 1$ and the theorem holds for any graph embeddable in a surface with Euler genus less than g. Refer to Figure 4.4. If the representativity of G is at least $4(r - 12g)$, then by Theorem 4.5 we can contract edges in G to obtain a partially triangulated $((r - 12g) \times (r - 12g))$-grid (with no additional edges) and we are done. Otherwise, the representativity of G is less than $4(r - 12g)$. In this case, the smallest noncontractible noose has vertex set S of size less than $4(r - 12g)$. Let G' be a splitting of G with respect to S as in Lemma 2.2. Recall that G' is now $(\emptyset, \emptyset)$-embeddable on a surface of Euler genus $g' \leq g - 1$.

By Lemma 4.6, the branchwidth of G' is at least the branchwidth of G minus $|S|$. Because $|S| \leq 4(r - 12g)$, we have that $\mathbf{bw}(G') \geq 4(r - 12g)(g + 1) - 4(r - 12g) = 4(r - 12g)g \geq 4(r - 12g)(g' + 1)$. By the induction hypothesis there exists some $(r - 12g', g')$-gridoid H', $(\emptyset, E)$-embeddable in $\mathbb{S}_0$ such that there exists some contraction mapping from G' to H with respect to their corresponding embeddings. From Lemma 4.4, there exists some $(r - 12g' - 12, g' + 1)$-gridoid H, $(\emptyset, E \cup \{\{e\}\})$-embeddable on the sphere such that there exists some contraction mapping from G to H with respect to their corresponding embeddings. Because $r - 12g' - 12 \geq r - 12g$

and $g' + 1 \leq g$, we are done. □

Now we have the conclusion of this section.

THEOREM 4.8. *If a graph G excludes all $(k - 12\mathbf{eg}(G), \mathbf{eg}(G))$-gridoids as contractions for some $k \geq 12\mathbf{eg}(G)$, then G has branchwidth at most $4k(\mathbf{eg}(G) + 1)$.*

By Lemma 2.1 we can obtain a treewidth-parameter bound as desired.

5. Algorithmic consequences. Define the *parameter* corresponding to an optimization problem to be the function mapping graphs to the solution value of the optimization problem. In particular, *deciding* a parameter corresponds to computing whether the solution value is at most a specified value k. A parameter is *contraction-bidimensional* if (1) its value does not increase when taking contractions and (2) its value on an $(r, O(1))$-gridoid is $\Omega(r^2)$.[4]

THEOREM 5.1. *Consider a contraction-bidimensional parameter P such that, given a tree decomposition of width at most w for a graph G, the parameter can be decided in $h(w) \cdot n^{O(1)}$ time. Then we can decide parameter P on a bounded-genus graph G in $h(O(\sqrt{k})) \cdot n^{O(1)} + 2^{O(\sqrt{k})} n^{3+\varepsilon}$ time.*

Proof. The algorithm proceeds as follows. First we approximately compute the treewidth and a corresponding tree decomposition of the graph G. Specifically, given a number ω, Amir's algorithm [Ami01] either reports that the treewidth of G is at least ω or produces a tree decomposition of width at most $(3 + \frac{2}{3})\omega$ in time $O(2^{3.698\omega} n^3 \omega^3 \log^4 n)$. We use this algorithm to check whether $\mathbf{tw}(G) = O(\sqrt{k})$ for a sufficiently large constant in the O notation (similar algorithmic results on treewidth that also work for our purposes can be found in [Lag96, Ree92, RS95]). If not, Theorem 4.8 tells us that the graph G has an $(O(\sqrt{k}), O(1))$-gridoid as a contraction. Property 2 of contraction bidimensionality tells us then that the parameter value is $\Omega(k)$. By choosing the constant in the O notation (in $\mathbf{tw}(G) = O(\sqrt{k})$) large enough, we can make the constant in the Ω notation greater than 1. Then we conclude that the parameter value is strictly greater than k (assuming k is at least some constant), so we can answer the decision problem negatively. On the other hand, if $\mathbf{tw}(G) = O(\sqrt{k})$, we apply the $h(\mathbf{tw}(G)) \cdot n^{O(1)}$ algorithm to the tree decomposition produced by Amir's algorithm. The overall running time is $h(O(\sqrt{k})) \cdot n^{O(1)} + 2^{O(\sqrt{k})} n^{3+\varepsilon}$. □

COROLLARY 5.2. *Vertex cover, minimum maximal matching, dominating set, edge dominating set, r-dominating set (for fixed r), and clique-transversal set can be solved on bounded-genus graphs in $2^{O(\sqrt{k})} n^{3+\varepsilon}$ time, where k is the size of the optimal solution. Feedback vertex set and connected dominating set can be solved on bounded-genus graphs in $2^{O(\sqrt{k} \log k)} n^{3+\varepsilon}$ time.*

Acknowledgments. Thanks go to Fedor Fomin for early collaboration on this project, and to the anonymous referees for their helpful feedback on the paper.

REFERENCES

[ABF+02] J. ALBER, H. L. BODLAENDER, H. FERNAU, T. KLOKS, AND R. NIEDERMEIER, *Fixed parameter algorithms for dominating set and related problems on planar graphs*, Algorithmica, 33 (2002), pp. 461–493.

[AFF+01] J. ALBER, H. FAN, M. R. FELLOWS, H. FERNAU, R. NIEDERMEIER, F. A. ROSAMOND, AND ULRIKE STEGE, *Refined search tree technique for Dominating Set on planar graphs*, in Proceedings of the 26th International Symposium on Mathematical Foundations of Computer Science (MFCS 2001), Lecture Notes in Comput. Sci. 2136, Springer-Verlag, Berlin, 2001, pp. 111–122.

[4] The requirement of $\Omega(r^2)$ can be weakened to allow any function $g(r)$, as in [DFHT04b, DFHT04a]; the only consequence is that $\sqrt{k}$ gets replaced by $g^{-1}(r)$.

[AFN04] J. Alber, H. Fernau, and R. Niedermeier, *Parameterized complexity: Exponential speed-up for planar graph problems*, J. Algorithms, 52 (2004), pp. 26–56.

[Ami01] E. Amir, *Efficient approximation for triangulation of minimum treewidth*, in Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI-2001), Morgan Kaufmann, San Francisco, 2001, pp. 7–15.

[CKL01] M.-S. Chang, T. Kloks, and C.-M. Lee, *Maximum clique transversals*, in Proceedings of the 27th International Workshop on Graph-Theoretic Concepts in Computer Science (WG 2001), Lecture Notes in Comput. Sci. 2204, Springer-Verlag, Berlin, 2001, pp. 32–43.

[DF99] R. G. Downey and M. R. Fellows, *Parameterized Complexity*, Springer-Verlag, New York, 1999.

[DFHT04a] E. D. Demaine, F. V. Fomin, M. Hajiaghayi, and D. M. Thilikos, *Bidimensional parameters and local treewidth*, SIAM J. Discrete Math., 18 (2004), pp. 501–511.

[DFHT04b] E. D. Demaine, F. V. Fomin, M. Hajiaghayi, and D. M. Thilikos, *Subexponential parameterized algorithms on graphs of bounded-genus and H-minor-free graphs*, in Proceedings of the 15th ACM-SIAM Symposium on Discrete Algorithms (SODA 2004), New Orleans, 2004, pp. 823–832.

[DFHT05] E. D. Demaine, F. V. Fomin, M. Hajiaghayi, and D. M. Thilikos, *Fixed-parameter algorithms for the (k, r)-center in planar graphs and map graphs*, ACM Trans. Algorithms, 1 (2005), pp. 33–47.

[DH04a] E. D. Demaine and M. Hajiaghayi, *Diameter and treewidth in minor-closed graph families, revisited*, Algorithmica, 40 (2004), pp. 211–215.

[DH04b] E. D. Demaine and M. Hajiaghayi, *Equivalence of local treewidth and linear local treewidth and its algorithmic applications*, in Proceedings of the 15th ACM-SIAM Symposium on Discrete Algorithms (SODA 2004), New Orleans, 2004, pp. 833–842.

[DH04c] E. D. Demaine and M. Hajiaghayi, *Fast algorithms for hard graph problems: Bidimensionality, minors, and local treewidth*, in Proceedings of the 12th International Symposium on Graph Drawing (Harlem, NY, 2004), Lecture Notes in Comput. Sci. 3383, Springer-Verlag, Berlin, 2004, pp. 517–533.

[DH05a] E. D. Demaine and M. Hajiaghayi, *Bidimensionality: New connections between FPT algorithms and PTASs*, in Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2005), Vancouver, 2005, pp. 590–601.

[DH05b] E. D. Demaine and M. Hajiaghayi, *Graphs excluding a fixed minor have grids as large as treewidth, with combinatorial and algorithmic applications through bidimensionality*, in Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2005), Vancouver, 2005, pp. 682–689.

[DHN+04] E. D. Demaine, M. Hajiaghayi, N. Nishimura, P. Ragde, and D. M. Thilikos, *Approximation algorithms for classes of graphs excluding single-crossing graphs as minors*, J. Comput. System Sci., 69 (2004), pp. 166–195.

[DHT05] E. D. Demaine, M. Hajiaghayi, and D. M. Thilikos, *Exponential speedup of fixed-parameter algorithms for classes of graphs excluding single-crossing graphs as minors*, Algorithmica, 41 (2005), pp. 245–267.

[FT03] F. V. Fomin and D. M. Thilikos, *Dominating sets in planar graphs: Branch-width and exponential speed-up*, in Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2003), Baltimore, 2003, pp. 168–177.

[GKLY05] G. Gutin, T. Kloks, C. M. Lee, and A. Yeo, *Kernels in planar digraphs*, J. Comput. System Sci., 71 (2005), pp. 174–184.

[KLL02] T. Kloks, C. M. Lee, and J. Liu, *New algorithms for k-face cover, k-feedback vertex set, and k-disjoint set on plane and planar graphs*, in Proceedings of the 28th International Workshop on Graph-Theoretic Concepts in Computer Science (WG 2002), Lecture Notes in Comput. Sci. 2573, Springer-Verlag, Berlin, 2002, pp. 282–295.

[KP02] I. Kanj and L. Perković, *Improved parameterized algorithms for planar dominating set*, in Proceedings of the 27th International Symposium on Mathematical Foundations of Computer Science, Lecture Notes in Comput. Sci. 2420, Springer-Verlag, 2002, pp. 399–410.

[Lag96] J. Lagergren, *Efficient parallel algorithms for graphs of bounded tree-width*, J. Algorithms, 20 (1996), pp. 20–44.

[Lov03] L. Lovász, Private communication, 2003.

[MT01] B. Mohar and C. Thomassen, *Graphs on Surfaces*, Johns Hopkins Stud. Math. Sci., The Johns Hopkins University Press, Baltimore, MD, 2001.

[Ree92] B. A. Reed, *Finding approximate separators and computing tree width quickly*, in Proceedings of the 24th Annual ACM Symposium on Theory of Computing (STOC 1992), ACM, New York, 1992, pp. 221–228.

[RS85] N. Robertson and P. D. Seymour, *Graph minors—a survey*, in Surveys in Combinatorics, I. Anderson, ed., Cambridge University Press, Cambridge, UK, 1985, pp. 153–171.

[RS86] N. Robertson and P. D. Seymour, *Graph minors.* II. *Algorithmic aspects of tree-width*, J. Algorithms, 7 (1986), pp. 309–322.

[RS88] N. Robertson and P. D. Seymour, *Graph minors.* VII. *Disjoint paths on a surface*, J. Combin. Theory Ser. B, 45 (1988), pp. 212–254.

[RS91] N. Robertson and P. D. Seymour, *Graph minors.* X. *Obstructions to tree-decomposition*, J. Combin. Theory Ser. B, 52 (1991), pp. 153–190.

[RS95] N. Robertson and P. D. Seymour, *Graph minors.* XIII. *The disjoint paths problem*, J. Combin. Theory Ser. B, 63 (1995), pp. 65–110.

[RS03] N. Robertson and P. D. Seymour, *Graph minors.* XVI. *Excluding a non-planar graph*, J. Combin. Theory Ser. B, 89 (2003), pp. 43–76.

[RST94] N. Robertson, P. D. Seymour, and R. Thomas, *Quickly excluding a planar graph*, J. Combin. Theory Ser. B, 62 (1994), pp. 323–348.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 372–394

CLASSIFICATION OF BIPARTITE BOOLEAN CONSTRAINT SATISFACTION THROUGH DELTA-MATROID INTERSECTION*

TOMÁS FEDER† AND DANIEL FORD‡

Abstract. Matroid intersection has a known polynomial time algorithm using an oracle. We generalize this result to delta-matroids that do not have equality as a restriction and give a polynomial time algorithm for delta-matroid intersection on delta-matroids without equality using an oracle. We note that when equality is present, delta-matroid intersection is as general as delta-matroid parity. We also obtain algorithms using an oracle for delta-matroid parity on delta-matroids without inequality, and for delta-matroid intersection where one delta-matroid does not contain either equality or inequality, and the second delta-matroid is arbitrary. Both these results also generalize matroid intersection. The results imply a dichotomy for bipartite Boolean constraint satisfaction problems using an oracle when one of the two sides does not contain equality, leaving open cases of delta-matroid parity when both sides have equality; the results also imply a full dichotomy for k-partite Boolean constraint satisfaction problems for $k \geq 3$. We then discuss polynomial cases of Boolean constraint satisfaction problems with two occurrences per variable through delta-matroid parity that cannot be obtained using the oracle approach.

Key words. Boolean constraint satisfaction, delta matroids, blossoms

AMS subject classifications. 68Q17, 68T20, 03C13

DOI. 10.1137/S0895480104445009

1. Introduction.

An instance of the Boolean constraint satisfaction problem consists of a collection of variables ranging over the Boolean domain and a collection of constraints on them. The aim is to assign value 0 or 1 to each variable so as to satisfy all the constraints. The Boolean constraint satisfaction problem is NP-complete. Schaefer [11] considered the restriction of Boolean constraint satisfaction problems to the case where the constraints used must each belong to a given collection of allowed constraint types. Schaefer then classified the Boolean constraint satisfaction problems as polynomial time solvable or NP-complete, depending on the choice of the collection of allowed constraint types. In the case where restricting a variable to take value 0 or to take value 1 is an allowed constraint, the Schaefer polynomial cases are conjunctions (1) of Horn clauses, (2) of dual-Horn clauses, (3) of 2-satisfiability clauses, and (4) of linear equations modulo 2.

The constraint satisfaction problem with a collection of allowed constraint types can be further restricted so that each variable is allowed to participate in only two constraints. While the polynomial cases of Schaefer's classification remain polynomial under this restriction, some of the NP-complete cases may become polynomial time solvable. Feder [5] showed that the NP-complete cases of Schaefer's classification remain NP-complete unless each allowed constraint type is a delta-matroid. In that case, the problem with two occurrences for each variable is the well-known delta-matroid parity problem [1], which generalizes matroid parity [10]. Only certain families of matroid and delta-matroid parity problems are known to be polynomial time solvable. The best known such problem is graph matching.

*Received by the editors July 2, 2004; accepted for publication (in revised form) September 14, 2005; published electronically May 3, 2006.
http://www.siam.org/journals/sidma/20-2/44500.html

†268 Waverley Street, Palo Alto, CA 94301 (tomas@theory.stanford.edu).
‡900 South East Baker Street, McMinnville, OR 97128 (dford@linfield.edu).

A further restriction of Boolean constraint satisfaction problems with two occurrences per variable requires the constraints in an instance to be partitioned into two sets, so that each variable participates in only one constraint from each set. This restricted problem is known as the bipartite Boolean constraint satisfaction problem. Again, Feder [5] showed that for the bipartite Boolean constraint satisfaction problem, the NP-complete cases of Schaefer's classification remain NP-complete unless each allowed constraint type is a delta-matroid. In that case, the bipartite constraint satisfaction problem is delta-matroid intersection, which generalizes matroid intersection, and in particular bipartite graph matching.

Since matroid intersection is polynomial time solvable by the algorithm of Edmonds [4], it is natural to ask whether delta-matroid intersection is polynomial time solvable. The main difficulty is that if the equality constraint is among the allowed constraint types, matroid intersection becomes as hard as matroid parity. In fact, a bipartite Boolean constraint satisfaction problem is more restrictive than the general Boolean constraint satisfaction problem with two occurrences per variable only if the equality constraint is not among the allowed constraints. In this paper, we thus consider delta-matroid intersection in the case where the delta-matroids do not contain the equality constraint as a restriction, and we give a polynomial time algorithm for the problem. This completes our first classification result for bipartite Boolean constraint satisfaction problems, which are assumed not to contain the equality constraint as an allowed constraint type, as polynomial time solvable or NP-complete.

In the model adopted, we impose no restriction on the size of constraints describing the two delta-matroids without equality to be intersected. We thus adopt the most general model, in which each of the delta-matroids is given by an oracle that can be queried in polynomial time to obtain a feasible assignment for the delta-matroid, or to determine whether a given assignment is feasible for the delta-matroid. We observe also that Schaefer's polynomial cases remain polynomial with a slightly more powerful oracle, which allows querying the oracle to determine whether a given partial assignment can be extended to a full assignment satisfying a given constraint. Both oracles have the same power in the case of delta-matroids.

In this general oracle model, we also show that delta-matroid parity for delta-matroids that do not have the inequality constraint as a restriction can also be solved in polynomial time. We further show that intersecting a delta-matroid that has neither the equality constraint nor the inequality constraint as a restriction, with an arbitrary delta-matroid, also has a polynomial time algorithm. As matroid intersection can be represented as the intersection of two delta-matroids that contain neither the equality constraint nor the inequality constraint as a restriction, all these results generalize matroid intersection. In fact all three results follow from a single more general algorithm for a class of delta-matroid parity problems.

This last result is then used to obtain a more general classification result for bipartite Boolean constraint satisfaction, in which the allowed constraint types may be different for both sides of the bipartition, and it is assumed that at least one side does not contain equality. If both sides contain equality, then both sides can be assumed to be the same, where the problems not yet classified are delta-matroid parity problems. We note that the polynomial cases in the classification are polynomial in the oracle model as well. See Table 1 for the classification. This also implies a dichotomy for k-partite Boolean constraint satisfaction with $k \geq 3$, where we have k sets of allowed constraint types and each variable is allowed to participate in only one constraint from each of k sets of constraints of the corresponding types.

TABLE 1.1

Classification of bipartite Boolean constraint satisfaction problems: Cases other than zebra are also polynomial with oracle.

- bipartite Boolean constraint satisfaction
  1. NP-complete cases
  2. Schaefer derived cases
     (a) Horn
     (b) dual-Horn
     (c) linear
     (d) 2-SAT
     (e) one side has only monadic constraints
     (f) upward 2-SAT in one side and constraints with 2-SAT downward closure in other side (and case interchanging upward and downward)
  3. one side has 2-SAT upward closure with delta-matroid downward closure and other side has 2-SAT downward closure with delta-matroid upward closure, each side is intersection of upward and downward closure, and a flat of the delta-matroid can intersect a 2-SAT clause in exactly one element only if the flat or the 2-SAT clause has only one element.
  4. delta-matroid derived cases
     (a) delta-matroid intersection without equality
     (b) delta-matroid intersection having one side without equality or inequality
     (c) upward delta-matroid in one side and constraints with delta-matroid downward closure in other side (and case interchanging upward and downward)
     (d) delta-matroid parity with equality
        i. local even or odd delta-matroid
        ii. A-local-zebra delta-matroid
        iii. linear-zebra delta-matroid
        iv. delta-matroid without inequality
        v. open cases
- k-partite Boolean constraint satisfaction for $k \geq 3$
  1. NP-complete cases
  2. polynomial cases using an oracle

The study is thus conducted in the full generality of the oracle model. On the other hand, the general case with two occurrences per variable cannot be solved in the general oracle model. In particular, matroid parity has an exponential lower bound due to Lovász [9] in the oracle model. We thus seek to study cases of delta-matroid parity where each of the constraints used is described explicitly. In this model, Feder [5] showed that coindependent delta-matroids have a polynomial time algorithm for delta-matroid parity. We show here that coindependent delta-matroid parity has an exponential lower bound when the coindependent delta-matroid is described by an oracle. We also introduce zebra delta-matroids, as a common generalization of coindependent delta-matroids and the delta-matroids from the general factor problem that was solved by Cornuejols [2]. We show that zebra delta-matroid parity can be solved in polynomial time when each of the zebra constraints is described explicitly. We also show how to recognize certain delta-matroids that can be represented through zebra delta-matroids, thus obtaining the class of zebra-compact delta-matroids generalizing the compact delta-matroids of Istrate [8] based on the general factor problem. Finally, for any class of even delta-matroids that has a polynomial time algorithm for delta-matroid parity, such as linear matroids with the algorithms of Lovász [9], Gabow and Stallman [6], linear delta-matroids with the algorithm of Geelen, Iwata, and Murota [7], or local delta-matroids with the algorithm of Dalmau and Ford [3], we define an associated class of delta-matroids that are not necessarily even, along the same line that defined zebra delta-matroids from certain even delta-matroids that can be obtained via graph matching. We show that these zebra like delta-matroids

associated with the given class of even delta-matroids also have a polynomial time algorithm for delta-matroid parity when the constraints used are described explicitly.

2. Definitions. A *delta-matroid* is a pair $M = (E, \mathcal{F})$, where E is a set and $\mathcal{F}$ is a set of subsets of E, satisfying the following axiom: for all $A, B \in \mathcal{F}$, and for all $x \in A\Delta B$, there exists a $y \in A\Delta B$ such that $A\Delta\{x, y\} \in \mathcal{F}$. Note that we may have $y = x$. The sets $A \in \mathcal{F}$ are called the *feasible sets* of the delta-matroid M.

A *restriction* of a delta-matroid $M = (E, \mathcal{F})$ is a delta-matroid $M_1 = (E_1, \mathcal{F}_1)$ with $E_1 \subseteq E$ such that for some $E_1' \subseteq E \setminus E_1$, we have $A \in \mathcal{F}_1$ if and only if $A \cup E_1' \in \mathcal{F}$. Given two delta-matroids $M_1 = (E_1, \mathcal{F}_1)$ and $M_2 = (E_2, \mathcal{F}_2)$ with $E_1 \cap E_2 = \emptyset$, the *direct sum* of M_1 and M_2 is the delta-matroid $M = (E, \mathcal{F})$ with $E = E_1 \cup E_2$ such that $A \in \mathcal{F}$ if and only if $A \cap E_1 \in \mathcal{F}_1$ and $A \cap E_2 \in \mathcal{F}_2$.

Let $M = (E, \mathcal{F})$ be a delta-matroid and $\mathcal{L}$ a partition of E into pairs. For every $u \in E$, its *mate* will be denoted by $\overline{u}$, that is, $\overline{u}$ is the only element in E such that $\{u, \overline{u}\} \in \mathcal{L}$.

Let $F \in \mathcal{F}$ be a feasible set. We will let $\mathcal{L}_F$ denote the subset of $\mathcal{L}$ containing those pairs $\{u, \overline{u}\} \in \mathcal{L}$ such that either both u and $\overline{u}$ are in F or neither u nor $\overline{u}$ is in F.

An instance of the *delta-matroid parity* problem consists of a delta-matroid $M = (E, \mathcal{F})$ and a partition $\mathcal{L}$ of E into pairs. The goal is to find a feasible set $F \in \mathcal{F}$ such that $\mathcal{L}_F$ is maximum, that is, at least as large as $\mathcal{L}_G$ for any other $G \in \mathcal{F}$.

The *delta-matroid intersection* problem is the special case of the delta-matroid parity problem where $M = (E, \mathcal{F})$ is the direct sum of $M_1 = (E_1, \mathcal{F}_1)$ and $M_2 = (E_2, \mathcal{F}_2)$, and every pair in $\mathcal{L}$ contains one element in E_1 and one element in E_2.

We consider two particular delta-matroids, the *equal* delta-matroid $M_= = (\{a, b\}, \{\emptyset, \{a, b\}\})$ and the *not-equal* delta-matroid $M_\neq = (\{a, b\}, \{\{a\}, \{b\}\})$. Note that every delta-matroid parity problem with $M, \mathcal{L}$ is equivalent to a delta-matroid intersection problem with $M', \mathcal{L}'$, where $M_1 = M$ and M_2 is the direct sum of $M_=$ delta-matroids, one for each pair in $\mathcal{L}$, where $\mathcal{L}'$ has the corresponding pairs $\{u, a\}$ and $\{\overline{u}, b\}$.

Thus delta-matroid intersection is a strict special case of delta-matroid parity only for delta-matroids M that do not have $M_=$ as a restriction. In an instance of delta-matroid parity or intersection, we are given an *oracle* for $M = (E, \mathcal{F})$ that can be queried to provide a particular feasible set in $\mathcal{F}$, and tested with some $A \subseteq E$ so that the oracle responds whether $A \in \mathcal{F}$, that is, whether A is feasible.

The following is known [5]. If $M = (E, \mathcal{F})$, $\mathcal{L}$ is an instance of delta-matroid parity, and $\mathcal{K}$ is a subset of pairs from $\mathcal{L}$, then we obtain a delta-matroid $M' = (E', \mathcal{F}')$ with E' consisting of the elements of E that are not in pairs in $\mathcal{K}$, and including in $\mathcal{F}'$ all sets $A \subseteq E'$ such that there exists a $B \in \mathcal{F}$ such that $B \cap E' = A$ and B is a feasible set for the delta-matroid M satisfying the pairings in $\mathcal{K}$, that is, $\mathcal{K}_B = \mathcal{K}$. We say that M' is the delta-matroid obtained from M, $\mathcal{L}$ by *contracting* $\mathcal{K}$. We can then let $\mathcal{L}' = \mathcal{L} \setminus \mathcal{K}$.

We give a polynomial time algorithm for any instance of delta-matroid parity on a delta-matroid M with pairing $\mathcal{L}$ such that no subset $\mathcal{K} \subset \mathcal{L}$ and pair $\{a, b\} \in \mathcal{L} \setminus \mathcal{K}$ are such that the delta-matroid M' obtained from M by contracting $\mathcal{K}$ has the not-equal delta-matroid $M_\neq$ on $\{a, b\}$ as a restriction. The bipartite Boolean constraint satisfaction classification will follow from this result.

3. Small delta-matroids. In this section we establish simple properties of delta-matroids with three or four elements.

LEMMA 3.1. *Let $M = (\{a,b,c\},\mathcal{F})$ be a delta-matroid with a feasible set F such that $F\Delta\{a,b,c\}$ is also feasible. Then one of $F\Delta\{a\}$, $F\Delta\{c\}$, $F\Delta\{a,c\}$ is also feasible.*

Proof. Let $A = F\Delta\{a,b,c\}$, $B = F$, and $x = b$ in the definition of delta-matroid. ∎

LEMMA 3.2. *Let $M = (\{a,b,c\},\mathcal{F})$ be a delta-matroid having feasible sets $F\Delta\{c\}$ and $F\Delta\{a,b\}$. Then one of of $F\Delta\{a\}$, $F\Delta\{a,c\}$, $F\Delta\{a,b,c\}$ is also feasible.*

Proof. Let $A = F\Delta\{c\}$, $B = F\Delta\{a,b\}$, and $x = a$ in the definition of delta-matroid. ∎

LEMMA 3.3. *Let $M = (\{a,b,c,d\},\mathcal{F})$ be a delta-matroid with a feasible set F such that $F\Delta\{a,b\}$ and $F\Delta\{a,b,c,d\}$ are also feasible. Then one of $F\Delta\{a\}$, $F\Delta\{a,c\}$, $F\Delta\{a,d\}$, $F\Delta\{c,d\}$ is also feasible.*

Proof. Consider $C = F\Delta\{a,c,d\}$. If C is feasible, let $A = F\Delta\{a,b\}$, $B = C$, and $x = b$ in the definition of delta-matroid. If C is not feasible, take $A = F\Delta\{a,b,c,d\}$, $B = F$, and $x = b$ in the definition of delta-matroid. ∎

LEMMA 3.4. *Let $M = (\{a,b,c,d\},\mathcal{F})$ be a delta-matroid with a feasible set F such that $F\Delta\{a,b\}$ and $F\Delta\{c,d\}$ are also feasible. Then one of $F\Delta\{a\}$, $F\Delta\{a,c\}$, $F\Delta\{a,d\}$, $F\Delta\{a,b,c,d\}$ is also feasible.*

Proof. Consider $C = F\Delta\{a,c,d\}$. If C is feasible, let $A = F\Delta\{a,b\}$, $B = C$, and $x = b$ in the definition of delta-matroid. If C is not feasible, take $A = F\Delta\{c,d\}$, $B = F\Delta\{a,b\}$, and $x = a$ in the definition of delta-matroid. ∎

4. Structure and algorithm for augmenting paths and blossoms. Let $M = (E,\mathcal{F})$, $\mathcal{L}$ be an instance of the delta-matroid parity problem. A *path* in M is an ordered collection $u_1,\ldots,u_n$ of different elements in E. Let $L \subseteq \mathcal{L}$ be any collection of pairs of $\mathcal{L}$. A path $u_1,\ldots,u_n$ is called *L-alternating* if (1) for every $1 \le 2j < n$, $\{u_{2j},u_{2j+1}\} \in L$, (2) $\{u_1,\overline{u_1}\} \notin L$, and (3) if n is even, then $\{u_1,u_n\},\{u_n,\overline{u_n}\} \notin L$, $u_n \neq \overline{u_1}$. Let $F \in \mathcal{F}$ be a feasible set. We say that a path $u_1,\ldots,u_n$ is an *F-augmenting path* (or simply an augmenting path when F is implicit) if (1) $F\Delta\{u_1,\ldots,u_{2j}\} \in \mathcal{F}$ for all $1 < 2j \le n$ and (2) $F\Delta\{u_1,\ldots,u_n\} \in \mathcal{F}$.

The basic intuition behind this definition is that if F is a feasible set such that $|\mathcal{L}_F|$ is not maximum, then there exists some F-augmenting $\mathcal{L}_F$-alternating path. This path can be used to obtain a new feasible set $G = F\Delta\{u_1,\ldots,u_n\}$ which increases the objective function that we intend to maximize, $|\mathcal{L}_G| > |\mathcal{L}_F|$. In fact if $|\mathcal{L}_F|$ is not maximum, then there exists an F-augmenting $\mathcal{L}_F$-alternating path that can be computed in time polynomial in $|E|$ given a $G \in \mathcal{F}$ with $|\mathcal{L}_G| > |\mathcal{L}_F|$; see, e.g., [3].

Given a feasible $F \in \mathcal{F}$, an *edge* is a pair $\{u,v\}$ of distinct elements in E such that $F\Delta\{u,v\} \in \mathcal{F}$, and a *special element* is a single element u in E such that $F\Delta\{u\} \in \mathcal{F}$.

THEOREM 4.1. *Suppose M has an F-augmenting $\mathcal{L}_F$-alternating path. Let $u_1,\ldots,u_s$ be a shortest such path. Then either* (1) *there exists an $\mathcal{L}_F$-alternating path $v_1,\ldots,v_n$ with $v_1 = u_1$, $v_n = u_s$, such that for $2 \le 2i \le n$, $\{v_{2i-1},v_{2i}\}$ is an edge, and v_n is a special element if n is odd, with each v_i among the u_j, or* (2) *there exists an $\mathcal{L}_F$-alternating path $w_1,\ldots,w_k$ with $w_1 = u_1$ and k odd, and a $2 \le 2r < k$ such that for every $2 \le 2j < k$, $\{w_{2j-1},w_{2j}\}$ is an edge, and $\{w_k,w_{2r-1}\}$ is also an edge, with each w_i among the u_j.*

The alternating path in case (2) is called a *blossom*.

Proof. Let $u_1,\ldots,u_n$ be a shortest augmenting path. We show that either (1) for each $2 \le 2j \le n$ there is an edge $\{u_{2i-1},u_{2j}\}$ for some $2 \le 2i \le 2j$, and if n is odd, then u_n is a special element, or (2) for some $2 \le 2k < n$ there is an edge $\{u_{2l+1},u_{2k+1}\}$ for some $0 \le 2l < 2k$, and for each $2 \le 2j \le 2k$ there is an edge

$\{u_{2i-1}, u_{2j}\}$ for some $2 \leq 2i \leq 2j$. In case (1), tracing back the edges from u_{2j} for $j = n$ or $j = n-1$ to u_{2i-1}, to the mate u_{2i-2}, then the edge joining u_{2i-2} to some $u_{2i'-1}$ with $2i' < 2i$, then to the mate $u_{2i'-1}$, and so on until $u_1 = u$ is reached, gives an alternating path from u_1 to u_n that alternates going to a mate and traversing an edge. In case (2), we get such a path from u_1 to u_{2k} to the mate u_{2k+1} with an edge to u_{2l+1}, with a similar alternating path from u_{2l+1} back to u_1, which at the point it meets the path from u_1 to u_{2k} completes the blossom.

For each $2j \leq n$ and each $0 \leq 2s < 2j$, we show that either there is a $2s < 2i \leq 2j$ such that $F\Delta\{u_1, \ldots, u_{2s}, u_{2i-1}, u_{2j}\}$ is feasible, or there is a $2s < 2i < 2j$ such that $F\Delta\{u_1, \ldots, u_{2s}, u_{2i-1}, u_{2j-1}\}$ is feasible, unless there is a blossom among elements $u_1, \ldots, u_{2s}, u_{2i-1}$. The proof is by induction with decreasing s. When we reach $s = 0$, we have either the edge $\{u_{2i-1}, u_{2j}\}$ or the edge $\{u_{2i-1}, u_{2j-1}\}$ as required.

The base case $2s = 2j - 2$ is verified since $F\Delta\{u_1, \ldots, u_{2j-2}, u_{2j-1}, u_{2j}\}$ is feasible by the definition of an augmenting path. Suppose the claim holds for $2s + 2$, and $F\Delta\{u_1, \ldots, u_{2s}, u_{2s+1}, u_{2s+2}, u_{2i-1}, u_t\}$ is feasible, where $t = 2j$ or $t = 2j - 1$. Let $G = F\Delta\{u_1, \ldots, u_{2s}\}$, and apply Lemma 3.3 with $a = u_{2s+1}$, $b = u_{2s+2}$, $c = u_{2i-1}$, $d = u_t$. We have that G, $G\Delta\{a, b\}$, and $G\Delta\{a, b, c, d\}$ are feasible. If $G\Delta\{a\}$ is feasible, we have a shorter augmenting path obtained from $G\Delta\{a\} = F\Delta\{u_1, \ldots, u_{2s}, u_{2s+1}\}$, contrary to assumption. If $G\Delta\{a, c\}$ is feasible, we have $G\Delta\{a, c\} = F\Delta\{u_1, \ldots, u_{2s}, u_{2s+1}, u_{2i-1}\}$ feasible, which inductively will give an edge $\{u_{2l+1}, u_{2k+1}\}$ as above with $2k+1 = 2i-1$ ant $2l+1 \leq 2s+1$, and thus a blossom. If $G\Delta\{a, d\}$ is feasible, we have $G\Delta\{a, d\} = F\Delta\{u_1, \ldots, u_{2s}, u_{2s+1}, u_t\}$ feasible, which inductively will give an edge $\{u_{2j-1}, u_t\}$ as above with $2j - 1 \leq 2s + 1$. If $G\Delta\{c, d\}$ is feasible, we have $G\Delta\{c, d\} = F\Delta\{u_1, \ldots, u_{2s}, u_{2i-1}, u_t\}$ feasible, which inductively will give an edge $\{u_{2r-1}, u_t\}$ as above with $2r - 1 \leq 2i - 1$.

It remains to show that u_n is a special element if n is odd. We show for each $0 \leq 2s < n$ that $F\Delta\{u_1, \ldots, u_{2s}, u_n\}$ is feasible inductively with s decreasing, unless there is a blossom. When we reach $s = 0$ we have u_n as a special element. The base case $2s = n-1$ is verified since $F\Delta\{u_1, \ldots, u_n\}$ is feasible. Suppose the claim holds for $2s+2$, and $F\Delta\{u_1, \ldots, u_{2s}, u_{2s+1}, u_{2s+2}, u_n\}$ is feasible. Let $G = F\Delta\{u_1, \ldots, u_{2s}\}$ and apply Lemma 3.1 with $a = u_{2s}$, $b = u_{2s+1}$, $c = u_n$. We have that G and $G\Delta\{a, b, c\}$ are feasible. If $G\Delta\{a\}$ is feasible, we have a shorter augmenting path obtained from $G\Delta\{a\} = F\Delta\{u_1, \ldots, u_{2s}, u_{2s+1}\}$, contrary to assumption. If $G\Delta\{c\}$ is feasible, we have $G\Delta\{c\} = F\Delta\{u_1, \ldots, u_{2s}, u_n\}$ feasible, proceeding with the induction for u_n. If $G\Delta\{a, c\}$ is feasible, we have $G\Delta\{a, c\} = F\Delta\{u_1, \ldots, u_{2s}, u_{2s+1}, u_n\}$ feasible, which inductively will give an edge $\{u_{2l+1}, u_{2k+1}\}$ as above with $2k+1 = n$ and $2l+1 \leq 2s+1$ and thus a blossom. □

We describe next an algorithm for finding an augmenting path or a blossom. Let u be such that $\{u, \overline{u}\} \notin \mathcal{L}_F$. Start a breadth first search at u that assigns levels to elements of E as follows. The element u is at level 1. If u_{2j-1} is at level $2j - 1$, then put at level $2j$ all elements u_{2j} not at levels up to $2j - 1$ such that $\{u_{2j-1}, u_{2j}\}$ is an edge. If u_{2j} is at level $2j$, then put its mate $u_{2j+1} = \overline{u_{2j}}$ at level $2j + 1$ if it is not at a level up to $2j$ and $\{u_{2j}, u_{2j+1}\} \in \mathcal{L}_F$. The mate $\overline{u}$ of $u = u_1$ is omitted from the depth first search.

The algorithm terminates in one of four situations: (1) two distinct elements u_{2j}, v_{2j} are mates, in which case a blossom has been found; (2) two distinct elements u_{2j+1}, v_{2j+1} have an edge $\{u_{2j+1}, v_{2j+1}\}$, in which case a blossom has also been found; (3) an element u_{2j} at level $2j$ is such that $\{u_{2j}, \overline{u_{2j}}\} \notin \mathcal{L}_F$, in which case an augmenting path has been found; (4) an element u_{2j+1} at level $2j + 1$ is a special element, in which case an augmenting path has been found.

THEOREM 4.2. *The claims about having found an augmenting path or a blossom by the breadth first search in the four cases are correct.*

Proof. In case (1) the two paths from u_1 to u_{2j} and v_{2j} plus the mates u_{2j} and v_{2j} complete a blossom. In case (2) the two paths from u_1 to u_{2j+1} and v_{2j+1} plus the edge $\{u_{2j+1}, v_{2j+1}\}$ complete a blossom.

In case (3) we have a path $u_1, \ldots, u_{2j}$. We claim that it is an augmenting path, inductively on j. Let $\mathcal{K}$ be the subset of $\mathcal{L}$ consisting of the pairs $\{u_{2i}, u_{2i+1}\}$ for each $2 \leq 2i \leq 2j-4$, and obtain M', $\mathcal{L}'$ by contracting $\mathcal{K}$. Let F' be the feasible set in M' corresponding to F in M. We apply Lemma 3.4 to F' with $a = u_1$, $b = u_{2j-2}$, $c = u_{2j-1}$, $d = u_{2j}$. We have that $F'\Delta\{a, b\}$ is feasible for M' since $F_1 = F\Delta\{u_1, u_2, \ldots, u_{2j-3}, u_{2j-2}\}$ is feasible inductively for M and $\mathcal{K}_{F_1} = \mathcal{K}$, by removing $\{u_{2j-2}, u_{2j-1}\}$ from $\mathcal{L}$ and adding new mates for u_{2j-2}, u_{2j-1}, so that $u_1, \ldots, u_{2j-2}$ will be an augmenting path by induction. Also $F'\Delta\{c, d\}$ is feasible for M' since $F_2 = F\Delta\{u_{2j-1}, u_{2j}\}$ is feasible for M by the definition of an edge, and $\mathcal{K}_{F_2} = \mathcal{K}$. If $F'\Delta\{a\}$ is feasible for M', then $F\Delta S_1$ is feasible for M with $u_1 \in S_1$ and $S_1 \subseteq \{u_1, u_2, \ldots, u_{2j-3}\}$, and $\mathcal{K}_{F\Delta S_1} = \mathcal{K}$, so there must be an augmenting path contained in S_1. This subset, however, does not have all the edges and special vertices needed to satisfy the conditions in Theorem 4.1 when there exists an augmenting path, as otherwise they would have been found in the breadth first search. If $F'\Delta\{a, c\}$ is feasible for M', then $F\Delta S_2$ is feasible for M with $u_1, u_{2j-1} \in S_2$ and $S_2 \subseteq \{u_1, u_2, \ldots, u_{2j-3}, u_{2j-1}\}$, and $\mathcal{K}_{F\Delta S_2} = \mathcal{K}$. In this case u_{2j-1} would have been reached earlier in the breadth first search. Similarly, if $F'\Delta\{a, d\}$ is feasible for M', then $F\Delta S_3$ is feasible for M with $u_1, u_{2j} \in S_3$ and $S_3 \subseteq \{u_1, u_2, \ldots, u_{2j-3}, u_{2j}\}$, and $\mathcal{K}_{F\Delta S_3} = \mathcal{K}$, so u_{2j} would have been reached earlier in the breadth first search. Therefore $F'\Delta\{a, b, c, d\}$ is feasible for M', and so $F\Delta S_4$ is feasible for M with $u_1, u_{2j-2}, u_{2j-1}, u_{2j} \in S_4$ and $S_4 \subseteq \{u_1, \ldots, u_{2j}\}$, and $\mathcal{K}_{F\Delta S_4} = \mathcal{K}$. Furthermore, S_4 must be equal to this subset, otherwise u_{2j} would have been reached earlier in the breadth first search. This proves we have obtained an augmenting path that replaces F with $F\Delta\{u_1, \ldots, u_{2j}\}$.

The proof in case (4) is analogous. We have a path $u_1, \ldots, u_{2j+1}$. We show again that it has the elements of some augmenting path. Let $\mathcal{K}$ be the subset of $\mathcal{L}$ consisting of the pairs $\{u_{2i}, u_{2i+1}\}$ for each $2 \leq 2i \leq 2j-2$, and obtain M', $\mathcal{L}'$ by contracting $\mathcal{K}$. Let F' be the feasible set in M' corresponding to F in M. We apply Lemma 3.2 to F' with $a = u_1$, $b = u_{2j}$, $c = u_{2j+1}$. We have that $F'\Delta\{c\}$ is feasible for M' since $F_3 = F\Delta\{u_{2j+1}\}$ is feasible for M because u_{2j+1} is a special element and $\mathcal{K}_{F_3} = \mathcal{K}$. Also $F'\Delta\{a, b\}$ is feasible for M' since $F_4 = F\Delta\{u_1, u_2, \ldots, u_{2j-1}, u_{2j}\}$ is feasible by the preceding case of an even length path. If $F'\Delta\{a\}$ is feasible for M', then $F\Delta S_5$ is feasible for M with $u_1 \in S_5$ and $S_5 \subseteq \{u_1, \ldots, u_{2j-1}\}$, and $\mathcal{K}_{F\Delta S_5} = \mathcal{K}$, so there must be an augmenting path contained in S_5, which does not have the edges in special vertices to satisfy the conditions in Theorem 4.1. If $F'\Delta\{a, c\}$ is feasible for M', then $F\Delta S_6$ is feasible for M with $u_1, u_{2j+1} \in S_6$ and $S_6 \subseteq \{u_1, \ldots, u_{2j-1}, u_{2j+1}\}$, and $\mathcal{K}_{F\Delta S_6} = \mathcal{K}$, which is not possible since u_{2j+1} would then have been reached earlier by the breadth first search. Therefore $F'\Delta\{a, b, c\}$ is feasible for M', and so $F\Delta S_7$ is feasible for M with $u_1, u_{2j}, u_{2j+1} \in S_7$ and $S_7 \subseteq \{u_1, \ldots, u_{2j+1}\}$, and $\mathcal{K}_{F\Delta S_7} = \mathcal{K}$. Furthermore, S_7 must be equal to this subset, otherwise u_{2j} would have been reached earlier in the breadth first search. This proves we have obtained an augmenting path that replaces F with $F\Delta\{u_1, \ldots, u_{2j+1}\}$. □

5. Delta-matroid intersection without equality. So far the argument has been carried in the full generality of arbitrary delta-matroids and the general par-

ity problem. The arguments usually become more difficult with the introduction of blossoms, which can contain other blossoms, and this can lead to requiring the delta-matroid to have a presentation that is not only by means of an oracle, or some other special structure, such as in the case of linear or local delta-matroids [7, 3]. In special cases with restrictions involving the equal delta-matroid $M_=$ and the not-equal delta-matroid $M_\neq$, blossoms can be more easily handled. This leads to our main result.

Suppose the algorithm of Theorem 4.2 found a blossom as in Theorem 4.1. Restrict the breadth search for an augmenting path to the elements $w_1, \ldots, w_{2r}, \ldots, w_k$ of the blossom. When we restrict the breadth first search further by excluding some w_{2i}, w_{2i+1} we may find a smaller blossom or an augmenting path as in Theorem 4.2. We may thus assume this does not happen for the blossom under consideration.

THEOREM 5.1. *There is a polynomial time algorithm using an oracle for any instance of delta-matroid parity on a delta-matroid M with pairing $\mathcal{L}$ such that no subset $\mathcal{K} \subset \mathcal{L}$ and pair $\{a, b\} \in \mathcal{L} \setminus \mathcal{K}$ are such that the delta-matroid M' obtained from M by contracting K has the not-equal delta-matroid $M_\neq$ on $\{a, b\}$ as a restriction.*

Proof. Let $w_1, \ldots, w_{2r}, \ldots, w_k$ be the blossom obtained above. Let $\mathcal{K}$ be the pairs $\{w_{2i}, w_{2i+1}\}$ for $2 \leq 2i \leq k-2$. Contracting $\mathcal{K}$ in M, we obtain M' with corresponding feasible set F'. Let $a = w_1$, $b = w_k$, $c = w_{k-1}$. The set $F'\Delta\{a, b\}$ is feasible by the augmenting path $w_1, \ldots, w_{2r-1}, w_k$ obtained after removing the pair $\{w_{k-1}, w_k\}$ from $\mathcal{L}$ and adding new mates for w_{k-1}, w_k, using Theorem 4.2. The set $F'\Delta\{a, c\}$ is feasible by the augmenting path $w_1, \ldots, w_{k-1}$ obtained also after removing the pair $\{w_{k-1}, w_k\}$ from $\mathcal{L}$ and adding new mates for w_{k-1}, w_k, using Theorem 4.2. Setting $G' = F'\Delta\{a\}$, we have that $G'\Delta\{b\}$ and $G'\Delta\{c\}$ are feasible, giving $M_\neq$ on $\{b, c\}$ as a restriction unless G' or $G'\Delta\{b, c\}$ is feasible.

If G' or $G'\Delta\{b, c\}$ is feasible, then there is an augmenting path involving a subset of $w_1, \ldots, w_k$, and this augmenting path cannot miss any $\{w_{2i}, w_{2i+1}\}$ by the choice of the blossom, so the elements $w_1, \ldots, w_k$ form an augmenting path in some order. The remaining case has $M_\neq$ on $\{b, c\}$ simulated by $G'\Delta\{b\}$ and $G'\Delta\{c\}$, contrary to assumption. □

We infer three results as special cases.

LEMMA 5.2. *Let $M = (\{a, b, c, d\}, \mathcal{F})$ be a delta-matroid that contracts to M' on $\{c, d\}$ using $\mathcal{K} = \{\{a, b\}\}$. If M' is either the $M_=$ or the $M_\neq$ delta-matroid, but M does not have M' on $\{c, d\}$ as a restriction, then $|F\Delta G|$ is even for all $F, G \in \mathcal{F}$.*

Proof. We have $M' = (\{c, d\}, \{F, F\Delta\{c, d\}\})$. By assumption on M, we have a feasible F such that $F\Delta\{a, b, c, d\}$ is feasible, but $F\Delta\{a, b\}$, $F\Delta\{c, d\}$ are not feasible, and furthermore $F\Delta\{c\}$, $F\Delta\{d\}$, $F\Delta\{a, b, c\}$, $F\Delta\{a, b, d\}$ are not feasible. This guarantees that M' is not a restriction of M and that contracting $\mathcal{K} = \{\{a, b\}\}$ gives M'.

If $F\Delta\{a\}$ is feasible, then taking $A = F\Delta\{a\}$, $B = F\Delta\{a, b, c, d\}$, and $x = b$ in the definition of delta-matroid yields a contradiction. If $F\Delta\{b\}$ is feasible, then taking $A = F\Delta\{b\}$, $B = F\Delta\{a, b, c, d\}$, and $x = a$ in the definition of delta-matroid yields a contradiction. If $F\Delta\{a, c, d\}$ is feasible, then taking $A = F\Delta\{a, c, d\}$, $B = F$, and $x = a$ in the definition of delta-matroid yields a contradiction. If $F\Delta\{b, c, d\}$ is feasible, then taking $A = F\Delta\{b, c, d\}$, $B = F$, and $x = b$ in the definition of delta-matroid yields a contradiction. Thus $|F\Delta G|$ is even for all feasible G. □

We shall make use of Wenzel's *strong exchange axiom* for even delta-matroids [12] $M = (E, \mathcal{F})$, which states that for all $A, B \in \mathcal{F}$ and $x \in A\Delta B$, there exists $y \in A\Delta B$ such that $A\Delta\{x, y\} \in \mathcal{F}$ and $B\Delta\{x, y\}$.

The first special case is as follows.

THEOREM 5.3. *There is a polynomial time algorithm using an oracle for delta-matroid parity on delta-matroids M that do not have the not-equal delta-matroid $M_{\neq}$ as a restriction.*

Proof. It suffices that if M' is obtained from M by contracting $\mathcal{K}$, then M' does not have $M_{\neq}$ as a restriction either, so that the algorithm of Theorem 5.1 applies. Suppose $M_{\neq}$ on $\{c,d\}$ is obtained as a restriction after contracting $\mathcal{K} = \{\{a,b\}\}$, but not before. We may restrict M to $\{a,b,c,d\}$ and apply Lemma 5.2. We thus have a feasible $F = \{c\}$ such that $|F\Delta G|$ is even for all feasible G and with $F\Delta\{a,b,c,d\}$ also feasible. By Wenzel's strong exchange axiom there are two other complementary feasible sets, say, $F\Delta\{a,c\}$ and $F\Delta\{b,d\}$. Restricting M to feasible sets that do not contain either b or d, we obtain two feasible sets $\{a\},\{c\}$ giving $M_{\neq}$ as a restriction on $\{a,c\}$ before contracting $\mathcal{K}$. □

The second special case is as follows.

THEOREM 5.4. *There is a polynomial time algorithm using an oracle for delta-matroid parity on a delta-matroid M with pairing $\mathcal{L}$ for which there exist two disjoint sets of elements S,T each containing one element from each pair in $\mathcal{L}$ such that if the not-equal delta-matroid $M_{\neq}$ is a restriction of M on $\{a,b\}$, then both a and b are in S, and if the equal delta-matroid $M_{=}$ is a restriction of M on $\{a,b\}$, then at least one of a,b is in S.*

Proof. It suffices that if M' is obtained from M by contracting $\mathcal{K} \subseteq \mathcal{L}$, then M' also satisfies the property in the theorem, so that M' does not have $M_{\neq}$ as a restriction on $\{a,b\}$ with at most one of a,b in S, and thus for $\{a,b\} \in \mathcal{L}$, and the algorithm of Theorem 5.1 applies. We show this by induction on $|\mathcal{K}|$.

Suppose $M_{=}$ or $M_{\neq}$ on $\{c,d\}$ is obtained as a restriction after contracting $\mathcal{K} = \{\{a,b\}\}$ with c and d in T. We may restrict M to $\{a,b,c,d\}$ and apply Lemma 5.2. We thus have a feasible F such that $|F\Delta G|$ is even for all feasible G, and with $F\Delta\{a,b,c,d\}$ also feasible. By Wenzel's strong exchange axiom there are two other complementary feasible sets, say, $F\Delta\{a,c\}$ and $F\Delta\{b,d\}$. If a is in T, then F and $F\Delta\{a,c\}$ give $M_{=}$ or $M_{\neq}$ as a restriction on $\{a,c\}$ with both a and c in T, which is not possible by inductive hypothesis. Otherwise b is in T, and then F and $F\Delta\{b,d\}$ give $M_{=}$ or $M_{\neq}$ as a restriction on $\{b,d\}$ with both b and d in T, which is not possible by inductive hypothesis.

Suppose $M_{\neq}$ on $\{c,d\}$ is obtained as a restriction after contracting $\mathcal{K} = \{\{a,b\}\}$, with c in S and d in T, but not before. We may restrict M to $\{a,b,c,d\}$ and apply Lemma 5.2. We thus have a feasible $F = \{c\}$ or $F = \{d\}$ such that $|F\Delta G|$ is even for all feasible G, and with $F\Delta\{a,b,c,d\}$ also feasible. By Wenzel's strong exchange axiom there are two other complementary feasible sets, say, $F\Delta\{a,c\}$ and $F\Delta\{b,d\}$. If $F = \{d\}$, then the two feasible sets $\{b\},\{d\}$ give a $M_{\neq}$ delta-matroid on $\{b,d\}$ with d in T, which is not possible by inductive hypothesis. If b is in T and $F = \{c\}$, then the two feasible sets F and $F\Delta\{b,d\}$ give an $M_{=}$ delta-matroid on $\{b,d\}$ with both b,d in T, which is not possible. Otherwise a is in T and $F = \{c\}$, and then the two feasible sets $\{a\},\{c\}$ give a $M_{\neq}$ delta-matroid on $\{a,c\}$ with a in T, which is not possible by inductive hypothesis. □

COROLLARY 5.5. *There is a polynomial time algorithm using an oracle for delta-matroid intersection on two delta-matroids M_1, M_2 where the delta-matroid M_1 is arbitrary, and the delta-matroid M_2 does not have either the equal delta-matroid $M_{=}$ or the not-equal delta-matroid $M_{\neq}$ as a restriction.*

Proof. Apply Theorem 5.4 with S consisting of the elements in M_1 and T consisting of the elements in M_2. □

The third special case is as follows.

THEOREM 5.6. *There is a polynomial time algorithm using an oracle for delta-matroid parity on a delta-matroid M with pairing $\mathcal{L}$ for which there exist two disjoint sets of elements S, T each containing one element from each pair in $\mathcal{L}$ such that if the not-equal delta-matroid $M_{\neq}$ is a restriction of M on $\{a, b\}$, then either both a and b are in S or both a and b are in T, and if the equal delta-matroid $M_{=}$ is a restriction of M on $\{a, b\}$, then one of a, b is in S and the other one is in T.*

Proof. It suffices that if M' is obtained from M by contracting $\mathcal{K} \subseteq \mathcal{L}$, then M' also satisfies the property in the theorem, so that M' does not have $M_{\neq}$ as a restriction on $\{a, b\}$ with one of a, b in S and the other one in T, and thus for $\{a, b\} \in \mathcal{L}$, and the algorithm of Theorem 5.1 applies. We show this by induction on $|\mathcal{K}|$.

Suppose $M_{\neq}$ on $\{c, d\}$ is obtained as a restriction after contracting $\mathcal{K} = \{\{a, b\}\}$, with c in S and d in T, but not before. We may restrict M to $\{a, b, c, d\}$ and apply Lemma 5.2. We thus have a feasible $F = \{c\}$ such that $|F \Delta G|$ is even for all feasible G, and with $F\Delta\{a, b, c, d\}$ also feasible. By Wenzel's strong exchange axiom there are two other complementary feasible sets, say, $F\Delta\{a, c\}$ and $F\Delta\{b, d\}$. If a is in T, then restricting M to feasible sets that do not contain either b or d, we obtain two feasible sets $\{a\}, \{c\}$ giving $M_{\neq}$ as a restriction on $\{a, c\}$ before contracting $\mathcal{K}$, which is not possible by inductive hypothesis. Otherwise b is in T, and restricting M to feasible sets that contain a and do not contain c, we obtain two feasible $\{a\}$ and $\{a, b, d\}$ giving $M_{=}$ as a restriction on $\{b, d\}$, which is not possible by inductive hypothesis.

Suppose $M_{=}$ on $\{c, d\}$ is obtained as a restriction after contracting $\mathcal{K} = \{\{a, b\}\}$, with c, d both in S, but not before. Say a is in S and b is in T. We may restrict M to $\{a, b, c, d\}$ and apply Lemma 5.2. We thus have a feasible $F = \emptyset$ or $F = \{a, b\}$ such that $|F \Delta G|$ is even for all feasible G, and with $F\Delta\{a, b, c, d\}$ also feasible. By Wenzel's strong exchange axiom there are two other complementary feasible sets, say, $F\Delta\{a, c\}$ and $F\Delta\{b, d\}$. If $F = \emptyset$, then restricting M to feasible sets that do not contain either b or d, we obtain two feasible sets $\emptyset, \{a, c\}$ giving $M_{=}$ as a restriction on $\{a, c\}$ with both a and c in S, which is not possible by inductive hypothesis. Otherwise $F = \{a, b\}$, and restricting M to feasible sets that contain a and do not contain c, we obtain two feasible sets $\{a, b\}$ and $\{a, d\}$ giving $M_{\neq}$ as a restriction on $\{b, d\}$ with b in T and d in S, which is not possible by inductive hypothesis. □

COROLLARY 5.7. *There is a polynomial time algorithm using an oracle for delta-matroid intersection on two delta-matroids M_1, M_2 that do not have the equal delta-matroid $M_{=}$ as a restriction. This generalizes matroid intersection, as matroids do not have the equal delta-matroid $M_{=}$ as a restriction.*

Proof. Apply Theorem 5.6 with S consisting of the elements in M_1 and T consisting of the elements in M_2. □

We note also that intersecting two matroids M_1 and M_2 is equivalent to intersecting M_1^- consisting of the independent sets of M_1, and M_2^+ consisting of the spanning sets of M_2. Furthermore both M_1^- and M_2^+ have neither $M_{=}$ nor $M_{\neq}$ as a restriction. Therefore all the results above generalize matroid intersection.

6. Bipartite Boolean constraint satisfaction. Corollary 5.7 also completes the classification of bipartite Boolean constraint satisfaction from [5, 11] as mentioned in the introduction. A *constraint* C on a set of Boolean variables X is a set of Boolean assignments x to the variables in X. A *restriction* of C by an assignment y to $Y \subseteq X$ is the constraint $C_{X,y}$ on the variables $X \setminus Y$ consisting of all assignments z such that if x is the assignment to X that agrees with y on Y and agrees with z on $X \setminus Y$, then x is in $C_{X,y}$. The *bipartite Boolean constraint satisfaction problem* on a set $\mathcal{C}$

of allowed constraints has an instance consisting of two sets of constraints S and T on subsets of a set of variables X, where each constraint in S or T corresponds to a constraint in $\mathcal{C}$ under some correspondence of variables, and each variable in X occurs in at most one constraint in S and at most one constraint in T. The aim is to assign Boolean values to the variables in X so as to satisfy the constraints in S and the constraints in T simultaneously. The bipartite case of Boolean constraint satisfaction differs from the general case with two occurrences per variable only when the equality constraint $\{00, 11\}$ is not an allowed constraint in $\mathcal{C}$. Let the inequality constraint be $\{10, 01\}$. A constraint is a delta-matroid if the collection of subsets of a set E with n elements defining a delta-matroid is viewed as a collection of assignments to n Boolean variables defining a constraint $\mathcal{C}$, where a 0 or 1 in a bit position corresponds to presence or absence of an element in the subset.

THEOREM 6.1. *Every bipartite Boolean constraint satisfaction problem, with a set of allowed constraints closed under restriction, and where equality is not an allowed constraint, is one of Schaefer's polynomial cases, or polynomial by delta-matroid intersection without $M_=$ as a restriction, or is NP-complete. The polynomial cases remain polynomial even when the two sides of the bipartition are given by an oracle that answers whether a restriction $S_{X,y}$ or $T_{X,y}$ of either side of the bipartition is nonempty. This oracle result holds in the general case where equality is an allowed constraint for Schaefer's polynomial cases and for delta-matroids that do not contain inequality.*

Proof. The classification is obtained by Feder [5], and the remaining open case of delta-matroid intersection without $M_=$ as a restriction is polynomial by Corollary 5.7 using an oracle. When $M_=$ is allowed in delta-matroids, forbidding $M_{\neq}$ gives polynomiality by Theorem 5.3 using an oracle. The polynomial cases of Schaefer [11] are the following: (1) each constraint is a conjunction of 2-satisfiability clauses, (2) each constraint is a conjunction of Horn clauses, (3) each constraint is a conjunction of dual-Horn clauses, and (4) each constraint is a conjunction of linear equations modulo 2. An oracle in (1) allows us to obtain all the 2-satisfiability clauses and solve the problem. An oracle in (2) (resp., (3)) allows us to obtain all the variables forced to value 1 (resp., value 0) by some clause, and once no variable is forced, the remaining variables can be assigned value 0 (resp., value 1) if a solution exists.

For (4), we consider a candidate assignment x to the variables X and if this candidate assignment does not satisfy one of the two oracles, we obtain an assignment y to a subset of variables $Y \subseteq X$ such that y is a restriction of x and does not satisfy the oracle, yet every restriction of y to $|Y| - 1$ variables in Y satisfies the oracle. This implies that the single linear equation involving precisely the variables in Y not satisfied by y must be satisfied by all assignments in the oracle. We then repeat the process for a candidate assignment x to the variables X satisfying this equation, and this assignment either satisfies both oracles or provides another equation. We proceed to add equations until a solution is found, or until the equations obtained so far are not satisfiable. Note that at most $n = |X|$ equations will be obtained, since each equation reduces by one the number of free variables, so the process terminates in polynomial time. □

We now proceed to the classification of bipartite Boolean constraint satisfaction problems in the case where the allowed constraint types may not be the same for both sides of the bipartition. Let $\mathcal{A}$ and $\mathcal{B}$ be two sets of constraint types. We say that $(\mathcal{A}, \mathcal{B})$ *simulates* constraint C on $\mathcal{A}$ if there exists an instance of bipartite Boolean constraint satisfaction with constraints from $\mathcal{A}$ in one side and constraints from $\mathcal{B}$ in the other side, with every variable constrained exactly once in the $\mathcal{A}$ side

and constrained at most once in the $\mathcal{B}$ side, such that the variables that are not constrained in the $\mathcal{B}$ side are the variables of C, and the set of assignments of values to variables in C for which there exists a solution to this instance is the same as C. We say that $(\mathcal{A},\mathcal{B})$ simulates constraint C on $\mathcal{B}$ if $(\mathcal{B},\mathcal{A})$ simulates constraint C on $\mathcal{B}$.

Let $\mathcal{A},\mathcal{B},\mathcal{A}',\mathcal{B}'$ be sets of constraint types. We say that $(\mathcal{A},\mathcal{B})$ *simulates* $(\mathcal{A}',\mathcal{B}')$ if $(\mathcal{A},\mathcal{B})$ simulates every constraint $C\in\mathcal{A}'$ on $\mathcal{A}$ and simulates every constraint $C\in\mathcal{B}'$ on $\mathcal{B}$. We say that $(\mathcal{A},\mathcal{B})$ is *closed under simulation* if whenever $(\mathcal{A},\mathcal{B})$ simulates $(\mathcal{A}',\mathcal{B}')$ we have $\mathcal{A}'\subseteq\mathcal{A}$ and $\mathcal{B}'\subseteq\mathcal{B}$.

THEOREM 6.2. *Let $\mathcal{A},\mathcal{B}$ be sets of constraint types such that $(\mathcal{A},\mathcal{B})$ is closed under simulation and both $\mathcal{A},\mathcal{B}$ contain the single variable constraints $\{0\}$, $\{1\}$, and $\{0,1\}$. Then the bipartite Boolean constraint satisfaction with constraints from $\mathcal{A}$ in one side and from $\mathcal{B}$ in the other has* (1) *polynomial cases derived from Schaefer's classification;* (2) *polynomial cases derived from delta-matroid intersection in the case where neither side has equality and in the case where one side has neither equality nor inequality;* (3) *a polynomial case that combines* 2*-satisfiability and delta-matroid intersection. If a problem is not in cases* (1), (2), *or* (3), *then either* (4) *$\mathcal{A}$ is the same as $\mathcal{B}$ and consists of delta-matroids including equality, so the problem is a delta-matroid parity problem, or* (5) *the problem is NP-complete. The polynomial cases* (1), (2), (3) *remain polynomial in the oracle model as in Theorem* 6.1.

We define some specific constraint types on variables x,y,z. Let $[x=y]$ be $\{00,11\}$. Let $[x\neq y]$ be $\{10,01\}$. Let $[x\leq y]$ be $\{00,01,11\}$. Let $[x\vee y]$ be $\{10,01,11\}$. Let $[x=y=z]$ be $\{000,111\}$. Let [1-3 x,y,z] be $\{100,010,001\}$. Let $[x\vee y\vee z]$ be $\{100,010,001,110,101,011,111\}$. Let $[x\leq y,z]$ be $\{000,001,010,011,111\}$. Let $[x+y+z=0]$ be $\{000,110,101,011\}$, and let $[x+y+z=1]$ be $\{100,010,001,111\}$. Let $[\approx x\vee y\vee z]$ be any constraint satisfying [1-3 x,y,z] $\subseteq[\approx x\vee y\vee z]\subseteq[x\vee y\vee z]$. Let $[\approx x\leq y,z]$ be any constraint satisfying $[x=y=z]\subseteq[\approx x\leq y,z]\subseteq[x\leq y,z]$. For these constraint types, we denote by $\overline{x}$ the complement of variable x, and by $\tilde{x}$ a literal that may be either x or $\overline{x}$. Feder [5] showed the following.

LEMMA 6.3. *For a given constraint C, we have that $(\{C\},\{\{0\},\{1\},\{0,1\}\})$ simulates* (1) *$[x\vee y]$ or $[x\neq y]$ if C is not Horn;* (2) *$[\overline{x}\vee\overline{y}]$ or $[x\neq y]$ if C is not dual-Horn;* (3) *$[x\leq y]$ or $[x\vee y]$ or $[\overline{x}\vee\overline{y}]$ if C is not linear;* (4) *some $[\approx\tilde{x}\vee\tilde{y}\vee\tilde{z}]$ if C is not* 2*-SAT; and* (5) *some $[\approx\tilde{x}\leq\tilde{y},\tilde{z}]$ if C is not a delta-matroid.*

If $X=x_1x_2\cdots x_k$ and $Y=y_1y_2\cdots y_k$ are k-bit vectors, write $X\leq Y$ if $x_i\leq y_i$ for all $1\leq i\leq k$, write $X<Y$ if $X\leq Y$ and $X\neq Y$, and let $d(X,Y)$ be the Hamming distance between X and Y, that is, the number of bits $1\leq i\leq k$ such that $x_i\neq y_i$.

LEMMA 6.4. *For a given constraint C,* (1) *if $(\{C\},\{\{0\},\{1\},\{0,1\}\})$ simulates neither $[x=y]$ nor $[x\leq y]$, then for every $X,Y\in C$ with $X\leq Y$ we have that every Z such that $X\leq Z\leq Y$ satisfies $Z\in C$;* (2) *if $(\{C\},\{\{0\},\{1\},\{0,1\}\})$ simulates neither $[x\neq y]$ nor $[\overline{x}\vee\overline{y}]$, then there exists $X\in C$ such that $Z\leq X$ for every $Z\in C$;* (3) *if $(\{C\},\{\{0\},\{1\},\{0,1\}\})$ simulates neither $[x\neq y]$ nor $[x\vee y]$, then there exists $X\in C$ such that $X\leq Z$ for every $Z\in C$.*

Proof. For (1), if $X\leq X'<Y'\leq Y$ with $X',Y'\in C$ and $d(X',Y')\geq 2$, then there exists $X'<Z'<Y'$ such that $Z'\in C$. Otherwise we can consider the restriction C' of C to bit vectors T such that $t_i=b_i$ if $x'_i=y'_i=b_i$, and select i,j such that $x'_i<y'_i$ and $x'_j<y'_j$, so that the condition defined by C' on bit positions i,j is $[x_i=x_j]$. Thus by induction there exist $X=X^0<X^1<\cdots<X^k=Y$ with $d(X^i,X^{i+1})=1$ and each $X^i\in C$. Consider the restriction C' of C to bit vectors T such that $t_i=b_i$ if $x_i=y_i=b_i$ and say X^i has $x^i_j=1$ for $1\leq j\leq i$ and $x^i_j=0$ for $i<j\leq k$. Assume inductively that if $X\leq T\leq X^i$, then $T\in C$. Suppose Z is

such that $X \leq Z \leq X^{i+1}$ and $Z \notin C$ with $d(Z, X^{i+1})$ minimum. Then $z_{i+1} = 1$, and choosing $1 \leq j \leq i$ such that $z_j = 0$, we have that the bit vectors T obtained from Z by changing z_j or z_{i+1} or both are in C, thus giving $[x_{i+1} \leq x_j]$, completing the induction and the proof of (1).

For (2), if the condition does not hold, then there exist $X, Y \in C$ such that there is no $T \in C$ with $X < T$ or $Y < T$, and $d(X, Y) \geq 2$. Choose $X, Y \in C$ such that if we consider the restriction C' of C to bit vectors T with $t_i = b_i$ if $x_i = y_i = b_i$, then there is not $T \in C'$ with $X < T$ or $Y < T$, and $d(X, Y) \geq 2$ is minimum with this property. The minimality of $d(X, Y)$ implies that if $Z \in C'$, then $Z \leq X$ or $Z \leq Y$; otherwise some $Z' \in C'$ with $Z' \geq Z$ is such that there is no $T \in C'$ with $Z' < T$ and $Z' \neq X, Y$, so that $2 \leq d(X, Z') < d(X, Y)$, contrary to minimality. Let i, j be bit positions such that $x_i = 0, x_j = 1, y_i = 1, y_j = 0$. Then there is no $Z \in C'$ such that $z_i = z_j = 1$, so the condition defined by C' on bit positions i and j is either $[x_i \neq x_j]$ or $[\overline{x_i} \vee \overline{x_j}]$, proving (2). The proof for (3) is the same as for (2). □

A constraint C is *upward closed* if for every $X \in C$, if $X \leq Z$, then $Z \in C$, and *downward closed* if for every $X \in C$, if $Z \leq X$, then $Z \in C$. The upward closure of a constraint C is the constraint $\mathrm{up}(C)$ consisting of the bit vectors Z such that there exists $X \in C$ with $X \leq Z$. The downward closure of a constraint C is the constraint $\mathrm{down}(C)$ consisting of the bit vectors Z such that there exists $X \in C$ with $Z \leq X$.

LEMMA 6.5. *Let $\mathcal{A}, \mathcal{B}$ be as in the statement of Theorem* 6.2*, and suppose every constraint in $\mathcal{A}$ can be decomposed into an upward closed constraint, and constraints $\{0\}$. Then the Boolean constraint satisfaction problem with constraints from $\mathcal{A}$ in one side and from $\mathcal{B}$ in the other is polynomial in cases* (1) *for every $C \in \mathcal{B}$,* $\mathrm{down}(C)$ *can be decomposed into $\{0\}$, $\{0,1\}$ constraints, or every $C \in \mathcal{A}$ can be decomposed into $\{0\}$, $\{1\}$, $\{0,1\}$ constraints;* (2) *the constraints in $\mathcal{A}$ are delta-matroids, and for every $C \in \mathcal{B}$,* $\mathrm{down}(C)$ *is a delta-matroid;* (3) *the constraints in $\mathcal{A}$ are 2-SAT, and for every $C \in \mathcal{B}$,* $\mathrm{down}(C)$ *is 2-SAT. If we are not in cases* (1), (2), (3)*, then $\mathcal{A}$ is NP-complete.*

Proof. We show that the problem reduces to the problem where $\mathcal{B}$ is replaced by $\mathrm{down}(\mathcal{B})$ consisting of the constraints $\mathrm{down}(C)$ for $C \in \mathcal{B}$. Given an instance of the problem with $\mathcal{B}$, if a constraint C in the $\mathcal{B}$ side has a variable x constrained by $\{0\}$ in $\mathcal{A}$, restrict C to bit vectors satisfying $x = 0$. If the resulting instance has a solution, then it also has a solution with each C in the $\mathcal{B}$ side replaced with $\mathrm{down}(C)$. If the resulting instance has a solution with each C in the $\mathcal{B}$ side replaced with $\mathrm{down}(C)$, then we may replace the X chosen from some $\mathrm{down}(C)$ with a $Y \in C$ such that $X \leq Y$. This gives a solution with C, since replacing X with $Y \geq X$ will still satisfy the constraints in the $\mathcal{A}$ side, because these are upward closed.

If $\mathrm{down}(C)$ in $\mathrm{down}(\mathcal{B})$ can always be decomposed into $\{0\}$ and $\{0,1\}$ constraints, then it suffices to test the corresponding restriction of the $\mathcal{A}$ side. The same argument holds if $C \in \mathcal{A}$ can be decomposed into $\{0\}, \{1\}, \{0,1\}$ constraints.

Suppose $\mathcal{A}$ is delta-matroid and $C \in \mathcal{A}$ cannot be decomposed into $\{0\}$, $\{1\}$, $\{0,1\}$ constraints. Then by (3) of Lemma 6.4 we can simulate $[x \neq y]$ or $[x \vee y]$ in the $\mathcal{A}$ side, and in fact we can simulate $[x \vee y]$ since $[x \neq y]$ is not upward closed. Given a constraint $D \in \mathcal{B}$ use for every variable x_i in D a corresponding condition $[x_i \vee y_i]$ in $\mathcal{A}$ to simulate a constraint C in $\mathcal{A}$ with variables y_i. The constraint C is obtained from $\mathrm{down}(D)$ by complementing all bits. Since $\mathcal{A}$ is delta-matroid and closed under simulation, it follows that C is delta-matroid and thus $\mathrm{down}(D)$ is delta-matroid. Once both sides are delta-matroids, the fact that the constraints in $\mathcal{A}$ are upward closed implies that they do not have $[x = y]$ or $[x \neq y]$ as a restriction, and

the intersection of a delta-matroid without equality or inequality with an arbitrary delta-matroid is polynomial by Corollary 5.5.

Suppose $\mathcal{A}$ is 2-SAT, and $C \in \mathcal{A}$ cannot be decomposed into $\{0\}, \{1\}, \{0,1\}$ constraints. Then as in the delta-matroid case we get $[x \vee y]$ in the $\mathcal{A}$ side, and so for every constraint $D \in \mathcal{B}$ we get the constraint C obtained by complementing all bits of down(B) in $\mathcal{A}$, so since $\mathcal{A}$ is 2-SAT and closed under simulation, it follows that down(D) is 2-SAT as well. The problem is thus reduced to 2-SAT and therefore polynomial.

In the remaining case, $\mathcal{A}$ is not delta-matroid or 2-SAT, and for some C in $\mathcal{B}$ we have that down(C) cannot be decomposed into $\{0\}, \{0,1\}$ constraints. Then by (3) of Lemma 6.4 we can simulate $[x \neq y]$ or $[\overline{x} \vee \overline{y}]$ in the $\mathcal{B}$ side, obtaining $[\overline{x} \vee \overline{y}]$ as the downward closure. By (4) of Lemma 6.3, we can simulate some $[\approx \tilde{x} \vee \tilde{y} \vee \tilde{z}]$ in the $\mathcal{A}$ side, and the only such constraint that is upward closed as required for the $\mathcal{A}$ side is $[x \vee y \vee z]$. By (5) of Lemma 6.3, we can simulate some $[\approx \tilde{x} \leq \tilde{y}, \tilde{z}]$ in the $\mathcal{A}$ side, and the only such constraint that is upward closed as required for the $\mathcal{A}$ side is $[\overline{x} \leq y, z]$, which we denote also by $[x \vee y, z]$.

We thus have $[x \vee y \vee z]$, $[x \vee y, z]$ in the $\mathcal{A}$ side. Combining these with conditions $[\overline{x} \vee \overline{x'}]$, $[\overline{y} \vee \overline{y'}]$, $[\overline{z} \vee \overline{z'}]$ on the $\mathcal{B}$ side gives corresponding $[\overline{x} \vee \overline{y} \vee \overline{z}]$, $[\overline{x} \vee \overline{y}, \overline{z}]$ in the $\mathcal{B}$ side. We do a reduction from 3-SAT. A 3-SAT clause that has both positive and negative literals can be decomposed into a clause that has only positive and a clause that has only negative literals, so that the two give the original 3-SAT clause by resolution. We already have positive and negative 3-SAT clauses simulated. By combining $[x \vee y \vee z]$ in $\mathcal{A}$ with $[\overline{z} \vee \overline{z'}]$ in $\mathcal{B}$ and with $[z' \vee z_1, z_2]$ in $\mathcal{A}$, we obtain $[x \vee y \vee z_1, z_2]$ in $\mathcal{A}$. This creates the multiple copies of variable z in the 3-SAT clause needed to combine with corresponding copies of $\overline{z}$ in 3-SAT clauses for $\mathcal{B}$, which can be obtained analogously. We thus have multiple copies of variables in clauses $[x \vee y \vee z]$ in $\mathcal{A}$ and clauses $[\overline{x} \vee \overline{y} \vee \overline{z}]$ in $\mathcal{B}$ as needed to complete the reduction and get NP-completeness. □

Note that the same result holds if we exchange upward closed with downward closed. A constraint C is Horn if every nonempty restriction C' of C has a least element. Thus in the first part of case (1) in Lemma 6.5, both $\mathcal{A}$ and $\mathcal{B}$ are dual-Horn. In the second part of case (1) the constraints in $\mathcal{A}$ decompose into monadic relations.

LEMMA 6.6. *Let $\mathcal{A}, \mathcal{B}$ be as in the statement of Theorem* 6.2*, and suppose some constraint in $\mathcal{A}$ cannot be decomposed into an upward closed constraint, and constraints $\{0\}$, and some constraint in $\mathcal{A}$ cannot be decomposed into a downward closed constraint, and constraints $\{1\}$. Then we have the polynomial cases where $\mathcal{A}$ and $\mathcal{B}$ are both Horn, both dual-Horn, both 2-SAT, both linear, or with at most one side having $[x = y]$ the polynomial case of delta-matroids. In the remaining cases, either both sides are delta-matroids with $[x = y]$ and $\mathcal{A} = \mathcal{B}$, corresponding to delta-matroid parity, or the $\mathcal{A}$ side is neither Horn, dual-Horn, 2-SAT, linear, or delta-matroid.*

Proof. If $\mathcal{A}$ is a delta-matroid and $\mathcal{B}$ is not a delta-matroid, then by (5) of Lemma 6.3 we have some $[\approx \tilde{x} \leq \tilde{y}, \tilde{z}]$ in $\mathcal{B}$. In the cases where $\mathcal{A}$ has one of $[t = t']$, $[t \neq t']$, $[t \leq t']$ or both $[t \vee t']$ and $[\overline{t} \vee \overline{t'}]$, then combining such conditions for t or t' being x, y, z and corresponding x', y', z' we get $[\approx \tilde{x} \leq \tilde{y}, \tilde{z}]$ in $\mathcal{A}$, which is not a delta-matroid, contrary to assumption. We thus have of these choices for t, t' either just $[t \vee t']$ or just $[\overline{t} \vee \overline{t'}]$. By cases (1), (2), (3) of Lemma 6.4, we have that $\mathcal{A}$ decomposes into constraints $\{0\}$, $\{1\}$, and just upward closed constraints or just downward closed constraints, contrary to assumption.

If both $\mathcal{A}$ and $\mathcal{B}$ are delta-matroids, and $\mathcal{A}$ does not have $[x = y]$, then either $\mathcal{B}$ does not have $[x = y]$ either and the problem is polynomial by Corollary 5.7, or $\mathcal{B}$ does have $[x = y]$, in which case $\mathcal{A}$ does not have $[x \neq x']$ since using also $[y \neq y']$ would give $[x' = y']$ in $\mathcal{A}$ as well, so the problem is polynomial by Corollary 5.5. If both sides have $[x = y]$, then every constraint in $\mathcal{A}$ can also be obtained in $\mathcal{B}$ and viceversa, so $\mathcal{A} = \mathcal{B}$ and we have a class of delta-matroid parity problems.

If $\mathcal{A}$ is 2-SAT and $\mathcal{B}$ is not 2-SAT, then by (4) of Lemma 6.3 we have some $[\approx \tilde{x} \vee \tilde{y} \vee \tilde{z}]$ in $\mathcal{B}$. In the cases where $\mathcal{A}$ has one of $[t = t']$, $[t \neq t']$, $[t \leq t']$, or both $[t \vee t']$ and $[\overline{t} \vee \overline{t'}]$, then combining such conditions for t or t' being x, y, z and corresponding x', y', z' we get $[\approx \tilde{x} \vee \tilde{y} \vee \tilde{z}]$ in $\mathcal{A}$, which is not 2-SAT, contrary to assumption. We thus have of these choices for t, t' either just $[t \vee t']$ or just $[\overline{t} \vee \overline{t'}]$. By cases (1), (2), (3) of Lemma 6.4, we have that $\mathcal{A}$ decomposes into constraints $\{0\}$, $\{1\}$, and just upward closed constraints or just downward closed constraints, contrary to assumption. If both sides are 2-SAT the problem is polynomial.

If $\mathcal{A}$ is linear and $\mathcal{B}$ is not linear, and $\mathcal{A}$ is not 2-SAT, then we have in $\mathcal{A}$ a linear constraint $[x + y + z = 0]$ or $[x + y + z = 1]$, and in $\mathcal{B}$ by (3) of Lemma 6.3 we have $[z \leq z']$, or $[z \vee z']$, or $[\overline{z} \vee \overline{z'}]$. combining these with $[x' + y' + z' = 0]$ or $[x' + y' + z' = 1]$ in $\mathcal{A}$ gives a constraint on x, y, x', y' that is not linear, contrary to assumption. If both sides are linear the problem is polynomial.

If $\mathcal{A}$ is Horn and $\mathcal{B}$ is not Horn, then in $\mathcal{B}$ we get $[x \neq y]$ or $[x \vee y]$ by (1) of Lemma 6.3. Since $\mathcal{A}$ does not decompose into $\{1\}$ constraints and downward closed constraints by assumption, we have by (1), (3) of Lemma 6.4 that $\mathcal{A}$ has $[x = y]$, or $[x \leq y]$, or $[x \neq y]$, or $[x \vee y]$, yet the last two are not Horn, so $\mathcal{A}$ must have $[x = y]$ or $[x \leq y]$. Combining these with $[x \neq y]$ or $[x \vee y]$ in $\mathcal{B}$ gives $[x' \vee y']$ in $\mathcal{A}$, which is not Horn, contrary to assumption. If both sides are Horn the problem is polynomial. The case of dual-Horn is identical. □

We now consider situations where the last case in the preceding lemma gives NP-completeness.

LEMMA 6.7. *Let $\mathcal{A}, \mathcal{B}$ be as in the statement of Theorem 6.2. Assume $\mathcal{A}$ is not Horn, dual-Horn, 2-SAT, linear, or delta-matroid. Assume also $\mathcal{B}$ contains either $[x = y]$ or $[x \leq y]$. Then the bipartite constraint satisfaction problem is NP-complete.*

Proof. If $\mathcal{B}$ contains $[x = y]$, then either we have $[x \neq y]$ in $\mathcal{A}$ which combines with a condition from (3) of Lemma 6.3 to give $[x \leq y]$ in $\mathcal{B}$, or we have both $[x \vee y]$ and $[\overline{x} \vee \overline{y}]$ by (1), (2) of Lemma 6.3 to also give $[x \leq y]$ in $\mathcal{B}$.

If $\mathcal{B}$ contains $[t \leq t']$, combining this with $[\approx \tilde{x} \vee \tilde{y} \vee \tilde{z}]$ from $\mathcal{A}$ by (4) of Lemma 6.3 gives $[\tilde{x} \vee \tilde{y} \vee \tilde{z}]$ in $\mathcal{B}$. Combining with $[\approx \tilde{x} \leq \tilde{y}, \tilde{z}]$ from $\mathcal{A}$ by (5) of Lemma 6.3 gives $[\tilde{x} \leq \tilde{y}, \tilde{z}]$ in $\mathcal{B}$. Since $\mathcal{A}$ contains either $[t \neq t']$ or both $[t \vee t']$ and $[\overline{t} \vee \overline{t'}]$ from (1) and (2) of Lemma 6.3, we also get $[\tilde{x} \vee \tilde{y} \vee \tilde{z}]$ and $[\tilde{x} \leq \tilde{y}, \tilde{z}]$ in $\mathcal{A}$. Using $[x \leq y]$ from $\mathcal{B}$ and $[t \neq t']$ or both $[t \vee t']$, $[\overline{t} \vee \overline{t'}]$ from $\mathcal{A}$ we get both $[x \vee y]$ and $[\overline{x} \vee \overline{y}]$ in $\mathcal{B}$, that is, all three kinds of 2-SAT clauses in $\mathcal{B}$. We can thus replace each $\tilde{x}$ in the conditions of $\mathcal{A}$ with any choice out of x or $\overline{x}$ using these clauses in $\mathcal{B}$. Since $\mathcal{A}$ contains $[\tilde{x} \vee \tilde{y} \vee \tilde{z}]$, we have that $\mathcal{A}$ contains $[x \vee y]$ or $[\overline{x} \vee \overline{y}]$. Say $\mathcal{A}$ contains $[\overline{x} \vee \overline{y}]$, and then using $[x \vee y \vee z]$ and $[x \vee y, z]$ in $\mathcal{B}$ gives NP-completeness as in the last part of the proof of Lemma 6.5. □

In the remaining case, neither $\mathcal{A}$ nor $\mathcal{B}$ contains either $[x = y]$ or $[x \leq y]$. By (1) of Lemma 6.4, this implies that every constraint C in $\mathcal{A}$ or $\mathcal{B}$ satisfies $C = up(C) \cap down(C)$. Furthermore $\mathcal{A}$ is not delta-matroid, so it contains a constraint $[\approx \tilde{x} \leq \tilde{y}, \tilde{z}]$, and by the property just stated this must be either $[x \vee y, z]$, $[x \vee y, z] \setminus \{111\}$, $[\overline{x} \vee \overline{y}, \overline{z}]$, $[\overline{x} \vee \overline{y}, \overline{z}] \setminus \{000\}$. Say by symmetry it is either $[x \vee y, z]$ or $[x \vee y, z] \setminus \{111\}$. Then $\mathcal{A}$

contains $[x \vee y]$. Since $\mathcal{B}$ is not dual-Horn, it contains either $[\overline{x} \vee \overline{y}]$ or $[x \neq y]$, and in this last case it contains $[\overline{x} \vee \overline{y}]$ as well by combination with $[x \vee y]$ from $\mathcal{A}$. Combining $[\overline{x} \vee \overline{y}]$ from $\mathcal{B}$ with either $[x \vee y, z]$ or $[x \vee y, z] \setminus \{111\}$ from $\mathcal{A}$, we get $[\overline{x} \vee \overline{y}, \overline{z}]$ in $\mathcal{B}$, and thus $[x \vee y, z]$ in $\mathcal{A}$. Furthermore, we get the complement of down(D) in $\mathcal{A}$ for every constraint D in $\mathcal{A}$, so by Lemma 6.5 the problem is NP-complete unless down(D) is 2-SAT. Similarly, we get the complement of up(C) in $\mathcal{B}$ for every constraint C in $\mathcal{A}$, and by Lemma 6.5 the problem is NP-complete unless up(C) is 2-SAT.

If up(D) is not delta-matroid, then by (5) of Lemma 6.3 it contains $[x \vee y, z]$ so D contains $[x \vee y, z]$ or $[x \vee y, z] \setminus \{111\}$. Then by the preceding argument exchanging $\mathcal{A}$ and $\mathcal{B}$, we have that the problem is NP-complete unless up(D) and down(C) are 2-SAT. In this last case, since $C = up(C) \cap down(C)$ and $D = up(D) \cap down(D)$, the whole problem is 2-SAT.

Thus in the remaining case $up(D)$ is a delta-matroid, and by the same argument $down(C)$ is a delta-matroid, while $up(C)$ and $down(D)$ are 2-SAT, for every C in $\mathcal{A}$ and D in $\mathcal{B}$. By complementing $\mathcal{A}$, this problem is more easily viewed as having an instance consisting of M such that $up(M)$ is a delta-matroid and $down(M)$ is 2-SAT, with the variables partitioned into pairs x, y that must satisfy $[x \neq y]$ in a solution.

Eliminate any variable x such that M has no bit-vector with $x = 1$ (resp., $x = 0$), while setting $y = 1$ (resp., $y = 0$) for the corresponding variable in the pair $[x \neq y]$. We may solve the $down(M)$ 2-SAT part with constraints $[x \neq y]$ and obtain a solution X if one exists. We may also solve the $up(M)$ delta-matroid part with constraints $[x \neq y]$ and obtain a solution Y if one exists, by Corollary 5.7. If X is in $up(M)$, then X is in $M = up(M) \cap down(M)$ and we are done. If X is not in $up(M)$, then there exists a $T \geq X$ with T not in $up(M)$ such that every $U > T$ is in $up(M)$. Let S be the set of variables x that have value 0 in T, called a *flat* of $up(M)$. There is no element of $up(M)$ that has all variables x in S with value 0. For every x in S, there is a least element of M that has $x = 1$ and with $y = 0$ for all other y in S.

We claim that X' obtained from X by changing $x = 0$ to $x = 1$ is also in $down(M)$. Otherwise X' fails to satisfy some 2-SAT clause involving x, say, $[\overline{x} \vee \overline{z}]$ for some z not in S. Then an element V of M with $x = 1$ and $y = 0$ for all other y in S also has $z = 0$. Restricting M to the variables in $S \cup \{z\}$, to obtain M', we have that $up(M')$ is a delta-matroid and contains V with a single variable $x = 1$. Let W be a least vector in M' having $z = 1$, and let y be some other variable that has $y = 1$ in W. Restrict M' by setting all variables other than x that have value 0 in W to value 0, thus obtaining M''. We have in M'' vectors with $yz = 00$ and vectors with $yz = 11$ but no vector with $yz = 01$ in M'', giving either $[z \leq y]$ or $[z = y]$, contrary to assumption. This proves the claim.

Note that the solution Y obtained above must have some x in S with value $x = 1$. Thus we may just obtain X' by trying all choices of variables x that have $x = 0$ in X and $x = 1$ in Y, until X' obtained by changing x does not fail to satisfy the 2-SAT clauses in $down(M)$. We may then change the mate that was linked to x by $[x \neq y]$ from $y = 1$ to $y = 0$ to obtain a new solution X'' in $down(M)$ which is closer to Y. Repeating the process, we eventually reach some X'' in $down(M)$ that is also in $up(M)$, since otherwise we keep getting closer to Y, and Y is in $up(M)$. We thus obtain from X and Y some X'' such that $X'' \in down(M) \cap up(M) = M$ and satisfies all conditions $[x \neq y]$ as well.

This algorithm completes the proof of Theorem 6.2.

THEOREM 6.8. *Let M be an upward closed delta-matroid, and let R be a collection of 2-SAT clauses $[\overline{x}]$, $[\overline{x} \vee \overline{y}]$ such that no flat of M with at least two elements meets*

a clause $[\overline{x} \vee \overline{y}]$ *in exactly one element. Then one can solve the constraints given by* M*,* R*, and a pairing with conditions* $[x \neq y]$ *in polynomial time.*

We now obtain a full classification for k-partite Boolean constraint satisfaction for $k \geq 3$.

THEOREM 6.9. *Let* $\mathcal{A}_1, \ldots, \mathcal{A}_k$ *be sets of constraint types each containing the single variable constraints* $\{0\}$*,* $\{1\}$*,* $\{0,1\}$ *and at least one constraint that does not decompose into these, with* $k \geq 3$*. Then the* k*-partite Boolean constraint satisfaction problem with constraints from* $\mathcal{A}_i$ *in part* i *and each variable participating in only one constraint from each part* i *is either polynomial time solvable using an oracle or NP-complete.*

Proof. If some $\mathcal{A}_i$ contains a constraint that is not a delta-matroid, say, $\mathcal{A}_1$, then the problem defined by $\mathcal{A}_1$ and $\mathcal{A}_2$ is either polynomial time solvable using an oracle or NP-complete by Theorem 6.2. The NP-completeness of this subproblem implies the NP-completeness of the entire problem, while if the subproblem is polynomial, then the algorithm simulates an oracle for the solutions of the subproblem, thus giving a new problem where the parts $\mathcal{A}_1$ and $\mathcal{A}_2$ have been combined into a single part $\mathcal{A}'$ that contains a constraint that is not a delta-matroid, thus reducing the analysis for k to $k-1$.

In the remaining case all $\mathcal{A}_i$ are delta-matroids. If at most one $\mathcal{A}_i$ contains $[x = y]$, say $\mathcal{A}_1$, then all $\mathcal{A}_j$ for $j > 1$ not containing $[x = y]$ also do not contain $[x \neq y]$ unless $\mathcal{A}_1$ does not contain $[x = y]$; otherwise $[x = y]$ could be simulated on $\mathcal{A}_j$ as well. We may thus intersect the delta-matroids from $\mathcal{A}_1$ and $\mathcal{A}_2$ by Corollaries 5.5 and 5.7, again giving a new problem where the parts $\mathcal{A}_1$ and $\mathcal{A}_2$ have been combined into a single part $\mathcal{A}'$ for which an oracle can be simulated, thus reducing the analysis for k to $k-1$.

If all $\mathcal{A}_i$ are delta-matroids, at least one $\mathcal{A}_i$ contains $[x = y]$, say, $\mathcal{A}_1$, and at least one $\mathcal{A}_i$ does not contain $[x = y]$, say, $\mathcal{A}_2$, then we may again combine $\mathcal{A}_1$ and $\mathcal{A}_2$ into a single part $\mathcal{A}'$ by Corollary 5.5, and this part is not a delta-matroid by a constraint on x, y, z given by $[x = y]$ in $\mathcal{A}_1$ and one of $[x \neq z]$, $[x \leq z]$, $[x \vee z]$, $[\overline{x} \vee \overline{z}]$ in $\mathcal{A}_2$, thus reducing the analysis to an earlier case from k to $k-1$.

Finally, if all $\mathcal{A}_i$ are delta-matroids and contain $[x = y]$, then combining $[x = y]$ in $\mathcal{A}_1$, $[y = z]$ in $\mathcal{A}_2$, and $[x = x'], [y = y'], [z = z']$ in $\mathcal{A}_3$ gives $[x' = y' = z']$ in $\mathcal{A}_3$, contrary to the assumption that $\mathcal{A}_3$ is a delta-matroid. ∎

7. Delta-matroid parity without oracle. For the remaining open cases of the general problem where equality is an allowed constraint, namely, cases where all constraints are given by delta-matroids, we note that not all known polynomial cases remain polynomial in the oracle model. The coindependent delta-matroid case from Feder [5] has an algorithm polynomial in $n2^k$, where n is the number of variables and k is the maximum number of variables per constraint, and has a lower bound exponential in k if a constraint on k variables is given by an oracle. There are other cases, such as local delta-matroids [3], that remain polynomial with an oracle.

We generalize the case of coindependent delta-matroids. A delta-matroid $M = (E, \mathcal{F})$ is a *zebra delta-matroid* if there exist integers $0 \leq r \leq s \leq |E|$ such that (1) for all feasible sets $F \in \mathcal{F}$, $r \leq |F| \leq s$; (2) for all $A \subseteq E$ with $|A| \in \{r, s\}$, A is a feasible set, that is, $A \in \mathcal{F}$; and (3) for all $A \subseteq E$ with $r < |A| < s$, either $A \in \mathcal{F}$, or for all $B \subseteq E$, if $|A \Delta B| = 1$, then $B \in \mathcal{F}$. Zebra delta-matroids generalize the delta-matroids arising in the general factor problem.

A zebra delta-matroid is a *coindependent delta-matroid* if $r \in \{0, 1\}$ and $s \in \{|E| - 1, |E|\}$. Coindependent delta-matroids were studied in [5].

THEOREM 7.1. *Delta-matroid parity on a coindependent delta-matroid with* $|E| = 2k$ *with oracle has a lower bound of* 2^k *on the number of queries to the oracle.*

Proof. Let $E = \{x_1, \ldots, x_{2k}\}$ and $\mathcal{L} = \{\{x_{2i-1}, x_{2i}\} : 1 \leq i \leq k\}$. Consider a set $A \subseteq E$ such that for all $1 \leq i \leq k$, $x_{2i-1} \in A$ if and only if $x_{2i} \in A$. Let $\mathcal{F}$ be the set of all subsets of E of odd cardinality plus A, which is of even cardinality. While an algorithm has queried fewer than 2^k sets $B \subseteq E$ such that for all $1 \leq i \leq k$, $x_{2i-1} \in B$ if and only if $x_{2i} \in B$, the oracle may answer that $B \notin \mathcal{F}$, and only when the 2^kth such B is queried set $A = B$, giving the answer to the problem. □

We prove a counterpart to this lower bound.

THEOREM 7.2. *Suppose* $M = (E, \mathcal{F})$ *with* $|E| = n$ *is the direct sum of zebra delta-matroids* $M_i = (E_i, \mathcal{F}_i)$ *with* $|E_i| \leq k$, $|\mathcal{F}_i| \leq f$. *Then delta-matroid parity on* M, $\mathcal{L}$, *can be solved in time* $O(n^3 f)$.

We successively simplify the problem.

LEMMA 7.3. *The problem reduces to the case where we have a feasible* F *and only a single* $\{a, b\} \in \mathcal{L}$ *such that* $\{a, b\} \notin \mathcal{L}_F$, *and one of the* M_i *has* $E_i = \{b\}$.

Proof. The problem reduces to finding an augmenting path. For each choice of an element a with which to start the augmenting path given $F \in \mathcal{F}$, so that $\{a, b\} \in \mathcal{L}$ and $\{a, b\} \notin \mathcal{L}_F$, let S be the set of elements c such that $\{c, d\} \in \mathcal{L}$ and $\{c, d\} \notin \mathcal{L}_F$ for some element d. For each $c \in S$, let $M'_c = (\{\overline{c}\}, \mathcal{F}'_c)$, where $\mathcal{F}'_c = \{\emptyset, \{\overline{c}\}\}$ if $c \neq a, b$, $\{\overline{b}\} \in \mathcal{F}'_b$ if and only if $b \in F$, $\emptyset \in \mathcal{F}'_b$ if and and only if $b \notin F$, $\{\overline{a}\} \in \mathcal{F}'_a$ if and only if $a \notin F$, and $\emptyset \in \mathcal{F}'_a$ if and only if $a \in F$. Let M' be the direct sum of the M_i and the M'_c. Extend the feasible F for M to a feasible F' for M' by including $\overline{c} \in S$ in F' if and only if $c \in F$ for $c \neq a$, and including $\overline{a}$ in F' if and only if $a \notin F$. Let $\mathcal{L}'$ consist of the pairs $\{c, d\} \in \mathcal{L}$ such that $c, d \notin S$, and the pairs $\{c, \overline{c}\}$ for $c \in S$. Thus the only $\{c, d\} \in \mathcal{L}'$ such that $\{c, d\} \notin \mathcal{L}'_{F'}$ is $\{c, d\} = \{a, \overline{a}\}$, and the augmenting paths for $M, \mathcal{L}$ started at a correspond to the augmenting paths for $M', \mathcal{L}'$, which must start at a. □

Consider the problem in the form of Lemma 7.3, with zebra delta-matroids $M_i = (E_i, \mathcal{F}_i)$ having corresponding r_i, s_i. Let $M'_i = (E_i, \mathcal{F}'_i)$, where $\mathcal{F}'_i$ consists of the sets $F'_i \subseteq E_i$ such that $r_i \leq |F'_i| \leq s_i$ and $F'_i \Delta (F \cap E_i)$ is of even size.

LEMMA 7.4. *The problem reduces to a problem where all but one of the* M_i *have been replaced by* M'_i *and the conditions of Lemma* 7.3 *are also met.*

Proof. The augmenting path starting at a must end in some $M_0 = (E_0, \mathcal{F}_0)$. Replace all $M_i \neq M_0$ with M'_i to obtain M'. Since every $M_i \neq M_0$ has an even number of elements in the augmenting path starting at a, it follows that this augmenting path is also an augmenting path in M'. Conversely, suppose we have an augmenting path in M', starting at a. Let the augmenting path be $a = x_1, \ldots, x_{2t+1}$ with $x_{2t+1} \in E_0$. Either this is also an augmenting path for M, or there exists an $1 \leq j \leq t$ such that $F \Delta \{x_1, \ldots, x_{2j-2}\}$ is feasible for M but $F \Delta \{x_1, \ldots, x_{2j}\}$ is not feasible for M. In this last case, since each M_i is a zebra delta-matroid, we have that x_{2j-1} is in some M_i with r_i, s_i and $(F \Delta \{x_1, \ldots, x_{2j-1}\}) \cap E_i$ has size $r_i \leq u \leq s_i$, so the fact that $F \Delta \{x_1, \ldots, x_{2j}\}$ is not feasible implies that $F \Delta \{x_1, \ldots, x_{2j-1}\}$ is feasible, giving an augmenting path $a = x_1, \ldots, x_{2j-1}$ for M. □

Consider the problem in the form of Lemma 7.4.

LEMMA 7.5. *The problem reduces to graph matching.*

Proof. So far, we have a single $M_0 = (E_0, \mathcal{F}_0)$ not of the form of the M'_i, with $|E_0| \leq k$. We may then consider each of the at most f feasible sets $F_0 \in \mathcal{F}_0$ such that $|(F \cap E_0) \Delta F_0|$ is odd and replace M_0 with $M'_0 = (E_0, \{F_0\})$, which decomposes into $|E_0| \leq k$ zebra delta-matroids, giving at most k unmatched pairs plus the pair $\{a, b\}$.

Now the problem has delta-matroids $M_i' = (E_i, \mathcal{F}_i')$ with $\mathcal{F}_i'$ consisting of all F_i' with $r \leq |F_i'| \leq s$ and both $t = s - r, |F_i'| - r$ even. Define a graph G consisting of a clique K on t vertices, an independent set I on r vertices, a complete bipartite graph with the $r + t$ vertices in $K \cup I$ in one side, and some additional $r + t$ vertices forming a set U in the other side. To match all vertices in $K \cup I \cup U$, we must have $r \leq r + 2j \leq r + t$ vertices in U matched to vertices in $K \cup I$, corresponding to a choice of $r + 2j$ elements from E_i forming some $F_i' \in \mathcal{F}_i'$. We may then join the sets $U = U_i$ for each M_i' with edges corresponding to the pairs $\{a, b\} \in \mathcal{L}$.

We may assume a is not in M_0. We look for an F'-augmenting path in the resulting graph starting at a for $F' = F\Delta((F\Delta F_0) \cap E_0)$. The augmenting path $a, \ldots, x_1$ will have $\overline{x_1} \in (F \cap E_0)\Delta F_0$, and either $(F \cap E_0)\Delta\{\overline{x_1}\}$ is feasible for M_0, in which case we replace F with $F' = F\Delta\{a, \ldots, x_1, \overline{x_1}\}$, or we replace F with $F'' = F'\Delta\{x_2\}$ for some $x_2 \in (F' \cap E_0)\Delta F_0$ such that $(F' \cap E_0)\Delta\{x_2\}$ is feasible for M_0, set $a = \overline{x_2}$, and proceed to look for an augmenting path starting at a. In the end, we either will have $F \cap E_0 = F_0$ or will have found a shorter augmenting path by Lemma 7.4. ∎

The graph of Lemma 7.5 has $O(n)$ vertices and $m = O(nk)$ edges and requires finding at most f augmentations in a graph if we go through the F_0 with $d_0 = |(F \cap E_0)\Delta F_0|$ in order of increasing d_0. Each augmentation can be done in time $O(m)$, giving a total time $O(mf) = O(nkf)$ for the problem of Lemma 7.4. This complexity can be reduced by only implicity maintaining the graph corresponding to of each M_i, so that only the $O(|E_i|)$ times that M_i is visited are counted, and the search for an augmenting path takes $O(n)$ time. Each of the sets in $\mathcal{F}_0$ is considered at most k times while finding augmenting paths, once for each element in E_0 to be included. Thus the problem of Lemma 7.4 is solved in $O(nf)$ time. The problem of Lemma 7.3 can be solved in time $O(n^2 f)$ by considering the at most n possible choices of M_0. Testing the at most n unmatched pairs to find a maximum number of augmentations for the original problem can be done in time $O(n^3 f)$, solving the original problem and proving Theorem 7.2.

Of course, for the general factor problem, which can be viewed as consisting of zebra delta-matroids such that if A is feasible then every B with $|B| = |A|$ is also feasible, we can let M_0' consist of the sets of a given size, and only two sizes need to be considered, namely, the sizes p, q such that $p < v = |F \cap E_0| < q$ that give the least odd values for $v - p, q - v$. This costs of a factor of k for at most k augmentations instead of f, giving a bound of $O(n^3 k)$ on the running time. See also Cornuejols [2] for a more efficient algorithm for the general factor problem.

Istrate [8] defined compact delta-matroids by combining the delta-matroids of the general factor problem in a star arrangement. More generally, we can combine zebra delta-matroids in a tree configuration. Formally, define $M = (E, \mathcal{F})$ to be a *zebra-compact delta-matroid* inductively if there exists a zebra delta-matroid $M_1 = (E_1 \cup \{a\}, \mathcal{F}_1)$ and a zebra-compact delta-matroid $M_2 = (E_2 \cup \{b\}, \mathcal{F}_2)$ such that M is obtained from M_1, M_2 by linking a and b and contracting $K = \{\{a, b\}\}$; also the direct sum of a zebra delta-matroid and a zebra-compact delta-matroid is a zebra-compact delta-matroid, and every zebra delta-matroid is a zebra-compact delta-matroid.

THEOREM 7.6. *A zebra-compact delta-matroid $M = (E, \mathcal{F})$ with $|E| = k$ can be recognized and decomposed into zebra delta-matroids in time $O(c^{k^2})$ for some constant c.*

Proof. We can test each of the 2^k possible decompositions $E_1 \subseteq E$, $E_2 = E \setminus E_1$, each in time $O(c^k)$ for some constant c. If there exist two elements $A \cup B, A' \cup B' \in \mathcal{F}$, where $A, A' \subseteq E_1$ and $B, B' \subseteq E_2$, such that $A \cup B', A' \cup B \notin \mathcal{F}$, then this determines

uniquely the feasible sets of M_1, M_2, namely, the possible choices of B'', B''' such that $A \cup B'', A' \cup B''' \in \mathcal{F}$ give feasible sets $B'' \cup \{b\}, B''' \in \mathcal{F}_2$ or $B'', B''' \cup \{b\} \in \mathcal{F}_2$, and similarly for $\mathcal{F}_1$. After verifying this decomposition, we proceed inductively.

Otherwise, unless M is a direct product of $M_1 = (E_1, \mathcal{F}_1)$ and $M_2 = (E_2, \mathcal{F}_\in$, we only have $A \cup B, A \cup B', A' \cup B' \in \mathcal{F}$ but $A' \cup B \notin \mathcal{F}$. Then we have $B \cup \{b\}, B' \in \mathcal{F}_2$, and possibly $B' \cup \{b\} \in \mathcal{F}_2$ (or equivalently $B, B' \cup \{b\} \in \mathcal{F}_2$ and possibly $B' \in \mathcal{F}_2$). All the possibly included subsets must be included either for $\mathcal{F}_1$ or for $\mathcal{F}_2$, say, for $\mathcal{F}_1$, and it can then be shown that including all or none for $\mathcal{F}_2$ will allow the decomposition to proceed if some subset of them allows the decomposition to proceed. However, we then may not get a zebra delta-matroid for each delta-matroid that is not further decomposed. It may at that point be decided whether to include the possibly included subsets in $\mathcal{F}_2$ for the resulting delta-matroid that is not further decomposed containing $\{b\}$. A similar situation arises for the case of a direct product of M_1 and M_2 where we still choose to decompose using elements a, b. In that case, for elements $A \cup B \in \mathcal{F}$, we must always include $B \in \mathcal{F}_2$ and possibly $B \cup \{b\} \in \mathcal{F}_2$ or always include $B \cup \{b\} \in \mathcal{F}_2$ and possibly $B \in \mathcal{F}_2$.

For the delta-matroids $M_i = (E_i, \mathcal{F}_i)$ that are not decomposed and must be zebra delta-matroids, we may choose what possibly included elements for $\mathcal{F}_i$ to exclude by choosing the corresponding $0 \leq r_i \leq s_i \leq |E_i|$ so as to satisfy the definition of a zebra delta-matroid, thus having to exclude all sets of size less than r_i or greater than s_i, while we may always choose to include sets of size from r_i to s_i.

There are thus d^k cases that take $O(c^k)$ time and reduce to a case for a smaller k, giving the recurrence $f(k) = d^k(c^k + f(k-1))$, $f(0) = 0$, on the time used for finding such a decomposition, that is time complexity of the order $O((cd)^{k^2})$. □

A delta-matroid $M = (E, \mathcal{F})$ is *even* if for all $F, G \in \mathcal{F}$, $|F \Delta G|$ is even. We consider any class $\mathcal{C}$ of even delta-matroids, closed under restriction and direct sum, such that there is a polynomial time algorithm for delta-matroid parity on matroids in $\mathcal{C}$. Examples of $\mathcal{C}$ include even local delta-matroids [3] and linear delta-matroids over a given field [7].

A delta-matroid $M = (E, \mathcal{F})$ is a *$\mathcal{C}$-zebra delta-matroid* if for every feasible set $A \in \mathcal{F}$, there exists a delta-matroid $M_A = (E, \mathcal{F}_A)$ in $\mathcal{C}$ such that (1) all sets $B \in \mathcal{F}$ such that $A \Delta B$ is of even size are also in $\mathcal{F}_A$; (2) if $B \in \mathcal{F} \cap \mathcal{F}_A$ and $C \in \mathcal{F}_A \setminus \mathcal{F}$ are such that $|B \Delta C| = 2$ and $|A \Delta C| = |A \Delta B| + 2$, then the two sets D such that $|B \Delta D| = |D \Delta C| = 1$ satisfy $C \in \mathcal{F}$.

We assume that a $\mathcal{C}$-zebra delta-matroid M is given together with appropriate presentations for the corresponding delta-matroids M_A in $\mathcal{C}$.

THEOREM 7.7. *Suppose $M = (E, \mathcal{F})$ with $|E| = n$ is the direct sum of $\mathcal{C}$-zebra delta-matroids $M_i = (E_i, \mathcal{F}_i)$ with $|E_i| \leq k$, $|\mathcal{F}_i| \leq f$. Then delta-matroid parity on M, $\mathcal{L}$, can be solved in time polynomial in n and f.*

Proof. The proof is analogous to the proof of Theorem 7.2. As in Lemma 7.3, the problem reduces to the case where we have a feasible F and only a single $\{a, b\} \in \mathcal{L}$ such that $\{a, b\} \notin \mathcal{L}_F$, and one of the M_i has $E_i = \{b\}$.

Consider the problem in this form with $\mathcal{C}$-zebra delta-matroids $M_i = (E_i, \mathcal{F}_i)$. For $A = F \cap E_i$, let $M_i' = (E_i, \mathcal{F}_i)$ be the delta-matroid $(M_i)_A$ in $\mathcal{C}$ in the definition of $\mathcal{C}$-zebra delta-matroids. As in Lemma 7.4, the problem reduces to a problem where all but one of the M_i have been replaced by M_i'. The key point as before is that given an augmenting path be $a = x_1, \ldots, x_{2t+1}$ with $x_{2t+1} \in E_0$ in M', either this is also an augmenting path for M, or we have $F \Delta \{x_1, \ldots, x_{2j-2}\}$ feasible for M and for M', but $F \Delta \{x_1, \ldots, x_{2j}\}$ feasible for M' and not feasible for M, and then

$F\Delta\{x_1,\ldots,x_{2j-1}\}$ is feasible for M by the definition of $\mathcal{C}$-zebra delta-matroids, giving an augmenting path for M as well. Finally, as in Lemma 7.5, we replace $M_0 = (E_0, \mathcal{F}_0)$ with $M_0' = (E_0, \{F_0\})$ for each $F_0 \in \mathcal{F}_0$ and find augmenting paths starting at a and ending in M_0' so that in the end, we either will have $F \cap E_0 = F_0$ or will have found a shorter augmenting path, using the algorithm for delta-matroid parity in $\mathcal{C}$, since any algorithm for delta-matroid parity can be used to find an augmenting path starting at a if it exists. □

An even delta-matroid $M = (E, \mathcal{F})$ is *local* if for every $F \in \mathcal{F}$ and every pairing $\mathcal{L}$, if $x_1,\ldots,x_{2k}$ is a path such that $F\Delta\{x_{2i-1}, x_{2i}\} \in \mathcal{F}$ for $2 \leq 2i \leq 2k$ and $\{x_{2i}, x_{2i+1}\} \in \mathcal{L}$ for $2 \leq 2i < 2k$, and there is no shorter path $x_1 = y_1,\ldots,y_{2l} = x_{2k}$ satisfying this property with $\{y_1,\ldots,y_{2l}\} \subset \{x_1,\ldots,x_{2k}\}$, then $F\Delta\{x_1,\ldots,x_{2k}\} \in \mathcal{F}$.

The algorithm of Dalmau and Ford [3] for local delta-matroid parity only requires in the case of even delta-matroids that this property hold, in an augmenting phase started at a feasible set $F \in \mathcal{F}$, for that particular feasible set F. We say that in that case M is F-local. The algorithm uses the fact that if $z_1,\ldots,z_{2r}$ is an F-augmenting, $\mathcal{L}_F$-alternating path, then by repeated application of Wenzel's strong exchange axiom for even delta-matroids, there exists a $\mathcal{L}_F$-alternating path $z_1 = x_1,\ldots,x_{2k} = z_{2r}$ such that $F\Delta\{x_{2i-1}, x_{2i}\} \in \mathcal{F}$ for $2 \leq 2i \leq 2k$. Such a path may be found by an augmentation in graph matching, by considering the graph consisting of all edges $\{x, y\}$ such that $F\Delta\{x, y\} \in \mathcal{F}$, plus edges $\{x, y\}$ in the given matching for $\{x, y\} \in \mathcal{L}_F$. The algorithm also attempts to find a shorter augmenting path, in the subgraph induced by $x_1,\ldots,x_{2i-1}, x_{2i+2},\ldots,x_{2k}$, for each $2 \leq 2i < 2k$. When such a shorter augmenting path is not found, the definition of F-local guarantees that $x_1,\ldots,x_{2k}$ is an F-augmenting path.

We generalize local-zebra delta-matroids by only requiring M_A to be A-local instead of local and still obtain a polynomial time algorithm for the corresponding A-local-zebra delta-matroid parity problem. We thus have polynomial time algorithms when the even delta-matroids M_A in the definition of $\mathcal{C}$-zebra are all linear over a given field, or all A-local. These classes are closed under direct sums. More generally, we may allow M_A to be obtained from M_B in the class of delta-matroids linear over a given field or the class of B-local delta-matroids by contracting $\mathcal{K}$ such that $\mathcal{K}_B = \mathcal{K}$ and A consists of the elements of B not in pairs from $\mathcal{K}$.

A main example of local-zebra delta-matroids is the class of delta-matroids $M = (E, \mathcal{F})$ that satisfy a stronger exchange property, namely, that for all feasible sets A, B and every element $x \in A\Delta B$, either $A\Delta\{x\}$ is feasible or for every $y \in A\Delta\{x\}\Delta B$ the set $A\Delta\{x, y\}$ is feasible. In this case the local delta-matroid $M_A = (E, \mathcal{F}_A)$ can be defined by letting $\mathcal{F}_A$ be the set of all $B \subseteq E$ such that $A\Delta B$ is of even size and $A\Delta B \subseteq A\Delta D$ for some $D \in \mathcal{F}$.

Indeed, condition (2) in the definition of local-zebra holds by the stronger exchange property applied to B and D such that $A\Delta C \subseteq A\Delta D$. It remains to show that M_A is a local delta-matroid. The subsets $D \in \mathcal{F}$ can be chosen without loss of generality with $A\Delta D$ of maximum size over $D \in \mathcal{F}$. Any two such D must satisfy $|D_1\Delta D_2| = 2$; otherwise, there exists $\{x, y\} \subset D_1\Delta D_2$ such that $|A\Delta(D_1\Delta\{x, y\})| = |A\Delta D_1| + 2$, and so $D_1\Delta\{x\}, D_1\Delta\{x, y\} \notin \mathcal{F}$, contrary to the stronger exchange property for D_1, D_2. There exists thus a set E such that $|D_i\Delta E| = 1$ for all such D_i, and if $|A\Delta D_i| > 1$, then E can be chosen so that $|A\Delta E| = |A\Delta D_i| + 1$. Furthermore, if there are at least two such feasible D_i, say, D_1 and D_2, then we may not have two D_i such that $|A\Delta E| = |A\Delta D_i| + 1$ that are not feasible, say, D_3 and D_4.

Otherwise, if $D_i = E\Delta\{x_i\}$, then the possible intersections of $A\Delta F$ for $F \in \mathcal{F}$ with $X = \{x_1, x_3, x_4\}$ include $\emptyset$ and X but do not include any other subset Y containing x_1 corresponding to some feasible G, by the stronger exchange property applied to D_1, G and x_1, contradicting M beign a delta-matroid by the exchange property for $\emptyset$, X and x_1.

Letting $E = D$ if there is only one such D_i, and letting $E = \emptyset$ if $|A\Delta D_i| \leq 1$, we have that $M_A = (E, \mathcal{F}_A)$ has $\mathcal{F}_A$ consisting of all sets G such that $A\Delta G$ is of even size and a subset of $A\Delta E$, or consisting of all these sets G except for a single G with $|A\Delta G| = |A\Delta E| - 1$. Suppose now $G \in \mathcal{F}_A$ and some path $x_1, \ldots, x_{2k}$ fails to satisfy the definition of G-local for M_A. If $k \geq 3$, then either $G\Delta\{x_1, x_4\}$ or $G\Delta\{x_1, x_6\}$ is in $\mathcal{F}_A$ since at most one even size subset of $G\Delta E$ fails to give a feasible set. Thus $x_1, \ldots, x_{2k}$ has a shortcut path as in the definition of G-local when $k \geq 3$. If $k = 2$, then either $G\Delta\{x_1, x_4\} \in \mathcal{F}_A$ giving again a shortcut, or $G\Delta\{x_1, x_2, x_3, x_4\} \in \mathcal{F}_A$ because at most one even size subset of $G\Delta E$ is missing, satisfying the definition of G-local. Thus M_A is G-local for all $G \in \mathcal{F}_A$, that is, M_A is local and therefore M is local-zebra.

8. Conclusions. We considered bipartite Boolean constraint satisfaction having two different sets of allowed constraint types for both sides of the bipartition. This case is properly bipartite only if at least one side does not contain equality. We obtain a classification for these problems if they do not have equality in both sides. All the algorithms work even in an oracle model, leaving open when equality is an allowed constraint in both sides cases of delta-matroid parity, which in general cannot be solved in oracle model. We also obtain a full classification for k-partite Boolean constraint satisfaction with $k \geq 3$.

Known polynomial cases of delta-matroid parity include local and linear delta-matroids, and delta-matroids obtained from these by simulation. All the remaining known polynomial cases are covered by cases studied in this paper. The first case is that of delta-matroids that do not contain inequality. The second case is obtained from any of the known polynomial cases $\mathcal{C}$ of even delta-matroids by considering the corresponding $\mathcal{C}$-zebra delta-matroids and adding closure under simulation as well. Here $\mathcal{C}$ may be the class of delta-matroids obtained by simulation from linear delta-matroids over a given field or the class of delta-matroids obtained by simulation from A-local delta-matroids.

REFERENCES

[1] A. Bouchet and B. Jackson, *Parity systems and the delta-matroid intersection problem*, Electron. J. Combin., 7 (2000).

[2] G. P. Cornuejols, *General factors of graphs*, J. Combin. Theory Ser. B, 45 (1988), pp. 185–198.

[3] V. Dalmau and D. Ford, *Generalized satisfiability with k occurrences per variable: A study through delta-matroid parity*, in Proceedings of the 28th International Symposium of Mathematical Foundations of Computer Science, 2003.

[4] J. Edmonds, *Matroid intersection*, Ann. Discrete Math., 14 (1979), pp. 39–49.

[5] T. Feder, *Fanout limitations on constraint systems*, Theoret. Comput. Sci., 255 (2001), pp. 281–293.

[6] H. N. Gabow and M. Stallman, *An augmenting path algorithm for linear matroid parity*, Combinatorica, 6 (1986), pp. 123–150.

[7] J. F. Geelen, S. Iwata, and K. Murota, *The linear delta-matroid parity problem*, J. Combin. Theory Ser. B, 88 (2003), pp. 377–398.

[8] G. Istrate, *Looking for a Version of Schaefer's Dichotomy Theorem when Each Variable Occurs At Most Twice*, Technical Report #652, Computer Science Deptartment, University of Rochester, New York, 1997.

[9] L. Lovász, *Matroid matching and some applications*, J. Combin. Theory Ser. B, 28 (1980), pp. 208–236.
[10] L. Lovász and M. Plummer, *Matching Theory*, North-Holland, Dordrecht, the Netherlands, 1986.
[11] T. J. Schaefer, *The complexity of satisfiability problems*, in Proceedings of the 10th ACM Symposium on Theory of Computing, 1978, pp. 216–226.
[12] W. Wenzel, *δ-matroids with the strong exchange conditions*, Appl. Math. Lett., 6 (1993), pp. 67–70.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 395–411

THE VOLUME OF THE GIANT COMPONENT OF A RANDOM GRAPH WITH GIVEN EXPECTED DEGREES*

FAN CHUNG† AND LINYUAN LU‡

Abstract. We consider the random graph model $G(\mathbf{w})$ for a given expected degree sequence $\mathbf{w} = (w_1, w_2, \ldots, w_n)$. If the expected average degree is strictly greater than 1, then almost surely the giant component in G of $G(\mathbf{w})$ has volume (i.e., sum of weights of vertices in the giant component) equal to $\lambda_0 \mathrm{Vol}(G) + O(\sqrt{n}\log^{3.5} n)$, where λ_0 is the unique nonzero root of the equation

$$\sum_{i=1}^{n} w_i e^{-w_i\lambda} = (1-\lambda)\sum_{i=1}^{n} w_i,$$

and where $\mathrm{Vol}(G) = \sum_i w_i$.

Key words. random graphs, expected degree sequences, giant connected component

AMS subject classification. 05C80

DOI. 10.1137/050630106

1. Introduction.

Among the many celebrated results of Erdős and Rényi on random graphs, one of the most well-known theorems is a sharp estimate for the size of the giant component. For the random graph $G(n,p)$, as introduced by Erdős and Rényi in 1959 [17], every pair of a set of n vertices is chosen to be an edge with probability p independently. Erdős and Rényi [17] showed that the size (i.e., the number of vertices) of the giant component of $G(n,p)$ satisfies the following.

THEOREM A. *If $d = np > 1$, a graph G of $G(n,p)$ almost surely contains a giant component with $(f(d)+o(1))n$ vertices, where $f(d)$ is given by*

$$f(d) = 1 - \frac{1}{d}\sum_{k=1}^{\infty} \frac{k^{k-1}}{k!}(de^{-d})^k. \tag{1}$$

In $G(n,p)$, every vertex has the same expected degree np. Although such a random graph model is useful in some applications, most real-world networks have degree distributions far from regular [1, 4, 5, 6, 20, 21, 22, 25, 26]. It is therefore not surprising that the random graph model $G(n,p)$ does not capture many behaviors of numerous networks [1, 2, 3, 9, 10, 11, 12, 13, 14, 15, 23].

Here we consider the random graph model $G(\mathbf{w})$ for a given expected degree sequence $\mathbf{w} = (w_1, w_2, \ldots, w_n)$, as introduced in [10, 11, 12, 13]. The edges are chosen independently and randomly as follows. The probability p_{ij} that there is an edge between v_i and v_j is proportional to the product $w_i w_j$ (as well as the loop at v_i with probability proportional to w_i^2). Namely,

$$p_{ij} = \frac{w_i w_j}{\sum_k w_k} = \frac{w_i w_j}{\mathrm{Vol}(G)}. \tag{2}$$

*Received by the editors April 27, 2005; accepted for publication (in revised form) December 13, 2005; published electronically May 3, 2006.
http://www.siam.org/journals/sidma/20-2/63010.html

†Department of Mathematics, University of California, San Diego, CA 92093 (fan@ucsd.edu). The work of this author was supported in part by NSF grants DMS 0457215, ITR 0205061, and ITR 0426858.

‡Department of Mathematics, University of South Carolina, Columbia, SC 29208 (lu@math.sc.edu).

Here the expected volume for a subset S of vertices, $\mathrm{Vol}(S)$, is defined as

$$\mathrm{Vol}(S) = \sum_{v_i \in S} w_i,$$

and $\mathrm{Vol}(G) = \mathrm{Vol}(V(G))$. The (actual) volume of S in a graph G is the sum of all degrees of vertices in S and is denoted by $\mathrm{vol}(S)$:

$$\mathrm{vol}(S) = \sum_{v_i \in S} d_i,$$

where d_i denotes the degree of vertex v_i. In order to avoid confusion when we deal with the graph G in a nonprobabilistic context, we can view w_i as a weight assigned to vertex v_i.

In [10], the following theorem was given concerning the giant components for graphs in the random graph model $G(\mathbf{w})$.

THEOREM B. *Suppose that G is a random graph in $G(\mathbf{w})$ with expected degree sequence $\mathbf{w}$. If the expected average degree d is strictly greater than 1, then the following hold:*

(1) *Almost surely G has a unique giant component. Furthermore, the volume of the giant component is at least $(1 - \frac{2}{\sqrt{de}} + o(1))\mathrm{Vol}(G)$ if $d \geq \frac{4}{e} = 1.4715\ldots$, and is at least $(1 - \frac{1+\log d}{d} + o(1))\mathrm{Vol}(G)$ if $d < 2$.*

(2) *The second-largest component almost surely has size at most $(1+o(1))\mu(d)\log n$, where*

$$\mu(d) = \begin{cases} \frac{1}{1+\log d - \log 4} & \text{if } d > \frac{4}{e}, \\ \frac{1}{d-1-\log d} & \text{if } 1 < d < 2. \end{cases}$$

Moreover, with probability at least $1 - n^{-k}$, the second-largest component has size at most $(k+1+o(1))\mu(d)\log n$ for any $k \geq 1$.[1]

In this paper, we will state a sharp asymptotic estimate for the volume of the giant component for a random graph in $G(\mathbf{w})$.

THEOREM 1. *If the expected average degree is strictly greater than 1, then almost surely the giant component in a graph G in $G(\mathbf{w})$ has volume $\lambda_0\mathrm{Vol}(G) + O(\sqrt{n}\log^{3.5} n)$, where λ_0 is the unique nonzero root of the following equation:*

$$\sum_{i=1}^{n} w_i e^{-w_i\lambda} = (1-\lambda)\sum_{i=1}^{n} w_i. \tag{3}$$

We remark that $\mathrm{Vol}(G)$ in the statement of Theorem 1 can be replaced by $\mathrm{vol}(G)$, since it was proved in [10] that with probability at least $1 - e^{-c}$,

$$|\mathrm{vol}(G) - \mathrm{Vol}(G)| \leq \sqrt{c\mathrm{Vol}(G)}.$$

Since the average degree is $\mathrm{vol}(G)/n$, the average degree can also be approximated by the average expected degree $\mathrm{Vol}(G)/n$.

The paper is organized as follows. Section 2 contains several facts concerning (3). In section 3, we show that the asymptotic formula for the volume of the giant components of a random graph in $G(\mathbf{w})$ is a generalization of Theorem A by Erdős

[1] The quantitative estimate of this probability is in the proof of Theorem 1 in [10].

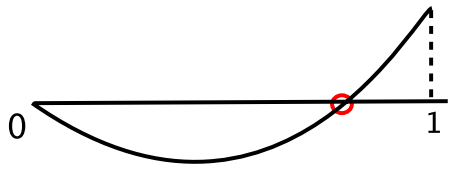


FIG. 1. *When $\tilde{d} > 1$, $f(x)$ has a unique positive root.*

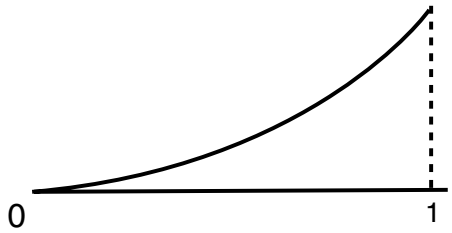


FIG. 2. *When $\tilde{d} < 1$, $f(x) > 0$ for all $x > 0$.*

and Rényi. Section 4 includes some improved lower bounds for the volume of the giant component of $G \in G(\mathbf{w})$ as a function of the expected average degree. In section 5, we give the complete proof of Theorem 1. In section 6, we derive a sharp estimate for the number of vertices in the giant components.

2. Preliminaries. Before we proceed, we examine some basic properties of the solutions to the equation in (3). The proof is quite straightforward and will be omitted here.

Let $\tilde{d}$ denote the expected second order average degree:

$$\tilde{d} = \frac{\sum_i w_i^2}{\sum_i w_i}.$$

LEMMA 1. *Suppose that the expected second order average degree satisfies $\tilde{d} > 1$. Define*

$$f(\lambda) = \sum_{i=1}^{n} w_i e^{-w_i \lambda} - (1 - \lambda) \sum_{i=1}^{n} w_i.$$

We have $f(0) = 0$, $f'(0) < 0$, and $f''(\lambda) > 0$. Hence $f(\lambda) = 0$ has a unique positive solution λ_0 (see Figure 1). In particular,

1. *if $f(x_1) \leq 0$ for some positive x_1, then $\lambda_0 \geq x_1$;*
2. *if $f(x_2) \geq 0$ for some positive x_2, then $\lambda_0 \leq x_2$;*
3. *$\lambda_0 < 1$ since $f(1) > 0$.*

When $\tilde{d} < 1$, we have $f(0) = 0$, $f'(0) > 0$, and $f''(\lambda) > 0$. Zero is the only nonnegative root for $f(x)$ (see Figure 2). This corresponds to the case in which there is no giant component.

The following fact is useful in the proof of the main theorem.

LEMMA 2. *Suppose that the expected average degree d satisfies*

$$d = \frac{1}{n} \sum_{i=1}^{n} w_i \geq 1 + \delta > 1$$

for some positive constant δ. Define $f(\lambda) = \sum_{i=1}^{n} w_i e^{-w_i\lambda} - (1-\lambda)\sum_{i=1}^{n} w_i$, and let λ_0 denote the unique nonzero root of $f(\lambda) = 0$. Then there is a positive constant $c = c(\delta)$ such that

$$f'(\lambda_0) \geq c \sum_{i=1}^{n} w_i.$$

Proof. Since $\tilde{d} \geq d > 1$, the unique root λ_0 of f exists. We have

$$f'(\lambda_0) = \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} w_i^2 e^{-w_i\lambda_0}.$$

Case 1. $\lambda_0 \geq \frac{1}{2}$. Since $xe^{-x\lambda_0}$ attains its maximum at $x = 1/\lambda_0$, we have

$$\begin{aligned} f'(\lambda_0) &= \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} w_i^2 e^{-w_i\lambda_0} \\ &\geq \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} w_i \frac{1}{e\lambda_0} \\ &= \left(1 - \frac{1}{e\lambda_0}\right) \sum_{i=1}^{n} w_i \\ &\geq \left(1 - \frac{2}{e}\right) \sum_{i=1}^{n} w_i. \end{aligned}$$

The statement holds for this case.

Case 2. $\lambda_0 < \frac{1}{2}$. We will utilize some convexity inequalities. First we will prove the following claim.

Now we consider the function $h(x) = (x^2 + \frac{x}{\lambda_0})e^{-\lambda_0 x}$. We have

$$\begin{aligned} h'(x) &= \left(\frac{1}{\lambda_0} + x - \lambda_0 x^2\right) e^{-\lambda_0 x}, \\ h''(x) &= -\lambda_0 x(3 - \lambda_0 x)e^{-\lambda_0 x}. \end{aligned} \tag{4}$$

We need the following facts, whose proofs will be given at the end of this section.

CLAIM A.

(i) $h(x)$ is concave downward over x in $(0, \frac{3}{\lambda_0})$. The maximum value of $h(x)$ for x in $[0, \infty)$ is reached at $x_0 = \frac{\sqrt{5}+1}{2\lambda_0}$.

(ii) $d < \frac{2}{e\lambda_0} < x_0$.

(iii) $\lambda_0 > 1 - \frac{1}{d}$.

Now, we consider the following function:

$$H(x) = \begin{cases} h(x), & 0 \leq x \leq x_0, \\ h(x_0), & x \geq x_0. \end{cases}$$

Using Claim A(i), $H(x)$ is concave downward and $H(x) \geq h(x)$ for all $x \geq 0$. We have

$$\begin{aligned}
f'(\lambda_0) &= \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} w_i^2 e^{-w_i\lambda_0} \\
&= \sum_{i=1}^{n} w_i + \frac{1}{\lambda_0}\sum_{i=1}^{n} w_i e^{-w_i\lambda_0} - \sum_{i=1}^{n} h(w_i) \\
&= \sum_{i=1}^{n} w_i + \frac{1}{\lambda_0}(1-\lambda_0)\sum_{i=1}^{n} w_i - \sum_{i=1}^{n} h(w_i) \\
&= \frac{1}{\lambda_0}\sum_{i=1}^{n} w_i - \sum_{i=1}^{n} h(w_i) \\
&\geq \frac{1}{\lambda_0}\sum_{i=1}^{n} w_i - \sum_{i=1}^{n} H(w_i) \\
&\geq \frac{1}{\lambda_0}\sum_{i=1}^{n} w_i - nH\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right) \\
&= \frac{1}{\lambda_0} nd - nH(d).
\end{aligned}$$

By Claim A(ii), we have $d < \frac{2}{e\lambda_0} < x_0$. Hence, $H(d) = h(d)$.

$$\begin{aligned}
f'(\lambda_0) &\geq \frac{1}{\lambda_0} nd - nh(d) \\
&= \frac{1}{\lambda_0} nd - n\left(d^2 + \frac{d}{\lambda_0}\right) e^{-\lambda_0 d} \\
&= nd\frac{1}{\lambda_0}(1 - (1+d\lambda_0)e^{-\lambda_0 d}) \\
&\geq nd(1 - (1+d\lambda_0)e^{-\lambda_0 d}).
\end{aligned}$$

The function $\psi(x) = 1 - (1+x)e^{-x}$ is increasing for x in $[0,\infty)$. For any $x > 0$, $\psi(x) > \psi(0) = 0$. Hence we have

$$\begin{aligned}
f'(\lambda_0) &\geq nd\psi(\lambda_0 d) \\
&\geq nd\psi(d-1) \\
&\geq cnd
\end{aligned}$$

by choosing $c = c(\delta) = \min\{\psi(\delta), 1 - 2/e\}$.

It remains to prove Claim A.

Proof of Claim A. (i) follows from (4).

To prove (ii), we use the facts that λ_0 is a root of f and $xe^{-\lambda_0 x}$ has its maximum value $\frac{1}{e\lambda_0}$ at $x = 1/\lambda_0$. Then

$$\begin{aligned}
(1-\lambda_0)nd &= (1-\lambda_0)\sum_{i=1}^{n} w_i \\
&= \sum_{i=1}^{n} w_i e^{-\lambda_0 w_i} \\
&\leq \sum_{i=1}^{n} \frac{1}{e\lambda_0} \\
&= \frac{n}{e\lambda_0}.
\end{aligned}$$

Thus,

$$\lambda_0(1-\lambda_0) \le \frac{1}{de}.$$

We have

$$\lambda_0 \le \frac{1}{2}\left(1-\sqrt{1-\frac{4}{de}}\right) \quad \text{or} \quad \lambda_0 \ge \frac{1}{2}\left(1+\sqrt{1-\frac{4}{de}}\right).$$

Then $\lambda_0 < \frac{1}{2}$ implies

$$\begin{aligned}\lambda_0 &\le \frac{1}{2}\left(1-\sqrt{1-\frac{4}{de}}\right)\\ &= \frac{2}{de}\frac{1}{1+\sqrt{1-\frac{4}{de}}}\\ &< \frac{2}{de}.\end{aligned}$$

Hence, we have $d < \frac{2}{e\lambda_0} < x_0$, as desired.

To prove (iii), we consider the function

$$g(x) = \begin{cases} xe^{-\lambda_0 x}, & 0 \le x \le \frac{1}{\lambda_0},\\ \frac{1}{e\lambda_0}, & x > \frac{1}{\lambda_0}.\end{cases}$$

We observe that $g(x)$ is concave downward and $g(x) \ge xe^{-\lambda_0 x}$ for all $x \ge 0$.

By the definition of λ_0, we have

$$\begin{aligned}(1-\lambda_0)nd &= (1-\lambda_0)\sum_{i=1}^{n} w_i\\ &= \sum_{i=1}^{n} w_i e^{-\lambda_0 w_i}\\ &\le \sum_{i=1}^{n} g(w_i)\\ &\le ng(d).\end{aligned}$$

By Claim A(ii), $d < \frac{2}{e\lambda_0}$. Thus, $g(d) = de^{-\lambda_0 d}$. We have

$$1-\lambda_0 \le e^{-\lambda_0 d}.$$

Note that $\phi(\lambda) = (1-\lambda) - e^{-\lambda d}$ is concave downward over $[0,\infty)$. Since $\phi(0) = 0$ and $\phi'(0) = d - 1 > 0$, $\phi(x)$ has a unique positive root, which we denote by s. We have $\phi(x) > 0$ for any $0 < x < s$. Since $\phi(\lambda_0) \le 0$ and $\lambda_0 \ne 0$, we have $\lambda_0 \ge s$.

Define $t = (1-s)d$; then we have

$$\frac{t}{d} = 1 - s = e^{-sd} = e^{-d+t}.$$

Thus t satisfies the following equation:

$$te^{-t} = de^{-d}. \tag{5}$$

The function xe^{-x} increases in $[0,1]$ and decreases in $[1,\infty]$. There is a unique $t<1$ satisfying (5).

We have

$$\lambda_0 \geq s = 1 - \frac{t}{d} > 1 - \frac{1}{d}.$$

The proof of Claim A is now finished, and therefore the proof of Lemma 2 is complete. □

3. Theorem 1 ⇒ Theorem A. In this section we want to show that the formula for the size of the giant component for a random graph in $G(n,p)$ as derived by Erdős and Rényi in Theorem A is a special case of Theorem 1. In other words, if we restrict the expected degree sequence to the case when all degrees are equal, then we recover the theorem of Erdős and Rényi.

THEOREM 2. *Theorem* 1 *implies Theorem* A *of Erdős and Rényi for* $G(n,p)$.

Proof. In $G(n,p)$, we have $w_1 = w_2 = \cdots = w_n = np = d$. Equation (3) becomes

$$e^{-d\lambda} = 1 - \lambda.$$

Let $\lambda = 1 - \frac{1}{d}z$. We have

$$e^{-d+z} = \frac{z}{d},$$

or equivalently,

$$z = de^{-d}e^{z}.$$

Here we use the following version of the well-known Lagrange inversion formula.

LAGRANGE INVERSION FORMULA. *Suppose that* z *is a function of* x *and* y *in terms of another analytic function* ϕ *as follows:*

$$z = x + y\phi(z).$$

Then z *can be written as a power series in* y *as follows:*

$$z = x + \sum_{k=1}^{\infty} \frac{y^k}{k!} D^{(k-1)}\phi^k(x),$$

where $D^{(t)}$ *denotes the* t*th derivative.*

We apply the above formula with $x=0$, $y = de^{-d}$, and $\phi(z) = e^z$. Then we have

$$\begin{aligned} z &= \sum_{k=1}^{\infty} \frac{y^k}{k!} D^{(k-1)} e^{kx}\big|_{x=0} \\ &= \sum_{k=1}^{\infty} \frac{k^{k-1}}{k!} y^k \\ &= \sum_{k=1}^{\infty} \frac{k^{k-1}}{k!} (de^{-d})^k. \end{aligned}$$

This is exactly (1) in Theorem A of Erdős and Rényi. □

4. Lower bounds. Theorem 1 gives an implicit formula for the volume of the giant component for a random graph with a given expected degree sequence. It is often useful to deduce some bounds which depend only on the expected average degree d. Of particular interest is the following question.

Among all random graphs $G(\mathbf{w})$ with the same expected average degree d, which degree distributions minimize or maximize the volume of the giant component?

One obvious example comes to mind. Almost surely $G(m,p)$ with $mp = \Omega(\log m)$ is connected. By adding $n-m$ vertices to $G(m,p)$ with weights zero, we get a random graph $G(\mathbf{w})$ with the expected average degree $d = \frac{mp}{n}$, which almost surely has a giant component with volume $\mathrm{Vol}(G)$.

One might be inclined to conjecture that the random graph with equal expected degrees generates the smallest giant component among all possible degree distributions with the same volume. The answer is "yes" for $1 < d \leq \frac{e}{e-1}$, and a surprising "no" if d is sufficiently large.

We will prove the following theorem.

THEOREM 3. *When $d \geq \frac{4}{e}$, almost surely the giant component of $G \in G(\mathbf{w})$ has volume at least*

$$\left(\frac{1}{2}\left(1+\sqrt{1-\frac{4}{de}}\right)+o(1)\right)\mathrm{Vol}(G).$$

We remark that $\frac{1}{2}(1+\sqrt{1-\frac{4}{de}}) = 1-\frac{1}{de}+O(\frac{1}{d^2})$ improves the bound in Theorem B. In fact, this bound is best possible as d approaches infinity, as shown by the following example.

Example. Let $m = \lfloor n^{3/4} \rfloor$ and $y = 1 + \frac{n}{m}(d-1) \approx (d-1)n^{1/4}$. We choose the expected degrees

$$w_1 = w_2 = \cdots = w_m = y, \qquad w_{m+1} = \cdots = w_n = 1.$$

The expected average degree of this random graph $G(\mathbf{w})$ is

$$\frac{my + (n-m)}{n} = d.$$

Let $x_0 = 1 - \frac{1}{de}$. To show that the giant component of G has volume at most $(x_0 + o(1))\mathrm{Vol}(G)$, it is sufficient to verify $f(x_0) \geq 0$. Here

$$f(\lambda) = \sum_{i=1}^{n} w_i e^{-w_i\lambda} - (1-\lambda)\sum_{i=1}^{n} w_i.$$

We have

$$\begin{aligned} f(x_0) &= \sum_{i=1}^{n} w_i e^{-w_i x_0} - (1-x_0)\sum_{i=1}^{n} w_i \\ &= mye^{-yx_0} + (n-m)e^{-x_0} - (1-x_0)nd \\ &\geq \frac{n}{e}(e^{\frac{1}{de}} - 1 - O(n^{-1/4})) \\ &\geq 0, \end{aligned}$$

as desired.

We are now ready to prove Theorem 3.

Proof of Theorem 3. We note that the function $g(z) = ze^{-z\lambda}$ reaches its maximum value at $z = \frac{1}{\lambda}$. We have

$$\begin{aligned} f(\lambda) &= \sum_{i=1}^{n} w_i e^{-w_i\lambda} - (1-\lambda)\sum_{i=1}^{n} w_i \\ &\le \sum_{i=1}^{n} \frac{1}{\lambda} e^{-1} - (1-\lambda)\sum_{i=1}^{n} w_i \\ &= \frac{n}{e\lambda}(1 - \lambda(1-\lambda)de). \end{aligned}$$

Since λ_0 is a solution of $f(\lambda) = 0$, we have

$$\lambda_0(1-\lambda_0) \le \frac{1}{de},$$

which implies either $\lambda_0 \le \frac{1}{2}(1 - \sqrt{1 - \frac{4}{de}})$ or $\lambda_0 \ge \frac{1}{2}(1 + \sqrt{1 - \frac{4}{de}})$.

We will show that $\lambda_0 \le \frac{1}{2}(1 - \sqrt{1 - \frac{4}{de}})$ is not true by proving $f(\frac{1}{2}) \le 0$.

We note that

$$\begin{aligned} f\left(\frac{1}{2}\right) &= \sum_{i=1}^{n} w_i e^{-w_i/2} - \frac{1}{2}\sum_{i=1}^{n} w_i \\ &\le 2ne^{-1} - \frac{1}{2}nd \\ &= \frac{n}{2}\left(\frac{4}{e} - d\right) \\ &\le 0. \end{aligned}$$

Thus we conclude that $\lambda_0 \ge \frac{1}{2}(1 + \sqrt{1 - \frac{4}{de}})$. □

When d is small and not in the range covered by Theorem 3, we can still derive the following lower bound.

THEOREM 4. *When $1 < d \le \frac{e}{e-1}$, then almost surely $G(\mathbf{w})$ has a giant component of size at least $(\lambda_1 + o(1))\mathrm{Vol}(G)$, where λ_1 is the nonzero root of the following equation:*

$$e^{-\lambda d} = 1 - \lambda. \tag{6}$$

In other words, among all random graphs $G(\mathbf{w})$ with fixed expected average degree d, the Erdős–Rényi random graph $G(n, \frac{d}{n})$ has the smallest giant component (measured in volume).

Proof. Consider the function

$$g(x) = \begin{cases} xe^{-\lambda_1 x}, & 0 \le x \le \frac{1}{\lambda_1}, \\ \frac{1}{e\lambda_1}, & x > \frac{1}{\lambda_1}. \end{cases}$$

We observe that $g(x)$ is concave downward and $g(x) \ge xe^{-\lambda_1 x}$ for all $x \ge 0$. We have

$$\begin{aligned} f(\lambda_1) &= \sum_{i=1}^{n} w_i e^{-\lambda_1 w_i} - (1-\lambda_1)nd \\ &\le \sum_{i=1}^{n} g(w_i) - (1-\lambda_1)nd \\ &\le ng\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right) - (1-\lambda_1)nd \\ &\le n(g(d) - (1-\lambda_1)d). \end{aligned}$$

Since λ_1 is an increasing function of d, $d\lambda_1$ is also an increasing function of d. When $d = \frac{e}{e-1}$, it is easy to verify that $\lambda = 1 - \frac{1}{e}$ is the other root of (6). Therefore, $d\lambda_1 \le 1$ when $d \le \frac{e}{e-1}$. In particular, we have

$$g(d) = de^{-\lambda_1 d}.$$

Hence

$$\begin{aligned} f(\lambda_1) &\le n(g(d) - (1-\lambda_1)d) \\ &= nd(e^{-\lambda_1 d} - (1-\lambda_1)) \\ &= 0. \end{aligned}$$

By Remark 1, we have $\lambda_0 \ge \lambda_1$, as desired. □

5. The proof of the main theorem. A central tool that we use in the proof of the main theorem is a relaxed version of the Azuma inequality (as seen in Theorem 1 of [12]), which can be described as follows.

Suppose that Ω is a probability space, and that $\mathcal{F}$ denote a σ-field on Ω (i.e., a collection of subsets of Ω, which contains $\emptyset$ and Ω and is closed under unions, intersections, and complementation). A *filter* $\mathbf{F}$ is an increasing chain of σ-subfields

$$\{0, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F}.$$

A martingale (obtained from) X is associated with a filter $\mathbf{F}$ and a sequence of random variables $X_0, X_1, \ldots, X_n$ satisfying $X_i = E(X \mid \mathcal{F}_i)$ and, in particular, $X_0 = E(X)$ and $X_n = X$. For undefined terminology on martingales, the reader is referred to [19].

For $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ a vector with positive entries, a martingale X is said to be c-Lipschitz if

$$|X_i - X_{i-1}| \le c_i \tag{7}$$

for $i = 1, 2, \ldots, n$.

If the c-Lipschitz condition is not satisfied, we can still consider the following relaxed version.

A martingale X is said to be near-c-Lipschitz with an exceptional probability η if

$$\sum_i \Pr(|X_i - X_{i-1}| \ge c_i) \le \eta. \tag{8}$$

THEOREM C (Theorem 1 as in [12]). *For nonnegative values, $c_1, c_2, \ldots, c_n$, a martingale X is near-c-Lipschitz with an exceptional probability η. Then X satisfies*

$$\Pr(|X - E(X)| < a) \le 2e^{-\frac{a^2}{2\sum_{i=1}^{n} c_i^2}} + \eta.$$

The idea for the proof of Theorem 1 is to first prove that the volume of giant component concentrates on its expected value $\mathrm{E}(\mathrm{Vol}(GCC))$ and then show that $\mathrm{E}(\mathrm{Vol}(GCC))/\mathrm{Vol}(G)$ can be approximated by the nonzero root of (3). To do so, we need to establish several useful facts.

LEMMA 3. *With probability at least* $1 - 2n^{-k}$*, a vertex with weight greater than* $\max\{8k, 2(k+1+o(1))\mu(d)\}\log n$ *is in the giant component of* $G(\mathbf{w})$.

Proof. Consider a vertex v_i with weight $w_i \geq \max\{8k, 2(k+1+o(1))\mu(d)\}\log n$. For a random graph G in $G(\mathbf{w})$, let d_i denote the degree of v_i in G. Then, d_i is the sum of independent 0-1 random variables with $\mathrm{E}(d_i) = w_i$. For any nonnegative value λ, we have

$$\Pr(d_i - \mathrm{E}(d_i) < -\lambda) \leq e^{-\frac{\lambda^2}{2\mathrm{E}(d_i)}}.$$

By choosing $\lambda = w_i/2$, we have

$$\Pr(d_i < w_i/2) \leq e^{-w_i/8} \leq n^{-k}.$$

With probability at least $1 - n^{-k}$, v_i is in a connected component of size at least $w_i/2$. If this connected component is not the giant component, then the second largest component must have size at least $w_i/2$. However, from Theorem B, this can happen with probability only at most n^{-k} because of the assumption that

$$w_i/2 \geq (k+1+o(1))\mu(d)\log n.$$

Hence, with probability at least $1 - 2n^{-k}$, a vertex with weight greater than $\max\{8k, 2(k+1+o(1))\mu(d)\}\log n$ is in the giant component. □

LEMMA 4. *For any* $k > 2$*, with probability at least* $1 - 6n^{-k+2}$*, we have*

$$|\mathrm{Vol}(GCC) - \mathrm{E}(\mathrm{Vol}(GCC))| \leq 2C_1(k+1)^2\sqrt{k-2}\sqrt{n}\log^{2.5} n,$$

for some positive constant C_1.

Proof. Let $L = L(k)$ be the set of vertices with weight greater than $\max\{8k, 2(k+1+o(1))\mu(d)\}\log n$. If $L \neq \emptyset$, we form a new graph G^* by adding a new vertex v_* to $G(\mathbf{w})$ and add edges from v_* to each vertex in L. $G(\mathbf{w})$ almost surely has a giant component, and so does G^*. Let X denote the volume of the giant component in G^*. (While computing the values for Vol of the giant component in G^*, we use the convention that the weight of v_* is zero.) If $L = \emptyset$, we simply let $X = \mathrm{Vol}(GCC)$.

We wish to show the concentration of the random variable X. It is sufficient to prove the following claim.

CLAIM B.

$$\Pr(|X - \mathrm{E}(X)| < \lambda) \leq 4n^{-k+2},$$

where $\lambda = 2C_1(k+1)^2\sqrt{k-2}\sqrt{n}\log^{2.5} n$.

We observe that X does not depend on whether $\{u, v\}$ is an edge if both u and v are in L. We list all pairs of vertices with at least one vertex not in L by $\{f_1, f_2, \ldots, f_m\}$, where $m = \binom{n}{2} - \binom{|L|}{2}$. (The order of edges in the list is arbitrarily chosen.) For $i = 0, 1, 2, \ldots, m$, let $\mathcal{F}_i$ denote the σ-field generated by exposing pairs $f_1, f_2, \ldots, f_i$. We apply Theorem C on the edge-exposing martingale X with $X_i = \mathrm{E}(X|\mathcal{F}_i)$ and $X_m = X$. We wish to find a good Lipschitz or near-Lipschitz bound for $|X_i - X_{i-1}|$. By definition, X_{i-1} is the conditional expectation of X_i. Choosing the

pair f_i as an edge can change X by at most the volume of a small component. Let v_i be a vertex of the pair f_i not in L. (If there is a tie, break arbitrarily.) Let G_{v_i} be the random graph obtained by deleting v_i from $G(\mathbf{w})$. The possible small component containing v before f_i is exposed can be broken into at most d_i largest connected components excluding the giant component in G_{v_i}.

First, we apply Theorem B to the random graph G_{v_i}. Note that the average degree of G_{v_i} is $(1+o(1))d$. Thus, with probability at least $1-n^{-k}$, all small components of G_{v_i} have size at most $(k+1+o(1))\mu(d)\log n$. Similarly, with probability at least $1-n^{-k}$, all small components of G_{v_i} have volumes at most $C(k+1)\log n$ for some positive constant depending only on d. Also, for any positive λ', the degree d_i of v_i can be upper bounded by

$$\Pr(d_i > w_i + \lambda') < e^{-\frac{\lambda^2}{2(w_i+\lambda'/3)}}.$$

By choosing $\lambda' = w_i + 2k\log n$, we have

$$\begin{aligned}\Pr(d_i > 2w_i + 2k\log n) &< e^{-\frac{\lambda^2}{2(w_i+\lambda/3)}}\\ &= e^{-\frac{(w_i+2k\log n)^2}{2(w_i+(w_i+2k\log n)/3)}}\\ &< n^{-k}.\end{aligned}$$

Thus, with probability at least $1-2n^{-k}$, we have

$$\begin{aligned}|X_i - X_{i-1}| &\le d_i \times (k+1)C\log n\\ &< (2w_i + 2k\log n)(k+1)C\log n\\ &< (10k\log n + 2(k+1+o(1))\mu(d)\log n)(k+1)C\log n\\ &< C_1(k+1)^2\log^2 n,\end{aligned}$$

where $C_1 = C(10+2\mu(d))$ is a bounded positive number.

Now we apply Theorem C on martingale X with $c_i = C_1(k+1)^2\log^2 n$ and $\eta \le \binom{n}{2}2n^{-k}$. For any positive λ, we have

$$\begin{aligned}\Pr(|X - \mathrm{E}(X)| > \lambda) &\le 2e^{-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}} + \eta\\ &\le 2e^{-\frac{\lambda^2}{2C_1^2(k+1)^4 n\log^4 n}} + 2n^{-k+2}.\end{aligned}$$

For $\lambda = 2C_1(k+1)^2\sqrt{k-2}\sqrt{n}\log^{2.5} n$, we have

$$Pr(|X - \mathrm{E}(X)| > \lambda) \le 4n^{-k+2},$$

as desired. □

Proof of Theorem 1. For any vertex v with weight w_v, the probability that v is not in the giant component of $G(\mathbf{w})$ can be estimated as follows. To simplify the notation, we write $C_k = \max\{8k, 2(k+1+o(1))\mu(d)\}$.

Case a. $w_v \ge C_k\log n$. By Lemma 3, we have

$$\Pr(v \notin GCC) \le \frac{2}{n^k}.$$

Case b. $w_v \le C_k\log n$. Let G_v be the random graph by removing v from G. Expose every pair of vertices in G_v. Let H be the giant component of G_v. Applying Lemma 4 to G_v, with probability at least $1-\frac{6}{(n-1)^{k-2}}$, we have

$$|\mathrm{Vol}(H) - \mathrm{E}(\mathrm{Vol}(H))| \le 2C_1(k+1)^2\sqrt{k-2}\sqrt{n}\log^{2.5} n.$$

Now we expose the pairs of vertices containing v. We have

$$\begin{aligned}\Pr(v \notin GCC|H) &= \prod_{v_j \in V(H)} (1 - w_v w_j \rho) \\ &= e^{-\sum_{v_j \in V(H)} w_v w_j \rho + \sum_{v_j \in V(H)} w_v^2 w_j^2 \rho^2} \\ &= e^{-w_v \mathrm{Vol}(H)\rho(1+O(w_v \tilde{d}\rho))}.\end{aligned}$$

The probability that v is not in the giant component can be estimated as follows:

$$\begin{aligned}\Pr(v \notin GCC) &= \mathrm{E}(\Pr(v \notin GCC|H)) + O(n^{-k+2}) \\ &= \mathrm{E}(e^{-w_v \mathrm{Vol}(H)\rho}) + O(n^{-k+2}) \\ &= e^{-w_v \mathrm{E}(\mathrm{Vol}(H))\rho + O(k^2 w_i \rho \sqrt{n} \log^{2.5} n)} + O(n^{-k+2}). \qquad (9)\end{aligned}$$

Note that GCC can be formed from H by joining at most d_v's small components. Thus, we have

$$\begin{aligned}|\mathrm{E}(GCC) - \mathrm{E}(H)| &\le \mathrm{E}(d_v)(k+1)C \log n + 2n^{-k} \\ &= w_v(k+1)C \log n + 2n^{-k} \\ &= O(w_v k \log n).\end{aligned}$$

By substituting $\mathrm{E}(H)$ by $\mathrm{E}(\mathrm{Vol}(GCC)) + O(w_v k \log n)$ in (9), we have

$$\begin{aligned}\Pr(v \notin GCC) &= e^{-w_v \mathrm{E}(\mathrm{Vol}(GCC))\rho + O(w_v^2 k \rho \log n) + O(k^2 w_v \rho \sqrt{n} \log^{2.5} n)} + O(n^{-k+2}) \\ &= (1 + O(k^3 \rho \sqrt{n} \log^{3.5} n)) e^{-w_v \mathrm{E}(\mathrm{Vol}(GCC))\rho} + O(n^{-k+2}).\end{aligned}$$

Putting these together, we have

$$\begin{aligned}&\mathrm{Vol}(G) - \mathrm{E}(\mathrm{vol}(GCC)) \\ &\quad= \sum_v w_v \Pr(v \notin GCC) \\ &\quad= \sum_{w_v < C_k \log n} w_v \Pr(v \notin GCC) + \sum_{w_v \ge C_k \log n} w_v \Pr(v \notin GCC) \\ &\quad= \sum_{w_v < C_k \log n} w_v \left[(1 + O(k^3 \rho \sqrt{n} \log^{3.5} n)) e^{-w_v \mathrm{E}(\mathrm{Vol}(GCC))\rho} + O(n^{-k+2}) \right] \\ &\qquad+ \sum_{w_v \ge C_k \log n} w_v O(2n^{-k}) \\ &\quad= \sum_{w_v < C_k \log n} w_v e^{-w_v \mathrm{E}(\mathrm{Vol}(GCC))\rho} + O(k^3 \sqrt{n} \log^{3.5} n).\end{aligned}$$

We choose k to be a constant large enough satisfying

$$C_k \ge \begin{cases} \frac{2}{(1 - \frac{2}{\sqrt{de}})} & \text{if } d > \frac{4}{e}, \\ \frac{2}{(1 - \frac{1 + \log d}{d})} & \text{if } 1 < d < 2. \end{cases}$$

By Theorem A, we have $C_k \mathrm{E}(\mathrm{Vol}(GCC))\rho \ge 2$. In particular, for any vertex v with $w_v \ge C_k \log n$, we have

$$e^{-w_v \mathrm{E}(\mathrm{Vol}(GCC))\rho} \le n^{-2}.$$

Thus,

$$\sum_{w_v \geq C_k \log n} w_v e^{-w_v \mathrm{E}(\mathrm{Vol}(GCC))\rho} = O(n^{-1}).$$

Therefore we have

$$\mathrm{Vol}(G) - \mathrm{E}(\mathrm{vol}(GCC)) = \sum_v w_v e^{-w_v \mathrm{E}(\mathrm{Vol}(GCC))\rho} + O(\sqrt{n}\log^{3.5} n).$$

Letting $x_0 = \frac{\mathrm{Vol}(GCC)}{\mathrm{Vol}(G)}$ and $f(x) = \sum_{i=1}^n w_i e^{-w_i x} - (1-x)\sum_{i=1}^n w_i$, we have

$$f(x_0) = O(\sqrt{n}\log^{3.5} n). \tag{10}$$

The equation $f(x) = 0$ has only two roots, $x = 0$ and $x = \lambda_0$. Note that $f(x)$ is concave upward with $|f'(0)| = n(d^2 - d)$, and $|f'(\lambda_0)| > cnd$. Consider a small interval I around 0 with diameter $O(\sqrt{n}\log^{3.5} n)$. The preimage $f^{-1}(I)$ has diameter at most $O(n^{-1/2}\log^{3.5} n)$. Since x_0 is bounded away from 0 by a small constant, we have $|x_0 - \lambda_0| = O(n^{-1/2}\log^{3.5} n)$. Therefore, almost surely the giant component has volume

$$\lambda_0 \mathrm{Vol}(G) + O(\sqrt{n}\log^{3.5} n).$$

Theorem 1 is proved. □

6. The complement of the giant component and its size. As we know, the giant component almost surely exists if the expected average degree $d > 1$. We consider the remaining graph G' after removing the giant component.

For a random graph G in the Erdős–Rényi model $G(n,p)$, where $p = d/n$, if $d > 1$, there is a unique $c < 1$ satisfying

$$ce^{-c} = de^{-d}.$$

We write $\lambda_0 = 1 - \frac{c}{d}$. For any vertex v, the probability that $v \in S$ is known [19] to be

$$e^{-\lambda_0 d} = e^{-d+c} = \frac{c}{d}.$$

Hence S has $(\frac{c}{d} + o(1))n$ vertices. After removing the giant component from $G(n,p)$, the remaining graph can be viewed as a random graph in $G(n',p)$, where $n' \approx \frac{c}{d}n$.

The above fact can be generalized to the random graph model $G(\mathbf{w})$. The following theorem is based on the proof of Theorem 1, and we omit the proof here.

THEOREM 5. *Suppose that the expected average degree d is strictly greater than 1. Let G' denote the remaining graph of a random graph G in $G(\mathbf{w})$ by removing the giant component. Then almost surely G' is an induced subgraph on a random subset S satisfying the following:*

1. *Any vertex v_i is contained in S with probability $e^{-\lambda_0 w_i}$, where λ_0 is as defined in (3).*
2. *For any $v_i, v_j \in S$, the probability that v_iv_j is an edge of G_S is $w_iw_j/\mathrm{Vol}(G)$. The induced subgraph G_S is a random graph with given expected degrees*

$$\{(1-\lambda_0)w_i\}_{v_i \in S}.$$

3. *$G' \setminus G_S$ consists of at most $O(\log n)$ components each with size $O(\log n)$.*

We further analyze the size of the giant component. The proof is similar and will be omitted.

THEOREM 6. *If the expected average degree is strictly greater than* 1*, then almost surely the giant component in a random graph of given expected degrees* w_i*,* $i = 1, \ldots, n$*, has* $n - \sum_{i=1}^{n} e^{-w_i\lambda_0} + O(\sqrt{n}\log^{4.5} n)$ *vertices and* $(\lambda_0 - \frac{1}{2}\lambda_0^2)\mathrm{Vol}(G) + O(\sqrt{\mathrm{Vol}(G)}\log^{3.5} n)$ *edges, where* λ_0 *is as defined in* (3)*.*

7. Comparing theoretical results with the data from the collaboration graph. To illustrate the effectiveness of our results, we use an example of the collaboration graph of the second kind. Based on the data of *Mathematics Review* [18], there are about 401,000 authors as vertices. Two vertices are joined by an edge if there is a paper by exactly two authors. There are about 284,000 edges. The giant component has 176,000 vertices and 248,000 edges. Suppose that we model this collaboration graph as a random graph with some given expected degrees w_i. Although we do not know the exact values of the w_i's, we can make the following deductions using the theorems in the previous section.

By Theorem 6, we have

$$\lambda_0(2 - \lambda_0) \approx \frac{\mathrm{Vol}(GCC)}{\mathrm{Vol}(G)} \approx \frac{248000}{284000}.$$

Solving the above equation, we have $\lambda_0 \approx 0.644$.

For a fixed vertex v_i, the degree of v_i follows the Poisson distribution with expected value w_i. Namely, for a fixed k, the probability that v_i has degree k is $\frac{w_i^k}{k!}e^{-w_i}$. Let n_k denote the number of vertices of degree k. Then by the linearity of the expectation, we have

$$\mathrm{E}(n_k) \approx \sum_{i=0} \frac{w_i^k}{k!} e^{-w_i}.$$

Theorem 6 implies that the size of the giant component satisfies

$$\begin{aligned}
|GCC| &\approx n - \sum_{i=1}^{n} e^{-\lambda_0 w_i} \\
&= n - \sum_{i=1}^{n} e^{(1-\lambda_0)w_i} e^{-w_i} \\
&= \sum_{k\geq 0} n_k - \sum_{i=1}^{n}\sum_{k=0}^{\infty} \frac{(1-\lambda_0)^k}{k!} w_i^k e^{-w_i} \\
&\approx \sum_{k\geq 0} n_k(1 - (1-\lambda_0)^k) \\
&= \sum_{k\geq 1} n_k(1 - (1-\lambda_0)^k). \qquad (11)
\end{aligned}$$

Here we estimate n_k by

$$n_k \approx \mathrm{E}(n_k) \approx \sum_{k=1}^{n} \frac{w_i^k}{k!} e^{-w_i}.$$

Grossman, Ion, and De Castro [18] have computed the n_k's as shown in Table 1.

TABLE 1
The degree sequence of the collaboration graph of the second kind.

n_0	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	$\cdots$
166381	145872	34227	16426	9913	6670	4643	3529	2611	2032	$\cdots$

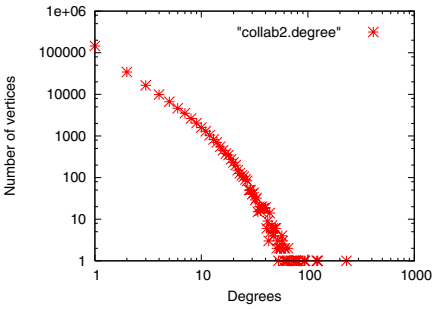


FIG. 3. *Degree distribution of the collaboration graph of the second kind.*

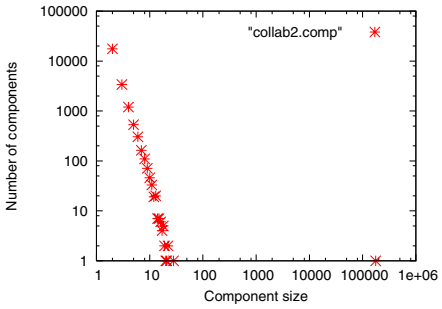


FIG. 4. *Size distribution of connected components of the collaboration graph of the second kind.*

By substituting the above n_k's into (11), the size of the giant component is estimated to be about $177,400$. This is rather close to the actual value $176,000$, within an error bound of less than 1%.

In Figures 3 and 4, we have plotted the degree distribution and the distribution of the sizes of connected components of the collaboration graph of the second kind.

Acknowledgment. The authors are grateful to the referees for numerous valuable comments and crucial corrections on the earlier draft of this paper.

REFERENCES

[1] W. AIELLO, F. CHUNG, AND L. LU, *A random graph model for massive graphs*, in Proceedings of the 32nd Annual ACM Symposium on Theory of Computing, Portland, OR, 2000, pp. 171–180.

[2] W. AIELLO, F. CHUNG, AND L. LU, *A random graph model for power law graphs*, Experiment. Math., 10 (2001), pp. 53–66.

[3] W. AIELLO, F. CHUNG, AND L. LU, *Random evolution in massive graphs*, in Handbook of Massive Data Sets, Vol. 2, J. M. Abello, P. M. Pardalos, and M. G. C. Resende, eds., Kluwer Academic Publishers, Norwell, MA, 2002, pp. 97–122.

[4] A.-L. BARABÁSI AND R. ALBERT, *Emergence of scaling in random networks*, Science, 286 (1999) pp. 509–512.

[5] A.-L. BARABÁSI, R. ALBERT, AND H. JEONG, *Scale-free characteristics of random networks: The topology of the World Wide Web*, Phys. A, 281 (2000), pp. 69–77.

[6] A. BRODER, R. KUMAR, F. MAGHOUL, P. RAGHAVAN, S. RAJAGOPALAN, R. STATA, A. TOMPKINS, AND J. WIENER, *Graph structure in the web*, in Proceedings of the WWW9 Conference, Amsterdam, 2000, pp. 309–320.

[7] B. BOLLABÁS AND O. RIORDAN, *Robustness and vulnerability of scale-free random graphs*, Internet Math., 1 (2003), pp. 1–35.

[8] F. CHUNG, *Spectral Graph Theory*, AMS, Providence, RI, 1997.

[9] F. CHUNG AND L. LU, *The diameter of random sparse graphs*, Adv. Appl. Math., 26 (2001), pp. 257–279.

[10] F. CHUNG AND L. LU, *Connected components in a random graph with given degree sequences*, Ann. Comb., 6 (2002), pp. 125–145.

[11] F. CHUNG AND L. LU, *The average distance in random graphs with given expected degrees*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 15879–15882.

[12] F. Chung and L. Lu, *Coupling online and offline analyses for random power law graphs*, Internet Math., 1 (2004), pp. 409–461.
[13] F. Chung, L. Lu, and V. Vu, *The spectra of random graphs with given expected degrees*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 6313–6318.
[14] C. Cooper and A. Frieze, *A general model of undirected web graphs*, Random Structures Algorithms, 22 (2003), pp. 311–335.
[15] C. Cooper, A. Frieze, and J. Vera, *Random deletions in a scale free random graph*, Internet Math., 1 (2004), pp. 463–483.
[16] P. Erdős and T. Gallai, *Gráfok előírt fokú pontokkal (Graphs with points of prescribed degrees)*, Mat. Lapok, 11 (1961), pp. 264–274 (in Hungarian).
[17] P. Erdős and A. Rényi, *On random graphs*. I, Publ. Math. Debrecen, 6 (1959), pp. 290–297.
[18] J. Grossman, P. Ion, and R. De Castro, *Facts about Erdős Numbers and the Collaboration Graph*, resource website at http://www.oakland.edu/enp/trivia.html.
[19] S. Janson, T. Łuczak, and A. Rucinski, *Random Graphs*, Wiley-Interscience, New York, 2000.
[20] H. Jeong, B. Tomber, R. Albert, Z. Oltvai, and A.-L. Babárasi, *The large-scale organization of metabolic networks*, Nature, 407 (2000), pp. 378–382.
[21] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal, *The web as a graph*, in Proceedings of the 19th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Dallas, TX, 2000, pp. 1–10.
[22] A. J. Lotka, *The frequency distribution of scientific productivity*, J. Washington Acad. Sci., 16 (1926), pp. 317–323.
[23] L. Lu, *The diameter of random massive graphs*, in Proceedings of the 12th ACM-SIAM Symposium on Discrete Algorithms, Washington, DC, 2001, SIAM, Philadelphia, pp. 912–921.
[24] C. McDiarmid, *Concentration*, in Probabilistic Methods for Algorithmic Discrete Mathematics, Algorithms Combin. 16, M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, eds., Springer-Verlag, Berlin, 1998, pp. 195–248.
[25] S. Milgram, *The small world problem*, Psychology Today, 2 (1967), pp. 60–67.
[26] M. Mitzenmacher, *A brief history of generative models for power law and lognormal distributions*, Internet Math., 1 (2004), pp. 226–251.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 412–427

ON THE SPANNING RATIO OF GABRIEL GRAPHS AND β-SKELETONS*

PROSENJIT BOSE†, LUC DEVROYE‡, WILLIAM EVANS§, AND DAVID KIRKPATRICK§

Abstract. The spanning ratio of a graph defined on n points in the Euclidean plane is the maximum ratio over all pairs of data points (u, v) of the minimum graph distance between u and v divided by the Euclidean distance between u and v. A connected graph is said to be an S-spanner if the spanning ratio does not exceed S. For example, for any S there exists a point set whose minimum spanning tree is not an S-spanner. At the other end of the spectrum, a Delaunay triangulation is guaranteed to be a 2.42-spanner [J. M. Keil and C. A. Gutwin, *Discrete Comput. Geom.*, 7 (1992), pp. 13–28]. For proximity graphs *between* these two extremes, such as Gabriel graphs [K. R. Gabriel and R. R. Sokal, *Systematic Zoology*, 18 (1969), pp. 259–278], relative neighborhood graphs [G. T. Toussaint, *Pattern Recognition*, 12 (1980), pp. 261–268], and β-skeletons [D. G. Kirkpatrick and J. D. Radke, *Comput. Geom.*, G. T. Toussaint, ed., Elsevier, Amsterdam, 1985, pp. 217–248] with $\beta \in [0, 2]$ some interesting questions arise. We show that the spanning ratio for Gabriel graphs (which are β-skeletons with $\beta = 1$) is $\Theta(\sqrt{n})$ in the worst case. For all β-skeletons with $\beta \in [0, 1]$, we prove that the spanning ratio is at most $O(n^\gamma)$, where $\gamma = (1 - \log_2(1 + \sqrt{1 - \beta^2}))/2$. For all β-skeletons with $\beta \in [1, 2)$, we prove that there exist point sets whose spanning ratio is at least $\left(\frac{1}{2} - o(1)\right)\sqrt{n}$. For relative neighborhood graphs [G. T. Toussaint, *Pattern Recognition*, 12 (1980), pp. 261–268] (skeletons with $\beta = 2$), we show that there exist point sets where the spanning ratio is $\Omega(n)$. For points drawn independently from the uniform distribution on the unit square, we show that the spanning ratio of the (random) Gabriel graph and all β-skeletons with $\beta \in [1, 2]$ tends to ∞ in probability as $\sqrt{\log n / \log\log n}$.

Key words. Gabriel graph, β-skeletons, spanners, proximity graphs, probabilistic analysis, computational geometry, geometric spanners

AMS subject classifications. Primary 68U05; Secondary 60D05, 68R10

DOI. 10.1137/S0895480197318088

1. Introduction.

Many problems in geometric network design, pattern recognition and classification, geographic variation analysis, geographic information systems, computational geometry, computational morphology, and computer vision use the underlying *structure* (also referred to as the *skeleton* or *internal shape*) of a set of data points revealed by means of a *proximity graph* (see, for example, [16, 13, 7, 9]). A proximity graph attempts to exhibit the relation between points in a point set. Two points are joined by an edge if they are deemed *close* by some proximity measure. It is the measure that determines the type of graph that results. Many different measures of proximity have been defined, giving rise to many different types of proximity graphs. An extensive survey on the current research in proximity graphs can be found in Jaromczyk and Toussaint [9].

*Received by the editors March 10, 1997; accepted for publication (in revised form) July 22, 2005; published electronically May 12, 2006.
http://www.siam.org/journals/sidma/20-2/31808.html

†School of Computer Science, Carleton University, Ottawa, ON, K1S 5B6, Canada (jit@scs.carleton.ca). This auhtor's research was supported by NSERC grant OGP0183877 and by a FIR grant.

‡School of Computer Science, McGill University, Montreal, PQ, H3A 2A7, Canada (luc@cs.mcgill.ca). This author's research was supported by NSERC grant A4456 and by FCAR grant 90-ER-0291.

§Department of Computer Science, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada (will@cs.ubc.ca, kirk@cs.ubc.ca). The research of these authors was supported by the NSERC.

We are concerned with the spanning ratio of proximity graphs. Consider n points in $\mathbb{R}^2$, and define a graph on these points, such as the Gabriel graph [8], or the relative neighborhood graph [16]. For a pair of data points (u, v), the length of the shortest path between u and v in the graph, where edge length is measured by Euclidean distance, is denoted by $L(u, v)$, while the direct Euclidean distance is $D(u, v)$. The *spanning ratio* of the graph is defined by

$$S \stackrel{\text{def}}{=} \max_{(u,v)} \frac{L(u, v)}{D(u, v)},$$

where the maximum is over all $\binom{n}{2}$ pairs of data points. Note that if the graph is not connected, the spanning ratio is infinite. In this paper, we will concentrate on connected graphs.

Graphs with small spanning ratios are important in some applications (see [7] for a survey on spanners). The history for the Delaunay triangulation is interesting. First, Chew [2, 3] showed that in the worst case, $S \geq \pi/2$. Subsequently, Dobkin, Friedman, and Supowit [5] showed that the Delaunay triangulation was a $((1 + \sqrt{5})/2)\pi \approx 5.08$ spanner. Finally, Keil and Gutwin [10, 11] improved this to $2\pi/(3\cos(\pi/6)$ which is about 2.42. It is conjectured that the spanning ratio of the Delaunay triangulation is $\pi/2$. The complete graph has $S = 1$, but is less interesting because the number of edges is not linear but quadratic in n. In this paper, we concentrate on the parametrized family of proximity graphs known as β-skeletons [12] with β in the interval $[0, 2]$. The family of β-graphs contains certain well-known proximity graphs such as the Gabriel graph [8] when $\beta = 1$ and the relative neighborhood graph [16] when $\beta = 2$. As graphs become sparser, their spanning ratios increase. For example, it is trivial to show that there are minimal spanning trees with n vertices for which $S \geq n - 1$, whereas the Delaunay triangulation has a constant spanning ratio.

In this note, we probe the expanse between these two extremes. We show that for any n there exists a point set in the plane whose Gabriel graph satisfies $S \geq c\sqrt{n}$, where c is a universal constant. We also show that for any Gabriel graph in the plane, $S \leq c'\sqrt{n}$ for another constant c'. For all β-skeletons with $\beta \in [0, 1]$, we prove that the spanning ratio is at most $O(n^\gamma)$, where $\gamma = (1 - \log_2(1 + \sqrt{1 - \beta^2}))/2$. For all β-skeletons with $\beta \in [1, 2)$, we prove that there exist point sets whose spanning ratio is at least $\left(\frac{1}{2} - o(1)\right)\sqrt{n}$. For relative neighborhood graphs, we show that there exist point sets where the spanning ratio is $\Omega(n)$. The second part of this paper deals with point sets drawn independently from the uniform distribution on the unit square. We show that the spanning ratio of the (random) Gabriel graph and all β-skeletons with $\beta \in [1, 2]$ tends to ∞ in probability as $\sqrt{\log n / \log\log n}$.

2. Preliminaries. We begin by defining some of the graph theoretic and geometric terminology used in this paper. For more details, see [1] and [15].

A *graph* $G = (V, E)$ consists of a finite nonempty set $V(G)$ of *vertices*, and a set $E(G)$ of unordered pairs of vertices known as *edges*. An edge $e \in E(G)$ consisting of vertices u and v is denoted by $e = uv$; u and v are called the *endpoints* of e and are said to be *adjacent* vertices or *neighbors*. A *path* in a graph G is a finite nonnull sequence $v_1 v_2 \ldots v_k$ with $v_i \in V(G)$ and $v_i v_{i+1} \in E(G)$ for all i. The vertices v_1 and v_k are known as the *endpoints* of the path. A graph is *connected* if, for each pair of vertices $u, v \in V$, there is a path with endpoints u and v (i.e., a path from u to v).

Intuitively speaking, a *proximity graph* on a finite set $P \subset \mathbb{R}^2$ is obtained by connecting pairs of points of P with line segments if the points are considered to be *close* in some sense. Different definitions of closeness give rise to different proximity

graphs. One technique for defining a proximity graph on a set of points is to select a geometric region defined by two points of P—for example, the smallest disk containing the two points—and then specifying that a segment is drawn between the two points if and only if this region contains no other points from P. Such a region will be referred to as a *region of influence* of the two points.

Given a set P of points in $\mathbb{R}^2$, the *relative neighborhood graph of* P, denoted by $RNG(P)$, has a segment between points u and v in P if the intersection of the open discs of radius $D(u,v)$ centered at u and v is empty. This region of influence is referred to as the *lune* of u and v. Equivalently, $u, v \in P$ are adjacent if and only if

$$D(u,v) \leq \max\{D(u,w), D(v,w)\} \text{ for all } w \in P, w \neq u, v.$$

The *Gabriel graph of* P, denoted by $GG(P)$, has as its region of influence the closed disk having segment $\overline{uv}$ as diameter. That is, two vertices $u, v \in P$ are adjacent if and only if

$$D^2(u,v) < D^2(u,w) + D^2(v,w) \text{ for all } w \in P, w \neq u, v.$$

A *Delaunay triangulation* of a set P of points in the plane, denoted by $DT(P)$, is a triangulation of P such that for each interior face, the triangle which bounds that face has the property that the circle circumscribing the triangle contains no other points of the graph in its interior. A set P may admit more than one Delaunay triangulation, but only if P contains four or more cocircular points. A list of properties of the Delaunay triangulation can be found in [15].

We describe another graph, a *minimum spanning tree*, which is not defined in terms of a region of influence. Given a set P of points in the plane, consider a connected straight-line graph G on P, that is, a graph having as its edge set E a collection of line segments connecting pairs of vertices of P. Define the *weight* of G to be the sum of all of the edge lengths of G. Such a graph is called a *minimum spanning tree of* P, denoted by $MST(P)$, if its weight is no greater than the weight of any other connected straight-line graph on P. (It is easy to see that such a graph must be a tree.) In general, a set P may have many minimum spanning trees (for example, if P consists of the vertices of a regular polygon).

The following relationships among the different proximity graphs hold for any finite set P of points in the plane.

LEMMA 1 (see [15]). $MST(P) \subseteq RNG(P) \subseteq GG(P) \subseteq DT(P)$.

A *β-skeleton* of a set P of points in the plane is a proximity graph in which the region of influence, $R(u,v,\beta)$, for two points $u, v \in P$ is a function of β:

1. For $\beta = 0$, $R(u,v,\beta)$ is the line segment $\overline{uv}$.
2. For $0 < \beta < 1$, $R(u,v,\beta)$ is the intersection of the two discs of radius $D(u,v)/(2\beta)$ passing through both u and v.
3. For $1 \leq \beta < \infty$, $R(u,v,\beta)$ is the intersection of the two discs of radius $\beta D(u,v)/2$ centered at the points $(1-\beta/2)u+(\beta/2)v$ and $(\beta/2)u+(1-\beta/2)v$.
4. For $\beta = \infty$, $R(u,v,\beta)$ is the infinite strip perpendicular to the line segment $\overline{uv}$.

The edge uv is in the β-skeleton of P if $R(u,v,\beta) \cap P \setminus \{u,v\} = \emptyset$. Notice that different values of the parameter β give rise to different graphs. Note also that different graphs may result for the same value of β if the regions of influence are constructed with open rather than closed discs; however, these boundary effects do not alter our results. When necessary, we will explicitly state whether the region of influence is open or closed. These graphs will be referred to as open β-skeletons and closed β-skeletons,

respectively. The closed 1-skeleton is the Gabriel graph and the open 2-skeleton is the relative neighborhood graph.

As the value of β increases, β-skeletons become sparser since each region of influence expands.

OBSERVATION 1. *If $\beta \leq \beta'$, then the β'-skeleton is a subset of the β-skeleton of a point set.*

β-skeletons with $\beta > 2$ may be disconnected, so we will concentrate on the interval $\beta \in [0, 2]$.

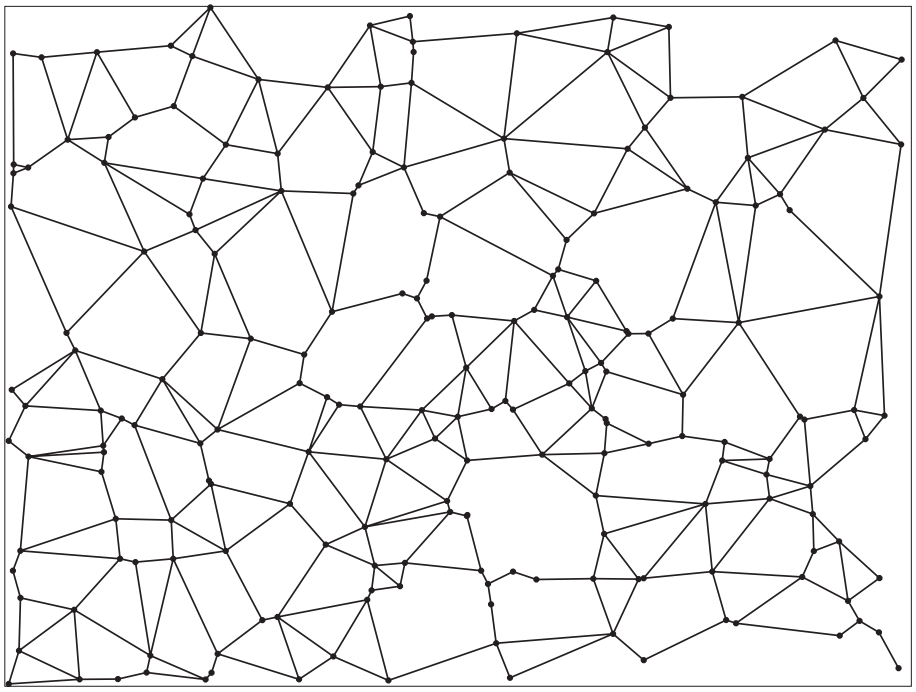


FIG. 1. *Gabriel graph for a random point set.*

3. Lower bounds. When $\beta = 0$, the β-skeleton of a point set has a spanning ratio of 1. When β is in the interval $(0, 1]$, Eppstein [6] presents an elegant fractal construction that proves a nonconstant lower bound on the spanning ratio. His result is summarized in the following theorem.

THEOREM 1 (see Eppstein [6]). *For any $n = 5^k + 1$, there exists a set of n points in the plane whose β-skeleton with $\beta \in (0, 1]$ has a spanning ratio of $\Omega(n^c)$, where $c = \log_5(5/(3 + 2\cos\theta))$ and $\theta < (2/3)\sin^{-1}\beta$.*

Our lower bounds apply to β-skeletons with $\beta \in [1, 2]$. The tower construction developed here in the proof of Theorem 2 is similar to the tower-like configuration we later use in lower bounding the spanning ratio of random Gabriel graphs.

THEOREM 2. *For any $n \geq 2$, there exists a set of n points in the plane whose β-skeleton with $\beta \in [1, 2]$ has a spanning ratio of*

$$S \geq \left(\frac{1}{2} - o(1)\right)\sqrt{n}.$$

Note that the closed 1-skeleton is the Gabriel graph and that all β-skeletons with

$\beta > 1$ are subgraphs of the Gabriel graph. Therefore, it suffices to prove the theorem for the Gabriel graph. Also, the $1/2 - o(1)$ factor can be improved to $2/3$.

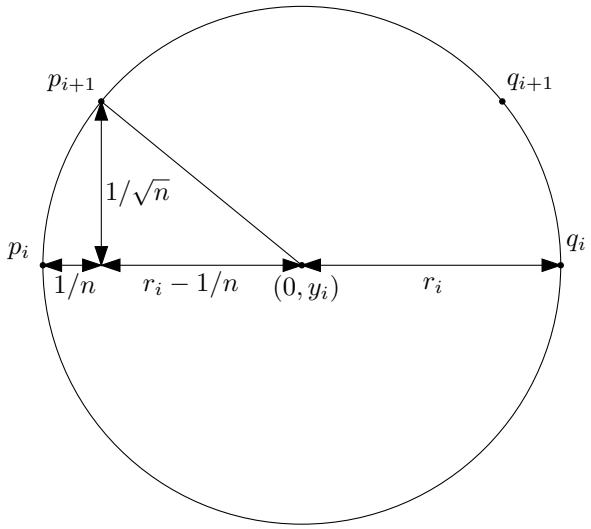


FIG. 2. *Illustration of one level in the Gabriel graph tower construction.*

Proof. Let $m = \lfloor n/2 \rfloor$. Place points p_i and q_i at locations $(-r_i, y_i)$ and (r_i, y_i), respectively, $(1 \le i \le m)$, where

$$r_i = 1 - (i-1)/n,$$
$$y_i = (i-1)/\sqrt{n}.$$

If n is odd place the remaining point at the same location as p_1.

We claim that for each pair p_i, q_i, the circle with diameter $p_i q_i$ contains the points p_{i+1} and q_{i+1} $(1 \le i \le m-1)$. Let d be the distance from the center of the circle with diameter $p_i q_i$ to the point p_{i+1}. For p_{i+1} to lie within this circle, d must be at most r_i. By construction,

$$d = \sqrt{(r_i - 1/n)^2 + 1/n}.$$

Thus we require $(r_i - 1/n)^2 + 1/n \le r_i^2$ or, equivalently, $r_i \ge 1/2 + 1/(2n)$, which holds for $1 \le i \le m-1$.

It follows that when $i \le j$, edge $p_i q_j$ does not belong to the Gabriel graph of these points (unless $i = j = m$), since p_{i+1} lies in or on the circle with diameter $p_i q_j$. Similarly, when $i > j$, edge $p_i q_j$ is precluded by point q_{j+1}.

The Euclidean distance between p_1 and q_1 is two. However, the shortest path from p_1 to q_1 using Gabriel graph edges is at least $2y_m$, which results in a spanning ratio of

$$S = y_m = (\lfloor n/2 \rfloor - 1)/\sqrt{n} = \left(\frac{1}{2} - o(1)\right)\sqrt{n}. \qquad \square$$

Note that for Gabriel graphs ($\beta = 1$), Eppstein's result (Theorem 1) implies a ratio of $\Omega(n^c)$ with $0.138 < c < 0.139$, while Theorem 2 provides a much stronger bound of $\Omega(\sqrt{n})$.

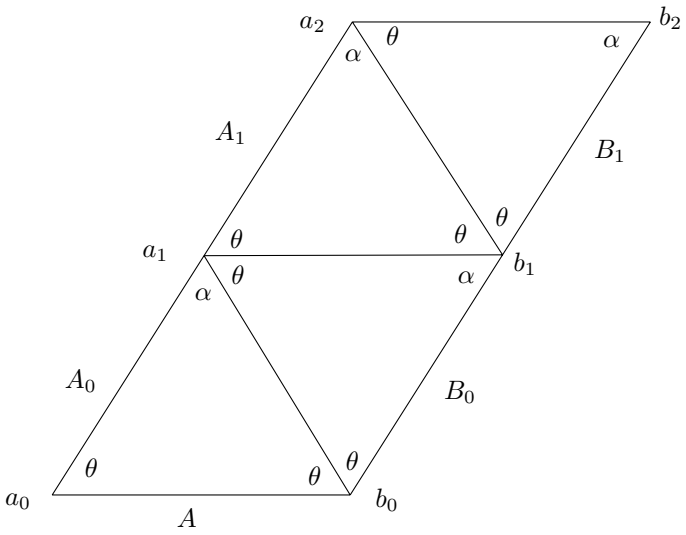


FIG. 3. *Relative neighborhood graph tower.*

For relative neighborhood graphs ($\beta = 2$), the lower bound is $\Omega(n)$.

THEOREM 3. *For any $n \geq 2$, there exists a set of n points in the plane whose relative neighborhood graph (open 2-skeleton) has a spanning ratio of $\Omega(n)$.*

Proof. Refer to Figure 3. Let $\theta = 60 - \epsilon$ and $\alpha = 60 + 2\epsilon$. We will fix ϵ later. Since $\alpha + 2\theta = \pi$, the points $a_0, a_1, \ldots, a_n$ are colinear. Similarly, the points $b_0, b_1, \ldots, b_n$ are colinear. The point a_{i+1} blocks the edge $a_i b_i$. An edge $a_i b_j$ for $i < j$ is blocked by a_{i+1} and an edge $a_i b_j$ for $i > j$ is blocked by b_{i+1}. Thus, the only edges in the relative neighborhood graph of these points are $a_i a_{i+1}$, $b_i b_{i+1}$, and $a_n b_n$. Let $A_i = \|a_{i+1} - a_i\|$. Let $B_i = \|b_{i+1} - b_i\|$.

Triangle(a_0, a_1, b_0) and triangle(a_1, b_1, b_0) are similar; therefore, $B_0 = A_0^2/A$. By the same argument, $A_1 = A_0^3/A^2$ and $B_1 = A_0^4/A^3$. In general, $A_i = A_0^{2i+1}/A^{2i}$ and $B_i = A_0^{2i+2}/A^{2i+1}$.

We choose an ϵ so that $A_0/A > (1/2)^{1/(2n)}$. Let L be the length of the path from a_0 to b_0. $L > \sum_{i=0}^{n-1} A_i + B_i = \sum_{i=0}^{2n-1} A_0 (A_0/A)^i$. Since $A_0/A > (1/2)^{1/(2n)}$, we have that $\sum_{i=0}^{2n-1} A_0 (A_0/A)^i > 1/2 \sum_{i=0}^{2n-1} A_0 = A_0 n$. Therefore, $L > A_0 n$. ∎

4. Upper bounds. We start with a straightforward upper bound that applies to all β-skeletons for $\beta \in [0, 2]$.

THEOREM 4. *For any $\beta \in [0, 2]$, the spanning ratio of the β-skeleton of a set of n points is at most $n - 1$.*

Proof. Let G be the β-skeleton of a set of n points P. Note that the minimum spanning tree $MST(P)$ is contained in G. Every edge in the unique path from u to v in $MST(P)$ has length at most $D(u, v)$, otherwise $MST(P)$ is not minimum. Therefore the shortest path in G from u to v has length at most $(n-1)D(u, v)$. ∎

The rest of this section establishes an upper bound for β-skeletons when $\beta \in [0, 1]$. The β-skeleton of a point set P for $\beta \in [0, 1]$ is a graph in which points x and y in P are connected by an edge if and only if there is no other point $v \in P$ such that $\angle xvy > \pi - \sin^{-1} \beta$.

To upper bound the spanning ratio of β-skeletons, we show that there exists a special walk $SW_\beta(x, y)$ in the β-skeleton between the endpoints of any Delaunay

edge xy. We upper bound the length $|SW_\beta(x,y)|$ of $SW_\beta(x,y)$ as a multiple of $D(x,y)$. We then combine this with an upper bound on the spanning ratio of Delaunay triangulations [10, 11] to obtain our result.

Let $DT(P)$ be the Delaunay triangulation of a points set P. In order to describe the walk between the endpoints of a Delaunay edge, we define the *peak* of a Delaunay edge.

LEMMA 2. *Let xy be an edge of $DT(P)$. For $\beta \in [0,1]$, either xy is an edge of the β-skeleton of P or there exists a unique z (called the peak of xy) such that triangle(xyz) is in $DT(P)$ and z lies in the β-region of xy.*

Proof. Suppose $xy \in DT(P)$ is not an edge in the β-skeleton of P. Then there exists a point $v \in P$ such that $\angle xvy > \pi - \sin^{-1}\beta$. Since xy is an edge of $DT(P)$, there exists a unique z on the same side of xy as v such that $\mathrm{disc}(xyz)$ is empty. This implies $\angle xzy \geq \angle xvy$ and thus z lies in the β-region of xy. Since $\beta \leq 1$, $\mathrm{disc}(xyz)$ contains that part of the β-region of xy which lies on the other side of xy from z. Since this circle is empty, z is unique. ∎

We now define the walk $SW_\beta(x,y)$ between the endpoints of the Delaunay edge xy. (Note that in a walk edges may be repeated; see Bondy and Murty for details [1].)

$$SW_\beta(x,y) = \begin{cases} xy & \text{if } xy \in \beta\text{-skeleton of } P, \\ SW_\beta(x,z) \cup SW_\beta(z,y) & \text{otherwise } (z \text{ is the peak of } xy). \end{cases}$$

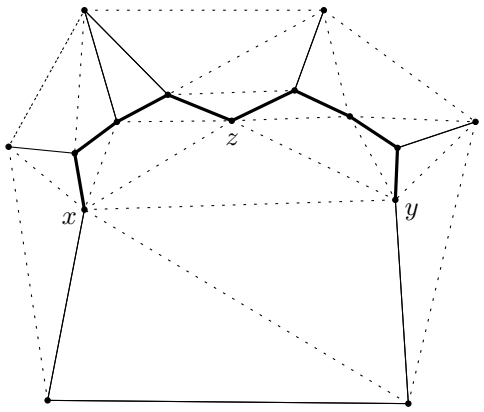


FIG. 4. *The solid lines form the Gabriel graph of the point set with $SW_1(x,y)$ in bold. All edges together form the Delaunay triangulation.*

LEMMA 3. *Given a set P of n points in the plane. If $xy \in DT(P)$ then the number of edges in $SW_\beta(x,y)$ is at most $6n-12$, for $\beta \in [0,1]$.*

Proof. Since a Delaunay edge is adjacent to at most two Delaunay triangles, an edge can occur at most twice in the walk $SW_\beta(x,y)$. Since there are at most $3n-6$ edges in $DT(P)$ by Euler's formula, $SW_\beta(x,y)$ can consist of at most $6n-12$ edges. ∎

LEMMA 4. *Let P be a set of n points in the plane. For any $\beta \in [0,1]$, for all $x,y \in P$, if $xy \in DT(P)$, then*

$$|SW_\beta(x,y)| \leq m^\gamma D(x,y),$$

where $\gamma = (1 - \log_2(1+\sqrt{1-\beta^2}))/2$ and m is the number of edges in $SW_\beta(x,y)$.

Proof.[1] The proof is by induction on the number of edges m in $SW_\beta(x, y)$. When $m = 1$, i.e., $SW_\beta(x, y)$ is simply the line segment from x to y, the lemma clearly holds.

If $m > 1$, then $|SW_\beta(x, y)| = |SW_\beta(x, z)| + |SW_\beta(z, y)|$ for z the peak of xy. Let k be the number of edges in $SW_\beta(x, z)$. Thus, $m - k$ is the number of edges in $SW_\beta(z, y)$. Let $a = D(x, y)$, $b = D(x, z)$, and $c = D(y, z)$. Since xz and zy are Delaunay edges, by induction, $|SW_\beta(x, z)| \le bk^\gamma$ and $|SW_\beta(z, y)| \le c(m - k)^\gamma$. Thus it suffices to prove that

$$bk^\gamma + c(m - k)^\gamma \le am^\gamma.$$

By the law of cosines, $a^2 = b^2 + c^2 - 2bc\cos A$, where A is the angle at the peak z. With this substitution for a, after dividing both sides by c and letting $\delta = b/c$, it remains to show

$$\delta k^\gamma + (m - k)^\gamma \le m^\gamma \le \sqrt{1 + \delta^2 - 2\delta\cos A},$$

where we may assume without loss of generality that $\delta \in [0, 1]$. As a function of k the left-hand side of the equation is maximized when $k = m/(1 + \delta^{-s})$, where $s = 1/(1 - \gamma)$. With this substitution for k, after factoring m^γ, it suffices to show

$$\frac{\delta + \delta^{-\gamma s}}{(1 + \delta^{-s})^\gamma} \le \sqrt{1 + \delta^2 - 2\delta\cos A} \qquad \text{when } \delta \in [0, 1].$$

We can simplify the left-hand side using the fact that $s = 1/(1 - \gamma)$:

$$\frac{\delta + \delta^{-\gamma s}}{(1 + \delta^{-s})^\gamma} = \frac{\delta(1 + \delta^{-s})}{(1 + \delta^{-s})^\gamma} = \delta(1 + \delta^{-s})^{1-\gamma} = (\delta^s + 1)^{1-\gamma}.$$

Thus, after squaring both sides of the inequality, it suffices to show

$$(1 + \delta^s)^{2/s} \le 1 + \delta^2 - 2\delta\cos A \qquad \text{when } \delta \in [0, 1].$$

The angle A is minimized (thus minimizing the right-hand side of the inequality) when z lies on the boundary of the β-region. For such z, $\cos A = -\sqrt{1 - \beta^2}$, and it remains to show

$$(1 + \delta^s)^{2/s} \le 1 + \delta^2 + 2\delta\sqrt{1 - \beta^2} \qquad \text{when } \delta \in [0, 1].$$

Let $L(\delta)$ be the left-hand side and $R(\delta)$ the right-hand side of this inequality. We want to show that $L(\delta) \le R(\delta)$ when $\delta \in [0, 1]$. The maximum of $L(\delta) - R(\delta)$ (for $\delta \in [0, 1]$) occurs at $\delta = 0$ or $\delta = 1$ or at some value δ with $L'(\delta) = R'(\delta)$. At $\delta = 0$, $L(0) = R(0) = 1$. At $\delta = 1$,

$$L(1) = 2^{2/s} \qquad \text{and} \qquad R(1) = 2 + 2\sqrt{1 - \beta^2}.$$

Since $\gamma = (1 - \log_2(1 + \sqrt{1 - \beta^2}))/2$, $s = 2/(1 + \log_2(1 + \sqrt{1 - \beta^2}))$, and $L(1) = R(1)$. The derivatives of $L(\delta)$ and $R(\delta)$ are

$$L'(\delta) = 2\delta^{s-1}(1 + \delta^s)^{2/s-1} \qquad \text{and} \qquad R'(\delta) = 2\delta + 2\sqrt{1 - \beta^2}.$$

[1] Thanks to Ansgar Grüne and Sébastien Lorenz at the University of Bonn for pointing out a flaw in an earlier proof.

For $\beta \in [0,1]$, $L'(0) \leq R'(0)$, and for our chosen value of γ, $L'(1) = R'(1)$. For $\beta \in [0,1]$, γ lies in $[0,1/2]$, which implies $s \in [1,2]$. Thus,

$$L'''(\delta) = 2(1+\delta^s)^{2/s-3}\delta^{s-3}(s-1)(s-2)(1-\delta^s) \leq 0$$

and the function $L'(\delta)$ is concave. Since $R'(\delta)$ is linear and $L'(1) = R'(1)$, there is at most one value of $\delta \in (0,1)$, where $L'(\delta) = R'(\delta)$. Since $L'(0) \leq R'(0)$, $L(\delta) - R(\delta)$ is a minimum at this value. Thus the maximum of $L(\delta) - R(\delta)$ is 0 for $\gamma = (1 - \log_2(1 + \sqrt{1-\beta^2}))/2$. □

THEOREM 5. *For $\beta \in [0,1]$, the spanning ratio of the β-skeleton of a set P of n points in the plane is at most*

$$\frac{4\pi(6n-12)^\gamma}{3\sqrt{3}},$$

where $\gamma = (1 - \log_2(1 + \sqrt{1-\beta^2}))/2$.

Proof. Given two arbitrary points x, y in P, let $M = e_1, e_2, \ldots, e_j$ represent the shortest path between x and y in $DT(P)$. Keil and Gutwin [10, 11] have shown that the length of P is at most $2\pi/(3\cos(\pi/6))$ times $D(x,y)$.

For each edge e_i in M, by Lemmas 3 and 4, we know there exists a path in the β-skeleton whose length is at most $(6n-12)^\gamma$ times the length of e_i. Therefore, the shortest path between x and y in the β-skeleton has length at most $2\pi(6n-12)^\gamma/(3\cos(\pi/6))$ times $D(x,y)$. The theorem follows. □

COROLLARY 1. *The spanning ratio of the Gabriel graph ($\beta = 1$) of an n-point set is at most*

$$\frac{4\pi}{3}\sqrt{2n-4}.$$

When β lies strictly between 0 and 1, there is a gap between the upper bound and lower bound on the spanning ratio of β-skeletons. As noted in section 3, the spanning ratio is at least $\Omega(n^c)$, where $c = \log_5(5/(3+2\cos\theta))$ and $\theta < (2/3)\sin^{-1}\beta$. We have shown here that the spanning ratio is at most $O(n^\gamma)$, where $\gamma = (1 - \log_2(1 + \sqrt{1-\beta^2}))/2$; refer to Figure 5 for a graph of the exponents of the upper and lower bound. For Gabriel graphs ($\beta = 1$), the lower bound construction given in section 3, together with the upper bound given here, show that the spanning ratio is indeed $\Theta(\sqrt{n})$.

5. Random Gabriel graphs. If n points are drawn uniformly and at random from the unit square $[0,1]^2$, the spanning ratio of the induced Gabriel graph grows unbounded in probability. In particular, we have the following theorem.

THEOREM 6. *If n points are drawn uniformly and at random from the unit square $[0,1]^2$, and S is the spanning ratio of the induced Gabriel graph, then*

$$\mathbf{P}\left\{S < c\sqrt{\frac{a\log n}{\log\log n}}\right\} \leq 2e^{-2n^{1-12a-o(1)}}$$

for constants c and $a < 1/12$. Thus, for $a < 1/12$, with probability tending exponentially quickly to one,

$$S \geq c\sqrt{a\log n/\log\log n}.$$

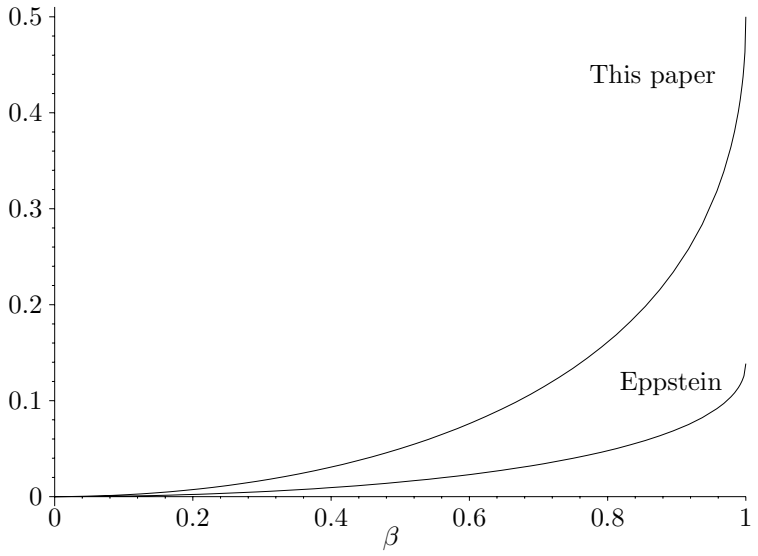


FIG. 5. *Exponents of n in the upper and lower bound on the spanning ratio of β-skeletons when* $\beta \in [0, 1]$.

Proof. The main idea is to show that a set of n points randomly distributed in the unit square contains many tower-like structures of size $c \log n / \log \log n$, each of which has spanning ratio approximately the square root of its size. We first define what a tower-like structure is and then show that the expected number of such structures is large.

A tower-like structure resembles the towers of section 3 but the points may be slightly perturbed. For $i = 1, ..., k$, let A_i and B_i be discs both of radius d/k (the constant d will be specified later) located at (r_i, y_i) and $(-r_i, y_i)$, respectively, where the sequences r_i and y_i are given below,

$$r_i = 1 - \frac{i-1}{2k},$$
$$y_i = (i-1)\sqrt{\frac{1/2 - (1+\sqrt{2})d}{k}\left(1 - \frac{1/2 - (1+\sqrt{2})d}{k}\right)}.$$

The value of d is chosen so that y_i is positive ($d < 1/(2 + 2\sqrt{2})$).

Let C be the smallest square enclosing the A_i and B_i within a border of width y_k. Typically, when k is large enough and the tower is taller than it is wide, C extends from $(-3y_k/2 - d/k, -y_k - d/k)$ to $(3y_k/2 + d/k, 2y_k + d/k)$; see Figure 7 for an example of such a square, and note that in this figure the discs A_i and B_i would be smaller than the dots used to represent points.

Assume that each of the A_i and B_i contain exactly one point and C contains no other data point beyond these $2k$ points. We claim that among the points in C, the only edges are those connecting A_1 with A_2, A_2 with A_3, and so forth, up to A_{k-1} and A_k. Then A_k connects with B_k, B_k with B_{k-1} and so forth down to B_1. The proof of this claim is rather technical and is deferred to the appendix. Note that the A_i's and B_i's are disjoint.

Let u and v be the points in A_1 and B_1, respectively. We have $D(u, v) \leq 2 + 2d/k$.

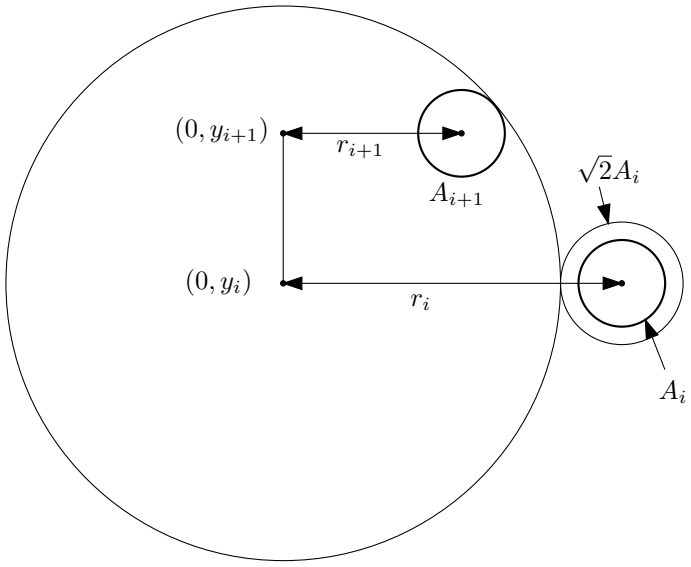


FIG. 6. *The construction of A_i and A_{i+1}.*

Also, any path from u to v entirely in C must be equal in length to the chain, which is longer than $2y_k$. If the path leaves C, then at least two edges leave C, and those edges have a length of at least $2y_k$, taken together. Thus, $L(u, v) \geq 2y_k$ and

$$S \geq \frac{L(u,v)}{D(u,v)} \geq \frac{y_k}{1 + d/k} \geq c\sqrt{k}$$

for sufficiently large k where c is a constant that depends on d.

Let bC denote the scaled down set $\{bx : x \in C\}$.

Divide $[0, 1]^2$ into n nonoverlapping *tiles* of size $1/\sqrt{n} \times 1/\sqrt{n}$. For $b = 1/(4\sqrt{kn})$, bC fits within one of these tiles. Thus we may place n nonoverlapping copies of bC within the unit square. For a given data set, we call a tile *tower-like* if it contains exactly $2k$ data points, one each for bA_i and bB_i, $1 \leq i \leq k$ within it. Let N be the number of tiles that are tower-like.

Clearly, since the distribution is uniform,

$$\mathbf{E}N = n\mathbf{P}\{\text{a tile is tower-like}\}.$$

Pick one tile and partition the n data points over the following disjoint sets: the bA_i's, the bB_i's, $bC - \cup bA_i \cup bB_i$, and $[0, 1]^2 - bC$. The cardinalities of these sets, taken together, form a multinomial random vector with probabilities given by the areas of the sets involved. For example, area $(bA_i) = b^2\pi d^2/k^2$. According to the formula for

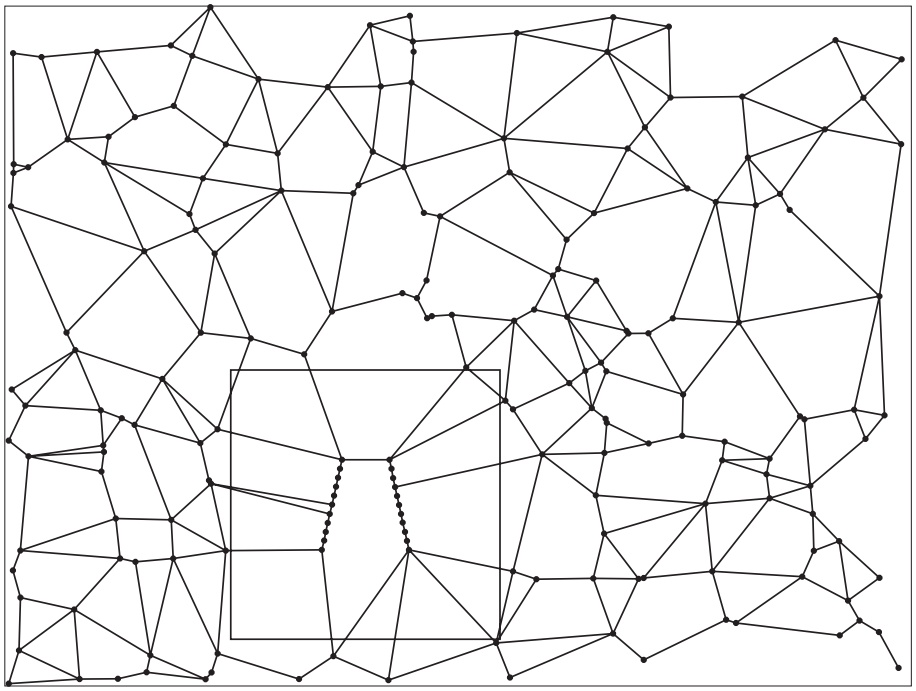


FIG. 7. *Gabriel graph with tower-like square.*

the multinomial distribution,

$$\begin{aligned}\mathbf{P}\{\text{a tile is tower-like}\} &= \frac{n!}{(n-2k)!}\left(\frac{b^2\pi d^2}{k^2}\right)^{2k}(1-1/n)^{n-2k}\\ &\geq (n-2k+1)^{2k}\left(\frac{\pi d^2}{16nk^3}\right)^{2k}(1-1/n)^{n}\\ &\geq \frac{1}{4}\left(\frac{(n-2k+1)\pi d^2}{16nk^3}\right)^{2k}\\ &\geq \frac{1}{4}\left(\frac{\pi d^2}{32k^3}\right)^{2k}\end{aligned}$$

provided that n is sufficiently large and $k < (n+2)/4$. We conclude that

$$\mathbf{E}N \geq \frac{n}{4}\left(\frac{\pi d^2}{32k^3}\right)^{2k}.$$

If $k = a\log n/\log\log n$ for a constant $a < 1/6$, then

$$\mathbf{E}N \geq n^{1-6a-o(1)} \to \infty.$$

For each one of these tower-like squares, there is a pair of data points for which the spanning ratio is at least

$$c\sqrt{k} \geq c\sqrt{\frac{a\log n}{\log\log n}}.$$

TABLE 1
Summary table of results on the spanning ratio of β-skeletons.

	$\beta = 0$	$0 < \beta < 1$	$\beta = 1$	$1 < \beta < 2$	$\beta = 2$	$\beta > 2$
Lower bound	1	$\Omega(n^c)$[6]	$\Omega(\sqrt{n})$	$\Omega(\sqrt{n})$	$\Omega(n)$	∞
Upper bound	1	$O(n^\gamma)$	$O(\sqrt{n})$	$O(n)$	$O(n)$	∞

$c = \log_5(5/(3 + 2\cos\theta))$ and $\theta < (2/3)\sin^{-1}\beta$.
$\gamma = (1 - \log_2(1 + \sqrt{1 - \beta^2}))/2$.

Change one of the n data points. That will change the number N by at most one. But then, by McDiarmid's inequality [14], we have

$$\mathbf{P}\{|N - \mathbf{E}N| \geq t\} \leq 2e^{-2t^2/n}.$$

In particular, for fixed $\epsilon > 0$,

$$\mathbf{P}\{|N - \mathbf{E}N| \geq \epsilon \mathbf{E}N\} \leq 2e^{-2\epsilon^2 n^{1-12a-o(1)}} \to 0$$

when $a < 1/12$. This shows that $N/\mathbf{E}N \to 1$ in probability for such a choice of a (and thus k), and thus that for every $\epsilon > 0$,

$$\mathbf{P}\{N < (1 - \epsilon)\mathbf{E}N\} \to 0.$$

As another application, we have

$$\begin{aligned}
\mathbf{P}\{S < c\sqrt{a \log n / \log\log n}\} &\leq \mathbf{P}\{N = 0\} \\
&= \mathbf{P}\{N - \mathbf{E}N \leq -\mathbf{E}N\} \\
&\leq 2e^{-2n^{1-12a-o(1)}} \\
&\to 0.
\end{aligned}$$

Note that this probability decreases exponentially quickly with n. □

We have implicitly shown several other properties of random Gabriel graphs. For example, a Gabriel graph partitions the plane into a finite number of polygonal regions. The outside polygon which extends to ∞ is excluded. Let D_n be the maximum number of vertices in these polygons. Then $D_n \to \infty$ in probability, because D_n is larger than the maximum size of any tower that occurs in the point set, and this was shown to diverge in probability. From what transpired above, this is bounded from below in probability by $\Omega(a \log n / \log\log n)$.

6. Conclusion. We studied the spanning ratio of β-skeletons with β ranging from 0 to 2. This class of proximity graphs includes the Gabriel graph and the relative neighborhood graph. Table 1 summerizes our results. For $\beta > 2$, β-skeletons lose connectivity; thus, their spanning ratio leaps to infinity. For points drawn independently from the uniform distribution on the unit square, we showed that the spanning ratio of the (random) Gabriel graph (and all β-skeletons with $\beta \in [1, 2]$) tends to ∞ in probability as $\sqrt{\log n / \log\log n}$.

Several open problems arise from this investigation. It would be interesting to close the gap between upper and lower bounds for β-skeletons in the ranges $0 < \beta < 1$ and $1 < \beta < 2$. Also, for random point sets, it would be interesting to try to find a matching upper bound for the spanning ratio.

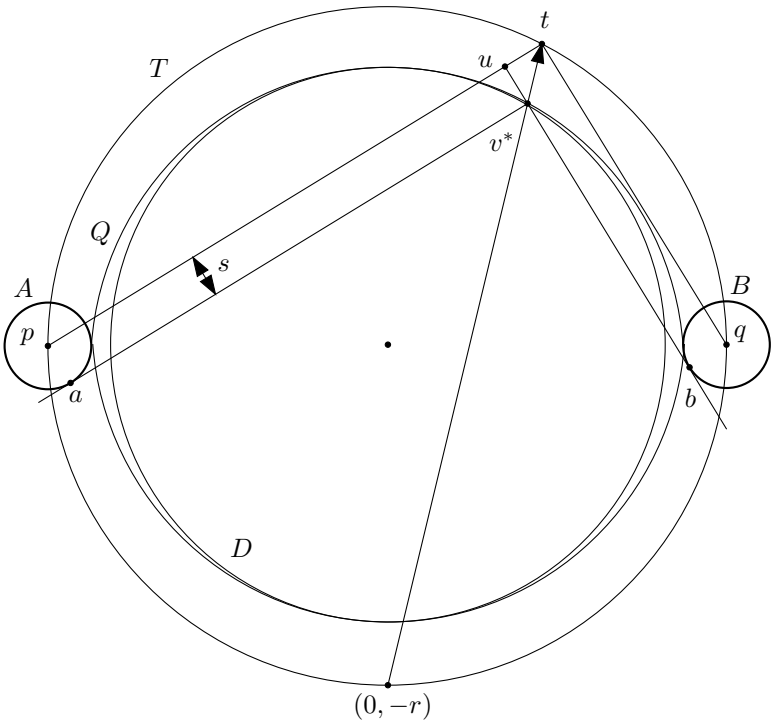


FIG. 8. *If the region Q contains a point, then for any $a \in A$ and $b \in B$ the edge ab is not in the Gabriel graph of the point set. We insure that there is a point in the disc $D \subset Q$.*

Appendix. Tower-like construction. The purpose of this section is to show that the only Gabriel edges among the points in the tower-like construction described in section 5 are between A_i and A_{i+1}, B_i and B_{i+1}, and A_k and B_k for $i = 1, \ldots, k-1$. Recall that the tower-like construction consists of $2k$ points, one in each of the discs A_i, B_i for $i = 1, \ldots, k$, where A_i and B_i are discs of radius d/k centered at (y_i, r_i) and $(y_i, -r_i)$, respectively. The definition of the sequences r_i and y_i is repeated here:

$$r_i = 1 - \frac{i-1}{2k},$$

$$y_i = (i-1)\sqrt{\frac{1/2 - (1+\sqrt{2})d}{k}\left(1 - \frac{1/2 - (1+\sqrt{2})d}{k}\right)}.$$

In a Gabriel graph, two points are connected by an edge if and only if the disc whose diameter is the segment joining those points is empty. In our construction, we do not have precise information as to the location of the points. We only know that a point lies within a small disc (whose location we do know). Thus a basic problem is, given two discs, what is the region that, if it contains a point, will forbid an edge between a point in one disc and a point in the other. After we have determined this region, we must show for any two discs in our construction between which we claim no edge exists, that there is a third disc contained within that pair's region.

Let A and B be two discs each of radius s, whose centers are at $p = (r, 0)$ and $q = (-r, 0)$. The region Q we are interested in is the intersection of all discs whose diameter has one endpoint a in A and the other endpoint b in B; see Figure 8.

We will determine the upper boundary of Q (the points with positive y coordinate). The lower boundary is symmetric. Consider a ray with origin $(0, -r)$ that intersects the segment pq. A point v on this ray is inside Q if and only if for all points $a \in A$ and $b \in B$, $\angle avb \geq \pi/2$. For each point v, the points $a \in A$ and $b \in B$ that minimize $\angle avb$ are the tangent points of the lines through v tangent to the A and B, respectively. (Strictly speaking, there are two tangent lines from a point to a disc, and a and b are defined by those tangent lines which form the minimum of the resulting four possible angles.) For v with positive y coordinate, this minimum angle is a continuous, decreasing function of the distance between v and $(0, -r)$. Thus the upper boundary of Q intersects the ray at a single point v^*, where $\min_{a \in A, b \in B} \angle av^*b = \pi/2$.

Let T be the circle whose diameter is the segment pq, and let t be the point (other than $(0, -r)$), where our chosen ray intersects T. We claim that v^* is the point on the ray that is distance $s\sqrt{2}$ from t. To show this, consider the lines from t to p and q. These lines are parallel to the tangent lines from v^* to a and b, respectively, where $a \in A$ and $b \in B$ minimize $\angle av^*b$. In order to establish that $\overline{v^*a}$ is parallel to $\overline{tp}$, drop a line perpendicular to $\overline{tp}$ from v^* to a point u on $\overline{tp}$. Since $\angle p, t, (0, -r) = \pi/4$, the triangle $\triangle tuv^*$ is a right, isosceles triangle. Its hypotenuse has length $s\sqrt{2}$ so its sides have length s. Thus $\overline{v^*a}$ is consistently distance s from $\overline{tp}$. The same argument applies to $\overline{v^*b}$ and $\overline{tq}$. Since $\angle ptq = \pi/2$, the claim is established. It is perhaps surprising that the region Q does not touch A or B.

For our tower-like construction we use a disc to approximate the region Q. The disc D centered at $(0, 0)$ with radius $r - s\sqrt{2}$ is contained within Q. (Note: The point v^* appears to lie on the boundary of D in Figure 8. This is misleading. The point v^* does not lie on the boundary of D except for v^* with x-coordinate equal to 0.) Thus if a point lies within D, there is no Gabriel edge between any two points $a \in A$ and $b \in B$. The tower-like construction insures that this is the case for any pair of discs $A = A_i$ and $B = B_i$, by placing the discs A_{i+1} and B_{i+1} within the disc D. Also the disc D for any pair $A = A_i$ and $B = B_j$ with $i \neq j$ contains either A_{i+1} if $i < j$ or B_{i+1} if $i > j$. Finally, for $A = A_i$ and $B = A_j$ with $i < j - 1$, the disc D contains A_k, where $i < k < j$. (This holds for B_i discs by symmetry.)

It remains to show that the remaining edges in the tower-like construction do exist. A similar argument to the one presented above establishes that the union of the discs with diameter ab with $a \in A$ and $b \in B$ is a region contained in the disc $\hat{D}$ of radius $r + s\sqrt{2}$ centered at the origin. This region is empty for each pair $A = A_i$ and $B = A_{i+1}$ since $r_i \geq 1/2$ while the distance between the centers of A and B is $O(1/\sqrt{k})$.

Acknowledgments. The authors would like to thank Ansgar Grüne and Sébastien Lorenz at the University of Bonn for pointing out a flaw in an earlier proof. We also thank the referees whose comments helped improve the presentation of the paper.

REFERENCES

[1] J. A. Bondy and U. S. R. Murty, *Graph theory with applications*, Elsevier, New York, 1976.
[2] L. P. Chew, *There is a planar graph almost as good as the complete graph*, in Proceedings of the 2nd Annual ACM Symposium on Computational Geometry, New York, 1986, pp. 169–177.
[3] L. P. Chew, *There are planar graphs almost as good as the complete graph*, J. Comput. System Sci., 39 (1989), pp. 205–219.
[4] L. Devroye, *The expected size of some graphs in computational geometry*, Comput. Math. Appl., 15 (1988), pp. 53–64.
[5] D. P. Dobkin, S. J. Friedman, and K. J. Supowit, *Delaunay graphs are almost as good as complete graphs*, in Proceedings of the 28th Annual Symposium on the Foundations

of Computer Science, Los Angeles, 1987, pp. 20–26. Also in Discrete Comput. Geom., 5 (1990), pp. 399–407.

[6] D. Eppstein, *Beta-skeletons have unbounded dilation*, Comput. Geom., 23 (2002), pp. 43–52.

[7] D. Eppstein, *Spanning trees and spanners*, Handbook of Computational Geometry, North–Holland, Amsterdam, 2000, pp. 425–461.

[8] K. R. Gabriel and R. R. Sokal, *A new statistical approach to geographic variation analysis*, Systematic Zoology, 18 (1969), pp. 259–278.

[9] J. W. Jaromczyk and G. T. Toussaint, *Relative neighborhood graphs and their relatives*, in Proceedings of the IEEE, 80 (1992), pp. 1502–1517.

[10] J. M. Keil and C. A. Gutwin, *The Delaunay triangulation closely approximates the complete Euclidean graph*, in Proceedings of the 1st Workshop Algorithms Data Struct., Ottawa, Canada, Lecture Notes in Computer Science 382, Springer-Verlag, 1989, pp. 47–56.

[11] J. M. Keil and C. A. Gutwin, *Classes of graphs which approximate the complete Euclidean graph*, Discrete Comput. Geom., 7 (1992), pp. 13–28.

[12] D. G. Kirkpatrick and J. D. Radke, *A framework for computational morphology*, Comput. Geom., G. T. Toussaint, ed., Elsevier, Amsterdam, 1985, pp. 217–248.

[13] D. W. Matula and R. R. Sokal, *Properties of Gabriel graphs relevant to geographic variation research and the clustering of points in the plane*, Geograph. Anal., 12 (1980), pp. 205–222.

[14] C. McDiarmid, *On the method of bounded differences*, in Surveys in Combinatorics, London Math. Soc. Lecture Note Ser. 141, Cambridge University Press, Cambridge, UK, 1989, pp. 148–188.

[15] F. P. Preparata and M. I. Shamos, *Computational geometry. An introduction*, Springer-Verlag, New York, 1985.

[16] G. T. Toussaint, *The relative neighborhood graph of a finite planar set*, Pattern Recognition, 12 (1980), pp. 261–268.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 428–443

FULL COLOR THEOREMS FOR $L(2,1)$-COLORINGS*

PETER C. FISHBURN† AND FRED S. ROBERTS‡

Abstract. The span $\lambda(G)$ of a graph G is the smallest k for which G's vertices can be $L(2,1)$-colored, i.e., colored with integers in $\{0,1,\ldots,k\}$ so that adjacent vertices' colors differ by at least 2, and colors of vertices at distance two differ. G is full-colorable if some such coloring uses all colors in $\{0,1,\ldots,\lambda(G)\}$ and no others. We prove that all trees except stars are full-colorable. The connected graph G with the smallest number of vertices exceeding $\lambda(G)$ which is not full-colorable is C_6. We describe an array of other connected graphs that are not full-colorable and go into detail on full-colorability of graphs of maximum degree four or less.

Key words. distance-two colorings, no-hole colorings, channel assignment problems

AMS subject classifications. 05C78, 05C15, 94C15

DOI. 10.1137/S0895480100378562

1. Introduction. Ordinary colorings of graphs that assign different integers to adjacent vertices have the property that the fewest colors needed can always lie in an interval all of whose integers are assigned to vertices. The present paper considers special integer colorings called $L(2,1)$-colorings, whose shortest interval containing all colors may necessarily have integers assigned to no vertex. The paper identifies a variety of graphs with this feature, along with conditions which imply that every color in a shortest interval can be assigned to some vertex. Our topic is related to so-called no-hole colorings, which are $L(2,1)$-colorings that use every integer in an interval of colors but do not presume that the interval is the shortest possible; see, for example, [3, 4, 7].

$L(2,1)$-colorings assign integers to vertices so that adjacent vertices' colors differ by at least 2 and so that the colors of any two nonadjacent vertices that are adjacent to a third vertex differ by at least 1. $L(2,1)$-colorings were first investigated extensively by Yeh [10] and Griggs and Yeh [5] as a generalization of T-colorings for the channel assignment problem, which were introduced by Hale [6] (see also Roberts [9]). By now, there are over 50 papers on the subject of $L(2,1)$-colorings. For a recent compilation of references, see [1]. K_2 is the smallest graph with the feature that every interval which contains the integers of an $L(2,1)$-coloring also contains an integer assigned to no vertex.

The rest of this introduction specifies our notation and assumptions, notes prior results for $L(2,1)$-colorings, and outlines our new results.

Let $\mathcal{G}^*$ be the set of finite simple graphs. An $L(2,1)$-*coloring* of $G=(V,E)$ in $\mathcal{G}^*$ is a vertex coloring, $f: V \to \mathbb{Z}$ for which

$$\begin{array}{ll} |f(u)-f(v)| \geq 2 & \text{for all } \{u,v\} \in E, \\ |f(u)-f(v)| \geq 1 & \text{whenever } u \neq v,\ \{u,v\} \notin E, \text{ and } \{u,t\},\ \{t,v\} \in E \\ & \text{for some } t \in V. \end{array}$$

*Received by the editors September 25, 2000; accepted for publication (in revised form) September 1, 2005; published electronically May 12, 2006.
http://www.siam.org/journals/sidma/20-2/37856.html
†AT&T Labs-Research, Florham Park, NJ 07932 (fish@research.att.com).
‡DIMACS, Rutgers University, Piscataway, NJ 08854 (froberts@dimacs.rutgers.edu). This author thanks the National Science Foundation for its support under grants NSF-SBR-9709134 and NSF-INT-0140431.

The *span* $\lambda = \lambda(G)$ of $G \in \mathcal{G}^*$ is the smallest $k \geq 2$ for which G has an $L(2,1)$-coloring $f : V \to \{0, 1, \ldots, k\}$. A *span coloring* is an $L(2,1)$-coloring $f : V \to \{0, 1, \ldots, \lambda\}$. We say that G is *full-colorable*, or *FC* for short, if some span coloring has $f(V) = \{0, 1, \ldots, \lambda\}$, i.e., so that all colors in $\{0, 1, \ldots, \lambda\}$ are used by f, and call such an f a *full coloring*. On the other hand, G is *NFC* if it is not full-colorable.

$L(2,1)$-colorings into $\{0, 1, \ldots, \lambda\}$ have natural duals. If f is an $L(2,1)$-coloring, its *dual* g is defined by $g(x) = \lambda - f(x)$. Since span colorings must use color 0, g is a span coloring if and only if f is, and g is full if and only if f is.

Henceforth, f denotes a span coloring, and all colorings considered are span colorings. As usual, P_n, C_n, K_n, and $K_{n,m}$ denote the n-vertex path, n-vertex cycle, complete n-vertex graph, and complete bipartite graph with part sizes n and m, respectively. Every K_n, $n \geq 2$, is NFC with $\lambda = 2(n-1)$ and $f(V) = \{0, 2, 4, \ldots, 2(n-1)\}$. Moreover [5, 10],

$$\lambda(P_2) = 2,\ \lambda(P_3) = \lambda(P_4) = 3,\ \lambda(P_n) = 4 \quad \text{for} \quad n \geq 5,$$
$$\lambda(C_n) = 4 \quad \text{for all} \quad n \geq 3.$$

Although $K_2 = P_2$, K_3, and P_3 are NFC, the two-component graphs $K_2 + K_3$ and $P_2 + P_3$ are FC, as verified by $(f(K_2), f(K_3)) = (\{1, 3\}, \{0, 2, 4\})$ and $(f(P_2), f(P_3)) = (\{0, 2\}, \{0, 1, 3\})$.

Full colorability is impossible when $|V| \leq \lambda$ because $|\{0, 1, \ldots, \lambda\}| = \lambda + 1$. We will note a variety of graphs that have $|V| \leq \lambda$, but will be more concerned with NFC graphs for which $|V| \geq \lambda + 1$. Pekeč [8] has observed that "almost all" graphs have $\lambda = |V| - 1$ and have all span colorings full. That is, if we choose a random graph $G_{n,.5}$ (a graph with n vertices and edges independently chosen with probability $1/2$), then the probability that $\lambda = n-1$ and the probability that all span colorings are full both approach 1 as n approaches ∞. In fact, this is true for a random graph $G_{n,p(n)}$ as long as $np(n)^2 - 2\log n$, $n^2[1 - p(n)]$, and $n[1 - p(n)] - \log n - \log\log n$ approach ∞ as $n \to \infty$.

We confine further attention to connected graphs and will base much of our analysis on $\Delta = \Delta(G)$, the *maximum degree* of a vertex in G. Accordingly, let

$$\mathcal{G} = \{G \in \mathcal{G}^* :\ G \text{ is connected}\},$$
$$\mathcal{G}_\Delta = \{G \in \mathcal{G} :\ G \text{ has maximum degree } \Delta\},$$
$$\mathcal{G}_\Delta(\lambda) = \{G \in \mathcal{G}_\Delta :\ G \text{ has span } \lambda\} \text{ for } \lambda \geq \Delta + 1,$$

where $\lambda \geq \Delta + 1$ is presumed because $\mathcal{G}_\Delta(\lambda)$ is empty if $\lambda \leq \Delta$. Also let $\mathcal{T}$, $\mathcal{T}_\Delta$, and $\mathcal{T}_\Delta(\lambda)$ be the subsets of trees in $\mathcal{G}$, $\mathcal{G}_\Delta$, and $\mathcal{G}_\Delta(\lambda)$, respectively. The following lemma and theorem from [5] involve the smallest possible span for Δ.

LEMMA 1.1. *If $G \in \mathcal{G}_\Delta(\Delta + 1)$, then $f(u) \in \{0, \Delta + 1\}$ for every vertex u of degree Δ in G.*

THEOREM 1.2. *If $T \in \mathcal{T}_\Delta$, then $\lambda(T) \in \{\Delta + 1, \Delta + 2\}$.*

Thus, every tree with maximum degree Δ has span $\Delta + 1$ or $\Delta + 2$. Chang and Kuo [2] present a polynomial time algorithm that determines $\lambda(T)$ for $T \in \mathcal{T}$. On the other hand, Griggs and Yeh [5] prove that determination of $\lambda(G)$ for $G \in \mathcal{G}$ is NP-complete.

Henceforth let $n = |V|$. An outline of our results for $n \geq \lambda + 1$ follows. We will also identify graphs in $\mathcal{G}_\Delta$ with $n \leq \lambda$.

We show first in section 2 that the only NFC graph in $\mathcal{G}_2$ with $n \geq \lambda + 1$ is C_6, where $f(C_6) = \{0, 2, 4\}$. This is followed in section 3 by the observation that all trees in $\mathcal{T}_\Delta$ with $\Delta \geq 3$ and $n \geq \lambda + 1$ are FC.

We then focus on $\mathcal{C}_\Delta(\lambda)$, the set of connected nontree graphs with maximum degree Δ and span λ, for $\Delta \geq 3$ and $\lambda \geq \Delta + 1$. A complete account of full coloring is given in section 4 for $\mathcal{C}_3(4)$. It has exactly three NFC graphs with $n \geq 5$. Two involve C_6 with appended vertices, and the third is C_9 with three appended vertices equally spaced around C_9. We also identify every graph in $\mathcal{C}_3(4)$ that has both a full coloring and a nonfull coloring.

Partial results are provided for $\mathcal{C}_4(5)$ and $\mathcal{C}_3(5)$ in sections 5 and 6. The only NFC graph in $\mathcal{C}_4(5)$ with $6 \leq n \leq 8$ is an 8-vertex graph. $\mathcal{C}_4(5)$ contains at least six NFC graphs with $n = 9$, and at least 18 with $n = 10$. $\mathcal{C}_3(5)$ also has no NFC graph for $n = 6$ and at least one for $n = 8$, three for $n = 9$, and two for $n = 10$. Moreover, $\mathcal{C}_3(5)$ contains an infinite number of NFC graphs.

Our final three results, given in section 7, involve NFC graphs for larger values of Δ and λ. The first says that $\mathcal{C}_\Delta(\lambda)$ has an NFC graph with $n \geq \lambda + 1$ whenever $\Delta \geq 3$ and $\Delta + 1 \leq \lambda \leq 2\Delta - 1$. The second says that every $\mathcal{C}_\Delta(\Delta + 2)$ for $n \geq 3$ has an NFC graph with $n \geq \lambda + 1 = \Delta + 3$, all of whose colorings omit two colors in $\{0, 1, \ldots, \Delta + 2\}$. The third notes that for every $\Delta \geq 3$ there is a graph in $\mathcal{C}_\Delta(2\Delta - 2)$ with $n \geq \lambda + 1 = 2\Delta - 1$ such that every coloring of the graph omits $\Delta - 2$ colors in $\{0, 1, \ldots, 2\Delta - 2\}$.

2. Maximum degree 1 or 2. Because $\mathcal{G}_1 = \{P_2\}$ with $\lambda(P_2) = 2$ and $f(P_2) = \{0, 2\}$, the only graph in $\mathcal{G}_1$ is NFC. The following theorem addresses $\Delta = 2$.

THEOREM 2.1. *$\mathcal{G}_2 = \{P_3, P_4, \ldots\} \cup \{C_3, C_4, \ldots\}$ with $\lambda(P_3) = \lambda(P_4) = 3$ and $\lambda(G) = 4$ for all other $G \in \mathcal{G}_2$. All graphs in $\mathcal{G}_2$ are FC except for P_3, C_3, C_4, and C_6. Moreover, $f(C_6) = \{0, 2, 4\}$.*

Proof. It is easily seen that $\mathcal{G}_2 = \{P_3, P_4, \ldots\} \cup \{C_3, C_4, \ldots\}$. The spans of these graphs were noted earlier [5]. The members of $\mathcal{G}_2$ with $n \leq \lambda$ are P_3, C_3, and C_4, and so they are NFC.

Suppose $G = P_4$ with $\lambda(P_4) = 3$. The $L(2,1)$-coloring 1302 of successive vertices shows that P_4 is FC. When $G = P_n$ for $n \geq 5$ with $\lambda(P_n) = 4$, the first n terms of $3024130241\ldots$ taken as successive vertex colors show that P_n is FC.

Suppose $G = C_6$. If a span coloring of C_6 uses color 1, then five successive vertex colors around C_6 must be 41302, and no color from $\{0, 1, 2, 3, 4\}$ for the sixth vertex is admissible for an $L(2,1)$-coloring. Hence $1 \notin f(C_6)$. Similarly (replace c by $4 - c$), $3 \notin f(C_6)$, and the successive colors around C_6 are 024024 or 042042.

Successive vertex colors 02413 for C_5 show that it is FC. The following successive vertex colors for the C_n with $n \geq 7$ show that they are FC:

$$\begin{aligned} n \equiv 0 \pmod 3 &: \quad 024\ldots024130413, \\ n \equiv 1 \pmod 3 &: \quad 024\ldots0240314, \\ n \equiv 2 \pmod 3 &: \quad 024\ldots02413. \end{aligned}$$ ∎

The only NFC graph in $\mathcal{G}_2$ with $n \geq \lambda + 1$ is C_6, where $f(C_6)$ contains neither 1 nor 3.

3. Trees with $\Delta \geq 3$. We assume henceforth that $\Delta \geq 3$. Our next theorem accounts for all trees that are not paths.

THEOREM 3.1. *The only NFC trees in $\mathcal{T}$ with $\Delta \geq 3$ are the $K_{1,\Delta}$ for $\Delta = 3, 4, \ldots$.*

Proof. We consider trees with $\Delta \geq 3$. By Theorem 1.2, $\lambda \in \{\Delta + 1, \Delta + 2\}$. Given Δ, the tree with the fewest vertices is $K_{1,\Delta}$, which is clearly NFC. We show that all

other trees with $\Delta \geq 3$ are FC. If tree T has $n \in \{\Delta+2, \Delta+3\}$ vertices, then T consists of $K_{1,\Delta}$ and one or two other vertices. It follows easily that $\lambda(T) = \Delta+1$ and that T is FC. Assume henceforth that $n \geq \Delta+4$. The rest of the proof divides into two cases, $\lambda(T) = \Delta+1$ or $\Delta+2$.

Case 1. $\lambda(T) = \Delta+2$. We show first that we can choose vertices $v_1, v_2, \ldots, v_{\Delta+3}$ so that $v_1, v_2, \ldots, v_{\Delta+1}$ form $K_{1,\Delta}$ with center v_1 and so that there is an $L(2,1)$-coloring of v_1 through $v_{\Delta+3}$ that uses exactly the colors $0, 1, \ldots, \lambda = \Delta+2$. The choice of $v_{\Delta+2}$ and $v_{\Delta+3}$ depends on adjacencies to v_1 through $v_{\Delta+1}$. Cases (a)–(c) below exhaust the possibilities.

(a) Suppose that a neighbor x of v_1 has degree 3 or more. Denote two non-v_1 neighbors of x by $v_{\Delta+2}$ and $v_{\Delta+3}$. Take $f(v_1) = 0$, $f(x) = \Delta+2$, $f(v_{\Delta+2}) = 1$, $f(v_{\Delta+3}) = 2$, and use colors 3 through $\Delta+1$ for the non-x neighbors of v_1.

(b) Suppose that (a) does not apply, but that there are degree-2 neighbors x and y of v_1. Let $v_{\Delta+2}$ and $v_{\Delta+3}$ be non-v_1 neighbors of x and y, respectively. Take $f(v_1) = 0$, $f(x) = \Delta+1$, $f(v_{\Delta+2}) = 1$, $f(y) = \Delta+2$, $f(v_{\Delta+3}) = 2$, and use colors 3 through Δ for the other neighbors of v_1.

(c) Suppose that neither (a) nor (b) applies. Then all but one of v_2 through $v_{\Delta+1}$ are leaves, i.e., terminal vertices, and the other, say v_2 for definiteness, has degree 2. Let $v_{\Delta+2}$ be the non-v_1 neighbor of v_2, and let $v_{\Delta+3}$ be a new vertex adjacent to $v_{\Delta+2}$. Take $f(v_1) = 0$, $f(v_2) = 2$, $f(v_{\Delta+2}) = \Delta+2$, $f(v_{\Delta+3}) = 1$, and use colors 3 through $\Delta+1$ for the terminal neighbors of v_1.

The tree structure in each case allows us to extend our analysis to an ordering v_1, $v_2, \ldots, v_n$ of all vertices so that v_j for $j = 2, \ldots, n$ is adjacent to exactly one vertex in $\{v_1, v_2, \ldots, v_{j-1}\}$. We now note that we can proceed greedily to extend the coloring of $\{v_1, v_2, \ldots, v_{\Delta+3}\}$ to a full $L(2,1)$-coloring of T by successively labeling $v_{\Delta+4}, \ldots, v_n$ with the smallest admissible color in $\{0, 1, \ldots, \Delta+2\}$ as in the first-fit greedy algorithm in [5]. Because each such v_j is adjacent to only one v_i for $i < j$, and because that v_i is adjacent to at most $\Delta-1$ others, at most $\Delta+2$ of the $\Delta+3$ colors are forbidden for v_j, and thus the greedy procedure can be completed. This completes the proof for $\lambda = \Delta+2$.

Case 2. $\lambda(T) = \Delta+1$. Given T, let M be the set of degree-Δ vertices in T. We say that distinct $x, y \in M$ are *M-neighbors* if the unique path between them has no interior vertex in M. Let s denote a distance between M-neighbors, and let s^* be the minimum such s. The following lemma provides a partial basis for the rest of the proof of Theorem 3.1.

LEMMA 3.2. *Suppose that there are NFC trees other than $K_{1,\Delta}$ with $\lambda = \Delta+1$. Let T be such a tree with the fewest vertices. Then*

(i) *every leaf of T is adjacent to a vertex in M,*

(ii) $|M| \geq 2$,

(iii) $s^* \geq 3$.

Proof. (i) Suppose that leaf l of T is adjacent to vertex $u \notin M$. Then $T-l$ is FC by the fewest-vertices condition for T. Let f be a full coloring of $T-l$. Whatever $f(u)$ equals, at least $(\Delta+2)-3 = \Delta-1$ colors are available for its neighbors, and because no more than $\Delta-2$ of these are used for the non-l neighbors of u, we can extend the full coloring f of $T-l$ to T. However, this contradicts the presumed NFC status of T.

(ii) This follows from (i) and $T \neq K_{1,\Delta}$.

(iii) If $s \in \{1, 2\}$ for M-neighbors, then their colors in a span coloring of T must be 0 and $\Delta+1$ by Lemma 1.1. It follows that T is FC, a contradiction. ∎

We suppose henceforth that $T = (V, E)$ is an NFC tree for $\Delta \geq 3$ and $\lambda = \Delta + 1$, as described in Lemma 3.2. By duality, we can assume that f is a coloring of T with $f(u) = 0$ for every $u \in M$, for if f takes on values of 0 and $\Delta + 1$ for vertices in M, then f would be a full coloring. By hypothesis, $f(V) = \{0, 2, 3, \ldots, \Delta + 1\}$. Let

$$Z = \{z \in V : f(z) = 0\},$$
$$X = V \setminus Z,$$

and define $f' : V \to \{0, 2, 3, \ldots, \Delta + 1\}$ by

$$f'(z) = 0 \quad \text{for all} \quad z \in Z,$$
$$f'(x) = \Delta + 3 - f(x) \quad \text{for all} \quad x \in X.$$

It is easily seen that f' is also a coloring of T. We say that x *is adjacent to* Z if some $z \in Z$ is adjacent to x. A similar definition applies below when Z is replaced by Z_1.

LEMMA 3.3. *Let f be a coloring such that $f(u) = 0$ for all $u \in M$ and $f(V) = \{0, 2, 3, \ldots, \Delta + 1\}$. Suppose $x \in X$. If $f(x) \in \{2, 3, \Delta, \Delta + 1\}$, then x is adjacent to Z. For $\Delta \geq 5$, if $f(x) \in \{4, \ldots, \Delta - 1\}$ and x is not adjacent to Z, then x is adjacent to $v, w \in X$ for which $f(v) = 2$ and $f(w) = \Delta + 1$.*

Proof. Suppose $f(x) \in \{2, 3\}$. If x is a leaf, it is adjacent to a $u \in M \subseteq Z$. Otherwise, with x adjacent to no $z \in Z$, a change of x's color to 1 yields a full coloring of T, a contradiction. If $f(x) \in \{\Delta, \Delta + 1\}$, we use f' to arrive at the same conclusion.

Suppose for $\Delta \geq 5$ that $4 \leq f(x) \leq \Delta - 1$ and x has no neighbor in Z. If x has no color-2 neighbor, we can change x's color to 1 to contradict the NFC status of T. If x has no neighbor v with $f(v) = \Delta + 1$, we use f' to arrive at the same conclusion. □

Lemma 3.3 does not exhaust the restrictions on f, but it suffices to derive the contradiction that T is in fact FC. To begin this part of the proof, let T_1 be a tree with $\Delta \geq 3$ and $\lambda = \Delta + 1$ that satisfies (i)–(iii) of Lemma 3.2 and has a coloring f that satisfies hypotheses and conclusions of Lemma 3.3 with Z_1 the vertices of T_1 with $f(z) = 0$. Obviously, T is such a T_1.

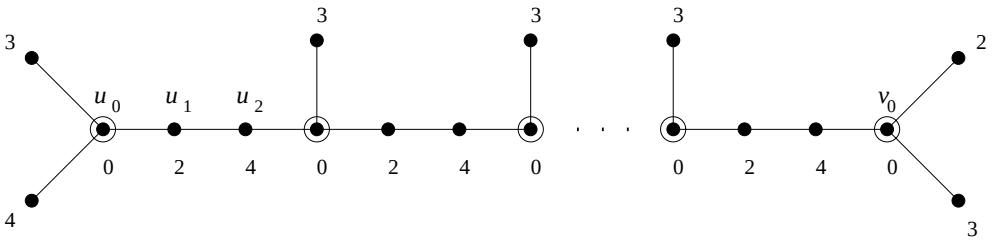
Let $u_0 \in M$ be such that all but one of its Δ neighbors are leaves. (It suffices to choose u_0 as one of a pair of M-vertices that are farthest apart.) Let u_1 be the nonleaf neighbor of u_0, and let u_2 be any other vertex adjacent to u_1 in T_1. We show that T_1 has a full coloring g for which

$$(g(u_0), g(u_1), g(u_2)) = (0, 3, 1).$$

We actually prove a slightly stronger result. Let T_2 be T_1 with leaves added to members of Z_1 so that all $z \in Z_1$ have degree Δ. It is easily seen that $\lambda(T_2) = \Delta + 1$. Henceforth X_2 is the set of vertices of T_2 not in Z_1. We argue that T_2 has a full coloring h with $(h(u_0), h(u_1), h(u_2)) = (0, 3, 1)$. Then we can take g for T_1 as the restriction of h to the vertices in T_1.

In coloring T_2, we work outward from u_0u_1 along paths to the rest of T_2. Every $z \in Z_1$ must get color 0 or $\Delta + 1$. The colors of the vertices in X_2 are subject only to the $L(2, 1)$ constraints. For visual convenience, vertices in Z_1 will be marked by $\odot$. We consider three subcases, $\Delta = 3$, $\Delta = 4$, and $\Delta \geq 5$.

Case 2.1. *Suppose* $\Delta = 3$. Then, under the restrictions of Lemmas 3.2 and 3.3, and with u_0 and v_0 maximum-distance vertices in M, f or f', extended from T_1 to T_2, must be similar to



To verify the diagram, we start at u_0 and go right. We must use colors 024024... or 042042... on the path between u_0 and v_0. We may assume the former without loss of generality. Then every vertex off that path must have color 3, as shown, and it must be a leaf, or else any other neighbor of it would have color 1. However, note that we can change the diagram's coloring to get a full coloring h with color sequence 0314024...4024 along the path between u_0 and v_0, with leaves colored appropriately.

Case 2.2. *Suppose* $\Delta = 4$. We construct h by starting with colors 0, 3, 1 for u_0, u_1, u_2 respectively. We continue along paths away from u_1. Every vertex of Z_1 must get color 0 or color $\Delta + 1$. Once we reach and color a vertex in Z_1, there are just enough feasible colors in $\{0, 1, \ldots, \Delta + 1\}$ to color its as-yet-uncolored successors in the path from u_0, since there are at most $\Delta - 1$ of these. We show that we can continue the coloring any time we have reached and colored a vertex $x \in X_2$. We have to be careful not to color a pair xy of adjacent vertices in X_2 going away from u_0 with color pairs 04, 05, 50, or 51 in that order, because then if a new neighbor of y is in Z_1, the h-coloring will be stymied at this point. We shall suppose that we have achieved a coloring up to $h(x)$ that never uses any of color pairs 04, 05, 50, or 51 on successive vertices ab for $a, b \in X_2$ and show that we can extend the coloring to X_2 neighbors y of x so that $h(x)h(y)$ is not one of these color pairs.

Since $x \in X_2$, it has at most two successors away from u_1. Let x' be its predecessor on a path from u_1. By checking all cases of $h(x)h(x')$, one can easily show that there are always two colors from $\{0, 1, 2, 3, 4, 5\}$ available for the at most two successors of x. If x' is in Z_1, then $h(x') = 0$ or 5. Moreover, if $h(x)$ is also 0 or 5, then we can avoid $h(x)h(y) = 04$, 05, 50, or 51 for y a successor of x by using colors 2 and 3 for the at most two successors of x.

If x' is in X_2, then $h(x')h(x) \neq 04$, 05, 50, 51, by hypothesis. By checking the cases of $h(x')h(x)$, one can easily see that there are always two colors available for the successors of x, one of them 0 or 5 to be used on a successor of x in Z_1, if there is such, and the other 2 or 3, thus allowing us to avoid 04, 05, 50, and 51 for $h(x)h(y)$ for y a successor of x in X_2. This gives us a full coloring h of T_2.

Case 2.3. *Suppose* $\Delta \geq 5$. By our definitions and Lemma 3.3, we have $f(x) = 0 \Leftrightarrow x \in Z_1$, $f(x) \in \{2, 3, \Delta, \Delta + 1\} \Rightarrow x$ is adjacent to Z_1, and if x with $4 \leq f(x) \leq \Delta - 1$ is not adjacent to Z_1, then it is adjacent to vertices of colors 2 and $\Delta + 1$. Under these conditions, we prove the following claim.

CLAIM 3.4. *Suppose* $x \in X_2$. *If* x *is not adjacent to* Z_1, *then degree*$(x) \leq \Delta - 3$. *If* x *is adjacent to* Z_1, *then degree*$(x) \leq \Delta - 1$.

Proof. Suppose $x \in X_2$ is not adjacent to Z_1. Then $f(x) \notin \{0, 2, 3, \Delta, \Delta + 1\}$, so $4 \leq f(x) \leq \Delta - 1$. The neighbors of x cannot have colors in $\{0, f(x) - 1, f(x), f(x) + 1\}$, so degree$(x) \leq |\{0, 2, 3, \ldots, \Delta + 1\}| - 4 = \Delta - 3$. Suppose that $x \in X_2$ is adjacent to Z_1 and $f(x) = 2$. Then the non-Z_1 neighbors of x can have colors in $\{4, 5, \ldots, \Delta + 1\}$, so x can have as many as $1 + (\Delta - 2) = \Delta - 1$ neighbors. Using f', the same thing is true if $f(x) = \Delta + 1$. On the other hand, if $x \in X_2$ is adjacent to Z_1 and $3 \leq f(x) \leq \Delta$, then x can have at most $1 + |\{2, 3, \ldots, \Delta + 1\}| - 3 = \Delta - 2$ neighbors. □

We begin h for T_2 as before with colors 0, 3, and 1 for u_0, u_1, and u_2, respectively, and continue along paths away from u_1. Every $z \in Z_1$ must get color 0 or $\Delta+1$, and once we color a $z \in Z_1$ there are just enough feasible colors in $\{0, 1, \ldots, \Delta+1\}$ to color its as-yet-uncolored $\Delta - 1$ neighbors. In addition, we can continue the coloring from vertices in X_2 without encountering a point where the coloring cannot be continued, so long as no color pair in $\{(0,\Delta), (0,\Delta+1), (\Delta+1,0), (\Delta+1,1)\}$ is ever used on a pair xy of adjacent vertices in X_2 going away from u_1. The problem with these four color pairs is that if y is followed by a $z \in Z_1$, then z cannot be colored.

To show that we can continue the h-coloring while avoiding the four noted color pairs on adjacent vertices in X_2, suppose that $x \in X_2$ has just been colored. Let c denote that color of x's predecessor on the path from u_1 to x, and suppose that none of $(0,\Delta), (0,\Delta+1), (\Delta+1,0), (\Delta+1,1)$ has been used thus far on adjacent vertices in X_2 going away from u_1. We consider the following three cases for coloring x's immediate successors:

Case I. x is not adjacent to Z_1;

Case II. the predecessor of x is in Z_1;

Case III. a successor of x is in Z_1.

We analyze these in turn. Let $S = \{0, 1, \ldots, \Delta+1\}$, so that $|S| = \Delta + 2$.

Case I. By Claim 3.4, x has at most $\Delta - 4$ successors. If $h(x) = 0$, the set of feasible colors for x's successors that avoids color pairs $(0,\Delta)$ and $(0,\Delta+1)$ for x and a successor is $S \setminus \{c, 0, 1, \Delta, \Delta+1\}$, and there are at least $\Delta-3$ such colors. A similar remark applies if $h(x) = \Delta+1$. And if $1 \le h(x) \le \Delta$, any $\Delta-4$ of the $\Delta-2$ colors in $S \setminus \{c, h(x)-1, h(x), h(x)+1\}$ can be used for x's successors.

Case II. Here c must be 0 or $\Delta+1$, so assume $c = \Delta+1$ without loss of generality. By Claim 3.4, x can have as many as $\Delta-2$ successors, all of which are in X_2. If $h(x) = 0$, the set of feasible colors for x's successors that avoid Δ (and $\Delta+1$, used for c) is $S \setminus \{0, 1, \Delta, \Delta+1\}$, and there are exactly $\Delta-2$ such colors. If $1 \le h(x) \le \Delta-1$, the feasible color set for x's successors is $S \setminus \{\Delta+1, h(x)-1, h(x), h(x)+1\}$, and again there are exactly $\Delta-2$ such colors.

Case III. By Claim 3.4, x can have no more than $\Delta-3$ non-Z_1 successors. If $h(x) = 0$, prior avoidance of $(\Delta+1, 0)$ for adjacent X_2-vertices implies $c \le \Delta$. Therefore $\Delta+1$ can be used to color the Z_1-successor of x, and the feasible color set for the non-Z_1 successors of x that avoids $(0,\Delta)$ and $(0,\Delta+1)$ is $S \setminus \{c, 0, 1, \Delta, \Delta+1\}$, which has as at least $\Delta-3$ colors. If $h(x) = 1$, prior avoidance of $(\Delta+1, 1)$ for adjacent X_2-vertices implies $c \le \Delta$, and so we use $\Delta+1$ to color the Z_1-successor of x. This leaves $S \setminus \{c, 0, 1, 2, \Delta+1\}$ for the feasible color set for x's non-Z_1 successors (we do not forbid $(1,\Delta)$ for an adjacent pair in X_2), and this set has $\Delta-3$ colors. Similar remarks apply if $h(x) = \Delta+1$ or if $h(x) = \Delta$. Further, if $2 \le h(x) \le \Delta-1$, we use one of 0 and $\Delta+1$ not equal to c to color x's Z_1-successor, which leaves feasible color set $S \setminus \{c, h(x)-1, h(x), h(x+1), (0 \text{ or } \Delta+1)\}$ with exactly $\Delta-3$ colors for the non-Z_1 successors of x.

The preceding three cases are exhaustive and show that h can be extended throughout T_2. This completes the proof of Theorem 3.1.

4. Analysis of $\mathcal{C}_3(4)$. Because Theorems 2.1 and 3.1 account for all trees with $\Delta \ge 2$, we focus henceforth on connected graphs that are not trees and let

$$\mathcal{C}_\Delta(\lambda) = \mathcal{G}_\Delta(\lambda) \setminus \mathcal{T}_\Delta(\lambda) \quad \text{for} \quad \Delta \ge 3 \quad \text{and} \quad \lambda \ge \Delta+1.$$

The only nonempty $\mathcal{C}_\Delta(\lambda)$ with $\Delta \ge 3$ for which we have complete results is $\mathcal{C}_3(4)$.

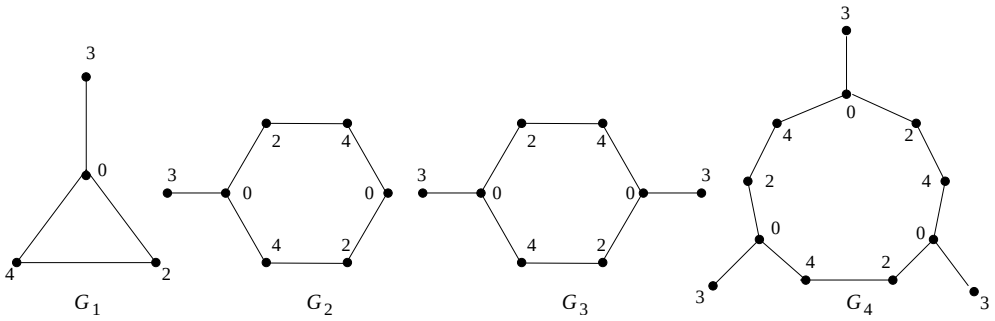


FIG. 1. *The NFC graphs in* $C_3(4)$. *Numbers shown are* $L(2,1)$*-colorings of span* 4.

THEOREM 4.1. *Every graph in* $\mathcal{C}_3(4)$ *is FC except* G_1 *through* G_4 *in Figure* 1.

To prove Theorem 4.1, we note that the labels on the vertices of G_1–G_4, giving $L(2,1)$-colorings of span 4, are unique up to duality. This is obvious from $f(C_6) = \{0,2,4\}$ in Theorem 2.1 for G_2 and G_3. It follows by contradiction for G_4 if we use Lemma 1.1 and note that if on C_9 a label 1 is used and a vertex of degree 3 gets label 0, then we may assume that the first four vertices clockwise from the top of C_9 are labeled 0314. Thus, G_1–G_4 are in $\mathcal{C}_3(4)$ and are NFC.

We assume that G is connected with $\Delta = 3$ and $\lambda = 4$ and, because it is not a tree, has a cycle. The only G with these properties and $n < 5$ is G_1 of Figure 1 (C_4 with diagonal has $\lambda > 4$), so assume henceforth that $n \geq 5$.

To complete the proof of Theorem 4.1, we show that every $G \in \mathcal{C}_3(4)\backslash\{G_1, G_2, G_3, G_4\}$ is FC. Our proof proceeds through a series of observations to a key lemma, Lemma 4.2, which guides its completion. As usual, all colorings are span colorings.

Step 1. Because $\lambda = 4$ for all cycles [5], every cycle uses colors 0 and 4, and so G uses these colors. By Lemma 1.1, every degree-3 vertex has color 0 or 4.

Step 2. Suppose that C_m in G is colored without using color 2. Then C_m's successive colors can be assumed to be

$$0314\ 0314\ 0314\ \ldots\ ,$$

and thus $m = 4t$ for some $t \geq 1$. Connectedness and $\Delta = 3$ require at least one pendant edge or chord for C_m. A chord gives a contradiction because it would have to join vertices labeled 0 and 3, 0 and 4, or 1 and 4. However, in the first case, 0 has a neighbor 3 on the cycle, which is impossible, and similarly for the other two cases. Hence C_m has a pendant vertex which must have color 2 because its edge-mate in C_m is either 0 (with adjacent colors 3 and 4) or 4 (with adjacent colors 0 and 1). Hence G is FC.

Step 3. Suppose that C_m in G is colored without using color 1 or, equivalently (by duality), without using color 3. Then C_m's successive colors can be assumed to be

$$024\ 024\ 024\ \ldots\ ,$$

and thus $m = 3t$ for some $t \geq 1$.

Suppose $t = 1$, and thus $m = 3$. Then either C_3 has exactly two pendant edges to new vertices, which must have colors 1 and 3, or, since G isn't G_1, C_3 has a P_2 off of a C_3 vertex whose two new vertices must have colors 1 and 3. Thus $m = 3$ implies a full coloring.

Suppose $t \geq 2$ with $m = 3t$. A chord in C_{3t} gives a contradiction, and so C_{3t} has at least one pendant edge to a new vertex. Suppose such a pendant vertex x has another neighbor y. If y is not adjacent to any vertex of C_{3t}, then the two new vertices x and y must have colors 1 and 3. If y is adjacent to a vertex of C_{3t}, let u be the vertex of C_{3t} adjacent to x, and v the vertex of C_{3t} adjacent to y. A contradiction occurs if u and v receive the same color. Thus, u and v are colored using 0 and 4, and x and y are colored using 1 and 3. In both situations, we have a full coloring.

Step 4. Steps 2 and 3 imply that the only way to avoid a full coloring of G is for G to consist of exactly one C_m with $m = 3t$ and $t \geq 2$ which is colored $024024\ldots$, has no chord, has one or more pendant edges to distinct new vertices, and has no other vertices. Suppose that G satisfies these restrictions. Number C_m's vertices successively as $1, 2, 3, \ldots, 3t$ and assume without loss of generality that the corresponding colors are $024024\ldots$ and that vertex 1 has degree 3. The pendant vertex adjacent to vertex 1 must have color 3. If another pendant edge goes off a color-4 vertex of C_{3t}, the pendant vertex there must have color 1, and we get a full coloring. We conclude that a full coloring can be avoided only if all other pendant edges (if any) besides the one off vertex 1 go off C_m vertices colored 0, i.e., vertices numbered 4, 7, 10, ... This brings us to the key lemma.

LEMMA 4.2. *Suppose $G \in \mathcal{C}_3(4)$ with $n \geq 5$. Then G is FC unless it has the following structure with feasible coloring as indicated:*

(i). *G has exactly one C_m, m must be in $\{3t : t \geq 2\}$, and the successive vertices of C_m are colored* $024024\ldots$.

(ii). *G has at least one other vertex besides the m of C_m, and each of its other vertices is a terminal vertex adjacent to a color-0 vertex of C_m (and has color 3).*

Step 5. Assume that G has the structure described by (i) and (ii) of Lemma 4.2 and that, with the vertices of C_m numbered successively as $1, 2, \ldots, m$, the non-C_m vertices of G for (ii) are adjacent to C_m vertices in $\{1, 4, 7, 10, \ldots, m-2\}$.

Suppose $m = 6$. Then G is G_2 or G_3, and both are NFC.

Suppose $m = 9$. If there are three other vertices besides the nine in C_9, the most allowed by (ii), we get G_4, which is NFC. If G has one or two vertices not in C_9, we recolor C_9's vertices successively as 031420314, arranged so that the one other vertex is adjacent to vertex 1 (the first vertex in the labeling), or the two other vertices are adjacent to vertices 1 and 4. Then G is FC.

Suppose $m \geq 12$. Recolor C_m's vertices successively as

$$031\ 402\ 413\ 024\ 024\ldots024.$$

Then G is FC because all pendant vertices are adjacent to vertices in positions $1, 4, 7, 10, \ldots, m-2$ which have colors in $\{0, 4\}$ and can therefore be colored to give an overall $L(2,1)$-coloring. This completes the proof of Theorem 4.1. □

Some FC graphs in $\mathcal{C}_3(4)$ must be fully colored by every span coloring, whereas others also have nonfull span colorings. The following result notes when the latter possibility arises.

THEOREM 4.3. *Suppose $G \in \mathcal{C}_3(4) \setminus \{G_1, G_2, G_3, G_4\}$. Then G has both a full coloring and a nonfull coloring if and only if it consists entirely of one C_{3t} for $t \geq 3$ plus one or more pendant edges to terminal vertices adjacent to vertices of C_{3t} that are spaced at distances that are multiples of* 3 *around the cycle.*

We omit the proof.

5. Aspects of $\mathcal{C}_4(5)$. The next cases for $\mathcal{C}$ are $\mathcal{C}_4(5)$ with $\lambda = \Delta + 1$, and $\mathcal{C}_3(5)$ with $\lambda = \Delta + 2$. We present partial results for these cases, then give examples of NFC graphs with $n \geq \lambda + 1$ and larger values of Δ and λ. The following theorem indicates that $\mathcal{C}_4(5)$ is already considerably more daunting than $\mathcal{C}_3(4)$.

THEOREM 5.1. *$\mathcal{C}_4(5)$ has four graphs with $n = 5$ (and all are NFC). The only NFC graph in $\mathcal{C}_4(5)$ with $6 \leq n \leq 8$ is the 8-vertex graph G_5 in Figure 2. $\mathcal{C}_4(5)$ has at least six NFC graphs with $n = 9$ and at least 18 NFC graphs with $n = 10$.*

Proof. The four graphs with $n = 5$ in $\mathcal{C}_4(5)$ are the nonisomorphic ways that one or more edges can be added between terminal vertices of $K_{1,4}$ that admit $L(2,1)$-colorings of span 5. Figure 2 shows $L(2,1)$-colorings of the other graphs of Theorem 5.1. A dashed line in the figure indicates an optional edge that is $L(2,1)$-compatible but is not needed to show that G is NFC. With and without the optional edges, there are six graphs for $n = 9$ in the figure. The options for $n = 10$ give 3, 3, 7, and 6 nonisomorphic graphs for I, II, III, and IV, respectively, but the two maximum-edge graphs in the upper $n = 10$ row are identical, and the number of distinct graphs shown for $n = 10$ is 18. The upper-right of the figure also shows an 8-point graph that is not NFC but plays a role for $n \in \{9, 10\}$.

We now verify that the graphs with $n \geq 6$ described here are in fact NFC.

In what follows, we begin with the coloring of $K_{1,4}$ shown at the top of a graph and refer to its four terminal vertices as the *key vertices*. Note that a new vertex cannot be adjacent to both key vertices 2 and 4, or 2 and 5.

We verify that G_5 in Figure 2 is NFC, prove that it is the only NFC graph in $\mathcal{C}_4(5)$ with $6 \leq n \leq 8$, and then note that the graphs of Figure 2 for $n \in \{9, 10\}$ are NFC.

Suppose $G = G_5$. Vertices x and y must be adjacent to key vertices 2 and 3, and 4 and 5. Then $f(x)$ must be 5, as shown on the upper-left of Figure 2, and $f(y) \in \{1, 2\}$. The bottom vertex must have color 0 or 1, but it is forced to have color 0 because of y, which then has $f(y) = 2$. This gives a unique (up to duality) coloring that is nonfull. Note that no edges can be added to G_5 without forcing $\lambda \geq 6$, since adding any edge violates an $L(2,1)$-coloring requirement.

We now prove that G_5 is the only NFC graph in $\mathcal{C}_4(5)$ with $6 \leq n \leq 8$. Let G consist of $K_{1,4}$ plus sixth, seventh, and eighth vertices x, y, and z, as needed. We use the usual coloring for $K_{1,4}$ with color 0 for its center and take x adjacent to one or more key vertices for connectedness. *Switch colors* means that colors $2, 3, 4, 5$ are replaced by $5, 4, 3, 2$, respectively. We consider $n = 6, 7, 8$ in turn.

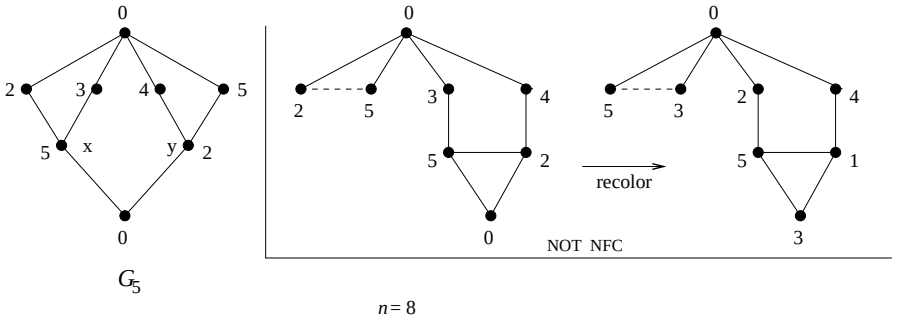
n = 6. If $f(x) = 1$, G is FC. If $f(x) \in \{2, 3\}$, change $f(x)$ to 1 for a full coloring. If $f(x) \in \{4, 5\}$, switch colors and then change $f(x)$ to 1.

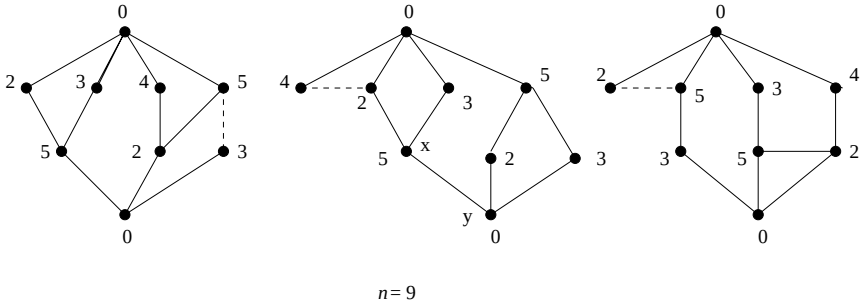
n = 7. Suppose that y as well as x is adjacent to a key vertex. Suppose that $\min\{f(x), f(y)\} > 1$, else G is FC. If $\{f(x), f(y)\} \cap \{2, 3\} \neq \phi$, change the smaller color for x and y to 1 to get a full coloring. If $\{f(x), f(y)\} \subseteq \{4, 5\}$, switch colors, then repeat the preceding step. Suppose that y is adjacent only to x, and $f(x) > 1$. If $f(x) \geq 3$, take $f(y) = 1$. If $f(x) = 2$, set $f(y)$ to 0, switch colors, and then change $f(y)$ to 1.

n = 8. Let $F = \{f(x), f(y), f(z)\}$ and suppose $1 \notin F$.

Suppose $0 \notin F$. If $F \cap \{2, 3\} \neq \phi$, change those in $\{x, y, z\}$ with F's smallest color to color 1 to get a full coloring. If $F \subseteq \{4, 5\}$, switch colors and repeat the preceding step.

Suppose $0 \in F$, and say $f(z) = 0$, so that z has edges only to $\{x, y\}$. If z has only one edge to $t \in \{x, y\}$, take $f(z) = 1$ if $f(t) \geq 3$. If $f(t) = 2$, switch colors





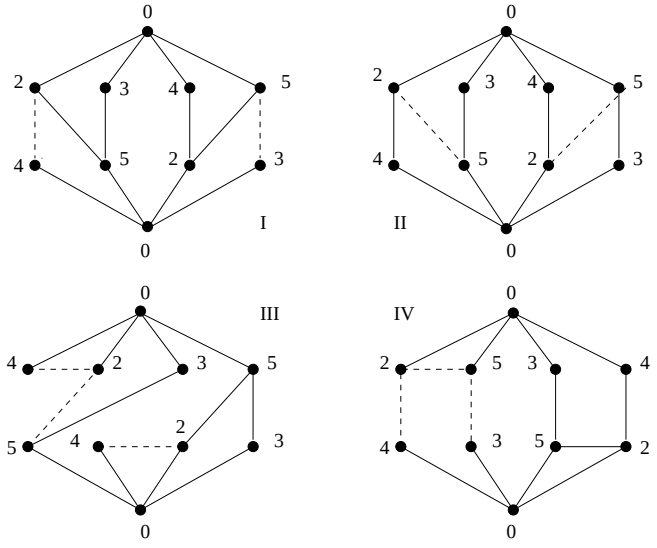


FIG. 2. *NFC graphs in $C_4(5)$ with $L(2,1)$-colorings of span 5 shown, and one non-NFC graph with a recoloring. Dashed edges can be included in the graph or omitted.*

and then set $f(z) = 1$ to get a full coloring. Assume henceforth that $\{x,z\},\{y,z\} \in E$ so that $f(x) \geq 2$ and $f(y) \geq 2$. If $\min\{f(x),f(y)\} \geq 3$, change $f(z)$ to 1; if $\max\{f(x),f(y)\} \leq 4$, switch colors and then change $f(z)$ to 1. It follows that a full coloring is impossible only if $\{f(x),f(y)\} = \{2,5\}$.

With $f(z) = 0$, $\{f(x),f(y)\} = \{2,5\}$, and $\{x,z\},\{y,z\} \in E$, suppose $\{x,y\} \notin E$. If y has no edge to a key vertex, we can presume that $f(x) = 5$ and change $(f(z),f(y))$ to $(1,3)$ to get a full coloring. To prevent this, assume that $f(y) = 2$ with y adjacent to one or both of the key vertices colored 4 and 5. If y is adjacent to only one key vertex, we can choose this key vertex to have color 5 and then change $(f(z),f(y))$ to $(1,3)$ to get a full coloring. It follows that the only way to prevent a full coloring is to have x adjacent to the key vertices with colors 2 and 3, with $f(x) = 5$, and to have y adjacent to the key vertices with colors 4 and 5, with $f(y) = 2$. This gives G_5. We have already proved that G_5 is NFC and that any edge additions to G_5 force $\lambda > 5$.

With $f(z) = 0$, $\{f(x),f(y)\} = \{2,5\}$, and $\{x,z\},\{y,z\} \in E$, suppose $\{x,y\} \in E$. Assume without loss of generality that $f(x) = 5$. Then x is adjacent only to the key vertex with color 3, and y is adjacent to no key vertex or to the key vertex with color 4. If y is adjacent to no key vertex, interchange key vertex colors 2 and 3, then change $(f(z),f(y))$ to $(1,3)$ to get a full coloring. If y is adjacent to the key vertex with color 4, we obtain the middle diagram on the top of Figure 2 and recolor as indicated there to get a full coloring. This completes the proof that G_5 is the only NFC graph in $\mathcal{C}_4(5)$ with $6 \leq n \leq 8$.

It remains to show that graphs for $n \in \{9,10\}$ in Figure 2 are in fact NFC. This is immediate from our analysis of G_5 for the leftmost diagram for $n = 9$ and diagram I for $n = 10$. The adjacencies of II for $n = 10$ force the nonfull coloring there.

It is easily seen that, up to duality, the two colorings of the 8-point graph on the upper-right of Figure 2 are the only $L(2,1)$-colorings of that graph. When a P_2 is added from the bottom vertex to an upper-left key vertex, as in the rightmost diagram for $n = 9$, the full coloring of the upper-right 8-point graph is not $L(2,1)$-feasible, and so we must use the top-middle coloring on the rightmost diagram for $n = 9$, and it is NFC. For a different reason, namely Lemma 1.1, diagram IV for $n = 10$ must also use the top-middle coloring, and so it too is NFC.

This leaves only the middle diagram for $n = 9$ and diagram III for $n = 10$. Careful analysis of the middle $n = 9$ diagram shows that if we delete its left $\{2,5\}$ edge, then the bottom vertex must have color 0 or 1. One can try different values for the rightmost vertex labeled 5 and show that only 0 or 1 works for y. However, the presence of the left $\{2,5\}$ edge forces the bottom vertex to have color 0, and no vertex can have color 1. For diagram III for $n = 10$, where the left $\{2,5\}$ edge is now optional, suppose as usual that the top vertex is colored 0. Lemma 1.1 forces the bottom vertex to be 0 or 5, but a careful analysis shows that 5 is not possible. If the bottom vertex is 0, then its neighbors do not include 1, so III never has color 1. □

6. Aspects of $\mathcal{C}_3(5)$. The following results for $\mathcal{C}_3(5)$ are like those for $\mathcal{C}_3(4)$ in that no graph in $\mathcal{C}_3(5)$ for $n = 6$ is NFC but at least one for $n = 8$ is NFC. It is true also that $\mathcal{C}_3(5)$ has no NFC graph for $n = 7$, but the proof is case-intensive, so we omit $n = 7$ from the following theorem. We also show that $\mathcal{C}_3(5)$ has arbitrarily large NFC graphs, and thus it has an infinite number of such graphs in distinction to the small finite number of Theorem 4.1 for $\mathcal{C}_3(4)$.

THEOREM 6.1. *$\mathcal{C}_3(5)$ has four graphs with $n \leq 5$ (and all are NFC). It has 17 graphs with $n = 6$, and all are FC. The number of NFC graphs in $\mathcal{C}_3(5)$ is at least one for $n = 8$, at least three for $n = 9$, and at least two for $n = 10$. Moreover, $\mathcal{C}_3(5)$*

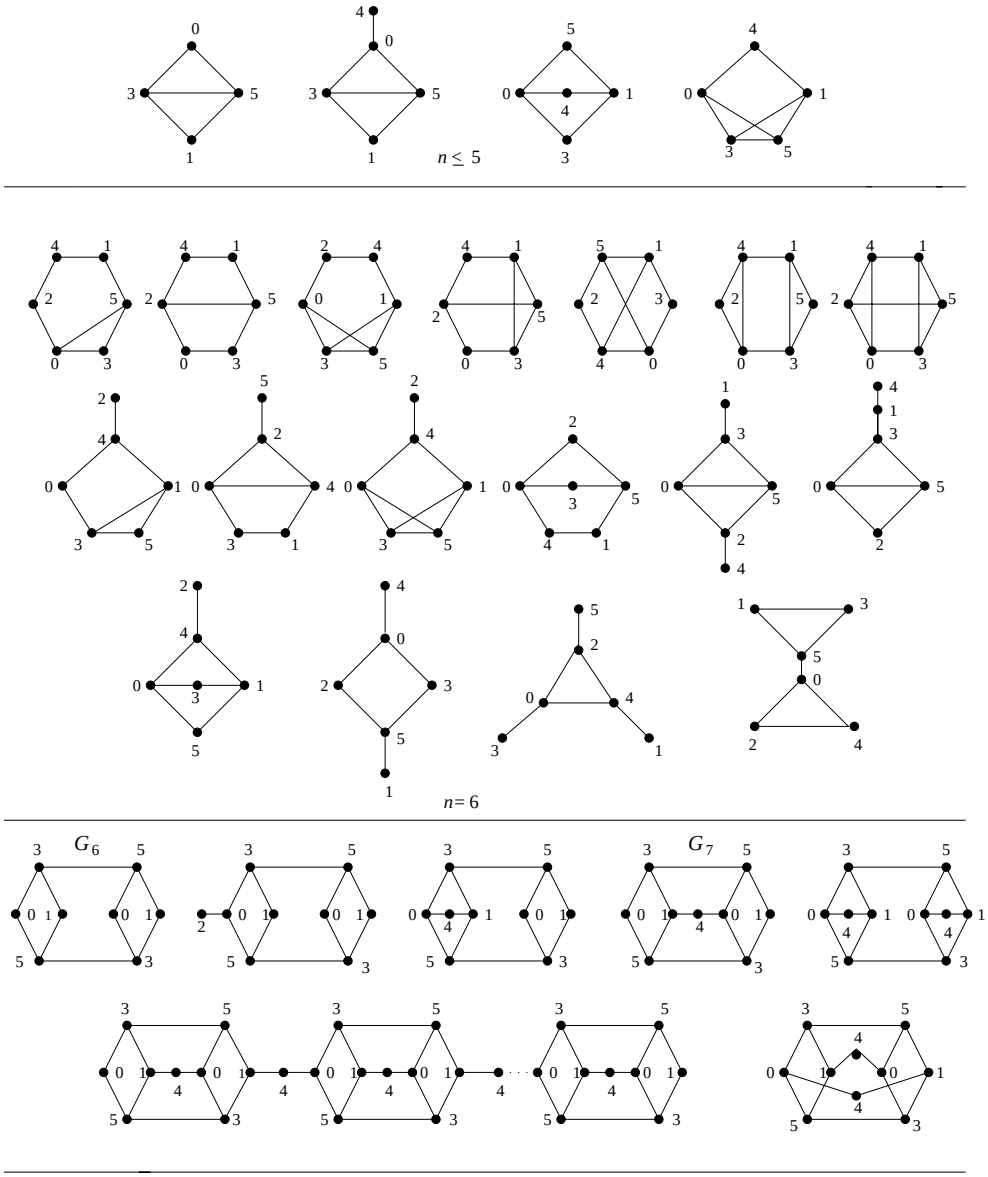


FIG. 3. *Graphs in $C_3(5)$ with full colorings for $n = 6$ and essentially unique span colorings for $n = 8$, 9, 10, and $10m - 1$.*

has at least one NFC graph for every $n \in \{10m - 1 : m = 2, 3, 4, \ldots\}$.

Proof. The four graphs in $\mathcal{C}_3(5)$ with $n \leq 5$ are shown at the top of Figure 3, and the 17 with $n = 6$ appear below them. The NFC graphs for $n \in \{8, 9, 10\}$ are in the lower part of the figure along with the pattern used to prove NFC for $n = 10m - 1$.

We begin the proof of these observations with comments that lead easily to the conclusion that the only members of $\mathcal{C}_3(5)$ with $n \leq 5$ are the four graphs in the top row of Figure 3. (Of course, all are NFC.) A triangle with two appended vertices and $\Delta = 3$ has $\lambda = 4$. The vertices of C_4 for $\lambda = 4$ are successively colored 0314, and

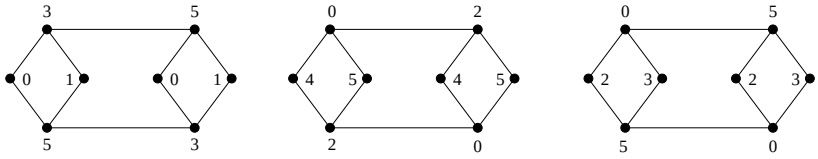


FIG. 4. *$L(2,1)$-colorings of graphs G_6 and G_7 of Figure 3 with $\lambda = 5$.*

thus the addition of a diagonal forces $\lambda = 5$. Two diagonals give K_4 with $\lambda = 6$. Addition of a vertex outside C_4 or outside C_4 plus a diagonal leads to the examples. The successive-vertex coloring of C_5 for $\lambda = 4$ is 03142. This allows one chord (0 to 4), but two disjoint chords require $\lambda = 5$. Two chords with a common vertex or more than two chords give $\Delta > 3$.

The graphs of Figure 3 for $n = 6$ are organized by cycle sizes. The only $\lambda = 4$ coloring of C_6 is 024024, and thus the addition of one or more chords forces $\lambda \geq 5$. Eight such additions have $\Delta = 3$. Seven appear in the second row of Figure 3. The eighth has three diagonal chords and $\lambda = 6$. The rest of Figure 3 for $n = 6$ shows that $\mathcal{C}_3(5)$ has four graphs with a C_5 and no C_6, two with a C_4 plus diagonal and no C_5 or C_6, two with a C_4 and no diagonal and no C_5 or C_6, and two with a C_3 and no C_4, C_5, or C_6. This yields 17 graphs in $\mathcal{C}_3(5)$ with $n = 6$, and all are FC.

The rest of Theorem 6.1 is based on the 8-vertex graph G_6 in the lower-left of Figure 3. It has exactly three span colorings with $\lambda = 5$; see Figure 4. These three result from a systematic examination of feasible colorings for the left diamond (C_4). Only the three of Figure 4 extend to the right diamond, and their extensions are unique. The original G_6 coloring yields the colorings for $n \in \{9, 10\}$ at the bottom of Figure 3. All are NFC.

Let G_7 be the 9-vertex graph on the lower halves of Figures 3 and 4. It has only two $L(2,1)$-colorings of span 5, shown in Figure 4, because the rightmost G_6 coloring in the upper part of Figure 4 does not accommodate the new vertex. Denote by $G_7(m)$ the graph in $\mathcal{C}_3(5)$ with m copies of G_7 linked in series by $m-1$ new vertices, as illustrated for $m = 3$ in the bottom row of Figure 3. It has $9m + m - 1 = 10m - 1$ vertices. A new vertex adjacent to a degree-2 vertex of G_7 must have color 2 or 4 if we use the left coloring of G_7, and must have color 1 or 3 if we use the right coloring of G_7 in Figure 4. Because $\{2,4\} \cap \{1,3\} = \phi$, all copies of G_7 in $G_7(m)$ have identical colorings. If the left G_7 coloring is used, all $m-1$ linking vertices must have color 4, and if its right dual coloring is used, all $m-1$ linking vertices must have color 1. Hence $G_7(m)$ is NFC. □

7. More NFC graphs. Our next two theorems follow the lead of G_6 on the lower-left of Figure 3 to obtain NFC graphs for a variety of other (Δ, λ) pairs.

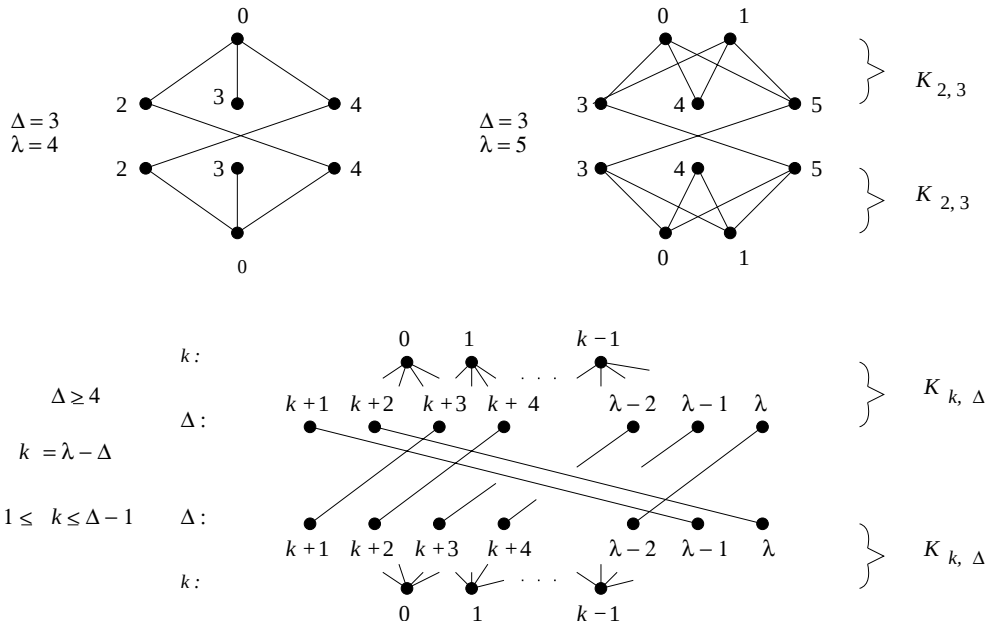


FIG. 5. *$L(2,1)$-colorings of graphs for proof of Theorem* 7.1.

THEOREM 7.1. *If $\Delta \geq 3$, and $\Delta + 1 \leq \lambda \leq 2\Delta - 1$, then $\mathcal{C}_\Delta(\lambda)$ contains an NFC graph with $n \geq \lambda + 1$.*

Proof. Let G be the graph of Figure 5 for $\Delta \geq 3$, $1 \leq k \leq \Delta - 1$, and $\lambda = \Delta + k$. G has 2λ vertices, which exceeds $\lambda + 1$. The colorings shown and the fact that $\lambda(K_{k,\Delta}) = k + \Delta$ verify that its span is in fact λ.

The colorings of Figure 5 are not full because they omit color k. It is easily seen that the only alternative span coloring of $K_{k,\Delta}$ is the dual of the one shown with colors $\Delta + 1$ through λ for row 1 (or 4) and 0 through $\Delta - 1$ for row 2 (or 3). This dual coloring for rows 1 and 2 is $L(2,1)$-incompatible with the original coloring of rows 3 and 4, and if the dual coloring is used for both pairs of rows, then color Δ is not used. It follows that G is NFC. □

The next theorem shows that there are graphs in $\mathcal{C}_\Delta(\Delta + 2)$ with $n \geq \lambda + 1$, whose span colorings must omit at least two colors in $\{0, 1, \ldots, \lambda\}$. Of course K_n colorings omit nearly half the colors in their span coloring sets, but they also have $\lambda = 2(n-1) > n$. Our final result, Theorem 7.3, notes that there are connected graphs with $n \geq \lambda + 1$ whose colorings omit nearly half the colors in their span coloring sets.

THEOREM 7.2. *If $\Delta \geq 3$ and $\lambda = \Delta + 2$, then $\mathcal{C}_\Delta(\Delta + 2)$ contains an NFC graph with $n = 2(\Delta + 1)$, all of whose colorings use exactly $\lambda - 1$ colors.*

Figure 6 presents a graph that satisfies the hypotheses and conclusions of Theorem 7.2. Its coloring omits colors $\Delta - 1$ and $\Delta + 1$. The dual coloring omits colors 1 and 3. It is easily seen that all colorings must omit at least two colors. We omit the proof that no other span coloring of G omits fewer than two colors.

THEOREM 7.3. *If $\Delta \geq 3$ and $\lambda = 2\Delta - 2$, then $\mathcal{C}_\Delta(2\Delta - 2)$ contains a graph with $n = 2\Delta + 1$, all of whose colorings omit exactly $\Delta - 2$ colors in $\{0, 1, \ldots, 2\Delta - 2\}$.*

Proof. Given $\Delta \geq 3$, let G consist of disjoint copies A and B of K_Δ and one more vertex, x, of degree 2 that has an edge to $a \in A$ and to $b \in B$. Every coloring of G assigns colors $0, 2, 4, \ldots, 2\Delta - 2$ to each of A and B, and an odd color between 0

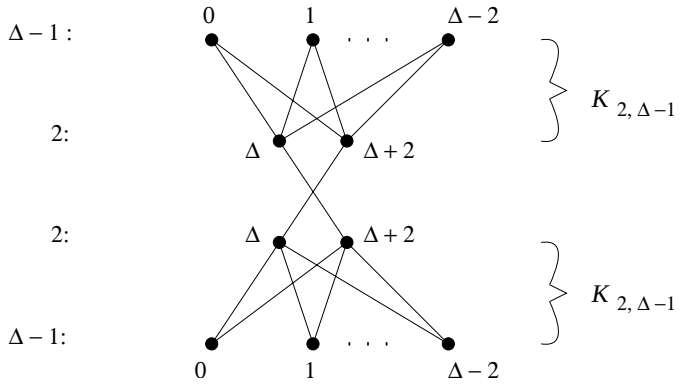


FIG. 6. *An $L(2,1)$-coloring for Theorem* 7.2.

and $2\Delta - 2$ to x. It follows that every coloring of G omits exactly $\Delta - 2$ colors in $\{0, 1, \ldots, 2\Delta - 2\}$. □

Acknowledgment. We are indebted to Jan Kratochvíl, Renu Laskar, Aleksandar Pekeč, Denise Sakai Troxell, and John Villalpando for discussions on $L(2,1)$-colorings, and to the referees for their help in improving the present paper.

REFERENCES

[1] G. J. CHANG, *References for $L(2,1)$-Labelings*, website at http://www.math.ntu.edu.tw/gjchang/courses/2005-02-graph-algorithm/2004-02-06-l21-ref-chu.pdf, accessed February 6, 2004.

[2] G. J. CHANG AND D. KUO, *The $L(2,1)$-labeling problem on graphs*, SIAM J. Discrete Math., 9 (1996), pp. 309–316.

[3] P. C. FISHBURN AND F. S. ROBERTS, *No-hole $L(2,1)$-colorings*, Discrete Appl. Math., 130 (2003), pp. 513–519.

[4] J. P. GEORGES, D. W. MAURO, AND M. A. WHITTLESEY, *Relating path coverings to vertex labelings with a condition at distance two*, Discrete Math., 135 (1994), pp. 103–111.

[5] J. R. GRIGGS AND R. K. YEH, *Labelling graphs with a condition at distance* 2, SIAM J. Discrete Math., 5 (1992), pp. 586–595.

[6] W. K. HALE, *Frequency assignment: Theory and application*, Proc. IEEE, 68 (1980), pp. 1497–1514.

[7] D. D.-F. LIU AND R. K. YEH, *On distance two labellings of graphs*, Ars Combin., 47 (1997), pp. 13–22.

[8] A. PEKEČ, *private communication*, Fuqua School of Business, Duke University, Durham, NC, 2000.

[9] F. S. ROBERTS, *T-colorings of graphs: Recent results and open problems*, Discrete Math., 93 (1991), pp. 229–245.

[10] R. K. YEH, *Labeling Graphs with a Condition at Distance Two*, Ph.D. thesis, Department of Mathematics, University of South Carolina, Columbia, SC, 1990.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 444–462

A LINEAR-TIME ALGORITHM FOR FINDING A MAXIMAL PLANAR SUBGRAPH*

HRISTO N. DJIDJEV†

Abstract. We construct an optimal linear-time algorithm for the maximal planar subgraph problem: given a graph G, find a planar subgraph G' of G such that adding to G' an extra edge of G results in a nonplanar graph. Our solution is based on a fast data structure for incremental planarity testing of triconnected graphs and a dynamic graph search procedure. Our algorithm can be transformed into a new optimal planarity testing algorithm.

Key words. planar graphs, planarity testing, incremental algorithms, graph planarization, data structures, triconnectivity

AMS subject classifications. 05C10, 05C85, 68R10, 68Q25, 68W40

DOI. 10.1137/S0895480197328771

1. Introduction. A graph is *planar* if it can be drawn in the plane so that no two edges intersect except at a common endpoint. Planar graphs arise naturally in many applications of graph theory, e.g., in circuit and VLSI design, network design and analysis, computational geometry, and are one of the most intensively studied classes of graphs [21]. Many problems that are computationally hard for arbitrary graphs have efficient solutions for the case of planar graphs; testing an n-vertex m-edge graph for planarity takes $O(n+m)$ time [16, 2].

If the graph is not planar, then often a problem arises of how to find a planar subgraph that is as close to the given graph as possible. A problem of this type is called a *graph planarization problem*. For any n-vertex graph G of genus g there exists a vertex set of size $O(\sqrt{ng})$ whose removal leads to a planar graph [9]. However, the linear-time implementation of the algorithm that finds such a planarizing set requires a genus-g embedding of G as input; the best algorithm that finds such an embedding [10] is polynomial in n, but doubly exponential in g. Another version of the graph planarization problem, the problem of finding the smallest number of edges whose removal leaves a planar graph, is known to be NP-complete [13].

Since finding a maximum planar subgraph is very hard, many researchers have investigated the problem of constructing, for a given n-vertex m-edge graph G, a planar subgraph G' of G such that adding to G' any edge of $E(G)-E(G')$ results in a nonplanar graph. Such a graph G' is called a *maximal planar subgraph* of G. This problem has been intensively investigated in relation to its applications to circuit layout [23, 20, 5, 22, 19]. More recently, Cai, Han, and Tarjan [4] developed an $O(m \log n)$ algorithm for the maximal planar subgraph problem based on the Hopcroft–Tarjan planarity testing algorithm. Their result improved (if $m = o(n^2/\log n)$) the best previous

*Received by the editors October 15, 1997; accepted for publication (in revised form) December 7, 2005; published electronically June 2, 2006. A preliminary version of this paper was presented at WADS '95 [8].

http://www.siam.org/journals/sidma/20-2/32877.html

†Los Alamos National Laboratory, Basic and Applied Simulation Science (CCS-5), Los Alamos, NM 87545 (djidjev@lanl.gov). This work was partially supported by NSF grant CCR-9409191 and by EPA grant R82-5207-01-0.

$O(n^2)$ algorithm from [19] (based on the PQ-tree technique [2]). An algorithm with the same complexity bound of $O(m \log n)$ can also be derived from the incremental planarity testing algorithm of Di Battista and Tamassia [6]. Using an approach similar to [6], Westbrook [25] described an algorithm that works in $O(n \log n + m\alpha(m, n))$ worst-case time plus an additional $O(n)$ expected time. La Poutré [24] recently gave an incremental planarity testing algorithm that takes $O(\alpha(m, n))$ amortized time per operation, which can be transformed into an $O(n + m\alpha(m, n))$ time algorithm for the maximal planar subgraph problem.

In this paper we describe a linear-time $O(n + m)$-time algorithm for the maximal planar subgraph problem. Our algorithm uses a tree-represented decomposition of a biconnected graph into triconnected components, a common feature of the incremental planarity testing algorithms [6, 7, 25, 24]. We use a variation of the decomposition tree of Di Battista and Tamassia; however, any of the alternative representations could be used instead. Our algorithm has the following structure: (i) it initially constructs a depth-first spanning tree of G (we can assume w.l.o.g. that G is connected) and uses it as an initial approximation of the maximal planar subgraph; (ii) it adds the edges one by one, making an online choice of the next edge to be added so that the testing time will be appropriately small.

Note that our ability to make a choice of the order in which to insert, while possible, the edges into the subgraph so that planarity is preserved is essential for achieving $O(1)$ amortized time per test and insert operation. As noted by Westbrook [25], there is an $\Omega(\alpha(m, n))$ lower bound on the amortized time per operation of any algorithm that maintains a decomposition of the triconnected components of a graph subject to *arbitrary* edge insertions, which gave rise to the conjecture that $O(\alpha(m, n))$ is the best possible time bound for the incremental planarity testing problem [24].

Another technique we use is maintaining in each bicomponent a special dynamic path of nodes of the decomposition tree such that all testing and updating operations are performed on nodes of that path. This makes it possible to implement data structures supporting set union and set split operations in a constant amortized time. Also, we develop a new efficient data structure used for incremental planarity testing of triconnected graphs, which works in $O(1)$ amortized time per operation, an improvement over the best previous $O(\alpha(m, n))$-time algorithms.

Independently of our result, Hsu [17] has constructed a linear-time algorithm for the maximal planar subgraph problem that is based on the modified version of the Hopcroft–Tarjan planarity testing algorithm.

Our algorithm for the maximal planar subgraph problem can be transformed into a linear-time algorithm for planarity testing based on an approach entirely different from the existing ones. The previous algorithms of Hopcroft and Tarjan [16] and Booth and Lueker [2] are based on the Jordan Curve Theorem, which states that any closed curve in the plane divides it into exactly two disjoint connected regions, while our planarity testing algorithm exploits the fact that any triconnected planar graph has a unique embedding in the plane.

This paper is organized as follows. In section 2 we give some definitions and review a dynamic data structure that maintains a decomposition of a connected graph into biconnected and triconnected components. In section 3 we develop an algorithm for online planarity testing in triconnected graphs in a constant amortized time, which we use as a subroutine in the main algorithm. In section 4 we give the overall structure of the algorithm as well as more details about individual data structures and update operations.

2. Preliminaries. In this section we give some basic definitions related to graph connectivity and graph orientation and describe briefly the data structure for maintaining the biconnected and triconnected components of a graph developed by Di Battista and Tamassia [6, 7].

2.1. Definitions. We use standard graph terminology [14]. An undirected graph G is *connected* if any two vertices of G are connected by a path. The maximal connected subgraphs of G are the *connected components* of G. A vertex v is a *cutvertex* if the removal of v increases the number of components. G is *biconnected* if G is connected and G has no cutvertices. The maximal biconnected subgraphs of G are called *bicomponents*. A pair v, w of vertices of G is a *separation pair* if the deletion of v and w disconnects G. G is *triconnected* if G has no cutvertex and no separation pair.

An essential property of triconnected graphs related to planarity is given in the next lemma. A *subdivision* of a graph H is a graph H' that can be obtained by H by replacing some of the edges of H by paths having at most their endpoints in common.

LEMMA 2.1 (see [21]). *A planar graph G has a unique embedding in the plane iff G is a subdivision of a triconnected graph.*

The *triconnected components (or tricomponents)* of G are produced by a recursive procedure that, if G has a separation pair v, w, divides G into two subgraphs G_1 and G_2 defined by the separation pair. Each of v and w is included in both G_1 and G_2. For the precise definition and a linear-time algorithm that finds the tricomponents of a graph, see [15].

An *st-graph* is a directed acyclic graph with exactly one source and exactly one sink. Any biconnected graph can be converted into an st-graph using the linear-time st-numbering algorithm of [11]. A *planar st-graph* is an st-graph that is embedded in the plane such that the source and the sink belong to the external face of the embedding.

Let G be an st-graph. A *split pair* $\{a, b\}$ of G is either a separation pair or a pair of adjacent vertices of G. A *split component* of a split pair $\{a, b\}$ is either an edge (a, b) or a maximal subgraph G' of G that is an st-graph with a source a and a sink b such that $\{a, b\}$ is not a split pair of G'. For instance, the graph G from Figure 2.1(a) has split pairs (3,8), (3,5), (2,7), and all pairs of adjacent vertices; its split components are the subgraphs G_2, G_4, and G_5 from Figure 2.1, as well as all edges of G. If there is no other split pair $\{a', b'\}$ such that $\{a, b\}$ is contained in a split component of $\{a', b'\}$, then $\{a, b\}$ is a *maximal split pair*.

2.2. Decompositions of biconnected graphs. First we consider the case where G is biconnected. Let n be the number of vertices and m be the number of edges of G.

We recall the definition of SPQR trees from [6]. An *SPQR* tree for G is a recursively defined tree T closely related to the decomposition of G with respect to its split pairs. T has four types of nodes S, P, Q, and R, and there is an st-graph, $skeleton(\mu)$, associated with each node μ of T. The skeletons of the internal nodes of T are in one-to-one correspondence with the tricomponents of G, and hence their number is $O(m)$. The endpoints of each edge e in the skeleton of the root of T correspond to a maximal split pair of G, and e represents the set of split components of that split pair (see Figure 2.1).

Formally, SPQR trees are defined as follows [6, 7]. Let G be an st-graph with source s and sink t. An SPQR tree T for G has four types of nodes S, P, Q, and R,

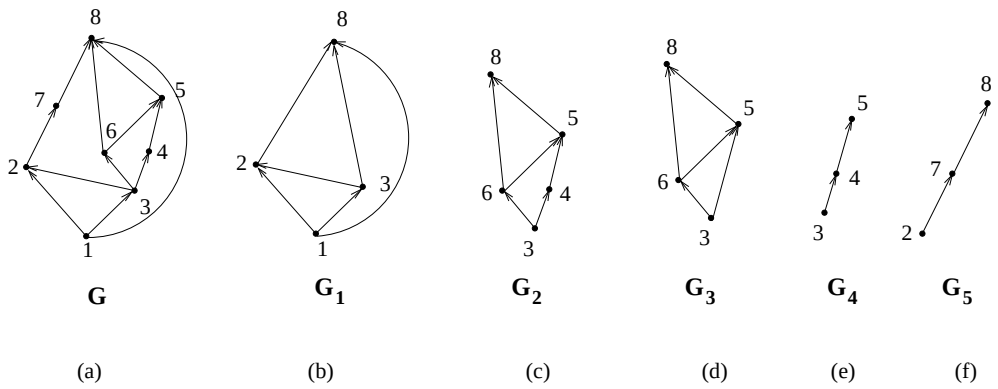


FIG. 2.1. *Illustration of the definition of SPQR trees:* (a) *A planar st-graph with source* 1 *and sink* 8. (b) *The skeleton associated with the root* μ *of the SPQR tree for* G *(*μ *is an R node).* (c) *The split component corresponding to edge* $(3,8)$ *of* G_1. (d) *The skeleton associated with node*(G_2). (e) *The split component corresponding to edge* $(3,5)$ *of* G_3. (f) *The split component corresponding to edge* $(2,8)$ *of* G_1.

and there is an *st*-graph, *skeleton*(μ), associated with each node μ of T. T is defined recursively as follows.

(i) (*trivial case*) If G is a single edge from s to t, then T consists of a single Q node μ (leaf), and *skeleton*(μ) is G.

(ii) (*parallel case*) If $\{s,t\}$ is a split pair with split components $G_1,\ldots,G_k$, then the root of T is a P node μ, and *skeleton*(μ) consists of vertices s and t joined by k parallel edges $e_1,\ldots,e_k$.

(iii) (*series case*) If G has cutvertices $x_1,\ldots,x_{k-1}$, $k\geq 1$, dividing it into components $G_1,\ldots,G_k$ in this order from s to t, then the root of T is an S node μ, and *skeleton*(μ) is the path $s=x_0,x_1,\ldots,x_{k-1},t=x_k$. We denote edge $e_i=(x_{i-1},x_i)$, for $i=1,\ldots,k$.

(iv) (*rigid case*) If none of the above cases applies, let $\{a_1,b_1\},\ldots,\{a_k,b_k\}$ be the maximal split pairs of G, and let G_i be the union of all split components of $\{a_i,b_i\}$. Then the root of T is an R node μ, and *skeleton*(μ) is the graph obtained by replacing in G each subgraph G_i by an edge between a_i and b_i with orientation compatible with the orientation of G_i (so that the resulting graph is an *st*-graph.)

In cases (ii)–(iv) μ has children nodes $\mu_1,\ldots,\mu_k$, which are the roots of the SPQR trees of $G_1,\ldots,G_k$. We denote *node*$(G)=\mu$ and *node*$(G_i)=\mu_i$. The edge e_i representing the skeleton of μ_i in G is called a *virtual edge* of μ_i.

A property of the SPQR trees that is relevant to planarity testing is that either the skeleton of any internal node μ of a SPQR tree has a unique planar embedding (if μ is an R node), or *any* two edges can be placed on the same face (if μ is a P, Q, or S node.) For a more detailed discussion of SPQR trees, see [6, 7]. Our next goal is to show how to reduce a planarity testing in a graph to planarity testing in skeletons of nodes of its SPQR tree.

2.2.1. Planarity testing using SPQR trees. An *allocation node* of a vertex v of G is a node μ such that *skeleton*(μ) contains v. A *proper allocation node* of v, denoted by *proper*(v), is the least common ancestor of all allocation nodes of v.

For any pair of vertices v_1 and v_2 we will define *projections* $pr(v_1)$ and $pr(v_2)$, so either can be a vertex or an edge of the skeleton of an appropriate node μ of the

SPQR tree. The relevant property of the projections of v_1 and v_2 is that v_1 and v_2 belong to the same face of G iff $pr(v_1)$ and $pr(v_2)$ will belong to the same face of the skeleton of μ.

Let v_1 and v_2 be two vertices of G, and let $proper(v_1)$ and $proper(v_2)$ be the proper allocation nodes of v_1 and v_2. Assume that $proper(v_1)$ and $proper(v_2)$ belong to a simple tree path to the root of the SPQR tree. (In our algorithm described in section 4.1 this will always be the case.) Assume w.l.o.g. that $proper(v_1)$ is an ancestor of $proper(v_2)$. Define $\mu = \mu(v_1, v_2)$ to be the nearest common ancestor of $proper(v_1)$ and $proper(v_2)$ ($proper(v_1)$ in our case). Call a *joining path* $p(v_1, v_2)$ the tree path in the SPQR tree between $proper(v_1)$ and $proper(v_2)$ excluding μ and its child. For $i = 1$ and $i = 2$, if $skeleton(\mu)$ contains v_i, then define $pr(v_i)$ to be the vertex v_i of $skeleton(\mu)$; otherwise define $pr(v_i)$ to be the virtual edge in $skeleton(\mu)$ corresponding to the subgraph containing v_i.

We define a *peripheral* vertex (resp., edge) of an st-graph to be a vertex (resp., edge) that appears on the external face of some st-planar embedding of the graph. A *peripheral node* is a node μ whose virtual edge is peripheral.

The following lemma relates incremental planarity testing in an arbitrary graph to incremental planarity testing in its tricomponents (assuming that $proper(v_1)$ is an ancestor of $proper(v_2)$).

LEMMA 2.2 (see [6]). *There exists a planar embedding of G such that v_1 and v_2 belong to the same face iff*

(i) *$pr(v_1)$ and $pr(v_2)$ are on the same face of some planar embedding of $skeleton(\mu(v_1, v_2))$,*

(ii) *all the nodes on the joining path $p(v_1, v_2)$ are peripheral, and*

(iii) *if $proper(v_1) \neq proper(v_2)$, then v_2 is a peripheral vertex of $proper(v_2)$.*

As an example, apply the lemma for vertices $v_1 = 2$ and $v_2 = 4$ of G (Figure 2.1). Their projections (in G_1) are $pr(v_1) = 2$ and $pr(v_2) = (3, 8)$. The joining path $p(v_1, v_2)$ consists of a single node $node(G_4)$ whose virtual edge $(3, 5)$ in G_3 is peripheral. By Lemma 2.2, edge $(2, 4)$ can be added to G while preserving planarity.

In the next sections we will describe data structures for answering queries of types (i), (ii), and (iii) from Lemma 2.2 in a constant time.

2.3. Decompositions of connected graphs. In order to handle connected graphs that are not necessarily biconnected we define the BC trees introduced in [7], which are extensions of the SPQR trees. To construct a BC tree of a connected graph G first find all bicomponents of G. Then construct a tree that contains a node of type B for any bicomponent b and a node of type C for any cutvertex c of G. Associate with each B node b an SPQR tree representing b. Connect a C node c and a B node b iff c belongs to b. Finally root the tree at an arbitrary B node. Call the nodes of B *level*-1 nodes and the nodes of the SPQR trees *level*-2 nodes.

Suppose that an edge (v_1, v_2) has to be added to G. If v_1 and v_2 belong to the same bicomponent b of G, then the BC tree of G is not changed after the insertion. In this case we use the SPQR tree associated with b and Lemma 2.2 to determine if (v_1, v_2) can be added while preserving the planarity and do the insertion by modifying the SPQR tree for b. Now assume that v_1 and v_2 belong to different bicomponents $b(v_1)$ and $b(v_2)$. Let $p = \{b_1 = b(v_1), c_1, b_2, c_2, \ldots, c_k, b_{k+1} = b(v_2)\}$ be the unique tree path between $b(v_1)$ and $b(v_2)$. If one of b_1 and b_{k+1} is an ancestor to the other and the edge (c_{i-1}, c_i) can be added to b_i while preserving planarity, then we call b_i a *peripheral* level-1 node (with respect to path p.) We have the following lemma [7].

LEMMA 2.3. *There exists a planar embedding of G such that v_1 and v_2 belong*

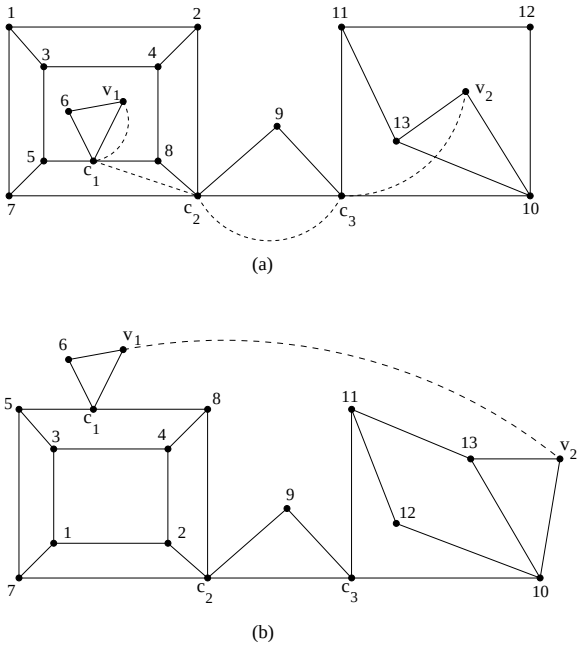


FIG. 2.2. *Illustration of Lemma* 2.3. (a) *A graph with* 4 *bicomponents.* (b) *The result after adding the edge* (v_1, v_2) *to the planar embedding.*

to the same face iff edges (v_1, c_1) *and* (c_k, v_2) *can be added to* G *while preserving planarity and nodes* $b_2, \ldots, b_k$ *are peripheral.*

See Figure 2.2 for illustration.

For edges (v_1, c_1) and (c_k, v_2) we do planarity testing using the corresponding SPQR trees, and for edges $(c_1, c_2), \ldots, (c_{k-1}, c_k)$ we use dynamically maintained maximal paths of edges (c_i, c_{i+1}) whose addition preserves planarity. We will give more details in section 4.

In order to use the above data structure for the maximal planar subgraph problem we also need algorithms for efficiently updating the data structure after the insertion of any edge. Before discussing the update operations we will describe the data structures for incremental planarity testing in triconnected graphs and give an outline of the whole algorithm.

3. The triconnected case. By Lemma 2.1 the maximal planar subgraph (MPS) problem is easier to solve if a planar spanning triconnected subgraph of the original graph is known. Accordingly, we will first describe a linear-time algorithm for the following restricted version of the MPS problem, which we call the *triconnected maximal planar subgraph* (TMPS) problem.

PROBLEM. *Let* G *be a planar triconnected graph and* E' *be a set of edges between vertices of* G. *Find a maximal set* $E'' \subset E'$ *such that* $G + E''$ *is still planar.*

No linear-time algorithm for this problem is known. We will use the solution of the TMPS problem, with some little modifications, for the problem of finding an MPS of a general graph.

Our solution is based on the fact that any triconnected planar graph has a unique planar embedding (Lemma 2.1). Thus an edge (v, w) can be added to a triconnected

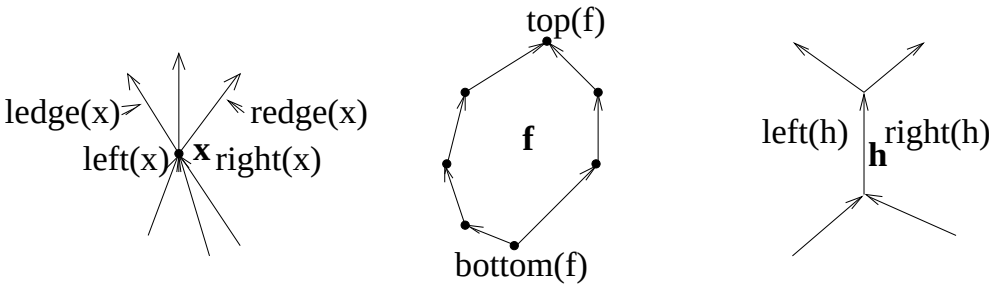


FIG. 3.1. *Illustration of the definition of left, right, top, and bottom for vertices and left and right for edges.*

embedded planar graph so that planarity is preserved, iff v and w belong to the same face of the embedding. To solve the TMPS problem we need a fast procedure that tests whether any arbitrary pair of vertices belong to any face of the embedding. We will use the following method of representing planar embeddings from [6].

3.1. Representation of planar *st*-graphs. Let G be a planar st-graph. For any vertex x, the incoming edges in x appear consecutively around x, and the edges outgoing from x also appear consecutively around x. Thus there is a single face, denoted by $left(x)$, that separates incoming and outgoing edges in a clockwise direction, and there is a single face, $right(x)$, that separates incoming and outgoing edges in a counterclockwise direction (see Figure 3.1). We also record for x a pair of outgoing edges $ledge(x)$ and $redge(x)$ which are incident to x and, respectively, to faces $left(x)$ and $right(x)$. Furthermore, the vertices from the boundary of each face f form two directed paths. The common start vertex of these paths will be denoted by $bottom(f)$, and the common endvertex will be denoted by $top(f)$. Furthermore, if h is an edge of G, we denote by $right(h)$ the face whose clockwise boundary contains h and by $left(h)$ the other face containing h. The values of $left$, $right$, $ledge$, $redge$, $bottom$, and top can be easily computed in linear time and space.

The next lemma concerns testing in a static graph.

LEMMA 3.1. *For any triconnected planar n-vertex graph G there exists a data structure for G that can be constructed in $O(n)$ time, uses $O(n)$ space, and that provides answers in $O(1)$ time to the following two types of queries:*

(a) *If v and w are vertices of G and v has degree $O(1)$, check if v and w belong to the same face of G.*

(b) *If v is a vertex and e is an edge of G, check if v and e belong to the same face of G.*

Queries of type (b) will be used in the algorithms described in section 4.

Proof. Consider first a query of type (a). Since the degree of v is $O(1)$, we can answer the query in a constant time by checking if, for some $f_v \in \{left(v), right(v)\}$ and some $f_w \in \{left(w), right(w)\}$, any of the following cases applies:

$$f_v = f_w; \tag{3.1}$$

$$top(f_v) = w \quad \text{or} \quad bottom(f_v) = w; \tag{3.2}$$

$$top(f_w) = v \quad \text{or} \quad bottom(f_w) = v; \tag{3.3}$$

$$\{v, w\} = \{top(f), bottom(f)\} \quad \text{for some face } f \text{ incident to } v. \tag{3.4}$$

Similarly, for queries of type (b), the problem is reduced to checking whether the following condition is satisfied:

(3.5) There exists $f \in \{left(e), right(e)\}$ such that $f \in \{left(v), right(v)\}$, or $top(f) = v$, or $bottom(f) = v$. □

Note that condition (3.4) is the only condition that we would not be able to check in constant time if there were no restrictions on the degree of v. Our next goal is to show that for solving the TMPS problem, an edge $(v, w) \in E'$ can always be chosen so that the number of faces f incident to either v or w that are "relevant" in certain context to the planarity testing is $O(1)$. The idea of our solution is related to the observation that any n-vertex planar graph G has no more than $3n - 3$ edges and thus there exists a vertex of G of degree less or equal to 6.

Define a graph G_{tb} with vertices $V(G)$ and where the set of edges consists of all pairs $(top(f), bottom(f))$, for each face f of the graph G. Note that G_{tb} is a planar graph (since any edge of G_{tb} can be drawn inside a distinct face of G), and its edges correspond to all pairs of vertices that satisfy condition (3.4) above. We define a graph G'_{tb} to be the subgraph of G_{tb} induced by the set of vertices incident to at least one edge of E'. From the definition of G'_{tb} the next lemma follows.

LEMMA 3.2. *For any edge $(v, w) \in E'$ vertices v and w belong to the same face f of G iff (v, w) is in G'_{tb} or at least one of the conditions* (3.1)–(3.3) *holds. Furthermore, G'_{tb} always contains a vertex of degree at most* 6.

In our algorithm for solving the TMPS problem described below we iteratively choose a new edge e of E' using information about the degrees of G'_{tb}. We add e to G if the planarity of the embedding is preserved, and we update G'_{tb}. The following procedure specifies the details.

ALGORITHM TRICONNECTED.

{Finds a maximal planar subgraph G of $G_{tr} + E'$ if G_{tr} is triconnected.}

1. Set initially $G = G_{tr}$, and construct the data structures of Lemma 3.1 for G.
2. Construct graphs G_{tb} and G'_{tb} for G.
3. For each vertex v of G construct the linked list of the edges of E' incident to v by a lexicographical sort. For each vertex v of G maintain the values $degree_1(v)$ of the number of edges from E' incident to v and $degree_2(v)$ of the degree of v in G'_{tb}. Also maintain a list $SmallDeg$ of all vertices whose degree in G'_{tb} is less than or equal to 6.
4. Repeat until $E' = \emptyset$.
   - 4.1. Pick any vertex x in $SmallDeg$ and choose any edge $(x, y) \in E'$.
   - 4.2. If any of the conditions of Lemma 3.1 is satisfied for $x = v$ and $y = w$, then add (x, y) to G and update variables $left, right, top$, $bottom$, and the graph G'_{tb} for G.
   - 4.3. Remove (x, y) from E' and update G'_{tb} and variables $degree_1, degree_2$, and $SmallDeg$.

The correctness of Algorithm Triconnected follows from Lemmas 2.1 and 3.2. Next we will discuss how the algorithm can be implemented in linear time.

By the definition of G'_{tb} and Lemma 3.2 the list $SmallDeg$ is nonempty iff $E' \neq 0$ and the conditions from step 4.2 of Algorithm Triconnected can be checked in $O(1)$ time. Furthermore, the initial construction of the data structures in steps 1 and 2 requires $O(n)$ time. Updating any of the data structures except $left$ and $right$ takes clearly $O(1)$ time.

Maintaining the $left$ and $right$ relations is more complex, because inserting a new edge in G splits a face f of G into two faces f_1 and f_2 which may require as many as

half of the vertices and edges on f to change their $left$ or $right$ pointer from f to f_1 or f_2. Next we show how to solve this problem making use of the microset technique of Gabow and Tarjan [12].

3.2. A find-split-insert data structure.

3.2.1. Formulation of the problem. We describe a data structure that maintains a partition of a set of edges of a graph G into a set $\mathcal{P}$ of edge-disjoint paths under the following operations:

$find(e)$: Return the label of the path containing edge e within it. (Each path is labeled by a distinct integer.)

$split(v, e)$: Split the path p containing edge $e = (v, w)$ into two paths, the path of the vertices from the start vertex of the path to v and the path from v to the last vertex of the path.

$insert\_vertex(v, (x, y))$: Replace edge (x, y) by edges (x, v) and (v, y) in the path containing (x, y).

$add\_vertex(x, (y, z))$: If edge (y, z) belongs to a path p and y is an endpoint of p, then add a new edge (x, y) to p making x an endpoint of p.

$insert\_edge(v, w, e)$: Transform the path p containing e into two paths, one consisting of the portion of p between v and w and the other containing the remaining two parts of p joined by an additional edge (v, w). Here it is assumed that v and w belong to p and divide it into exactly three nonempty open paths.

$new\_path(x, y)$: Create a new path consisting of a single new edge (x, y). It is assumed that no edge (x, y) belongs to a current path in $\mathcal{P}$.

Each operation will take $O(1)$ amortized time on a random access machine with unit cost measure and $O(\log n)$ bits machine word. The technique we use is similar to the one developed by Gabow and Tarjan [12] for a variation of the set union problem. The same technique was used also by Imai and Asano [18] for the problem of maintaining a partition of a sorted sequence of integers under a sequence of split, find, and insert operations.

3.2.2. The data structure. The edges currently in paths of $\mathcal{P}$ are partitioned into subsets of edges occupying consecutive memory locations called *mezzosets*. Each mezzoset contains at most $ln = \lceil \log n \rceil$ edges so that the total number of mezzosets is $O(n/\log n)$ ($n > 2$ is the total number of edges of G). Each mezzoset is partitioned into smaller blocks called *microsets* of at most $\lambda(n) = \lceil \log \log n \rceil$ edges each so that the total number of microsets is $O(n/\log \log n)$.

Each path is labeled by a distinct integer. Each mezzoset and each microset will have a label that can be a name of a path in $\mathcal{P}$ or nil. A path will consist of a subset of the edges belonging to at most $\lceil n/ln \rceil$ mezzosets with the label of p, at most $\lambda(n)$ additional microsets with the label of p, and at most 2 microsets with label nil. Intuitively, mezzosets and microsets with the label of p will be "internal" for p; i.e., they will contain some edge of p but will contain neither the first nor the last edge of p. (We will make sure that no more than one such path exists for any microset so the labels of microsets are uniquely defined.)

A microset containing only internal edges will be called a *labeled microset*, and a mezzoset containing only identically labeled microsets will be called a *labeled mezzoset*. The other microsets and mezzosets will be called *unlabeled*; see Figure 3.2 for an illustration. The label of any path p is recorded in variables associated with the first and the last edge of p as well.

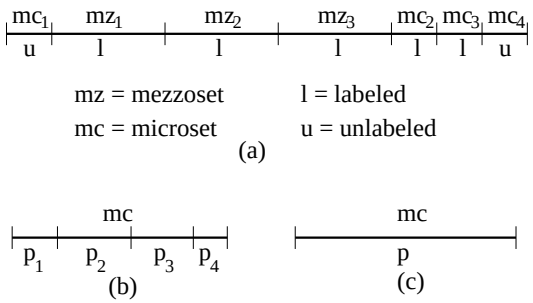


FIG. 3.2. *Examples:* (a) *A (long) path consisting of* 3 *mezzosets and* 4 *microsets.* (b) *An unlabeled microset containing (short) paths* p_2 *and* p_3 *and portions of paths* p_1 *and* p_4. (c) *A labeled microset containing only internal edges of a path* p.

We will use different strategies for maintaining partitions and implementing the query and update operations on levels of (sequences of) labeled mezzosets or microsets and on the level of individual microsets. To execute an update operation involving an edge e, we locate the microset μ_1 and the mezzoset μ_2 containing e. Then we perform the necessary updates on μ_1 using the algorithm for microsets. If μ_1 is unlabeled, we are done. If μ_1 is labeled, we apply the algorithm for labeled microsets on μ_1. Finally, if μ_2 is labeled, we apply the algorithm for labeled mezzosets on μ_2. Next we will describe the algorithms applied in each case.

On levels of labeled mezzosets we use the "relabel the smaller half" technique [12]. By this technique we store with each mezzoset its label and maintain the sequence of labeled mezzosets in any path as a doubly linked list. When we split a path the labels of the labeled mezzosets in the smaller half are updated. For a set of k mezzosets this algorithm requires totally $O(k \log k)$ time for all splits and $O(1)$ time for any query [1]. Since $k = O(n/\log n)$, this yields an $O(n)$ bound on the time for operations on the level of labeled mezzosets. A similar method and the same bound apply for maintaining the labels of the labeled microsets.

For maintaining the partition of the edges in the individual microsets we use the following table lookup method. The edges in each microset are arranged in doubly linked lists corresponding to the partition defined by paths in $\mathcal{P}$. With each edge e of a microset μ we keep the addresses of the previous and the next edge in the list containing e, if any. The addresses are computed relative to the beginning of the memory block corresponding to μ, and therefore each address occupies $\lceil \log \lambda(n) \rceil$ bits. The whole microset can be recorded in one computer word of $O(\lambda(n) \log \lambda(n)) = O(\log n)$ bits.

Next we will explain how this information can be used to answer queries and implement updates in constant time.

Note. The main difference between our implementation and those of Gabow and Tarjan [12] and Imai and Asano [18] is that we maintain linked list data structures within each microset, while in previous cases the data structures maintained are arrays. We need lists in order to implement the *insert_edge* operation.

3.2.3. The *find* operation on microset level. To answer a $find(e)$ query one can follow the backward or forward pointers to the first or the last edge of the path, if any of these edges is in the same microset. Recall that information about the name of the path is associated with the first and the last edge. If both the first and the last edge are contained in other microsets, then the current microset must be a labeled

one, and its label gives the name of the path containing e.

In order to do these computations faster, we use the table lookup method. The idea of the table lookup approach is to precompute the results of all possible $find$ operations for any possible structure of the microset and record them in a table. Due to the small size of the microsets, the preprocessing will take only linear time and space, as illustrated bellow.

We compute a $first$ and $last$ table defined as follows. If e is an edge in a microset μ, if i is the position of e in μ, and if μ is encoded by an integer m, then $first(m,i)$ (resp., $last(m,i)$) denotes the address of the first (resp., last) edge of the path containing e that is also in μ. The number of entries of $first$ is $2^{\lambda(n)\log\lambda(n)}\lambda(n)$, and each entry can be computed in $O(\lambda(n))$ time. Since $\lambda(n) = \lceil \log\log n \rceil$, then

$$2^{\lambda(n)\log\lambda(n)}\lambda(n)^2 = 2^{(\lambda(n)+2)\log\lambda(n)} = 2^{o(\log n)} = o(n).$$

Thus the entire $first$ table occupies $O(n)$ space and can be constructed in $O(n)$ time. Similarly we compute the $last$ table.

3.2.4. The update operations. Note that updating a pointer in the representation of a microset μ is equivalent to changing a digit in the radix $\lambda(n)$ representation of μ. Thus the computer word representing μ can be updated in a constant time if one precomputes the values of $\lambda(n)^k, k = 1, \ldots, \lambda(n) - 1$, and stores them in a table. We implement the update operations on a microset level as follows:

- $split(v,e)$: Let μ be the microset containing e. Splitting along v the doubly linked list representing μ requires updating the corresponding $O(1)$ pointers, which takes a constant time.
- $insert\_vertex(v,(x,y))$: Update the corresponding pointers in the doubly linked list. If as a result of the insertion the size of the microset, μ, becomes larger than $\lambda(n)$, divide μ into two microsets of sizes at most $\lambda(n)/2 + 1$ so that at most one path does not belong to a single microset and compute the encodings of the resulting microsets. The time required for one division is $O(\lambda(n))$. The total time for all divisions is easily shown to be $O(n)$. Similarly divide a mezzoset if its size exceeds ln.
- $add\_vertex(x,(y,z))$: The implementation of this operation is similar to $insert\_vertex$.
- $insert\_edge(v,w,e)$: Recall that for this operation v and w must belong to the same path p determined by e. If v and w belong to the same microset, then just update the corresponding pointers in $O(1)$ time. If v and w belong to different microsets, then split p at v and w by changing the pointers. Add the edge (v,w) to the corresponding path in one of the microsets. If that microset becomes too large, divide it into two microsets. Recall that the doubly linked lists of labeled microsets and mezzosets representing the two resulting paths are updated at the corresponding higher levels (levels of labeled microsets and mezzosets).
- $new\_path(x,y)$: Choose an arbitrary microset μ and add the path consisting of the edge $e = (x,y)$ to μ. This is done by adding in μ the $next$ and $previous$ pointers from e to itself and storing with e a pointer to its location in microset μ.

We need also to explain how the above update operations affect the labels of mezzosets and microsets. Recall that a microset or a mezzoset is labeled if for some path it contains only internal edges of that path. Consider a *split* in a microset μ.

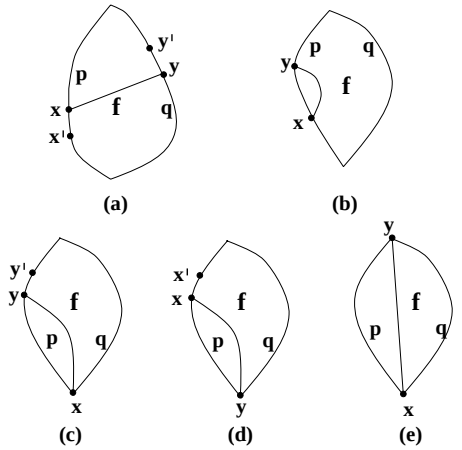


FIG. 3.3. *Cases for inserting edge* (x, y) *in* f.

After the operation the labels of μ and the mezzoset containing μ are set to *nil*. After an *insert_vertex* in μ, if μ has not been divided because of its larger size, then no label is changed. If the microset is divided, we have the following two cases: If the original microset has a non-nil label, then copy the same label to both resulting microsets. If the original microset has a *nil* label, label any of the resulting microsets that consists of edges from only one path p but contains no endpoint of p (if there is such a microset) with the label of p, or otherwise keep the *nil* label. A similar rule applies when a mezzoset is divided into two mezzosets because of a larger size. *Add_vertex* operation is similar to *insert_vertex*. Finally, for *insert_edge*, if the microset and mezzoset are not divided because of a larger size, we keep the original label (or *nil*). If any of them is divided, we assign labels to the resulting microsets using the rules described for *insert_vertex*.

We summarize the main result from this subsection in the following theorem.

THEOREM 1. *Any sequence consisting of at most* k *insert_vertex, add_vertex, insert_edge, and new_path operations and at most* l *find and split operations on a set of edge-disjoint paths* $\mathcal{P}$ *can be implemented in* $O(k + l + m)$ *time and* $O(k + m)$ *space, where* m *is the original number of edges in the paths of* $\mathcal{P}$.

3.2.5. Maintaining the *left* and *right* relations. Now let us consider the original problem of maintaining the $left$ and $right$ relations in Algorithm Triconnected. Let us consider the $right$ relation for vertices (the $right$ relation for edges and the $left$ relations are maintained similarly). We maintain the set of paths $\mathcal{P}$ representing the original triconnected planar st-graph G. Specifically, for any face f, we keep in $\mathcal{P}$ the simple directed path p of all edges e such that $right(e) = f$. We associate with p a variable containing the name of f. Denote by q the other directed path on f. Let $x \in p$ and y be two vertices on f. The addition of a new edge $e = (x, y)$ to the embedding can be implemented as a sequence of $O(1)$ of the above type operations on paths in $\mathcal{P}$ depending on the locations of x and y in p and q (see Figure 3.3). Consider the following cases. By x' and y' we denote two vertices adjacent, respectively, to x and y, as denoted on Figure 3.3.

(a) $x \notin q$, $y \in q$, $y \notin p$ (Figure 3.3(a)). Then we add e to the embedding by performing the following sequence of operations: $split(x, (x, x'))$, $add\_vertex(y, (x, x'))$.

(b) $x \notin q$, $y \in p$, $y \notin q$ (Figure 3.3(b)). Then $add\_edge(x, y)$.
(c) $x \in q, y \in p$, $y \notin q$ (Figure 3.3(c)). Then $split(y, (y, y'))$, $add\_vertex(x, (y, y'))$.
(d) $x \notin q$, $y \in p$, $y \in q$ (Figure 3.3(d)). Then $split(x, (x, x'))$, $add\_vertex(y, (x, x'))$.
(e) $x \in p, q$, $y \in p, q$ (Figure 3.3(e)). In this case we just do $new\_path(x, y)$.

In all cases (a)–(e) we update, if necessary, the variables associated with each of the resulting paths that contain the name of the adjacent face to the right of the path.

For implementing $right$ queries we use the variable $redge(x)$ defined for any vertex x (see Figure 3.1). Then $right(x)$ is the face associated with path $find(redge(x))$ and can be found in $O(1)$ time.

We showed that adding an edge requires $O(1)$ time, and any $left$ or $right$ query requires $O(1)$ time. Since any planar n-vertex graph has $O(n)$ edges, the total number of edge additions will be $O(n)$. Clearly, the number of $find$ operations is $O(|E'|)$. Thus the total time for all operations connected with maintaining $left$ and $right$ relations for vertices will be $O(n + m)$.

Note that we did not use $insert\_vertex$ for maintaining $left$ and $right$ relations. We will use this operation in our algorithms given in the next sections.

This concludes our discussion of the triconnected case. We proved the following theorem.

THEOREM 2. *Let G be a planar triconnected n-vertex graph and E' be a set of m edges between vertices of G. A maximal set $E'' \subset E'$ such that $G + E''$ is planar can be constructed in $O(n + m)$ time.*

4. Finding a maximal planar subgraph of an arbitrary graph.

4.1. Outline of the algorithm. Our algorithm uses the decomposition tree described in section 2 to represent the decomposition of the current planar subgraph. Recall that we can assume w.l.o.g. that the input graph is connected, because otherwise we can apply the same algorithm to each connected component. For maintaining the embeddings of skeletons and for answering queries at each node of a SPQR tree we use a procedure similar to Algorithm Triconnected. At each iteration the algorithm chooses a new edge and checks if it is possible to add it to the subgraph so that planarity is preserved. The efficiency of our algorithm essentially depends on the order in which the edges are tested for insertion in the subgraph. Another feature of the algorithm is that it maintains a dynamic set $Upaths$ of paths in the decomposition tree called *update paths* which will be our "working" paths; i.e., all information we could currently need will be associated with nodes in these paths, and all updates will be done on nodes in paths from $Upaths$. By using properties of these paths we will be able to make queries and do updates more efficiently.

ALGORITHM MAXPLANAR.
(Outline)
Input: A connected n-vertex m-edge graph G.
Output: A maximal planar subgraph G' of G.

1. Construct a depth-first spanning tree T of G. Associate a BC tree B with T whose root is the root of T. Let $E' = E(G) - E(T)$, $E^* = E(T)$, where E' denotes the set of edges of G not examined yet and E^* denotes the set of edges of the current approximation of the maximal planar subgraph.
2. Initialize for the skeleton of each level-2 node of B the data structures for online planarity testing from steps 1, 2, and 3 of Algorithm Triconnected.
3. Use a variation of a postorder search to visit the nodes of B. Denote the current level-1 node by λ^* and the current level-2 node by μ^*.

{*Comment:* The postorder will guarantee that any level-2 node that either is a descendant of μ^* or belongs to a level-1 node which is a proper descendant of λ^* is marked. (A level-2 node μ will be marked if no vertex of $skeleton(\mu)$ is incident to any edge from E'.)}

3.1. Update $Upaths$ and the associated data structures if a new node has been examined in the previous step (details to be given below.)

3.2. Pick any vertex x of the skeleton of μ^* belonging to the $SmallDeg$ list of μ^*.

3.3. Pick an edge $(x, y) \in E'$ and update $E' := E' - \{(x, y)\}$. If no vertex remains in $skeleton(\mu^*)$ that is incident to an edge of E', then mark μ^*. Denote by ν^* the proper allocation node $proper(y)$ of y. Check if (x, y) can be added to E^* by considering the following cases:

3.3.1. If $\nu^* = \mu^*$, then test whether y belongs to any of the faces incident to x by checking whether either μ^* is a P, S, or Q node, or the conditions of Lemma 3.2 for $v = x$ and $w = y$ are satisfied. If the answer is "yes," then add (x, y) to E^* and update the data structures associated with μ^*.

3.3.2. If ν^* is a proper ancestor of μ^*, then let edge d be the projection $pr(x)$ of x on the skeleton of ν^*. By Lemma 2.2, we have to check if y and d belong to the same face of $skeleton(\nu^*)$ and if all nodes on the joining path $p(x, y)$ are peripheral. If ν^* is a P, S, or Q node or if condition (3.5) of the proof of Lemma 3.1 is satisfied for $v = y$ and $e = d$, then y and d belong to the same face of the skeleton of ν^*. Using information stored in $Upaths$ we decide whether all nodes on $p(x, y)$ are peripheral. If both queries give positive answers, then we add (x, y) to E^* and update the data structures, as will be described in the next subsection.

3.3.3. If ν^* belongs to a level-1 node κ^* that is a proper ancestor of λ^*, then, by Lemma 2.3, we perform on the SPQR trees corresponding to κ^* and λ^* two test operations of the type described in step 3.3.2 and (using precomputed information stored with $Upaths$) also check if all other level-1 nodes on the tree path π^* between κ^* and λ^* are peripheral. If all answers are positive, we add (x, y) to E^* and merge into the SPQR tree of κ^* the SPQR trees of all other level-1 nodes of π^*.

Next we give more details about the implementation of some of the steps. The search of B in step 3 is essentially a postorder search applied at two levels: first with respect to level-1 nodes and then, when the level-1 node is chosen, with respect to the level-2 nodes in its SPQR tree. Another feature is that the search is applied online to dynamic trees. Thus, if in steps 3.3.2 or 3.3.3 a path of two or more level-1 or level-2 nodes has been shrunk, then the resulting node, μ, might have unmarked children (which must be visited now before continuing with μ), even if μ^* may have had no marked children before the shrinking.

In the next subsection we describe the data structures associated with $Upaths$ that will allow us to determine in a constant time whether a certain path of level-1 or level-2 nodes contains only peripheral nodes. We will also show that it is possible to do all updates on B and on the planar st-graphs associated with its nodes after shrinking of paths of level-1 and level-2 nodes in $O(n)$ total time.

4.2. Updating the data structures. Our update algorithms are simpler and

more efficient than the algorithms of [6, 25, 24] because of our use of *Upaths*.

4.2.1. The update paths. *Upaths* are a set of paths called *update paths* that includes the current path π of level-1 nodes of B from λ^* to the root of the tree and a path of level-2 nodes in the SPQR tree of each node in π. Let $\pi = \{b_1, c_1, b_2, c_2, \ldots, c_{k-1}, b_k\}$, where the path of level-2 nodes in node $b_1 = \lambda^*$ is called the *top update path* and b_k is the root of B. For λ^*, the corresponding path of level-2 nodes is from μ^* to the root of the SPQR tree of λ^*. The update path in b_i, for $2 \leq i \leq k$, is the path from the allocation node of c_{i-1} to the root of the SPQR tree for b_i. At any iteration we will use only π and the top update path. The other paths in *Upaths* will be needed when we backtrack to an ancestor node during the postorder search. We prove the following important property of *Upaths*.

LEMMA 4.1. *Let x be a vertex in μ^* and $(x, y) \in E'$ be the edge chosen in step 3.3 of Algorithm Maxplanar. Then the proper allocation node proper(y) of y belongs to the update path of λ^* or to the SPQR tree of a node in π different from λ^*.*

Proof. Because the tree T constructed in step 1 is a depth-first spanning tree, then any nontree edge joins two vertices of T, one of which is a descendant to the other. By the definition of a BC tree, such a property will also hold for the tree B constructed in step 1 of Algorithm Maxplanar; i.e., any nontree edge with respect to T joins vertices from two level-1 nodes, one of which is a descendant of the other. For any of the following iterations, we prove that the endpoints of any edge not in B have allocation nodes such that either one of them is an ancestor of the other, or they belong to the SPQR trees of level-1 nodes one of which is a proper ancestor of the other. This follows by induction from the fact that each modification of B consists of shrinking a path of either level-1 or level-2 nodes to a single node.

Assume that *proper*(x) and *proper*(y) belong to the same SPQR tree. Then, by the above observation, either μ^* is a descendant of *proper*(y), or vice versa. However, since the nodes of B are visited in postorder, all descendants of μ^* have already been visited and marked. Hence *proper*(y) is an ancestor of μ^*, and it belongs to the update path of λ^* (the top update path).

The proof in the case when *proper*(x) and *proper*(y) belong to different SPQR trees is similar. In this case the node corresponding to the SPQR tree containing *proper*(y) is a proper ancestor of λ^*, and by definition that node belongs to π. □

The update paths change during the computation when a new node of B is visited and when a subpath of B is contracted. This requires that some information associated with *Upaths* be dynamically updated, as described below.

4.2.2. Update algorithms. We need to dynamically maintain the following types of information:

- the proper allocation nodes of the vertices of G;
- for any level-1 or level-2 node μ, the nearest ancestor of μ that is not peripheral;
- a planar embedding of a triconnected planar st-graph G' (a skeleton of a level-2 R node ν) with respect to the structures described in section 3, subject to the operation replacement of an edge of G' with a planar st-graph (the skeleton of a child of ν.)

1. *Proper allocation nodes.* Information about the proper allocation nodes of the vertices of G is needed in step 3.3 of Algorithm Maxplanar. We store for each vertex v of G a pointer to the representative of v in the skeleton of *proper*(v). Furthermore, we dynamically maintain for each level-2 node μ belonging to a path in *Upaths* a set of all vertices x of G such that *proper*$(x) = \mu$. This can be done by using the

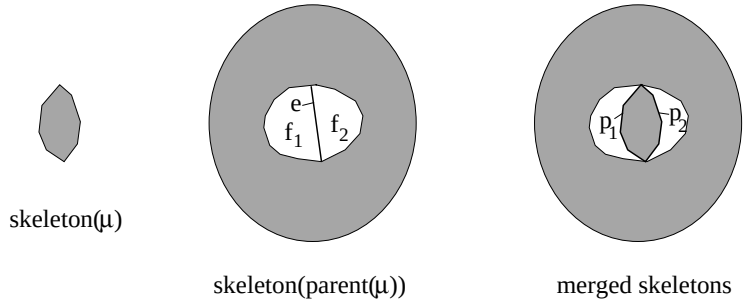


FIG. 4.1. *Merging a skeleton of a level-2 node into the skeleton of its parent.*

linear-time incremental set union algorithm of Gabow and Tarjan [12]. Hence one can find the proper allocation node μ of any vertex of G in $O(1)$ amortized time, provided that μ belongs to a path of $Upaths$. By the description of Algorithm Maxplanar and Lemma 4.1, the proper allocation nodes of the vertices x and y examined in step 3.3 will always belong to a path of $Upaths$.

2. *Nonperipheral ancestors.* For any level-2 node ν of a path from $Upaths$ we store in a variable $nearest(\nu)$ the value of the nearest nonperipheral ancestor of ν. Whenever a new level-2 node μ is examined in step 3 of Algorithm Maxplanar, we add μ to the top update path and check if μ is peripheral. Depending on the value of *nearest* for the parent of μ (if μ is not the root of B), the value of $nearest(\mu)$ is determined. Note that when a subpath of level-2 nodes is contracted, that subpath always includes the endvertex (i.e., the most recently added vertex) of the top update path. Thus in this case at most one value of *nearest* needs to be updated which takes $O(1)$ time. The information about the nearest nonperipheral ancestors of level-1 nodes is maintained in a similar way.

3. *Merging skeletons.* Merging the skeleton of a level-2 node μ of B with the skeleton of its parent $parent(\mu)$ requires a replacement of an edge e of a planar st-graph G_e (the skeleton of $parent(\mu)$) with another st-graph (the skeleton of μ.)

Consider the case where both μ and $parent(\mu)$ are R nodes (the nontrivial case) and e is an internal edge. The set of the faces of the resulting planar st-graph is the union of all internal faces of the skeleton of $parent(\mu)$ and all faces of the skeleton of μ, where the two faces, say f_1 and f_2, incident to e, are modified as follows. In each of f_1 and f_2 we replace e by a path using a sequence of $insert\_vertex$ operations. Let p_1 and p_2 be the resulting paths (Figure 4.1). The time needed for these $insert\_vertex$ operations will be proportional to the sum of the lengths of p_1 and p_2. However, for the whole execution of the algorithm, the time needed for such insertions will be $O(n)$, since any edge of p_1 or p_2 becomes internal and cannot be inserted again.

If e is on the periphery of G_e, we apply the same algorithm. But now the above argument for bounding the number of edges that need to be inserted does not apply because some edges may need to be reinserted more than once. To handle this case we modify our data structure for maintaining the planar embeddings of the skeletons of level-2 nodes by using the same label *outer* to denote the outer face of *any* skeleton of a level-2 node. For example, if x is a vertex on the left boundary path of such a skeleton μ incident to an internal face f, then $right(x) = f$ and $left(x) = outer$. Thus, the value of $left(x)$ does not need to be changed if x becomes a boundary vertex of another skeleton as a result of a merge.

Testing if the addition of an edge (x, y) preserves planarity requires checking whether $left(x) = left(y) = outer$ and whether x and y have the same proper allocation node.

If x becomes an internal edge of a skeleton after some merge, then we set $left(x)$ to the (actual) internal face containing x and maintain its value using the original algorithm.

We can summarize our main result as follows.

THEOREM 3. *Given any n-vertex m-edge graph G, a maximal planar subgraph of G can be found in $O(n + m)$ time.*

5. Conclusion. We can also adapt our technique to find a maximal *outerplanar* subgraph of an n-vertex graph. Create an additional vertex z, and join z to all vertices of G. Then find a maximal planar subgraph of the resulting graph by a procedure similar to Algorithm Maxplanar; however, the initial tree constructed in step 1 is the star spanning graph with root z. This guarantees that the maximal planar graph constructed by the modified algorithm will contain all edges incident to z. Removing at the end z and all incident edges clearly will result in a maximal outerplanar graph. We need to show that the time complexity of this algorithm is still $O(n+m)$, since our initial subgraph is not a depth-first tree as in Algorithm Maxplanar, and the analysis of the new algorithm (e.g., Lemma 4.1) cannot be directly applied. In this case, however, we do not need to use the update paths since each level-1 or level-2 node can have at most one ancestor and condition (ii) of Lemma 2.2 and the condition from Lemma 2.3 can be directly checked in a constant time. Thus we have the following theorem.

THEOREM 4. *Given any n-vertex m-edge graph G, a maximal outerplanar subgraph of G can be found in $O(n + m)$ time.*

Also we note that our linear-time algorithm for the MPS problem yields a linear-time algorithm for planarity testing. Given an n-vertex m-edge graph G, we can test in $O(n + m)$ time whether G is planar by finding a maximal planar subgraph G' of G. Then G is planar iff $G = G'$. This result is interesting because the new algorithm is based on an approach entirely different from the existing ones. The linear-time algorithms of Hopcroft and Tarjan [16] and Booth and Lueker [2] (and their modifications) essentially use the Jordan Curve Theorem which states that any closed curve in the plane divides it into exactly two connected regions. In contrast, our algorithm is based on the uniqueness of the planar embedding of any triconnected planar graph. It will be of theoretical and practical interest to refine our technique in order to construct a new practical algorithm for planarity testing whose performance is comparable to the algorithms of [16] and [2].

As another approach to the graph planarization problem, other researchers have constructed approximation algorithms for the maximum planar subgraph problem. The algorithm in [3] constructs in $O(m^{3/2} n \log^6 n)$ time a "maximum triangular structure," a planar graph whose bicomponents are single edges or triangles, and prove approximation ratio 2/5. Although the approximation ratio corresponding to a maximal planar subgraph in the worst case cannot be proved to be better than 1/3, it seems that in most cases our algorithm produces larger subgraphs, e.g., for planar or almost planar graphs, for sparse graphs (e.g., with less than $(5/2)n - 5$ edges), for any bipartite graphs. For practical purposes, probably the best algorithm for constructing large planar subgraphs will be a combination of both approaches: first a planar subgraph is constructed by the approximation algorithm guaranteeing a good approximation ratio, and then the subgraph is augmented to a maximal planar subgraph.

Acknowledgment. The author would like to thank the anonymous referees for their helpful comments.

REFERENCES

[1] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[2] K. Booth and G. Lueker, *Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithm*, J. Comput. System Sci., 13 (1976), pp. 335–379.

[3] G. Călinescu, C. G. Fernandes, U. Finkler, and H. Karloff, *A better approximation algorithm for finding planar subgraphs*, in Proceedings of the 7th Annual ACM-SIAM Symposium on Discrete Algorithms, Atlanta, GA, 1996, SIAM, Philadelphia, pp. 16–25.

[4] J. Cai, X. Han, and R. E. Tarjan, *An O(m log n)-time algorithm for the maximal planar subgraph*, SIAM J. Comput., 22 (1993), pp. 1142–1162.

[5] T. Chiba, I. Nishioka, and I. Shirakawa, *An algorithm of maximal planarization of graphs*, in Proceedings of the IEEE International Symposium on Circuits and Systems, 1979, IEEE Press, Piscataway, NJ, pp. 649–652.

[6] G. Di Battista and R. Tamassia, *Incremental planarity testing*, in Proceedings of the IEEE Symposium on Foundations of Computer Science, 1989, IEEE Press, Piscataway, NJ, pp. 436–441.

[7] G. Di Battista and R. Tamassia, *On-line graph algorithms with SPQR trees*, in Proceedings of the International Colloquium on Automata, Languages and Programming, Warwick, UK, Lecture Notes in Comput. Sci. 443, Springer-Verlag, New York, 1990, pp. 598–611.

[8] H. Djidjev, *A Linear algorithm for the maximal planar subgraph problem*, in Proceedings of WADS'95, Kingston, ON, Lecture Notes in Comput. Sci. 955, Springer-Verlag, Berlin, 1995, pp. 369–380.

[9] H. N. Djidjev, *On some properties of nonplanar graphs*, C.R. Acad. Bulgare Sci., 37 (1984), pp. 1183–1184.

[10] H. N. Djidjev and J. Reif, *An efficient algorithm for the genus problem with explicit construction of forbidden subgraphs*, in Proceedings of the Annual ACM Symposium on Theory of Computing, 1991, ACM, New York, pp. 337–347.

[11] S. Even and R. E. Tarjan, *Computing an st-numbering*, Theoret. Comput. Sci., 2 (1976), pp. 339–344.

[12] H. Gabow and R. E. Tarjan, *A linear-time algorithm for a special case of disjoint set union*, J. Comput. System Sci., 30 (1985), pp. 209–221.

[13] M. R. Garey and D. S. Johnson, *Algorithms and Intractability: A Guide to the Theory of NP Completeness*, Freeman, San Francisco, 1979.

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

[15] J. E. Hopcroft and R. E. Tarjan, *Dividing a graph into triconnected components*, SIAM J. Comput., 2 (1973), pp. 135–158.

[16] J. E. Hopcroft and R. E. Tarjan, *Efficient planarity testing*, J. ACM, 21 (1974), pp. 549–568.

[17] W.-L. Hsu, *A linear time algorithm for finding maximal planar subgraphs*, in ISAAC'95, Lecture Notes in Comput. Sci. 1004, Springer-Verlag, Berlin, 1995, pp. 352–361.

[18] H. Imai and T. Asano, *Dynamic orthogonal segment intersection search*, J. Algorithms, 8 (1987), pp. 1–18.

[19] R. Jayakumar, K. Thulasiraman, and M. N. S. Swamy, *$O(n^2)$ algorithms for graph planarization*, in Lecture Notes in Comput. Sci. 344, Springer-Verlag, Berlin, 1989, pp. 352–377.

[20] M. Marek-Sadowska, *Planarization algorithm for integrated circuits engineering*, in Proceedings of the IEEE International Symposium on Circuits and Systems, 1979, IEEE Press, Piscataway, NJ, pp. 919–923.

[21] T. Nishizeki and N. Chiba, *Planar Graphs: Theory and Algorithms*, North–Holland, Amsterdam, 1988.

[22] T. Ozawa and H. Takahashi, *A graph-planarization algorithm and its applications to random graphs*, in Graph Theory and Algorithms, Lecture Notes in Comput. Sci. 108, Springer-Verlag, London, 1981, pp. 95–107.

[23] K. Pasedach, *Criterion and algorithms for determination of bipartite subgraphs and their application to planarization of graphs*, in Graphen-Sprachen und Algorithmen in Graphen, Hanser, Munich, 1976, pp. 175–183.

[24] J. A. La Poutré, *Alpha-algorithms for incremental planarity testing*, in Proceedings of the Annual ACM Symposium on Theory of Computing, 1994, ACM, New York, pp. 706–715.
[25] J. Westbrook, *Fast incremental planarity testing*, in Proceedings of the International Colloquium on Automata, Languages, and Programming, Lecture Notes in Comput. Sci. 623, Springer-Verlag, Berlin, 1992, pp. 342–353.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 463–501

SPARSE SOURCEWISE AND PAIRWISE DISTANCE PRESERVERS*

DON COPPERSMITH† AND MICHAEL ELKIN‡

Abstract. We introduce and study the notions of *pairwise* and *sourcewise preservers*. Given an undirected N-vertex graph $G = (V, E)$ and a set $\mathcal{P}$ of pairs of vertices, let $G' = (V, H)$, $H \subseteq E$, be called a *pairwise preserver of G with respect to $\mathcal{P}$* if for every pair $\{u, w\} \in \mathcal{P}$, $dist_{G'}(u, w) = dist_G(u, w)$. For a set $\mathcal{S} \subseteq V$ of *sources*, a pairwise preserver of G with respect to the set of all pairs $\mathcal{P} = \binom{\mathcal{S}}{2}$ of sources is called a *sourcewise preserver of G with respect to $\mathcal{S}$*. We prove that for every undirected possibly weighted N-vertex graph G and every set $\mathcal{P}$ of $P = O(N^{1/2})$ pairs of vertices of G, there exists a *linear-size* pairwise preserver of G with respect to $\mathcal{P}$. Consequently, for every subset $\mathcal{S} \subseteq V$ of $S = O(N^{1/4})$ sources, there exists a *linear-size* sourcewise preserver of G with respect to $\mathcal{S}$. On the negative side we show that neither of the two exponents (1/2 and 1/4) can be improved even when the attention is restricted to unweighted graphs. Our lower bounds involve constructions of dense convexly independent sets of vectors with small Euclidean norms. We believe that the link between the areas of *discrete geometry* and *spanners* that we establish is of independent interest and might be useful in the study of other problems in the area of low-distortion embeddings.

Key words. graph theory, spanners, distance preservation

AMS subject classifications. 05C12, 05C85, 68R05

DOI. 10.1137/050630696

1. Introduction.

For a graph $G = (V, E)$, its sparse subgraph $G' = (V, H)$, $H \subseteq E$, is called a *spanner* of G if the metric space that is defined by G' is close in some respect to the metric space that is defined by G.

Graph spanners were introduced in a pioneering paper of Peleg and Schäffer [26], and since then have been used as an underlying combinatorial structure for many applications, mostly in the areas of graph algorithms and distributed computing. Among the most prominent applications of spanners are algorithms for computing almost shortest paths [3, 14, 17], routing algorithms [27, 5, 28], and algorithms for constructing synchronizers [4, 6] and for network design [23]. There are also indirect applications for distance labeling and distance oracles [25, 19, 29]. Significant research efforts were also invested in the problem of devising *efficient* algorithms for *constructing* spanners [3, 14, 16, 17, 10]. Despite this extensive study of the *algorithmic aspects* of spanners, so far relatively little attention has been devoted to their *combinatorial properties*. The study of these properties is the subject of the current paper.

The first result of this kind was due to Peleg and Schäffer, who have shown [26] that for any unweighted undirected N-vertex graph $G = (V, E)$, and a positive integer parameter $\kappa = 1, 2, \ldots$, there exists a subgraph $G' = (V, H)$, $H \subseteq E$, with $N^{1+O(1/\kappa)}$ edges satisfying that for every pair of vertices $u, w \in V$, the distance between them in G', denoted $dist_{G'}(u, w)$, is at most κ times greater than the distance between them in G, $dist_G(u, w)$. It was also shown in [26] that this trade-off is optimal up

*Received by the editors May 5, 2005; accepted for publication (in revised form) January 30, 2006; published electronically June 2, 2006. A preliminary version of this paper appeared in *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms* [15]. This work was supported by the DoD University Research Initiative (URI) administered by the Office of Naval Research under grant N00014-01-1-0795.

http://www.siam.org/journals/sidma/20-2/63069.html

†IBM Research, Yorktown Heights, NY 10598 (dcopper@idaccr.org).

‡Department of Computer Science, P.O.B. 653, Ben-Gurion University of Negev, Beer-Sheva, 84105, Israel (elkin@cs.bgu.ac.il).

to the constants hidden by the O-notation. The next result of this flavor was due to Dor, Halperin, and Zwick [16], who have shown that for every unweighted undirected N-vertex graph $G = (V, E)$ there exists a subgraph $G' = (V, H)$, $H \subseteq E$, with $O(N^{3/2} \cdot \log N)$ edges such that for every pair of vertices $u, w \in V$, $dist_{G'}(u, w) \leq dist_G(u, w) + 2$ (such a subgraph is called an *additive* 2-*spanner*). A lower bound of $\Omega(N^{3/2})$ on the size of an additive 2-spanner follows directly from [26]. The gap of $\log N$ was closed (by improving the upper bound) in [18]. Very recently Baswana et al. [9] proved the existence of additive 6-spanners with $O(n^{4/3})$ edges for every (undirected unweighted) graph.

Further, Elkin and Peleg [18] have shown that for every $\epsilon > 0$, $\kappa = 1, 2, \ldots$, there exists $\beta = \beta(\epsilon, \kappa)$ such that for every unweighted undirected N-vertex graph $G = (V, E)$ there exists a subgraph $G' = (V, H)$ with $O(N^{1+1/\kappa})$ edges such that for every pair of vertices $u, w \in V$, $dist_{G'}(u, w) \leq (1 + \epsilon) \cdot dist_G(u, w) + \beta$.

Finally, recently Bollobas, Coppersmith, and Elkin [12] have shown that for every unweighted undirected N-vertex graph $G = (V, E)$, and a positive integer parameter D, there exists a subgraph $G' = (V, H)$ with $|H| = O(N^2/D)$ edges that preserves all the distances between pairs of vertices that are at a distance of D or more from one another in the graph G. It is also shown in [12] that this result is optimal up to a small constant factor.

In this paper we continue this line of research and introduce the following notions. Given an N-vertex graph $G = (V, E)$ and a set $\mathcal{P}$ of pairs of vertices, let $G' = (V, H)$, $H \subseteq E$, be called a *pairwise preserver of* G *with respect to* $\mathcal{P}$ if for every pair $\{u, w\} \in \mathcal{P}$, $dist_{G'}(u, w) = dist_G(u, w)$. For a set $\mathcal{S} \subseteq V$ of vertices, called *sources*, a pairwise preserver of G with respect to the set of all pairs $\mathcal{P} = \binom{\mathcal{S}}{2}$ of sources is called a *sourcewise preserver of* G *with respect to* $\mathcal{S}$. We prove the following results.

THEOREM 1.1.

1. *For every undirected weighted or unweighted N-vertex graph $G = (V, E)$, and a set $\mathcal{P}$ of $P = O(N^{1/2})$ pairs of vertices of G, there exists a* linear-size *pairwise preserver $G' = (V, H)$, $H \subseteq E$, of G with respect to the set $\mathcal{P}$.*
   Consequently, for every graph G and subset $\mathcal{S} \subseteq V$ of $S = O(N^{1/4})$ vertices, there exists a linear-size *sourcewise preserver $G' = (V, H)$, $H \subseteq E$, of G with respect to the set $\mathcal{S}$.*
2. *For every value α, $1/2 < \alpha < 2$, for every sufficiently large N and $P = \Theta(N^\alpha)$ there exists an unweighted undirected N-vertex graph $G = (V, E)$, and a set $\mathcal{P}$ of P pairs of vertices of G such that any pairwise preserver $G' = (V, H)$, $H \subseteq E$, of G with respect to the set $\mathcal{P}$ contains $\omega(N + P)$ edges.*
3. *For every value α, $1/4 < \alpha < 9/16$, for every sufficiently large N and $S = \Theta(N^\alpha)$ there exists an unweighted undirected N-vertex graph $G = (V, E)$, and a subset $\mathcal{S} \subseteq V$ of S sources such that any sourcewise preserver $G' = (V, H)$, $H \subseteq E$, of G with respect to the set $\mathcal{S}$ contains $\omega(N + S^2)$ edges.*

Note that the lower bounds (assertions 2 and 3) apply even for *unweighted undirected* graphs.

We remark that these results are special cases of the, far more general, theorems that we will prove. The detailed exposition of the latter theorems is deferred to section 1.2.

1.1. Motivation. We believe that the problem of constructing sparse pairwise and sourcewise preservers is an important basic combinatorial problem. Our results on it enable one to gain deep insight into the metric properties of graphs. Further, we believe that the ultimate resolution of this problem is essential for resolving other

major open problems in the area of spanners, such as the question of existence or nonexistence of sparse additive spanners (see [16, 12]).

From a broader perspective, spanners are currently widely recognized as one of the topics in the area of low-distortion embeddings (see, e.g., the section on spanners in the recent survey paper by Indyk and Matousek [20]). The latter area is currently one of the most intensively studied subdisciplines of theoretical computer science. For many fundamental existential results in this area (such as Bourgain's embeddings [13], Johnson–Lindenstrauss dimension reduction [22], average-stretch tree embeddings of Alon et al. [1]) there were found multiple important algorithmic applications, sometimes many years after these existential results were proven. We believe that our research of pairwise and sourcewise preservers will also bear algorithmic fruits.

We remark that in this paper we do not explore the *algorithmic* potential of the pairwise and sourcewise preservers, but instead focus on their *combinatorial* properties. We hope that their potential will be fully explored in subsequent work. Another promising direction that we did not study is the *approximate* variants of the pairwise and sourcewise preservers. In our opinion it is very likely that these approximate variants will be useful in the design of improved algorithms for fast distance estimation. However, we feel that the study of these approximate variants would be premature before gaining a thorough understanding of the *exact* variants of these problems. We believe that this paper is a major step towards achieving such an understanding.

1.2. Our results. The specific behavior of our lower bound on the size of pairwise preservers for unweighted undirected graphs is parameterized by the "dimension" parameter d and has the following form: for $d = 2, 3, \ldots$, for $\Omega(N^{2\cdot\frac{d^2-d-1}{(d-1)(d+2)}}) = P = O(N^{2\cdot\frac{d^2+d-1}{d(d+3)}})$, the lower bound is $|H| = \Omega(N^{\frac{2d}{d^2+1}} \cdot P^{\frac{d(d-1)}{d^2+1}})$. (Observe that if we denote $f(d) = 2 \cdot \frac{d^2-d-1}{(d-1)(d+2)}$, then the condition on P is of the form $\Omega(N^{f(d)}) = P = O(N^{f(d+1)})$.) Note that this result directly implies assertion 2 of Theorem 1.1; that is, this lower bound is *superlinear in* $N + P$ for the entire feasible range of P, $\omega(\sqrt{N}) = P = o(N^2)$. For undirected *weighted* graphs we show an even stronger lower bound of $|H| = \Omega((N \cdot P)^{2/3})$, and this lower bound is also *superlinear in* $N + P$ in the same range of P.

We also show that there are unweighted undirected N-vertex graphs $G = (V, E)$, and subsets $\mathcal{S} \subseteq V$ of S vertices, such that any *sourcewise* preserver $G' = (V, H)$ of G with respect to $\mathcal{S}$ contains $|H| = \Omega(\max\{N^{9/11} \cdot S^{6/11}, N^{10/11} \cdot S^{4/11}\})$ edges. This lower bound is *superlinear in* $N + S^2$ for $\omega(N^{1/4}) = S = o(N^{9/16})$ (i.e., it implies assertion 3 of Theorem 1.1). This result cannot be extended for $S = O(N^{1/4})$ in view of our upper bound of Theorem 1.1(1). For undirected *weighted* graphs we show a slightly stronger lower bound of $|H| = \Omega(N^{6/7} \cdot S^{4/7})$. This lower bound is *superlinear in* $N + S^2$ in a slightly wider range $\omega(N^{1/4}) = S = o(N^{3/5})$.

Finally, we show two upper bounds. First, we show that for every undirected possibly weighted N-vertex graph $G = (V, E)$, and a set $\mathcal{P}$ of P pairs of vertices, there exists a pairwise preserver $G' = (V, H)$ of G with respect to $\mathcal{P}$ with $|H| = O(N + \sqrt{N} \cdot P)$ edges. Note that this upper bound implies assertion 1 of Theorem 1.1. Second, we show an analogous upper bound of $|H| = O(\sqrt{P} \cdot N)$ that applies even to the most general case of weighted directed graphs. See Tables 1 and 2, and Figures 1 and 2, for summaries of these results.

Note that our results do not rule out the possibility that for any unweighted undirected N-vertex graph $G = (V, E)$ and any subset $\mathcal{S}$ of $S = \Omega(N^{9/16})$ vertices, there exists a sourcewise preserver $G' = (V, H)$, $H \subseteq E$, of G with respect to $\mathcal{S}$ with

TABLE 1

A summary of our results on pairwise preservers. The columns correspond to different types of graphs. The upper (resp., lower) bounds appear in the first (resp., second) row. See Figure 1 *for a graphical illustration of these results.*

	Undirected unweighted graphs	Undirected weighted graphs	Directed weighted graphs
Upper bound	$O(\min\{N \cdot \sqrt{P}, \sqrt{N} \cdot P\})$	$O(\min\{N \cdot \sqrt{P}, \sqrt{N} \cdot P\})$	$O(N \cdot \sqrt{P})$
Lower bound	$\Omega\left(\max\{N^{\frac{2d}{d^2+1}} \cdot P^{\frac{d(d-1)}{d^2+1}} : d = 2, 3, \dots\}\right)$	$\Omega((NP)^{2/3})$	$\Omega((NP)^{2/3})$

TABLE 2

A summary of our results on sourcewise preservers. See also Figure 2 *for a graphical illustration.*

	Undirected unweighted graphs	Undirected weighted graphs
Upper bound	$O(\min\{\sqrt{N} \cdot S^2, N \cdot S\})$	$O(\min\{\sqrt{N} \cdot S^2, N \cdot S\})$
Lower bound	$\Omega\left(\max\{N^{9/11} \cdot S^{6/11}, N^{10/11} \cdot S^{4/11}\}\right)$	$\Omega(N^{6/7} \cdot S^{4/7})$

only $O(S^2)$ edges. To prove or disprove this statement is a challenging open problem. Also, as Figures 1 and 2 suggest, there are some gaps between the curves of the upper and lower bounds. Closing these gaps is a very interesting open problem as well. We hope that our paper will trigger future research on these fundamental problems.

1.3. Our techniques. The lower bounds constitute the technically more involved part of the paper, and the techniques that are used for proving them are mainly from the area of discrete geometry.

Specifically, we consider certain sets of points that belong to high-dimensional integer lattices, and we build graphs whose sets of vertices are those sets. Next, for each of those points we build a convex polytope, whose extreme points are also vertices of the graph, and connect this point to all the extreme points of its polytope via edges. This way the edgesets of our graphs are constructed. In some cases additional vertices and edges that have no geometric interpretation are added to the graph, and the proofs combine geometric and combinatorial techniques.

The main polytope that is used for constructing the edgesets is the convex hull of the set of integer points of a ball with a large radius $R \gg 1$. This polytope is an important object of study in discrete geometry, and our proofs make use of the most recent advances in the study of this object. Specifically, we use the results of Barany and Larman [8] and Balog and Barany [7] that analyze the number of vertices and faces of this polytope. We believe that introducing the techniques from the area of discrete geometry to the study of spanners is our important technical contribution.

To motivate the use of geometric (Euclidean) graphs, we remark that we are not aware of other constructions of *dense* graphs in which the structure of shortest paths is well understood and relatively simple. Designing significantly simpler constructions of graphs with these properties is a challenging open problem.

The proofs of our upper bounds are conceptually simpler and are based on double-counting of appropriate combinatorial quantities.

Structure of the paper. Section 3 is devoted to the lower bound on the cardinality of pairwise preservers for weighted graphs. In section 4 we turn to the lower bounds for unweighted graphs. In sections 5 and 6 we describe our lower bound for

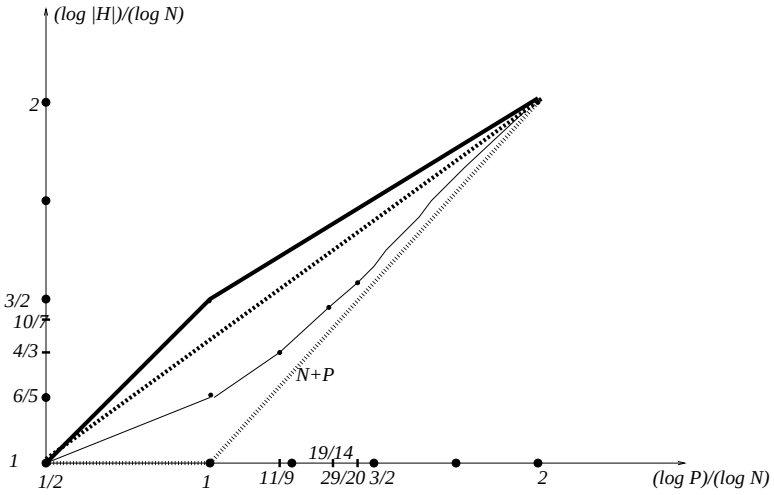


FIG. 1. *The x-axis of this graph corresponds to* $\log_N P$*, and the y-axis to* $\log_N |H|$*. The trivial lower bound of* $N+P$ *is depicted by the thick dashed line. The lower bound for undirected unweighted graphs is depicted by the thin solid line. The lower bound for weighted graphs is depicted by the thick dotted line, and, finally, the upper bound is depicted by the thick solid line. Note that the lower bound for undirected unweighted graphs is a piecewise linear curve with infinitely many linear segments. It is above the trivial lower bound in the entire feasible range of* P *and asymptotically converges to it when the exponent of* P *tends to* 2.

Particularly, for $P = N$ *(*$\log_N P = 1$*), the trivial lower bound is* $\Omega(N)$*, our lower bound for unweighted undirected graphs is* $\Omega(N^{6/5})$*, our lower bound for weighted directed graphs is* $\Omega(N^{4/3})$*, and our upper bound is* $O(N^{3/2})$*. See also Table* 1.

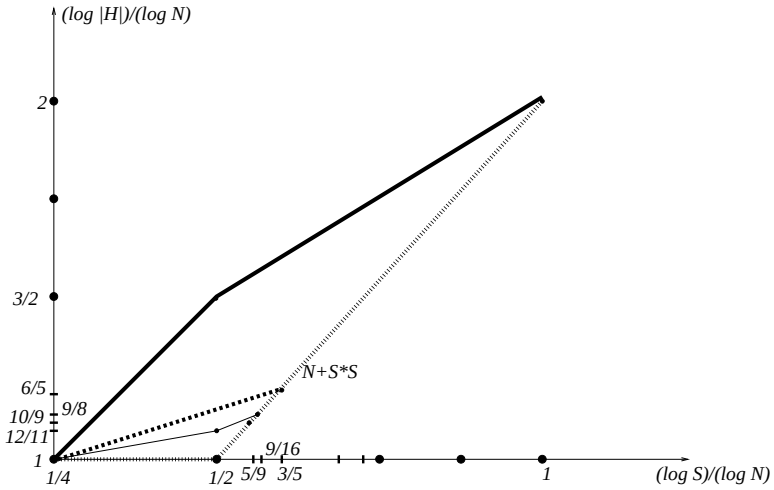


FIG. 2. *The x-axis of this graph corresponds to* $\log_N S$*, and the y-axis to* $\log_N |H|$*. The types of lines are consistent with those on Figure* 1.

Particularly, for $S = N^{1/2}$*, the trivial lower bound is* $\Omega(N)$*, our lower bound for unweighted undirected graphs is* $\Omega(N^{12/11})$*, our lower bound for weighted undirected graphs is* $\Omega(N^{8/7})$*, and the upper bound is* $O(N^{3/2})$*. See also Table* 2.

sourcewise preservers for unweighted and weighted graphs, respectively. Section 7 is devoted to our upper bounds. In section 6 we present a three-dimensional construction that yields slight improvements on the lower bounds of section 5. Section 6.4 is devoted to our somewhat stronger lower bounds on the sourcewise preservers that, however, apply only to weighted graphs.

2. Preliminaries.

2.1. Numbers and sets.

1. Let $\mathbb{N}$ (resp., $\mathbb{Z}$; $\mathbb{Q}$; $\mathbb{R}$; $\mathbb{R}^+$) denote the set of all *natural* (resp., *integer*; *rational*; *real*; *nonnegative real*) numbers. For a positive integer number k, let $[k]$ denote the set $\{1, 2, \ldots, k\}$, and $[(k)]$ denote the set $\{0, 1, \ldots, k-1\}$.
2. For a set C and a positive integer $k \leq |C|$, let $\binom{C}{k}$ (resp., $(\binom{C}{k})$) denote the set of all unordered (resp., ordered) k-tuples of different elements of C, and let C^k denote the set of all ordered k-tuples of not necessarily different elements of C.
3. For a function $f : A \to B$ between two sets A and B, and a subset $C \subseteq A$, let $f|_C$ denote the *restriction* of the function f to the set C.
4. For a pair of integer numbers a, b, $a \neq 0$, we say that a is a *divisor* of b, and denote $a|b$, if b/a is an integer. For a pair of integer numbers a, b, the greatest positive integer divisor of both a and b is denoted $\texttt{gcd}(a, b)$. If $\texttt{gcd}(a, b) = 1$, we say that a and b are *relatively prime*.
   The following basic fact is needed in our proof.
   LEMMA 2.1. *Let $(a, b) \in \mathbb{Z}^2$ such that $\texttt{gcd}(a, b) = 1$. Then for any real number $\rho > 1$ such that $(\rho a, \rho b) \in \mathbb{Z}^2$, we have $\rho \in \mathbb{Z}$.*
   Proof. Since $\texttt{gcd}(a, b) = 1$, there exists numbers $i, j \in \mathbb{Z}$ such that $a \cdot i + b \cdot j = 1$. Hence, $\rho = \rho \cdot (a \cdot i + b \cdot j) = i \cdot (\rho \cdot a) + j \cdot (\rho \cdot b) \in \mathbb{Z}$ since $\rho \cdot a, \rho \cdot b \in \mathbb{Z}$. □
5. Let $\mu : \mathbb{N} \to \{-1, 0, 1\}$ denote the *Mobius function*. For an argument d that is divisible by an integer square $k^2 \neq 1$, $\mu(d)$ is defined as 0; $\mu(1)$ is defined as 1; and for $d = p_1 \cdot \ldots \cdot p_k$, p_i are distinct primes different from 1, $\mu(d)$ is defined as $(-1)^k$.
   We will need the following two simple facts.
   LEMMA 2.2. *For a positive integer R,*

$$\frac{6}{\pi^2} - \sum_{d=1}^{R} \frac{\mu(d)}{d^2} \;=\; O\left(\frac{1}{R}\right).$$

   Proof. By definition of the Mobius function,

$$\sum_{d=1}^{\infty} \frac{\mu(d)}{d^2} = \Pi_p \left(1 - \frac{1}{p^2}\right),$$

   where the product ranges over all prime numbers p.
   As $(1 - 1/p^2)^{-1} = 1 + 1/p^2 + 1/p^4 + \cdots$, it follows that

$$\begin{aligned} \Pi_p \left(1 - \frac{1}{p^2}\right)^{-1} &= \Pi_p \left(1 + \frac{1}{p^2} + \frac{1}{p^4} + \cdots\right) \\ &= \sum_{n=1}^{\infty} \frac{1}{n^2} \;=\; \zeta(2) \;=\; \frac{\pi^2}{6}, \end{aligned}$$

where $\zeta(\cdot)$ denotes the Riemann zeta function.
Consequently,

$$\sum_{d=1}^{\infty} \frac{\mu(d)}{d^2} \;=\; \frac{6}{\pi^2}.$$

Moreover,

$$\left| \sum_{d=R+1}^{\infty} \frac{\mu(d)}{d^2} \right| \;\leq\; \sum_{d=R+1}^{\infty} \frac{1}{d^2} \;\leq\; \int_R^{\infty} \frac{1}{x^2} dx = \frac{1}{R}.$$

The lemma follows. □
An (elementary) proof of the following fact can be found in [2, Thm. 2.1, p. 25].
LEMMA 2.3. *If $n \geq 1$, then*

$$\sum_{d|n} \mu(d) = \left\lfloor \frac{1}{n} \right\rfloor.$$

6. For a real number r, let the *sign of* r, denoted *sign*(r), be defined as $+$ if r is positive, as $-$ if r is negative, and as 0 if $r = 0$.
7. For a vector $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$, let $||v|| = ||v||_2 = \sqrt{\sum_{i=1}^{d} v_i^2}$ denote its *Euclidean norm*, and let $||v||_\infty = \max\{|v_i| : i \in [d]\}$ denote its *ℓ_∞-norm*.
8. LEMMA 2.4. *Let B be a convex two-dimensional body containing the origin. Let A denote its area and D denote its diameter ($D = \max\{||u - v|| : u, v \in B\}$). Let M denote the number of integer points $(x, y) \in B$ with $\gcd(x, y) = 1$. Then $|M - A \cdot \frac{6}{\pi^2}| = O(A/D + D \cdot \log D)$.*
Proof. For a given positive real q, let an outer body B_q^+ consist of B and all the points at a distance of at most q from B, and let an inner body B_q^- consist of all points at a distance of at least q from the complement of B. The areas of B_q^+ and B_q^- differ from the area of B by at most $O(Dq)$ (because the perimeter of B is at most πD, and the area of $B_q^+ \setminus B_q$ is, consequently, at most $\pi D \cdot q$). Tile the plane with squares centered on lattice points (x, y), where $q|\gcd(x, y)$; each of these squares has area q^2. If such a lattice point is inside B, then the whole square is inside B_q^+. Hence, the number $M(B, q)$ of such lattice points is at most $(A + O(Dq))/q^2$. If a lattice point is outside B, then the whole square is outside B_q^-. Hence, $M(B, q)$ is at least $(A - O(Dq))/q^2$. It follows that the number of lattice points (x, y) with $\gcd(x, y) = 1$ is

$$\begin{aligned} M = \sum_{(k,\ell) \in B} \left\lfloor \frac{1}{\gcd(k, \ell)} \right\rfloor \;&=\; \sum_{(k,\ell) \in B} \sum_{d | \gcd(k,l)} \mu(d) \\ &= \sum_{d \leq D} \mu(d) \cdot M(B, d). \end{aligned}$$

(The second equality follows from Lemma 2.3. The last equality holds because the diameter of B is at most D, and B contains the origin.)
As we have seen,

$$\frac{A}{d^2} - O\left(\frac{D}{d}\right) \;\leq\; M(B, d) \;\leq\; \frac{A}{d^2} + O\left(\frac{D}{d}\right).$$

Hence,

$$\left| M - A \cdot \sum_{d \le D} \frac{\mu(d)}{d^2} \right| = \left| \sum_{d \le D} \mu(d) \cdot M(B,d) - A \cdot \sum_{d \le D} \frac{\mu(d)}{d^2} \right| = O(D) \cdot \left| \sum_{d \le D} \frac{\mu(d)}{d} \right|.$$

By Lemma 2.2,

$$\frac{6}{\pi^2} - \sum_{d \le D} \frac{\mu(d)}{d^2} \;\le\; O\left(\frac{1}{D}\right).$$

Also, $\sum_{d \le D} \mu(d)/d \le \sum_{d \le D} 1/d = O(\log D)$. Hence

$$\left| M - A \cdot \frac{6}{\pi^2} \right| = O(A/D) + O(D \cdot \log D). \qquad \square$$

2.2. Graphs.

1. An *undirected unweighted graph* $G = (V, E)$ is an ordered pair in which the first element V is the set of elements called *vertices*, and $E \subseteq \binom{V}{2}$. A *directed unweighted graph* $G = (V, E)$ is an ordered pair in which V is the set of vertices, and $E \subseteq \left(\binom{V}{2}\right)$. A *weighted undirected* (resp., *directed*) *graph* $G = ((V, E), wt)$, $wt : E \to \mathbb{R}^+$, is an unweighted undirected (resp., directed) graph with a nonnegative real weight function wt attached to it.
   We will occasionally refer to a weighted graph $G = ((V, E), wt)$ as $G = (V, E)$, omitting the weight function from the notation.
2. A sequence $\Pi = (v_0, v_1, \ldots, v_\ell)$ of distinct vertices of the graph $G = (V, E)$ is called a *path* if $\{v_i, v_{i+1}\} \in E$ (or $(v_i, v_{i+1}) \in E$ if G is a directed graph) for every index $i \in [(\ell)]$. The *length* of the path Π, denoted $L(\Pi)$, is defined as ℓ if the graph is unweighted, as $\sum_{i=1}^{\ell} wt(\{v_{i-1}, v_i\})$ if it is weighted and undirected, and as $\sum_{i=1}^{\ell} wt(\langle v_{i-1}, v_i \rangle)$ if it is weighted and directed. For a pair of vertices $u, w \in V$, let $\hat{\Pi}_{u,w}$ denote the set of paths between u and w in G (or from u to w if G is a directed graph).
   For a pair of vertices $u, w \in V$, the *distance* between u and w (resp., from u to w) in an undirected (resp., directed) weighted or unweighted graph G with vertex set V and edgeset E, denoted $dist_G(u, w)$ or $dist_E(u, w)$, is the length of the shortest path between u and w (resp., from u to w) in G, that is,

$$dist_G(u, w) = \min\{L(\Pi) : \Pi \in \hat{\Pi}_{u,w}\}.$$

3. For an undirected (resp., directed) graph G with vertex set V and edgeset E (weighted or unweighted) and a set $\mathcal{P} \subseteq \binom{V}{2}$ (resp., $\mathcal{P} \subseteq \left(\binom{V}{2}\right)$) of pairs of vertices, the subgraph $G' = (V, H)$, $H \subseteq E$, is called a *pairwise preserver of G with respect to the set $\mathcal{P}$* if for every pair $\{u, w\} \in \mathcal{P}$ (resp., $(u, w) \in \mathcal{P}$), $dist_G(u, w) = dist_{G'}(u, w)$.
   For a subset $\mathcal{S} \subseteq V$ of vertices, also called *sources*, the pairwise preserver with respect to the set $\binom{\mathcal{S}}{2}$ is also called the *sourcewise preserver of G with respect to the set $\mathcal{S}$*.
4. For a directed (weighted or unweighted) graph $G = (V, A)$, its *underlying undirected graph* $G' = (V, E')$ is defined by $E' = \{\{u, w\} : \langle u, w \rangle \in A\}$.
   An *arborescence* is a directed graph whose underlying undirected graph is a tree.

For a directed graph $G = (V, A)$, and a vertex $v \in V$, a subgraph $\tau = (V, E_\tau)$ is called a *shortest-path in-arborescence* (resp., *out-arborescence*) *rooted at* v if it is a spanning arborescence of G, and for every vertex $w \in V$, $dist_\tau(w, v) = dist_G(w, v)$ (resp., $dist_\tau(v, w) = dist_G(v, w)$).

5. For an undirected graph $G = (V, E)$, and a vertex $v \in V$, let $\deg_G(v)$, or $\deg_E(v)$, denote the *degree* of the vertex v in G, that is, $\deg_G(v) = |\{w : \{v, w\} \in E\}|$.

2.3. Polytopes and convexity. For a set $\{\gamma_1, \gamma_2, \ldots, \gamma_k\} \in \mathbb{R}^d$ of vectors, $\sum_{i=1}^{k} a_i\gamma_i$ is called a *convex combination* of them if $a_1, a_2, \ldots, a_k \in \mathbb{R}^+$, and $\sum_{i=1}^{k} a_i = 1$.

The *convex hull* of the set $\{\gamma_1, \ldots, \gamma_k\} \subset \mathbb{R}^d$ of vectors is the set of all convex combinations of these vectors. A *polytope* is the convex hull of some set of vectors. It is known that a polytope can also be defined as the set of feasible solutions of n linear inequalities, $n > d$, and these two definitions are equivalent (see, e.g., [24]). A *polygon* is a two-dimensional polytope.

The set $\{\gamma_1, \ldots, \gamma_k\} \subset \mathbb{R}^d$ is called a *convexly independent set* (henceforth, *CIS*) if for every index $i \in [k]$, $i \in [k] = \{1, 2, \ldots, k\}$, the vector γ_i cannot be expressed as a convex combination of the vectors $\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_k$.

For $k = 0, 1, \ldots, d$, a *k-face* of the polytope P is the set of points of the polytope in which some $d - k$ inequalities hold as equalities. The *extreme points* or *vertices* are 0-faces of the polytope, and the $(d-1)$-faces are called the *facets* of the polytope. Equivalently, an extreme point of the polytope P can be defined as a vector $v \in P$ that cannot be expressed as a convex combination of other vectors of P.

2.4. Lattices.

DEFINITION 2.5. *For a pair of vectors $\alpha, \beta \in \mathbb{R}^3$, the set $L = \{a \cdot \alpha + b \cdot \beta : a, b \in \mathbb{Z}\}$ is called a* two-dimensional lattice. *The pair $\{\alpha, \beta\}$ of vectors is called a* basis *of the lattice L.*

Note that the same lattice may have more than one basis.

DEFINITION 2.6. *For a lattice $L = \{a \cdot \alpha + b \cdot \beta : a, b \in \mathbb{Z}\}$ let $\mathcal{H} = \{x \cdot \alpha + y \cdot \beta : x, y \in \mathbb{R}\}$ be the plane that contains the lattice L. The area of the parallelogram whose corners are the origin, α, $\alpha + \beta$, and β (this parallelogram is contained in the plane $\mathcal{H}$), is called the* cell area, *or the* determinant, *of the lattice L and is denoted* $\det L$.

It is a well known fact (see, e.g., [24]) that the determinant of a lattice does not depend on the choice of its basis. We also need the following standard geometric fact.

LEMMA 2.7 (see [24]). *Consider the lattice $L = \{(x_1, x_2, x_3) : \langle a, x\rangle = 0\}$ for some vector $a \in \mathbb{Z}^3$, $a = (a_1, a_2, a_3)$, with $\gcd(a_1, a_2, a_3) = 1$. The determinant of this lattice is $\det L = ||a||$.*

The next lemma follows directly from Lemma 2.7.

LEMMA 2.8. *Consider a three-dimensional polytope P, all of whose extreme points belong to the integer lattice $\mathbb{Z}^3$. The area of a facet of P whose integer normal (in its reduced form) is $a = (a_1, a_2, a_3)$ is at least half the norm of the normal, that is, $||a||/2$.*

Proof. Consider the two-dimensional lattice that is spanned by the extreme points of the facet and is contained in the plane that contains the facet. By Lemma 2.7, and since the facet contains at least one half of a cell of the lattice, the lemma follows. ∎

3. Lower bounds: Pairwise preservers for weighted graphs. In this section we show lower bounds on the cardinalities of pairwise preservers. Specifically, we show that for any sufficiently large positive integer numbers N and P, $\Omega(\sqrt{N}) = P = O(N^2)$, there exist weighted undirected graphs $G = ((V, E), wt)$, and sets $\mathcal{P} \subseteq \binom{V}{2}$

of cardinality $|\mathcal{P}| = P$, such that any pairwise preserver of the graph G with respect to the set $\mathcal{P}$ requires $|E| = \Omega((N \cdot P)^{2/3})$ edges. This lower bound is superlinear in $N + P$ for the range $\omega(\sqrt{N}) = P = o(N^2)$. We remark that in view of our linear upper bound of Corollary 7.8 in section 7.2, this lower bound applies for the entire feasible range of P. In section 4 we will show similar, though weaker, lower bounds that are applicable for *unweighted undirected* graphs.

We next construct a fairly dense weighted graph $G = ((V, E), wt)$, $wt : E \rightarrow \mathbb{R}^+$, and a set $\mathcal{P} \subseteq \binom{V}{2}$ of pairs of vertices. We will show that for any subgraph $G' = ((V, H), wt)$, $H \subset E$ $(H \neq E)$, there exists a pair $p = \{u, w\} \in \mathcal{P}$ so that $dist_{G'}(u, w) > dist_G(u, w)$.

Consider a square portion of the two-dimensional integer lattice $\mathbb{Z}^2$, with real dimensions $\sqrt{N} \times \sqrt{N}$, where N is the number of vertices in the graph G that we construct. In other words, let $V = \{(i, j) : i, j \in [(\sqrt{N})]\}$. (We assume that $\sqrt{N}$ is integral; all the nonintegrality issues affect only lower-order terms of our results, and are, henceforth, ignored.)

Let T, $T \leq \sqrt{N}/10$, be a positive integer parameter of the construction that will be fixed later. For a pair of vertices $(i, j), (\ell, k) \in V$, $\{(i, j), (\ell, k)\}$ is an edge if the Euclidean distance $||(i, j) - (\ell, k)|| = \sqrt{(\ell - i)^2 + (k - j)^2}$ between (i, j) and (ℓ, k) is at most T, and $\texttt{gcd}(|\ell - i|, |k - j|) = 1$.

The weight function wt is defined by $wt(e) = ||(i, j) - (\ell, k)||$ for every edge $e \in E$. This completes the construction of the graph $G = ((V, E), wt)$.

The *boundary frame set* BF is defined by

$$BF = \{(i, j) : (i \in [(T)]) \text{ or } (i \in [(\sqrt{N})] \setminus [(\sqrt{N} - T)]) \\ \text{or } (j \in [(T)]) \text{ or } (j \in [(\sqrt{N})] \setminus [(\sqrt{N} - T)])\},$$

and the *boundary set* B is defined by

$$B = \{(i, j) : (i = 0) \text{ or } (i = \sqrt{N} - 1) \text{ or } (j = 0) \text{ or } (j = \sqrt{N} - 1)\}.$$

For a point $(i, j) \in V$, and an edge $e = \{(i, j), (\ell, k)\}$ that is adjacent to (i, j), consider the (straight) line $\mathcal{L}$ that passes through the two points (i, j) and (ℓ, k) in the Euclidean plane. Let $L = \mathcal{L} \cap V$. Let L' be the subset of L that contains those points of L that are farther from (i, j) than from (ℓ, k) (in terms of the Euclidean distance). Let $(i', j') \in L'$ be the farthest (in terms of the Euclidean distance) point from (i, j). This point is called the *antipodal point to* (i, j) *in the direction of* (ℓ, k) and is denoted $(i', j') = A((i, j), (\ell, k))$. See Figure 3. The set of pairs $\mathcal{P}$ is now defined by

$$\mathcal{P} = \{\{(i, j), (i', j')\} : (i, j) \in BF, (i', j') = A((i, j), (\ell, k)), \{(i, j), (\ell, k)\} \in E\}.$$

By Lemma 2.4, every point $(i, j) \in V$ has $\Theta(T^2)$ neighbors, and thus

$$|E| = \Theta(N \cdot T^2). \tag{1}$$

Also, since $|BF| = \Theta(\sqrt{N} \cdot T)$, it follows that

$$P = |\mathcal{P}| = \Theta(\sqrt{N} \cdot T^3). \tag{2}$$

We next argue that all the antipodal points belong to the boundary frame BF.

LEMMA 3.1. *For an edge* $\{(\ell, k), (\ell', k')\} \in E$, $A((\ell, k), (\ell', k')) \in BF$.

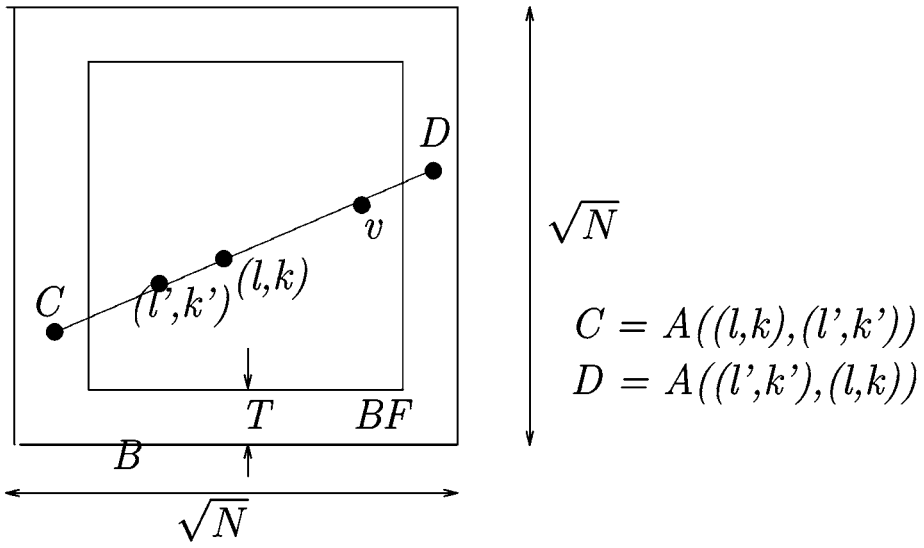


FIG. 3. *A graphical illustration of the instance. The pair* $\{C, D\}$ *belongs to the set* $\mathcal{P}$ *of pairs.*

Proof. Let $\mathcal{L}$ denote the line in the Euclidean plane that passes through the two points (ℓ, k) and (ℓ', k'), and consider the ray $\mathcal{R} \subseteq \mathcal{L}$ that starts in (ℓ', k') and does not contain the point (ℓ, k). Let S be the boundary of the square Sq whose corners are (in counterclockwise order) the origin, $(\sqrt{N}-1, 0)$, $(\sqrt{N}-1, \sqrt{N}-1)$, $(0, \sqrt{N}-1)$, that is, the union of the segments that connect every consecutive pair of these corners, and the origin, to $(0, \sqrt{N}-1)$. Let (x, y) be the intersection of the ray $\mathcal{R}$ with the boundary S of this square. If $\frac{x-\ell}{\ell'-\ell}$ is an integer, then, obviously, $(x, y) = A((\ell, k), (\ell', k'))$, and since $(x, y) \in BF$, we are done. Otherwise, let (x', y') be the point in L' closest to (x, y).

We next show that $(x', y') \in BF$. Let I denote the segment of the ray $\mathcal{R}$ that connects the points (x, y) and $(x - (\ell' - \ell), y - (k' - k))$. Obviously, $(x', y') \in I$. Note, however, that the distance between (x', y') and the boundary B of the square Sq is at most $\max\{|\ell' - \ell|, |k' - k|\} \leq T$. (Note that this is the distance between a point and a line that is parallel to one of the axes.) It follows that $(x', y') \in BF$, as required. ∎

The next lemma follows directly from the construction of the set $\mathcal{P}$ and from Lemma 3.1.

LEMMA 3.2. *For an edge* $\{(\ell, k), (\ell', k')\} \in E$, *the pair* $\{A((\ell, k), (\ell', k')), A((\ell', k'), (\ell, k))\}$ *belongs to the set* $\mathcal{P}$ *of pairs.*

Proof. By Lemma 3.1, both points $C = A((\ell, k), (\ell', k'))$ and $D = A((\ell', k'), (\ell, k))$ belong to BF. Furthermore, it is easy to see that all four points C, (ℓ, k), (ℓ', k'), and D belong to the same line $\mathcal{L}$. Consider the segment $\mathcal{J}$ of the line $\mathcal{L}$ that connects the points C and D. Let $v \in \mathcal{J} \cap L$ be the (unique) point that belongs to the segment and such that the edge $\{D, v\}$ belongs to the edgeset E. Obviously, $A(D, v) = C$, and so the pair $\{C, D\}$ belongs to $\mathcal{P}$. (See Figure 4 for an illustration.) ∎

For a graph $G = ((V, E), wt)$, and an edge $e \in E$, let $G \setminus e$ denote the graph $((V, E \setminus \{e\}), wt|_{(E \setminus \{e\})})$.

We next argue that no edge of the graph G can be removed without increasing the distance between some pair of points (i, j), (i', j') such that $\{(i, j), (i', j')\} \in \mathcal{P}$.

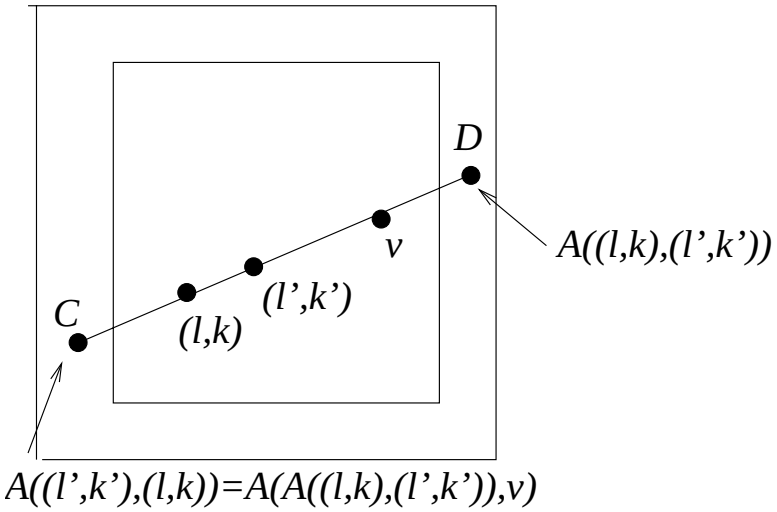


FIG. 4. *An illustration for the proof of Lemma* 3.2.

LEMMA 3.3. *For an edge* $e = \{u, w\}$, $u = (\ell, k)$, $w = (\ell', k')$, *let* $x = A(u, w)$, $y = A(w, u)$. *Then* $dist_{G \setminus e}(x, y) > dist_G(x, y)$.

Remark. The proof of this lemma is based on the observation that, intuitively, the shortest distance between x and y is attained (uniquely) by the straight line that connects them, and if this line is interrupted, the distance becomes longer.

Proof. Consider the line $\mathcal{L}$ that goes through the points (ℓ, k) and (ℓ', k'). Let $(i', j') = A((\ell, k), (\ell', k'))$ and $(i, j) = A((\ell', k'), (\ell, k))$. By Lemma 3.2 $\{(i, j), (i', j')\} \in \mathcal{P}$. By construction, the distance between these two points in G, $dist_G((i, j), (i', j'))$, is equal to the Euclidean distance $\sqrt{(i - i')^2 + (j - j')^2}$ between them, since they lie on the same line, and for each integer $0 \le h \le \frac{j'-j}{k'-k} - 1$, the edge $\{(i + h(\ell' - \ell), j + h(k' - k)), (i + (h+1)(\ell' - \ell), j + (h+1)(k' - k))\}$ belongs to the graph G. Let $\Pi(\mathcal{L})$ be the shortest path in G between (i, j) and (i', j') that uses these edges.

Any path Π in the graph G corresponds to a piecewise linear trajectory $\mathcal{C}$ that is formed by replacing each edge of Π with the linear segment that connects its endpoints in the Euclidean plane and by concatenating these segments. Let $\cdot$ denote the concatenation. For such a trajectory $\mathcal{C}$, let $\mathcal{C} = \mathcal{L}_1 \cdot \mathcal{L}_2 \cdot \ldots \cdot \mathcal{L}_t$, $t = 1, 2, \ldots$, be its (unique) representation as a concatenation of segments such that $\mathcal{L}_i \cdot \mathcal{L}_{i+1}$ is not a linear segment for every index $q \in [t - 1]$, and $(i, j) \in \mathcal{L}_1$. Henceforth, such a representation will be referred to as the *concatenating representation* of the path Π. (See Figure 5.)

It is easy to see that in the graph G, the length of any simple path Π is equal to the Euclidean length of the concatenating representation of Π. The unique shortest curve joining x and y is the straight line between them. Also, since every edge of G corresponds to a vector with relatively prime coordinates, for any path Π other than $\Pi(\mathcal{L})$ joining x and y, its concatenating representation necessarily contains a segment of slope which differs from that of $\mathcal{L}$. Consequently, the concatenating representation of Π is longer than that of $\Pi(\mathcal{L})$, and so Π is longer than $\Pi(\mathcal{L})$. Hence the unique shortest path between x and y in G contains e. □

COROLLARY 3.4. *For infinitely many positive integer numbers* N *and* T, $T \le \sqrt{N}/10$, *there exist weighted* N*-vertex graphs* $G = ((V, E), wt)$ *with* $|E| = \Theta(N \cdot$

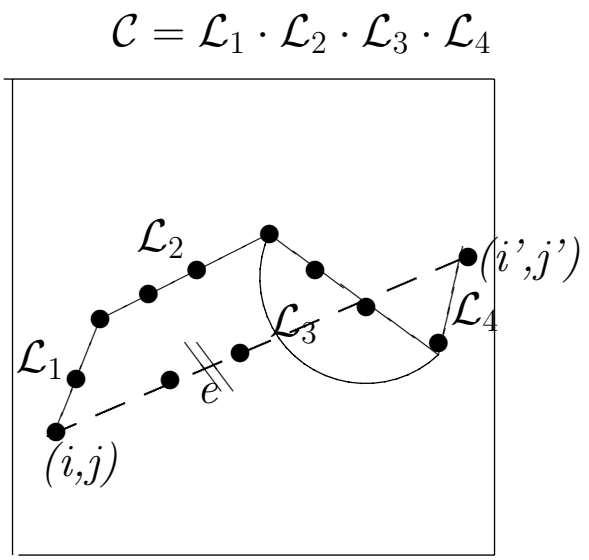


FIG. 5. *The direct path of length* $||(i',j')-(i,j)||$ *that uses the edge* e *is depicted by the dashed line. The path* Π *and its concatenating representation are depicted by the solid piecewise linear curve.*

T^2) *edges, and a collection* $\mathcal{P}$ *with* $P = O(\sqrt{N} \cdot T^3)$ *pairs of vertices, such that every subgraph* G' *that preserves all the distances between the pairs of vertices from* $\mathcal{P}$ *contains all the edges of* G.

Proof. The graph G and the collection $\mathcal{P}$ of pairs were defined above. The assertion follows from Lemma 3.3 and from (1) and (2). □

Note that (1) and (2) imply the lower bound of

$$|E| \;=\; \Omega((NP)^{2/3}), \tag{3}$$

which is the main result of this section. Note that the lower bound is superlinear in $P+N$ whenever $\omega(\sqrt{N}) = P = o(N^2)$. For $P = O(\sqrt{N})$ no superlinear lower bound is possible due to our upper bound from Corollary 7.8.

This construction generalizes readily to any constant dimension $d = 3, 4, \ldots$, but the obtained lower bounds are (weakly) inferior to that of (3).

4. Lower bounds: Pairwise preservers for unweighted graphs. In this section we present a more elaborate construction that enables us to show lower bounds on the cardinalities of pairwise preservers for *unweighted* undirected graphs. These lower bounds are somewhat weaker than the lower bounds for *weighted* graphs that are given by (3), but they are also superlinear in $N+P$ in the *entire feasible range* of P, that is, $\omega(\sqrt{N}) = P = o(N^2)$.

For a fixed large positive integer T, and a fixed small positive integer $d = 2, 3, \ldots$, consider the set $Ball_d(T) \cap \mathbb{Z}^d$, $Ball_d(T) = \{x \in \mathbb{R}^d : ||x|| \leq T\}$, of all the points of the integer lattice $\mathbb{Z}^d$ that belong to the d-dimensional ball of radius T centered at the origin.

Consider the convex hull $CH = CH(Ball_d(T) \cap \mathbb{Z}^d)$ of the set $Ball_d(T) \cap \mathbb{Z}^d$. Obviously, this convex hull is symmetric around the origin, i.e., if $\alpha \in CH$, then $(-\alpha) \in CH$ as well. Let $VH = VH(Ball_d(T) \cap \mathbb{Z}^d)$ be the set of the *extreme points* (or *vertices*) of the convex hull CH. This set is also symmetric around the origin. It

is known [8] that the cardinality of the set VH is $\Theta(T^{d-2+\frac{2}{d+1}})$, where the constant hidden by the Θ-notation depends only on the dimension d. (For the dimension $d = 2$ this result was shown in [7]. However, for our purposes any construction of a large CIS of vectors with small Euclidean norm is sufficient. Particularly, for the dimension $d = 2$ a much simpler classical construction of [21], instead of the construction of [8] or [7], can be plugged in our proof. For the most general version of our proof, though, the d-dimensional construction of [8] is required. The construction of [21] will be discussed in greater detail in section 5.)

We next construct the unweighted graph $G = (V, E)$, and the set of pairs $\mathcal{P} \subseteq \binom{V}{2}$, that will be used for our lower bound. Let N (the number of vertices of the graph) be a fixed large positive integer. The vertex set V is the set of all points $x = (x_1, \ldots, x_d)$ of the integer lattice $\mathbb{Z}^d$ with nonnegative coordinates and such that $||x||_\infty = \max\{x_i : i \in [d]\} \leq N^{1/d} - 1$. In other words,

$$V \;=\; \{(x_1, \ldots, x_d) \in \mathbb{Z}^d : 0 \leq x_i \leq N^{1/d} - 1 \quad \forall i \in [d]\}.$$

For every point $x = (x_1, \ldots, x_d) \in V$, let $\Gamma(x) = \{(x + y) \in V : y \in VH\}$ be the Minkowski sum of x and VH, where $T < \frac{1}{5}N^{\frac{1}{d}}$ is a positive integer parameter to be fixed later. Note that since the set VH is symmetric around the origin, $z \in \Gamma(x)$ if and only if $x \in \Gamma(z)$. The edgeset E is defined by $E = \{\{x, z\} : z \in \Gamma(x)\}$. Let the *boundary frame* set $BF \subseteq V$ be the set of all points $x = (x_1, \ldots, x_d) \in V$ so that at least one of the coordinates x_i, $i \in [d]$, satisfies either $0 \leq x_i \leq T - 1$ or $N^{1/d} - T \leq x_i \leq N^{1/d} - 1$, and let the *boundary* set $B \subseteq V$ be the set of all the points $x \in V$ such that at least one of the coordinates is either 0 or $N^{1/d} - 1$.

For each point $x \in V$ and each neighbor $z \in \Gamma(x)$, we define $A(x, z)$ (the *antipodal point to x in the direction $z - x$* (or *in the direction z*)) in the following way. Let $\mathcal{L}$ be the line in the d-dimensional Euclidean space that passes through the points x and z. Let $L = \mathcal{L} \cap V$ be the set of the vertices of the graph G that belong to this line, and let $L' \subseteq L$ be the subset of L that contains only vertices that are closer to the point z than to the point x, in terms of the Euclidean distance. The point $A(x, z)$ is defined as the farthest (in terms of the Euclidean distance) point from x that belongs to the set L'. Finally, similar to the way that the set $\mathcal{P}$ of pairs was constructed in section 3, we now define $\mathcal{P} = \{\{x, A(x, z)\} : x \in BF, z \in \Gamma(x)\}$. Denote $P = |\mathcal{P}|$.

Note that for every vertex $x = (x_1, \ldots, x_d) \in V \setminus BF$, the vertex x has precisely $|VH|$ neighbors in the graph. In other words, the edgeset E has cardinality at least $|E| \geq |V \setminus BF| \cdot |VH| = \Omega((N^{1/d} - 2T)^d \cdot T^{d-2+\frac{2}{d+1}}) = \Omega(NT^{d-2+\frac{2}{d+1}})$. (Hereafter all the asymptotic notation may hide dependence on the dimension d, but not on the number of vertices N or the parameter T.)

Observe also that $|BF| = \Theta(N^{1-\frac{1}{d}}T)$ and $P = O(|BF| \cdot |VH|) = O(N^{1-\frac{1}{d}} \cdot T^{d-1+\frac{2}{d+1}})$.

We next argue that no edge can be removed from the graph G without increasing the distance between some pair of vertices $\{x, y\} \in \mathcal{P}$. To this end, consider an edge $e = \{u, w\} \in E$. Let $y = A(u, w)$, and $x = A(w, u)$. It is easy to see that the proof of Lemma 3.1 generalizes readily to the d-dimensional space with $d > 2$, and thus, $x, y \in BF$, and, furthermore, $\{x, y\} \in \mathcal{P}$.

The statement of the next lemma is analogous to Lemma 3.3. Its proof is, however, more involved, since the "Euclidean lengths" of the edges of the graph are no longer uniform.

LEMMA 4.1. *$dist_{G \setminus e}(x, y) > dist_G(x, y)$.*

Proof. Let $\Pi_{x,y} = (\{x, x+(w-u)\}, \{x+(w-u), x+2(w-u)\}, \ldots, \{x+(k-1)(w-u), x+k(w-u)\})$, $x+k(w-u) = y$, be the path in the graph G that connects the vertices x and y, and all its edges lie on the line $\mathcal{L}$ that passes through the points u and w. (We say that an edge $(z_1, z_2) \in E$ lies on a line $\mathcal{L}'$ if both points z_1 and z_2 belong to this line.)

Observe that the length of this path is $k = \frac{y_j - x_j}{w_j - u_j}$, where x_j, y_j, u_j, and w_j are the jth coordinates of the vectors x, y, u, and w, respectively, and $j \in [d]$ is one of the indices that satisfy $w_j - u_j \neq 0$. (Note that since the points x, y, u, and w are colinear, the expression $\frac{y_j - x_j}{w_j - u_j}$ is independent of the choice of the index $j \in [d]$, as far as $w_j - u_j \neq 0$. Since $w - u \in VH$, it follows that $w - u \neq 0$.)

Consider some path $\Pi'_{x,y}$ of length at most k that connects the vertices x and y in the graph G. Let $\Pi'_{x,y} = (e^{(1)}, e^{(2)}, \ldots, e^{(\ell-1)})$, $\ell \leq k+1$, $e_i = \{x^{(i)}, x^{(i+1)}\}$, $i \in [\ell - 1]$, $x^{(1)} = x$, $x^{(\ell)} = y$.

For an index $i \in [\ell - 1]$, let $d^{(i)} = x^{(i+1)} - x^{(i)}$, where $e^{(i)} = \{x^{(i)}, x^{(i+1)}\}$. It follows that $y = x^{(\ell)} = x + \sum_{i=1}^{\ell-1} d^{(i)}$. Hence, $y - x = \sum_{i=1}^{\ell-1} d^{(i)}$, $\ell - 1 \leq k$. Note also that $y - x = k \cdot (w - u)$. Hence,

$$w - u = \frac{1}{k} \sum_{i=1}^{\ell-1} d^{(i)}.$$

Let $\{d'^{(1)}, \ldots, d'^{(\ell')}\}$, $\ell' \leq \ell - 1$, be the set of distinct terms that appear in the sum $\sum_{i=1}^{\ell-1} d^{(i)}$, and let α_i, $i \in [\ell']$, denote the number of times that the element $d'^{(i)}$ appears in this sum. It follows that $w - u = \frac{1}{k} \sum_{i=1}^{\ell'-1} \alpha_i d'^{(i)}$ and $\sum_{i=1}^{\ell'} \alpha_i = \ell - 1 \leq k$. Hence,

$$w - u \;=\; \sum_{i=1}^{\ell'} \beta_i d'^{(i)}, \tag{4}$$

where $\beta_i = \frac{\alpha_i}{k} > 0$, $\sum_{i=1}^{\ell'} \beta_i = \frac{1}{k} \sum_{i=1}^{\ell'} \alpha_i = \frac{\ell-1}{k} \leq 1$.

Recall that $\{u, w\}$ is an edge of the graph G, and so the vector $w - u$ belongs to the set VH. Analogously, for each index $i \in [\ell']$, there exists an index $j \in [\ell - 1]$ such that $d'^{(i)} = d^{(j)}$, and $d^{(j)} = x^{(j+1)} - x^{(j)}$, where $e^{(j)} = \{x^{(j)}, x^{(j+1)}\}$ is an edge of the graph G. Hence, the vector $d'^{(i)} = d^{(j)} = x^{(j+1)} - x^{(j)}$ also belongs to the set VH. In other words, $w - u, d'^{(1)}, d'^{(2)}, \ldots, d'^{(\ell')} \in VH$, and (4) is satisfied. However, since VH is the set of *extreme points* of the convex hull of a set of points in $\mathbb{R}^d$, it follows that a vector $w - u \in VH$ can be represented uniquely as a convex combination of vectors from VH, specifically, as $w - u = \beta_1(w - u)$ with $\beta_1 = 1$. It follows that $\ell' = 1$, and $d'_1 = w - u$, and so the path $\Pi'_{x,y}$ coincides with the path $\Pi_{x,y}$.

Hence, the path $\Pi_{x,y}$ is the unique shortest path between the vertices x and y in the graph G, and this path uses the edge e. Hence, $dist_{G \setminus e}(x, y) > dist_G(x, y)$. ∎

Recall that $|E| = \Omega(N \cdot T^{d-2+\frac{2}{d+1}})$ and $P = O(N^{1-\frac{1}{d}} \cdot T^{d-1+\frac{2}{d+1}})$, for $T \leq \frac{1}{5} N^{\frac{1}{d}}$. A straightforward calculation shows that

$$|E| \;=\; \Omega\left(N^{\frac{2d}{d^2+1}} \cdot P^{\frac{d(d-1)}{d^2+1}}\right), \tag{5}$$

and this lower bound is applicable to $\Omega(N^{1-\frac{1}{d}}) = P = O(N^{2-\frac{2}{d+1}})$. For $d = 2, 3, \ldots$, let E_d denote the right-hand side of (5). By comparing these lower bounds for different

values of d it follows that for $d = 2, 3, \ldots$, in the range $\Omega(N^{2\cdot\frac{d^2-d-1}{(d-1)(d+2)}}) = P = O(N^{2\cdot\frac{d^2+d-1}{d(d+3)}})$, the lower bound is

$$E = \Omega(E_d) = \Omega\left(N^{\frac{2d}{d^2+1}} \cdot P^{\frac{d(d-1)}{d^2+1}}\right). \tag{6}$$

5. Lower bounds: Sourcewise preservers for unweighted graphs. In this section we show lower bounds on the cardinalities of sourcewise preservers.

5.1. Constructing a large CIS. We start by describing our variant of the Jarnik construction [21] of a large CIS of two-dimensional vectors of norm at most R for some fixed parameter R. We will use this construction, and some of its properties, for our lower bound.

Let t be an even integer parameter to be fixed later. Let $Z = \{(a, b) : a, b \in [t], \texttt{gcd}(a, b) = 1\}$. By Lemma 2.4, $|Z| = \Theta(t^2)$. Sort all the elements of Z by the ratio b/a, starting from the pair (a, b) with the largest ratio b/a, and ending with the smallest one. (Note that if two ratios b_1/a_1, b_2/a_2 are equal, then the two vectors (a_1, b_1), (a_2, b_2) are colinear, and since both of them are integer vectors, one of them is a rational multiple of the other. However, since both have gcd equal to 1, it follows that the two vectors are equal, which is a contradiction.)

Let $(a_1, b_1), \ldots, (a_k, b_k)$ be the sorted sequence of the vectors of Z. Let $A = \sum_{i=1}^{k} a_i$, $B = \sum_{i=1}^{k} b_i$. Let

$$w_0 = (A, 0),\ w_1 = w_0 + (-a_1, b_1),$$

$$w_2 = w_1 + (-a_2, b_2), \ldots, w_k = w_{k-1} + (-a_k, b_k) = \left(A - \sum_{i=1}^{k} a_i, \sum_{i=1}^{k} b_i\right) = (0, B).$$

Generally, for $j = 0, 1, \ldots, k$,

$$w_j = \left(A - \sum_{i=1}^{j} a_i, \sum_{i=1}^{j} b_i\right) = \left(\sum_{i=j+1}^{k} a_i, \sum_{i=1}^{j} b_i\right),$$

where $\sum_{i=k+1}^{k}(\cdot)$ is defined as 0. Denote $W = \{w_0, w_1, \ldots, w_k\}$. Note that $|W| = k = |Z| = \Theta(t^2)$. The norms of the vectors w_j, $j = 0, 1, \ldots, k$, are at most $\sqrt{2}\cdot k \cdot t = O(t^3)$. It is also not hard to see (see [21] for the formal proof) that the set W is a CIS, or, in other words, that the vectors $w_0, w_1, \ldots, w_k$ are the extreme points of the convex hull of the set $\{w_0, w_1, \ldots, w_k\}$, $CH(w_0, w_1, \ldots, w_k)$.

Furthermore, consider the set U of vectors given by

$$U = \{u = (x, y) : \exists \sigma_x, \sigma_y \in \{-1, 1\} \text{ s.t. } (\sigma_x x, \sigma_y y) \in W\}.$$

It is easy to see that this set is a CIS as well. Consider the convex hull $CH(U)$ of the set U. Let $(v_0, v_1, \ldots, v_{4k-1})$ be the sequence of vectors of U ordered counterclockwise, starting with $v_0 = w_0$. By construction, every edge of the convex hull $CH(U)$ has slope b/a, where $|a|, |b| \le t$. (The *slope* of the line $\mathcal{L} = \{(x_1, y_1) + \alpha \cdot (x_2, y_2) : \alpha \in \mathbb{R}\}$ is defined as y_2/x_2 if $x_2 \neq 0$, and as ∞ otherwise. In this paper, however, all lines have finite slopes.) Setting $R = t^3$ we derive the following theorem.

THEOREM 5.1 (see [21]). *For infinitely many positive integer numbers R there exist CISs of vectors $(v_0, v_1, \ldots, v_{4k-1}) \in \mathbb{R}^2$, ordered counterclockwise, with $k = \Theta(R^{2/3})$ of norm $||v_j|| \le R$ for every $j \in \{0, 1, \ldots, 4k-1\}$. Furthermore, for every index j in this range, $||v_{j+1} - v_j||_\infty = O(R^{1/3})$, where $j+1$ is shorthand for the remainder of the division of $j+1$ by $4k$.*

5.2. Constructing supporting lines. We next describe the construction of a collection of supporting lines of the polygon $CH(U)$.

Fix a positive integer parameter T. Let C be the CIS that satisfies the properties guaranteed by Theorem 5.1 with $R = T$. (Consequently, $t = T^{1/3}$.) Let P be the boundary of the convex hull $\hat{P} = CH(C)$. Note that P is a convex polygon whose extreme points are the vectors of C.

We need an additional piece of notation. For a convex polygon Q, let $Ex(Q)$ denote the set of its extreme points. Note that $Ex(P) = C$.

For each vertex v of the polygon P, let $v^{(1)}$ and $v^{(2)}$ be the two vertices of P so that there are two facets of the polygon that connect $v^{(1)}$ to v and v to $v^{(2)}$. Let $b^{(1)}/a^{(1)}$ and $b^{(2)}/a^{(2)}$ be the two slopes of these two facets (in their lowest terms), and assume without loss of generality that $b^{(2)}/a^{(2)} > b^{(1)}/a^{(1)}$. (Note that a facet of a polygon is a segment, and so its slope is well defined.) Assume also that $a^{(1)}$, $a^{(2)}$ are both positive (other cases are similar). Then, by construction, $v^{(2)} - v = (a^{(2)}, b^{(2)})$, $v - v^{(1)} = (a^{(1)}, b^{(1)})$, and $|a^{(i)}|, |b^{(i)}| \leq T^{1/3}$ for $i = 1, 2$. Furthermore, there exists no vector $(a^{(3)}, b^{(3)}) \in \mathbb{Z}^2$ with $(a^{(3)})^2 + (b^{(3)})^2 \leq T^{2/3}$, and $\texttt{gcd}(a^{(3)}, b^{(3)}) = 1$ so that $b^{(1)}/a^{(1)} < b^{(3)}/a^{(3)} < b^{(2)}/a^{(2)}$. It follows that the vector $(a^{(1)} + a^{(2)}, b^{(1)} + b^{(2)})$ has an intermediate slope

$$b^{(1)}/a^{(1)} \;<\; (b^{(1)} + b^{(2)})/(a^{(1)} + a^{(2)}) \;<\; b^{(2)}/a^{(2)},$$

and that the numbers $a^{(1)} + a^{(2)}$ and $b^{(1)} + b^{(2)}$ are relatively prime. Furthermore, $||(a^{(1)} + a^{(2)}, b^{(1)} + b^{(2)})|| = \sqrt{(a^{(1)} + a^{(2)})^2 + (b^{(1)} + b^{(2)})^2} = O(T^{1/3})$.

For each vector $v \in C$, let $e(v)$ denote the vector $(a^{(1)} + a^{(2)}, b^{(1)} + b^{(2)})$.

We also need to guarantee that each $e(v)$ will have norm $\Omega(T^{1/3})$. As the number of pairs of relatively prime integers (a, b) with $a, b \leq t$ is $\Omega(t^2)$, it follows that $\sum_{i=0}^{4k-1} ||(a_i, b_i)|| = \Omega(t^3)$. Consequently, $\sum_{i=0}^{4k-1} ||(a_i + a_{i+1}, b_i + b_{i+1})|| = \Omega(t^3)$, where $i+1$ is shorthand for the remainder of the division of $i+1$ by $4k-1$. It follows that for at least $\Omega(k)$ indices i, $||(a_i + a_{i+1}, b_i + b_{i+1})|| = \Omega(t^3/k) = \Omega(t) = \Omega(T^{1/3})$. Moreover, since the construction is symmetric around the axes, the same applies for the vectors in the first orthant.

Let C' denote the subset of the CIS C that contains only those vectors $v \in C$ that satisfy the condition that $||(a_i + a_{i+1}, b_i + b_{i+1})||$ is $\Omega(T^{1/3})$, where (a_i, b_i), (a_{i+1}, b_{i+1}) are the facets of the polygon $CH(C)$ that are adjacent to the vertex v. It follows that $|C'| = \Theta(T^{2/3})$.

DEFINITION 5.2. *For a set $Z \subseteq \mathbb{R}^2$ of points, the line $\mathcal{L} = \{v + \alpha \cdot u : \alpha \in \mathbb{R}\}$, $u = (a, b) \in \mathbb{R}^2$, $v \in Z$, is called a* supporting line of the set Z in the point $v \in Z$ *if the following condition holds: For every two points $x, x' \in Z \setminus \{v\}$, the signs of the inner products $\langle x - v, u^{\perp} \rangle$ and $\langle x' - v, u^{\perp} \rangle$ are the same, where $u^{\perp} = (-b, a)$.*

LEMMA 5.3. *Let $\mathcal{L}$ be the line (in the Euclidean plane) that passes through the point v and is parallel to the vector $e(v)$. Then the line $\mathcal{L}$ is a supporting line of the polygon $\hat{P}$ in the point v.*

Proof. We prove the lemma for the special case when the vector v is in the first quadrant (i.e., has nonnegative coordinates). The proof for the case that one of the coordinates of v is negative is symmetric to the proof of this case.

Let $v = w_j = (\sum_{i=j+1}^{k} a_i, \sum_{i=1}^{j} b_i)$ for some $j = 0, 1, \ldots, k$. We will prove that for every extreme point $x \neq v$ of the polygon P, $\langle x - v, u^{\perp} \rangle < 0$ for $u = e(v)$. Since every other point $y \neq v$ of the polygon $\hat{P}$ is a nontrivial convex combination of the extreme points of the polygon, this is sufficient for proving the lemma.

Observe that $w_{j+1} - w_j = (-a_{j+1}, b_{j+1})$ and $w_j - w_{j-1} = (-a_j, b_j)$. Hence, $u = (-(a_j + a_{j+1}), (b_j + b_{j+1}))$, and consequently, $u^\perp = (b_j + b_{j+1}, a_j + a_{j+1})$. Let x be an extreme point of the polygon $\hat{P}$, $x \neq v$. The proof splits into several cases.

1. $x = w_h$ for $j < h \leq k$.
   Then $x - v = w_h - w_j = \sum_{i=j+1}^{h}(-a_i, b_i)$, and

$$\langle x - v, u^\perp \rangle \;=\; -(b_j + b_{j+1}) \sum_{i=j+1}^{h} a_i + (a_j + a_{j+1}) \sum_{i=j+1}^{h} b_i.$$

   Hence,

$$sign(\langle x - v, u^\perp \rangle) \;=\; sign\left(\frac{a_j + a_{j+1}}{b_j + b_{j+1}} - \frac{\sum_{i=j+1}^{h} a_i}{\sum_{i=j+1}^{h} b_i} \right).$$

   Note that

$$\frac{a_j + a_{j+1}}{b_j + b_{j+1}} < \frac{a_{j+1}}{b_{j+1}} < \frac{\sum_{i=j+1}^{h} a_i}{\sum_{i=j+1}^{h} b_i},$$

   and thus $sign(\langle x - v, u^\perp \rangle) ='' -''$.
2. $x = w_h$, $0 \leq h \leq j$.
   This case is symmetric to case 1, and

$$sign(\langle x - v, u^\perp \rangle) \;=\; sign\left(\frac{\sum_{i=h+1}^{j} a_i}{\sum_{i=h+1}^{j} b_i} - \frac{a_j + a_{j+1}}{b_j + b_{j+1}} \right).$$

   Note that

$$\frac{a_j + a_{j+1}}{b_j + b_{j+1}} > \frac{a_j}{b_j} > \frac{\sum_{i=h+1}^{j} a_i}{\sum_{i=h+1}^{j} b_i},$$

   and so $sign(\langle x - v, u^\perp \rangle) ='' -''$.
3. For $h = 0, 1, \ldots, k$, let $w'_h = (-\sum_{i=h+1}^{k} a_i, \sum_{i=1}^{h} b_i)$. (Recall that $w_h = (\sum_{i=h+1}^{k} a_i, \sum_{i=1}^{h} b_i)$, and so w'_h is a reflection of w_h with respect to the x-axis.)
   If $h = j$, then $x - v = w'_j - w_j = (-2\sum_{i=j+1}^{k} a_i, 0)$, and $\langle x - v, u^\perp \rangle = -2(\sum_{i=j+1}^{k} a_i)(b_j + b_{j+1}) < 0$, as required.
   If $x = w'_h$, $h > j$, then $x - v = w'_h - w_j = (-(\sum_{i=j+1}^{h} a_i + 2\sum_{i=h+1}^{k} a_i), \sum_{i=j+1}^{h} b_i)$.
   Hence,

$$\begin{aligned} sign(\langle x - v, u^\perp \rangle) &= sign\left(-(b_j + b_{j+1}) \left(\sum_{i=j+1}^{h} a_i + 2 \sum_{i=h+1}^{k} a_i \right) \right. \\ &\qquad \left. + (a_j + a_{j+1}) \sum_{i=j+1}^{h} b_i \right) \\ &= sign\left(\frac{a_j + a_{j+1}}{b_j + b_{j+1}} - \frac{\sum_{i=j+1}^{h} a_i + 2\sum_{i=h+1}^{k} a_i}{\sum_{i=j+1}^{h} b_i} \right) \;=\; ''-'' \end{aligned}$$

because

$$\frac{a_j + a_{j+1}}{b_j + b_{j+1}} - \frac{\sum_{i=j+1}^{h} a_i + 2\sum_{i=h+1}^{k} a_i}{\sum_{i=j+1}^{h} b_i} \;<\; \frac{a_j + a_{j+1}}{b_j + b_{j+1}} - \frac{\sum_{i=j+1}^{h} a_i}{\sum_{i=j+1}^{h} b_i} \;<\; 0.$$

For $x = w'_h$, $h < j$, the computation is symmetric (the roles of the indices h and j are switched).

4. For $\bar{x} = -x$ such that $\langle x - v, u^\perp \rangle < 0$, it follows that $\langle -x, u^\perp \rangle \le 0 \le \langle x, u^\perp \rangle \le \langle v, u^\perp \rangle$, since both vectors x and $u^\perp$ have only nonnegative coordinates.
5. $x = w''_h = (\sum_{i=h+1}^{k} a_i, -\sum_{i=1}^{h} b_i)$.
   If $h = j$, then $\langle x - v, u^\perp \rangle = -2(a_j + a_{j+1}) \sum_{i=1}^{j} b_i < 0$.
   If $h > j$, then

$$\begin{aligned} x - v = w''_h - w_j \;&=\; \left(\sum_{i=h+1}^{k} a_i, -\sum_{i=1}^{h} b_i \right) - \left(\sum_{i=j+1}^{k} a_i, \sum_{i=1}^{j} b_i \right) \\ &= \left(-\sum_{i=j+1}^{h} a_i, -2\sum_{i=1}^{j} b_i - \sum_{i=j+1}^{h} b_i \right). \end{aligned}$$

Hence

$$\langle x - v, u^\perp \rangle \;=\; -\left\langle \left(\sum_{i=j+1}^{h} a_i, 2\sum_{i=1}^{j} b_i + \sum_{i=j+1}^{h} b_i \right), (b_j + b_{j+1}, a_j + a_{j+1}) \right\rangle \;<\; 0,$$

as required.

Finally, the case $h < j$ is symmetric (the roles of the indices h and j are switched). □

5.3. Constructing the instance. We next describe the construction of the graph $G = (V, E)$, and a subset $\mathcal{S} \subseteq V$ of sources that will be used in the proof of our lower bound for the sources problem.

Let $N > 10 \cdot T^2$ be another positive integer parameter. Enlarge the polygon P by a factor $\sqrt{N}/T$. In other words, let $P' = (\sqrt{N}/T) \cdot P = \{(\sqrt{N}/T) \cdot u : u \in P\}$. Note that the polygon P' is contained in a disc of radius $\sqrt{N}$, and therefore, it contains $O(N)$ points of the integer lattice $\mathbb{Z}^2$.

For each extreme point v of the polygon P that belongs to the set C' (that is, has norm $\Omega(T^{1/3})$), consider the line that passes through the origin and is perpendicular to the vector $e(v)$. Let A_v be one of the two points on this line (chosen arbitrarily) that are at (Euclidean) distance exactly $\sqrt{N}/2$ from the origin. Let I_v be a segment of length $\sqrt{N}$ parallel to the vector $e(v)$ that passes through the point A_v, and, furthermore, A_v is its center. Let C_v and D_v be its endpoints. Let I'_v be a segment of the same length as I_v (that is, $\sqrt{N}$), parallel to I_v, and such that its center A'_v is located on the line that connects the origin to A_v, at distance exactly T from the point A_v. Let C'_v, D'_v be its endpoints. Let B_v be the rectangle $C_v D_v D'_v C'_v$. (See Figure 6.)

Let $F = \frac{\sqrt{N}}{T}$, and let B'_v be the box B_v displaced by $F \cdot v$. In other words, $B'_v = \{x + F \cdot v : x \in B_v\}$. Note that by the triangle inequality, the box B'_v is contained in a disc of radius $2\sqrt{N}$.

For each integer point $x \in B_v \cap \mathbb{Z}^2$, insert the edges $\{x, x + v\}, \{x + v, x + 2v\}, \ldots, \{x + (F - 1) \cdot v, x + F \cdot v\}$ into the edgeset E of the graph $G = (V, E)$ that we construct.

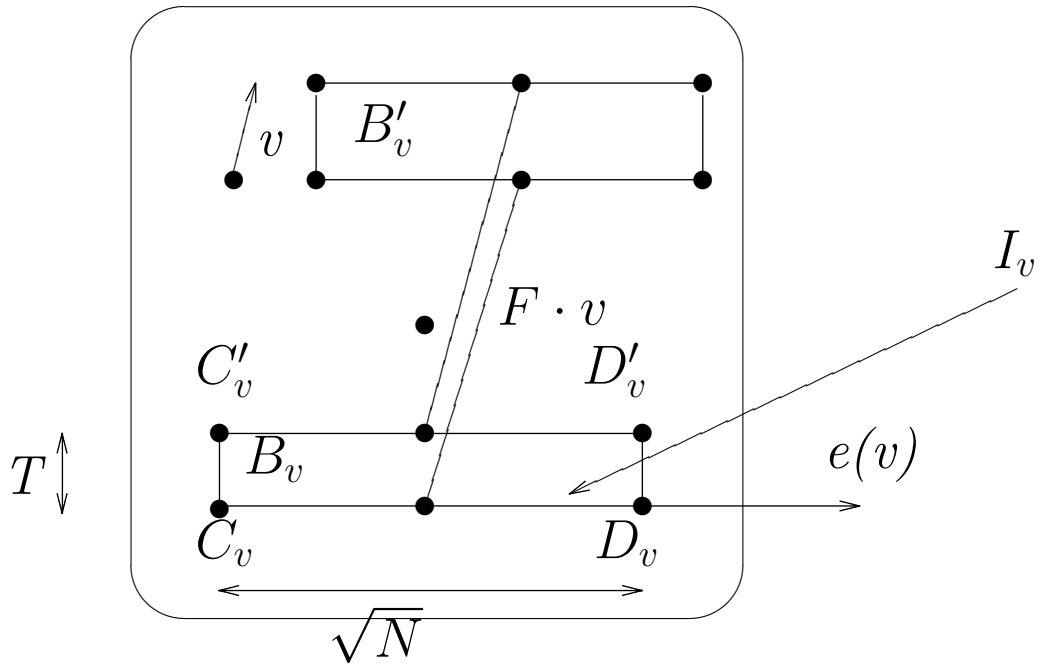


FIG. 6. *The boxes B_v and B'_v have real dimensions $\sqrt{N} \times T$, with the long edge of length $\sqrt{N}$ aligned in parallel to the vector $e(v)$.*

DEFINITION 5.4. *For the box B_v, a segment I_v is called an* aligned segment *of the box B_v if it contains at least one integer point x of the box B_v, it is parallel to the vector $e(v)$ (and, consequently, to the long edge of the box B_v), and it contains all the integer points of the box B_v of the form $x + \alpha \cdot e(v)$, $\alpha \in \mathbb{Z}$, and its endpoints lie on the boundary of the box B_v.*

For an aligned segment I_v , let Z_v denote the set of integer points of the segment I_v. The set Z_v will be referred to also as an integer aligned segment of the box B_v.

Fix the vector v for the rest of this section. Note that the real length of each aligned segment I_v is $\sqrt{N}$. Note also that the set of all the integer points of the box B_v decomposes into the disjoint union of integer aligned segments Z_v of B_v. For an integer aligned segment Z_v of the box B_v, let $Z'_v = Z_v + F \cdot v = \{x + F \cdot v : z \in Z_v\}$. By definition, the set Z'_v is an integer aligned segment of the box B'_v.

Let $m(v) = |Z_v|$ denote the cardinality of some set Z_v. The notation suggests that this cardinality does not depend on the choice of the particular integer aligned segment of the box B_v. In fact, this is not the case, and it may happen that for two different aligned segments Z_v and $\tilde{Z}_v$, $|Z_v| = \tilde{Z}_v \pm 1$. However, this difference of 1 has no effect on our analysis, and we will henceforth ignore it, and assume that $|Z_v| = m(v)$ for every integer aligned segment of the box B_v. As v is fixed, let m serve as a shortcut for $m(v)$.

For each integer aligned segment Z_v of B_v, order the points of Z_v according to the order in which they appear on the line (there are two such orders; choose either one of them arbitrarily). Let o_v be this ordering, and write $Z_v = (x_1, \ldots, x_m)$ with $x_i <_{o_v} x_j$ if and only if $i < j$. Order Z'_v in the same order, i.e., $Z'_v = (x'_1, \ldots, x'_m)$, $x'_i = x_i + F \cdot v$, $i \in [m]$.

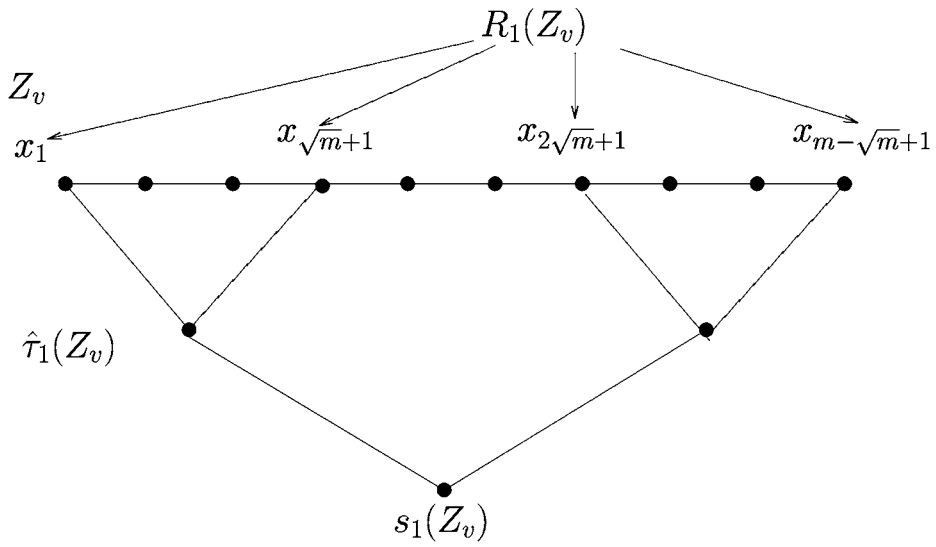


FIG. 7. *An illustration of an auxiliary tree.*

Arrange the numbers $(1, 2, \ldots, m)$ as an $\sqrt{m} \times \sqrt{m}$ matrix M (ignoring possible nonintegrality of $\sqrt{m}$; it affects only the lower-order terms of the analysis), with $M_{ij} = \sqrt{m}(i-1) + j$, for $i, j \in [\sqrt{m}]$.

Let R_i (resp., C_i), $i \in [\sqrt{m}]$, be the set of numbers that appear on the ith row (resp., column) of this matrix, i.e., $R_i = \{\sqrt{m}(i-1)+1, \sqrt{m}(i-1)+2, \ldots, (\sqrt{m})i\}$ (resp., $C_i = \{i, \sqrt{m}+i, \ldots, (\sqrt{m}-1)\sqrt{m}+i\}$). Let $R_i(Z_v) = \{x_\ell \in Z_v : \ell \in R_i\}$, and, analogously, $C_i(Z_v) = \{x_\ell \in Z_v : \ell \in C_i\}$. Obviously, for every pair of indices $i, j \in [\sqrt{m}]$, $|R_i \cap C_j| = 1$, and consequently, $|R_i(Z_v) \cap C_j(Z_v)| = 1$. For each index $i \in [\sqrt{m}]$, a new *source* $s_i(Z_v)$ is introduced, that is, a vertex of the graph G that belongs to the set $\mathcal{S}$ of sources. The source vertices do not correspond to points of the Euclidean plane. In addition, $\sqrt{m} - 2$ (assuming that $\sqrt{m}$ is a power of 2) new auxiliary vertices are introduced for each source. (If $\sqrt{m}$ is not a power of 2, then $2^t - 2$ vertices are introduced, where t satisfies $2^{t-1} < \sqrt{m} \le 2^t$.) These auxiliary vertices are used to form a complete binary tree rooted in $s_i(Z_v)$. This tree $\hat{\tau}$ has $\sqrt{m}/2$ leaves, all of which are new auxiliary vertices. Each of these leaves is connected to exactly two vertices of the set $R_i(Z_v)$, and each vertex of $R_i(Z_v)$ has exactly one neighbor in $\hat{\tau}$.

Consider the obtained complete binary tree $\hat{\tau}_i(Z_v)$, rooted in the source $s_i(Z_v)$. (See Figure 7 for an illustration.) This tree has depth $d = \log \sqrt{m}$, and its leaves are the $\sqrt{m}$ vertices of the set $R_i(Z_v)$. The vertices of this tree are assigned *levels* in the following way. Each leaf z is assigned level $\ell(z) = 0$, and each vertex w whose children are all assigned level ℓ is assigned level $\ell(w) = \ell + 1$. An edge $e = \{u, w\} \in \hat{\tau}_i(Z_v)$ that connects a vertex u to its parent w in the tree is assigned the level of the child u, i.e., $\ell(e) = \ell(u)$. Next, each edge e of the tree $\hat{\tau}_i(Z_v)$ is replaced with a path of length $20 \cdot 2^{\ell(e)} \cdot \frac{\sqrt{m}}{T^{2/3}}$, which is formed using new auxiliary vertices. The resulting tree is denoted $\tau_i(Z_v)$, and the tree $\hat{\tau}_i(Z_v)$ is called its *skeleton*.

Consider the edges e of the skeleton tree $\hat{\tau}_i(Z_v)$ that have a fixed level $\ell(e) = p$. The number of such edges is 2^{d-p}. Each edge e of $\hat{\tau}_i(Z_v)$ with $\ell(e) = p$ is replaced with a path of length $20 \cdot 2^p \cdot \frac{\sqrt{m}}{T^{2/3}}$ in the tree $\tau_i(Z_v)$, and so all edges e of level p

contribute together $O(\frac{\sqrt{m}}{T^{2/3}} \cdot 2^d) = O(\frac{m}{T^{2/3}})$ vertices to $\tau_i(Z_v)$, and thus, overall, $\tau_i(Z_v)$ contains $O(\frac{m \log m}{T^{2/3}})$ vertices.

An almost symmetrical tree is constructed for the vertices of Z'_v. Specifically, a new source $s'_i(Z'_v)$ is introduced, along with a bunch of new auxiliary vertices. These vertices are used to build a complete binary tree rooted in the source $s'_i(Z'_v)$. This tree has $\sqrt{m}/2$ leaves, and each of them is connected to two vertices of the set $C_i(Z'_v)$ such that each vertex of the set $C_i(Z'_v)$ is connected to exactly one such leaf. (Note that the difference is that $s'_i(Z'_v)$ is connected to vertices of $C_i(Z'_v)$, while $s_i(Z_v)$ is connected to vertices of $R_i(Z_v)$.) Finally, every edge e of the obtained tree $\hat{\tau}'_i(Z'_v)$ by a path of appropriate length (specifically, $20 \cdot 2^{\ell(e)} \cdot \frac{\sqrt{m}}{T^{2/3}}$, where $\ell(e)$ is the level of the edge), and this path is formed using new auxiliary vertices. The obtained tree is denoted $\tau'_i(Z'_v)$.

This is done separately for every extreme point $v \in Ex(P)$, for every integer aligned segment Z_v of B_v, and Z'_v of B'_v, and for every index $i \in [\sqrt{m(v)}]$.

5.4. The analysis of the construction. In this section we analyze the instance $G = (V, E)$, $\mathcal{S} \subseteq V$, that was constructed in the previous section. Note that for each vector v, each integer aligned segment Z_v, and each pair of sources $s_i(Z_v)$, $s'_j(Z'_v)$, $i, j \in [\sqrt{m(v)}]$, there exists a unique vertex $x \in Z_v$ such that x serves as a leaf of the tree $\tau_i(Z_v)$, and $x' = x + F \cdot v$ serves as a leaf of the tree $\tau'_j(Z'_v)$. Furthermore, for each vertex $x \in Z_v$, there exists a unique pair of sources $s_i(Z_v)$, $s'_j(Z'_v)$, as above. Therefore, for each vertex $x \in Z_v$, the shortest path between x and x' is indispensable; i.e., no edge of this path can be removed without increasing the distance between some pair of sources.

The involved way in which the trees $\tau_i(Z_v)$, $\tau'_j(Z'_v)$ are constructed is dictated by the necessity of balancing two contradictory requirements. On the one hand, we need to prevent these trees from affecting the geometric properties of the graph, that is, from making a distance in the graph between some pair of points of the Euclidean plane shorter than the Euclidean distance between them (divided by a scaling factor). On the other hand, we have to keep the number of vertices as small as possible, in order to achieve a stronger lower bound.

We next provide a few bounds on the number of vertices, edges, and sources of this instance.

LEMMA 5.5. *The number N' of vertices in the graph $G = (V, E)$ is $O(N + N^{3/4} \cdot T^{5/6} \cdot \log N)$, the number of edges is $|E| = O(N \cdot T^{2/3} + N^{3/4} \cdot T^{5/6} \cdot \log N)$, and the number of sources is $S = O(N^{1/4} \cdot T^{11/6})$.*

Proof. First, observe that all the vertices of the graph G that are contained in the different boxes B_v, B'_v for different extreme points v of the polygon P, as well as the vertices of the paths that connect these boxes, are all contained in the set of integer points of a disc of radius $2\sqrt{N}$, and therefore, the number of these vertices is at most $O(N)$. Hence, we need only argue that the number of auxiliary vertices is $O(N^{3/4} \cdot T^{5/6} \cdot \log N)$, and the number of sources is $O(N^{1/4} \cdot T^{11/6})$.

Consider a fixed box B_v for some extreme point $v \in Ex(P)$. Recall that the real dimensions of the box B_v are $\sqrt{N} \times T$, with the long edge (the one of length $\sqrt{N}$) parallel to the vector $e(v)$. Consider an integer aligned segment Z_v of B_v. Recall that the slope of the vector $e(v)$, a/b, satisfies $\Omega(T^{1/3}) = |a|, |b| \leq 2 \cdot T^{1/3}$, where the fraction a/b is in its lowest terms. It follows that $|Z_v| = \Theta(\sqrt{N}/T^{1/3})$. Observe that each aligned segment I_v of the box B_v has the same number $m(v) = \Theta(\sqrt{N}/T^{1/3})$ of integer points. Note that $m(v)$ does actually depend on $v \in Ex(P)$, but for every

$v \in Ex(P)$,

$$m(v) = \Theta(\sqrt{N}/T^{1/3}). \tag{7}$$

Let m denote $\Theta(\sqrt{N}/T^{1/3})$.

Recall that every set Z_v is partitioned into $\sqrt{m(v)}$ subsets $R_1(Z_v), R_2(Z_v), \ldots, R_{\sqrt{m(v)}}(Z_v)$ of equal size $\sqrt{m(v)}$. For each such subset, an auxiliary tree with

$$\Theta\left(\frac{m(v)\log m(v)}{T^{2/3}}\right) \;=\; \Theta\left(\frac{m\cdot\log m}{T^{2/3}}\right)$$

vertices is constructed. Hence, altogether these $\sqrt{m(v)}$ binary trees contain

$$\Theta\left(\frac{m^{3/2}\log m}{T^{2/3}}\right)$$

vertices.

The box B_v contains $\Theta(\sqrt{N}\cdot T)$ integer points, and since each aligned segment of B_v contains $m(v)$ vertices, it follows that the box B_v contains

$$\Theta\left(\frac{\sqrt{N}\cdot T}{m}\right)$$

aligned segments. Hence, altogether for all these segments, $\Theta(\sqrt{N}\cdot T^{1/3}\sqrt{m}\cdot\log m)$ vertices are created. Hence, overall there are $\Theta(\sum_v(\sqrt{N}\cdot T^{1/3}\sqrt{m}\cdot\log m)) = \Theta(N^{3/4}\cdot T^{5/6}\log N)$ auxiliary vertices (since there are $O(T^{2/3})$ vertices $v \in Ex(P)$).

To count the number of edges, note that for each extreme point $v \in Ex(P)$, the boxes B_v, B'_v are connected by $\Theta(\sqrt{N}\cdot T)$ paths of length $F = \sqrt{N}/T$ each. This sums up to at most $O(N)$ edges.

Let $\mathcal{H}_v$ denote the set of these paths. The norm of v is $\Theta(T)$, and the depth of the boxes B_v and B'_v is T. Hence, there is a subset $\mathcal{H}'_v \subseteq \mathcal{H}_v$ of these paths that contains at least a constant fraction of the paths of $\mathcal{H}_v$ (i.e., $\frac{|\mathcal{H}'_v|}{|\mathcal{H}_v|} = \Omega(1)$), and such that the paths in $\mathcal{H}'_v$ are pairwise disjoint. It follows that the number of edges in the paths of $\mathcal{H}_v$ is actually $\Theta(N)$.

Since there are $\Theta(T^{2/3})$ extreme points $v \in Ex(P)$, inequality (7) implies that there are $\Theta(N\cdot T^{2/3})$ edges in all these paths. The auxiliary trees contribute additional $\Theta(N^{3/4}\cdot T^{5/6}\cdot\log N)$ edges.

For the number of sources, note that for every extreme point v, and for every aligned segment of one of the boxes B_v or B'_v, $\sqrt{m(v)} = \Theta(\sqrt{m})$ sources are formed. Since each box contains

$$\Theta\left(\frac{\sqrt{N}\cdot T}{m}\right)$$

aligned segments, and there are $\Theta(T^{2/3})$ extreme points, overall we have

$$\Theta\left(\sqrt{m}\cdot\frac{\sqrt{N}\cdot T}{m}\cdot T^{2/3}\right) = \Theta\left(\frac{\sqrt{N}\cdot T^{5/3}}{\sqrt{m}}\right) = \Theta(N^{1/4}\cdot T^{11/6})$$

sources. ∎

Intuitively, the next lemma shows that the auxiliary trees do not affect the geometric properties of the graph.

LEMMA 5.6. *For an extreme point $v \in Ex(P)$, and a pair of points u, w that belong to the same integer aligned segment Z_v of the box B_v, any path Π between them that is contained entirely in some auxiliary tree $\tau_i(Z_v)$ can be replaced with a sequence $(u = u_0, u_1, \ldots, u_L = w)$, $L = |\Pi|$, of points in the plane satisfying that for every index $j \in [L]$, the vector $u_j - u_{j-1}$ is contained in the convex hull $\hat{P}$ of C.*

Proof. Consider the box B_v and an integer aligned segment Z_v of B_v. Consider a pair of vertices $u, w \in Z_v$, and let $\mathcal{L}$ be the line that passes through u and w. Suppose that there exists an auxiliary tree $\tau = \tau_i(Z_v)$, for some index $i \in [\sqrt{m(v)}]$, that contains both vertices u and w as leaves. Then u and w are also leaves of the skeleton tree $\hat{\tau} = \hat{\tau}_i(Z_v)$. Let z be the closest vertex of $\hat{\tau}$ that is an ancestor of both u and w in the tree $\hat{\tau}$. Note that the shortest path between u and w in the tree τ passes through the vertex z. Let $\ell \geq 1$ denote the distance between u and z in the skeleton tree $\hat{\tau}$. Then, by construction, the distance between them in the tree τ is

$$20 \cdot (\sqrt{m(v)}/T^{2/3}) \sum_{i=1}^{\ell-1} 2^i \;=\; 20 \cdot (\sqrt{m(v)}/T^{2/3}) \cdot (2^\ell - 1).$$

Hence, the distance between u and w in the tree τ is

$$40 \cdot (2^\ell - 1)(\sqrt{m(v)}/T^{2/3}) \geq 20 \cdot 2^\ell \cdot (\sqrt{m(v)}/T^{2/3}).$$

Consider the ordering o_v of the points of Z_v on the aligned segment I_v. By construction, since the closest common ancestor of u and w is at distance ℓ from each of them in the skeleton tree $\hat{\tau}$, it means that at most $(2^\ell - 2)$ points of Z_v that serve as leaves of $\hat{\tau}$ appear between u and w in the ordering o_v of the points of Z_v. In other words, at most $2^\ell - 2$ integer points of the set $R_i(Z_v)$ appear between u and w on the line $\mathcal{L}$. Each pair of consecutive points of $R_i(Z_v)$ is separated by at most $\sqrt{m(v)} - 2$ points of Z_v that appear between them on the line $\mathcal{L}$. Each pair of consecutive points of Z_v are at a Euclidean distance of at most $3T^{1/3}$. Hence, the Euclidean distance between u and w is at most $2^\ell \cdot \sqrt{m(v)} \cdot 3 \cdot T^{1/3}$.

In other words, starting from a vertex u on the Euclidean plane, and travelling for at least $q = 20 \cdot 2^\ell \cdot \frac{\sqrt{m(v)}}{T^{2/3}}$ edges on the tree $\tau_i(Z_v)$, brings us to another vertex w on the plane which is at a Euclidean distance of at most $r = 3 \cdot 2^\ell \cdot \sqrt{m(v)} \cdot T^{1/3}$ from the vertex u. Consider the sequence of points $(u, u + \frac{w-u}{q}, u + 2 \cdot \frac{w-u}{q}, \ldots, u + q \cdot \frac{w-u}{q} = w)$ in the plane (the points $u + i \cdot \frac{s-u}{q}$ are not necessarily integer). The Euclidean distance between each pair of consecutive points of this sequence is, naturally, $||w - u||/q = r/q \leq (3/20)T$. Observe that the convex hull $\hat{P} = CH(C)$ contains the disc of radius $(3/20)T$ centered at the origin. (In fact, it contains the disc of radius $T - o(T)$ centered at the origin, but for the current proof this much weaker statement is sufficient.) It follows that the vector $\frac{w-u}{q}$ is contained in the convex hull $\hat{P}$ of C. ∎

Obviously, the same lemma applies to the trees $\tau'_j(Z'_v)$.

Furthermore, this lemma readily generalizes to a pair of integer points u, w in the plane that belong to the vertex set V of the graph, but do not necessarily belong to the same aligned segment. Suppose that for such a pair u, w of points there exists a path Π' that is contained entirely in the union $\mathcal{J}$ of auxiliary trees τ and τ'. To see that Lemma 5.6 is applicable to such a pair of points, note that each such path Π' is a concatenation of one or more subpaths Π that satisfy the assumptions

of Lemma 5.6. Applying the lemma to each of these subpaths provides the desired generalized statement.

The proof of the next lemma was outlined in the beginning of this section. On the intuitive level, it uses Lemma 4.1 to argue that the way our construction uses CISs makes shortcutting impossible.

LEMMA 5.7. *No edge e of the graph $G = (V, E)$ can be removed without increasing the distance between at least one pair of sources $s, s' \in \mathcal{S}$.*

Proof. The edgeset E of the graph G contains two types of edges. First, we have the edges of the paths between u and u', that is, $\{u, u+v\}, \{u+v, u+2v\}, \ldots, \{u+(F-1)\cdot v, u+F\cdot v\}$, for some $v \in Ex(P)$, and $u \in B_v$. Let E_1 denote the collection of these edges. Second, we have the edges of the auxiliary trees τ and τ'. Let E_2 denote the collection of these edges, i.e., $E_2 = E \setminus E_1$. We start with proving the statement of the lemma for the edges $e \in E_1$.

Fix an edge $e = \{u + j \cdot v, u + (j+1) \cdot v\}$, $v \in Ex(P)$, $u \in B_v$, $j \in [(F)]$. Let u and u' be the endpoints of the corresponding path ($u' = u + F \cdot v$). Let Z_v be the integer aligned segment of B_v to which the vertex u belongs. Consequently, $u' \in Z'_v$. Let $m = m(v)$ denote $|Z_v| = |Z'_v|$. Recall that there exists a unique pair of indices $i, j \in [\sqrt{m}]$ such that $u \in R_i(Z_v) \cap C_j(Z_v)$. It follows that for this pair of indices $u \in R_i(Z_v)$, $u' \in C_j(Z'_v)$. We first argue that $dist_{G\setminus e}(u, u') > F$.

LEMMA 5.8. *The unique path between u and u' of length at most F in the graph G is $(\{u, u+v\}, \{u+v, u+2\cdot v\}, \ldots, \{u+(F-1)\cdot v, u+F\cdot v\})$, and this path uses the edge $e = \{u + j \cdot v, u + (j+1) \cdot v\}$.*

Proof. Consider some path Π' in G of length at most F between u and u'. Let Π' represent a concatenation, $\Pi' = \Pi'_1 \cdot \Pi'_2 \cdot \ldots \cdot \Pi'_\ell$, of the subpaths Π'_j, $j \in [\ell]$, $\ell \leq F$, such that every subpath Π'_j is contained entirely in either E_1 or E_2, and such that for every index $j \in [\ell - 1]$, $\Pi'_j \subseteq E_1$ if and only if $\Pi'_{j+1} \subseteq E_2$, and vice versa. Note that this representation is unique.

For each subpath Π'_j that is contained in E_2, observe that its endpoints w, w' are points in the plane. Hence (the generalized version of) Lemma 5.6 can be applied to each such subpath Π'_j, and so there exists a sequence $(w = w_0, w_1, \ldots, w_{L_j} = w')$ of points in the plane, with $L_j = |\Pi'_j|$, and for every index $i \in [L_j]$, the vector $w_i - w_{i-1}$ is contained in the convex hull $\hat{P}$ of the set C. We next construct a sequence Π'' of points in the following way. For each index $j \in [\ell]$ such that $\Pi'_j \subseteq E_1$, replace the subpath $(\{w_0^{(j)}, w_1^{(j)}\}, \{w_1^{(j)}, w_2^{(j)}\}, \ldots, \{w_{L_j-1}^{(j)}, w_{L_j}^{(j)}\})$ with the sequence $\Pi''_j = (w_0^{(j)}, w_1^{(j)}, \ldots, w_{L_j}^{(j)})$ of points in the plane. For each index $j \in [\ell]$ such that $\Pi'_j \subseteq E_2$, replace the subpath with the sequence of points $\Pi''_j = (w_0^{(j)}, w_1^{(j)}, \ldots, w_{L_j}^{(j)})$ that is obtained by applying Lemma 5.6 to this subpath. Finally, concatenate the sequences $\Pi'' = \Pi''_1 \cdot \Pi''_2 \cdot \ldots \cdot \Pi''_\ell$. (A concatenation of two sequences $(a_1, a_2, \ldots, a_t)$, $(a_t, a_{t+1}, \ldots, a_q)$ is defined as $(a_1, a_2, \ldots, a_q)$.)

Observe that the obtained sequence $\Pi'' = (u = u_0, u_1, \ldots, u_r = u')$, $r \leq F$, satisfies that for every index $j \in [r]$ the vector $u_j - u_{j-1}$ belongs to the convex hull $\hat{P}$ of the set C (as either an extreme point or an internal point). It follows that $F \cdot v = u' - u = \sum_{j=1}^{r}(u_j - u_{j-1})$. The argument is now identical to the one that was used in the proof of Lemma 4.1.

Specifically, let $\{d_1, \ldots, d_p\}$ be the set of distinct vectors from the collection $\{(u_j - u_{j-1}) : j \in [r]\}$, and let α_i, $i \in [p]$, denote their multiplicities (i.e., $\alpha_i = |\{j : u_j - u_{j-1} = d_i\}|$). It follows that $F \cdot v = \sum_{i=1}^{p} \alpha_i d_i$, $\sum_{i=1}^{p} \alpha_i = r \leq F$, and $\{d_1, \ldots, d_p\} \subseteq \hat{P}$. In other words, $v = \sum_{i=1}^{p} \frac{\alpha_i}{F} d_i$, $\sum_{i=1}^{p} (\frac{\alpha_i}{F}) \leq 1$, and for every

index $i \in [p]$, $\frac{\alpha_i}{F} \geq 0$, and $d_i \in \hat{P}$. Since the vector v is an extreme point of the polygon P, it follows that $p = 1$, $\alpha_1 = F$, and $d_1 = v$, proving that the unique path between u and u' of length at most F in the graph G is the path $(\{u, u+v\}, \{u+v, u+2v\}, \ldots, \{u+(F-1)\cdot v, u+F\cdot v\})$, and this path uses the edge e. Hence, $dist_{G\setminus e}(u, u') > F$. □

We next provide a lower bound on the distance in the graph G between a pair of "nonmatching" vertices $u \in Z_v$ and $w \in Z'_v$.

LEMMA 5.9. *For a vertex $u \in Z_v$, and a vertex $w' \in Z'_v$ such that $w' \neq u'$, $dist_G(u, w') > F$.*

Proof. Consider the polygon $u + F \cdot P = \{u + w : w \in F \cdot P\}$, where $F \cdot P = \{F \cdot z : z \in P\}$. Note that $u' = u + F \cdot v$ is an extreme point of this polygon (because v is an extreme point of the polygon P). Also, both vertices u' and w' belong to the integer aligned segment Z'_v. Hence, both points u' and w' lie on the aligned segment I'_v. By construction, the segment I'_v is parallel to the vector $e(v)$. Hence, the segment I'_v passes through the extreme point $u' = u + F \cdot v$ of the polygon $(u + F \cdot P)$ and is parallel to the vector $e(v)$. By Lemma 5.3, the segment I'_v is contained in a supporting line of the polygon $(u + F \cdot P)$ in the point u'. By the definition of the supporting line (Definition 5.2), for every point $z \neq u'$ on this line, z does not belong to the convex hull of the polygon $(u + F \cdot P)$, that is, to $(u + F \cdot \hat{P})$. Hence, $w' \notin (u + F \cdot \hat{P})$.

The rest of the proof is analogous to that of Lemma 5.8. Suppose for contradiction that there exists a path Π in the graph G of length $r \leq F$ between the vertices u and w'. Let $\Pi'' = (u = u_0, u_1, \ldots, u_r = w')$ be the sequence of points in the plane obtained from the path Π via the same transformation as that used in the proof of Lemma 5.8 for obtaining the sequence Π'' from the path Π'. This sequence satisfies that for every index $j \in [r]$, the vector $u_j - u_{j-1}$ belongs to the convex hull $\hat{P}$ of the set C.

Hence, $w' - u = \sum_{j=1}^{r}(u_j - u_{j-1})$. Let $\{d_1, \ldots, d_p\}$ be the set of distinct vectors from the collection $\{(u_j - u_{j-1}) : j \in [r]\}$, and let α_i, $i \in [p]$, denote their multiplicities. It follows that $w' - u = \sum_{i=1}^{p} \alpha_i d_i$, $\sum_{i=1}^{p} \alpha_i = r \leq F$, and $\{d_1, \ldots, d_p\} \subseteq \hat{P}$. Hence, $(w'-u)/F = \sum_{i=1}^{p}(\alpha_i/F)\cdot d_i$, $\sum_{i=1}^{p}(\alpha_i/F) \leq 1$, and for every index $i \in [p]$, $\alpha_i/F \geq 0$. It follows that the vector $(w' - u)/F$ belongs to the convex hull $\hat{P}$ of the polygon P. Hence, $w' - u \in F \cdot \hat{P}$, and so $w' \in (u + F \cdot \hat{P})$, which is a contradiction.

Hence, there is no path in G of length at most F between the vertices u and w' for $w' \neq u'$. Hence, $dist_G(u, w') > F$. □

We now return to the proof of Lemma 5.7. Consider the pair of vertices $u \in R_i(Z_v)$, $u' \in C_j(Z'_v)$. Recall that $s_i(Z_v)$ is the source that serves as the root of the auxiliary tree $\tau_i(Z_v)$ whose set of leaves is equal to $R_i(Z_v)$, and $s'_j(Z'_v)$ is the source that serves as the root of the auxiliary tree $\tau'_j(Z'_v)$ whose set of leaves is equal to $C_j(Z'_v)$. Observe that

$$dist_G(s_i(Z_v), s'_j(Z'_v)) \;=\; dist_G(s_i(Z_v), u) + dist_G(u, u') + dist_G(u', s'_j(Z'_v)).$$

By construction,

$$dist_G(s_i(Z_v), u) \;=\; dist_G(u', s'_j(Z'_v)) \;=\; 20 \cdot \frac{\sqrt{m}}{T^{2/3}}(2^{\ell} - 1), \tag{8}$$

where $\ell = \log\sqrt{m} = \frac{1}{2}\log m$. Hence the right-hand side of (8) is equal to $20\sqrt{m}(\sqrt{m}-1)/T^{2/3}$. Denote this expression by M. Hence, $dist_G(s_i(Z_v), s'_j(Z'_v)) = 2M + F$. We next show that $dist_{G\setminus e}(s_i(Z_v), s'_j(Z'_v)) > 2M + F$.

Observe that by construction,

$$dist_{G\setminus e}(s_i(Z_v), s'_j(Z'_v)) = \min\{dist_{G\setminus e}(s_i(Z_v), x) + dist_{G\setminus e}(x, y) + dist_{G\setminus e}(y, s'_j(Z'_v)) : x \in R_i(Z_v),\ y \in C_j(Z'_v)\}. \tag{9}$$

Since $dist_{G\setminus e}(s_i(Z_v), x) \geq dist_G(s_i(Z_v), x) = M$ for every $x \in Z_v$, and, analogously, $dist_{G\setminus e}(y, s'_j(Z'_v)) \geq M$ for every $y \in Z'_v$, it follows that $dist_{G\setminus e}(s_i(Z_v), s'_j(Z'_v)) \geq 2M + \min\{dist_{G\setminus e}(x, y) : x \in R_i(Z_v),\ y \in C_j(Z'_v)\}$. So, it remains to argue that for every pair of vertices $x \in R_i(Z_v)$, $y \in C_j(Z_v)$, $dist_{G\setminus e}(x, y) > F$.

Let u denote the unique vertex such that $\{u\} = R_i(Z_v) \cap C_j(Z_v)$. Let $u' = u + F \cdot v$. Note that $u' \in C_j(Z'_v)$. If $x = u$, $y = u'$, then by Lemma 5.8, $dist_{G\setminus e}(x, y) > F$. Otherwise, $y \neq x'$, and thus, by Lemma 5.9, $dist_G(x, y) \geq dist_{G\setminus e}(x, y) > F$. This proves the statement of the lemma for the case $e \in E_1$.

Consider an edge $e \in E_2$. There exists an auxiliary tree $\tau_i(Z_v)$ (the case that e belongs to a tree $\tau'_j(Z'_v)$ is symmetric) such that the edge e belongs to the edgeset of this tree. Let $u \in R_i(Z_v)$ be one of the leaves of the tree $\tau_i(Z_v)$ so that the unique path in the tree between the root $s_i(Z_v)$ and the leaf u uses the edge e. Let $j \in [m]$ be the unique index such that $\{u\} = R_i(Z_v) \cap C_j(Z_v)$. Consider the pair $s_i(Z_v)$, $s'_j(Z'_v)$ of sources. Following the argument that starts with (9), it is easy to see that in this case too,

$$dist_{G\setminus e}(s_i(Z_v), s'_j(Z'_v)) > 2M + F = dist_G(s_i(Z_v), s'_j(Z'_v)). \qquad \square$$

To summarize, we have proved that there exist infinitely many values of positive integer parameters N and T for which there exists a graph $G = (V, E)$ with $N' = O(N + N^{3/4} \cdot T^{5/6} \cdot \log N)$ vertices, $|E| = \Theta(N \cdot T^{2/3} + N^{3/4} \cdot T^{5/6} \cdot \log N)$ edges, and a subset $\mathcal{S} \subseteq V$ of $O(N^{1/4} \cdot T^{11/6})$ sources, so that removal of any edge e from the graph results in increasing the distance between some pair of sources.

Direct calculation shows that for $T \leq \frac{N^{3/10}}{\log^{6/5} N}$, $N' = O(N)$, and the lower bound of $|E| = \Omega(N^{10/11} \cdot S^{4/11})$ follows. (Up to this point the analysis requires only a weaker constraint $N > 10 \cdot T^2$. Hence, particularly, all previous calculations are applicable for T in the more narrow range $T \leq \frac{N^{3/10}}{\log^{6/5} N}$.) This lower bound is superlinear in $N + S^2$ for $\omega(N^{1/4}) = S = o(N^{5/9})$. We proved the following theorem.

THEOREM 5.10. *There exists infinitely many positive integer numbers N and S, $\Omega(N^{1/4}) = S = o(N^{5/9})$, for which there exist N-vertex unweighted undirected graphs $G = (V, E)$ and subsets $\mathcal{S} \subseteq V$ of S sources, with $|E| = \Omega(N^{10/11} \cdot S^{4/11})$ edges such that removal of any edge e from the graph results in increasing the distance between some pair of sources.*

6. The three-dimensional construction. In this section we devise a variant of our construction from section 5 that is based on a large CIS in the Euclidean *three-dimensional* space. This enables us to extend the range of values of S to which the lower bound of Theorem 5.10 applies. Specifically, while the lower bound of Theorem 5.10 is superlinear for $\omega(N^{1/4}) = S = o(N^{5/9})$, the lower bound that is based on the three-dimensional construction is superlinear in the range $\omega(N^{1/4}) = S = o(N^{9/16})$. Furthermore, the new lower bound is stronger than the one of Theorem 5.10 for $\omega(\sqrt{N}) = S = o(N^{9/16})$.

The d-dimensional variants of our construction, for $d = 4, 5, \ldots$, yield no improvement to these results, but require a more complicated analysis. Hence, we restrict our attention to the dimension $d = 3$.

6.1. Constructing a dense CIS. For some fixed positive integer parameter T, consider the polytope $\hat{P}$ defined as the convex hull of the set of integer points of the ball of radius T centered at the origin. By [8], for a sufficiently large T, this polytope has $\Theta(T^{3/2})$ extreme points and $\Theta(T^{3/2})$ facets. Let P denote the boundary of this polytope. Observe that the surface area of P is no greater than the surface area of the three-dimensional sphere of radius T, that is, $\Theta(T^2)$.

Note that since the facet f passes through three linearly independent integer points (the extreme points of the polytope P), it follows that it has an integer normal vector. We define $a(f)$ to be this integer normal in its reduced form. Formally, $a(f) = (a_1, a_2, a_3)$ satisfies $\mathtt{gcd}(a_1, a_2, a_3) = 1$. (We define $\mathtt{gcd}(0, x, y) = \mathtt{gcd}(x, y)$, $\mathtt{gcd}(0, 0, x) = x$, etc.)

LEMMA 6.1. *There exists a subset $V'' \subseteq V$ of the extreme points of the polytope P that satisfies the following:*

1. $|V''| \geq \frac{4}{5}|V|$ *(and hence* $|V''| = \Theta(T^{3/2})$*).*
2. *Every extreme point $v \in V''$ has $O(1)$ adjacent facets, and all these facets $f \in F$ have normals $a(f)$ of norm at most $||a(f)|| = O(T^{1/2})$.*

Proof. Let $|V| = c_V T^{3/2}$ denote the number of extreme points of the polytope P, $|F| = c_F T^{3/2}$ denote the number of two-dimensional facets of P, and $A = c_A T^2$ denote its surface area, where c_V, c_F, and c_A are positive real constants.

It follows that the number of facets $f \in F$ with normal $a(f)$ that satisfies $||a(f)|| > \max\{\frac{20 \cdot c_A}{c_F}, \frac{60 \cdot c_A}{c_V}\} \cdot T^{1/2}$ is at most

$$\frac{A}{10(c_A/c_F)T^{1/2}} \;=\; \frac{1}{10}c_F T^{3/2} \;=\; \frac{1}{10}|F|.$$

Let $c_0 = \max\{\frac{20 \cdot c_A}{c_F}, \frac{60 \cdot c_A}{c_V}\}$. Furthermore, those facets contain together at most

$$\frac{3 \cdot A}{||a(f)||/2} \;\leq\; \frac{1}{10}|V|$$

integer points (because if we divide the facet f into triangles, then each triangle has an area of at least $||a(f)||/2$ and contains three integer points).

Let $F' = \{f \in F : ||a(f)|| \leq c_0 T^{1/2}\}$. It follows that $|F'| \geq \frac{9}{10}|F| = \frac{9}{10}c_F T^{3/2}$. Let $V' \subseteq V$ be the subset of extreme points of the polygon P that contains only extreme points that are adjacent only to facets $f \in F'$. Since all the extreme points of P have integer coordinates, it follows that $|V'| \geq \frac{9}{10}|V| = \frac{9}{10}c_V T^{3/2}$.

Consider the following bipartite graph $G'_P = (V', F', E'_P)$, with $(v, f) \in E'_P$ whenever $v \in f$, $v \in V'$, $f \in F'$. (Analogously, let $G_P = (V, F, E_P)$, $E_P = \{(v, f) : v \in f, v \in V, f \in F\}$.) We next argue that $|E'_P| = O(|V|)$. To see it, note that a facet that is adjacent to ℓ vertices is an ℓ-gon and is adjacent to ℓ one-dimensional facets (edges) of P. Each such edge is shared by exactly two facets. So the facet contributes ℓ edges to the graph G_P, and $\ell/2$ one-dimensional facets to the polytope P. Hence, denoting by R the number of the one-dimensional facets of the polytope P, we get $R = |E_P|/2$. By Euler's formula, $|V| - R + |F| = 2$, and so $|E_P|/2 = R = |V| + |F| - 2$, implying $|E_P| = 2|V| + 2|F| - 4 = O(|V|)$ (since for a three-dimensional polytope, $|F| = O(|V|)$). Hence, $|E'_P| \leq |E_P| = O(|V|)$.

Let c denote the constant such that $|E'_P| \leq c|V|$. Disregard all the extreme points with at least $10c$ adjacent facets. There are at most $(1/10)|V|$ such extreme points. Let $V'' \subseteq V'$ be the subset of all the extreme points of V' with at most $10c$ adjacent facets. It follows that $|V''| \geq \frac{4}{5}|V|$. ∎

For each extreme point $v \in V$, let $e(v)$ be the plane that passes through v and that is normal to the vector

$$a(v) \;=\; \sum\{a(f) : (v,f) \in E'_P\}, \tag{10}$$

where $a(f)$ is the normal of the facet f.

Observe that since each $v \in V''$ has only $O(1)$ adjacent facets, $||a(v)|| = O(1) \cdot \max\{||a(f)|| : (v,f) \in E'_P\}$. Also, since for every vector $v \in V''$ and facet f such that $(v,f) \in E'_P$, $||a(f)|| = O(T^{1/2})$, it follows that $||a(v)|| = O(T^{1/2})$ as well.

Note that Definition 5.2 of supporting lines naturally generalizes to supporting planes.

LEMMA 6.2. *The plane $e(v)$ is a supporting plane of the polytope P.*

Proof. Consider some point $x \in \hat{P}$. We need to show that $\langle x - v, a(v)\rangle < 0$, for every $x \in \hat{P}$, $x \neq v$. Since $x \in \hat{P}$, and $\hat{P}$ is a convex polytope, x is a convex combination of the extreme points of $\hat{P}$. That is, $x = \sum_{i=1}^{\ell} \alpha_i \cdot v_i$, $\sum_{i=1}^{\ell} \alpha_i = 1$, $\alpha_i > 0$, for every $i \in [\ell]$, and $\ell \geq 1$. Hence,

$$\begin{aligned}\left\langle \sum_{i=1}^{\ell} \alpha_i v_i, a(v) \right\rangle - \langle v, a(v)\rangle &= \sum_{i=1}^{\ell} \alpha_i \left\langle v_i - v, \sum_{(v,f)\in E_P} a(f) \right\rangle \\ &= \sum_{i=1}^{\ell} \sum_{(v,f)\in E_P} \alpha_i \langle v_i - v, a(f)\rangle. \end{aligned} \tag{11}$$

The inner products $\langle v_i - v, a(f)\rangle$ are always nonpositive because $v_i \in \hat{P}$ and the facet f contains the vertex v. It can be equal to zero only when $v_i \in f$. Hence, for the sum (11) to degenerate, *all* the vertices $v_1, v_2, \ldots, v_\ell$ must belong to all the facets $f \in F$ such that $(v,f) \in E_P$. In other words, it follows that $v_1 = v_2 = \cdots = v_\ell = v$, as required. ∎

To summarize, we have constructed a CIS V'' of $\Theta(T^{3/2})$ integer vectors of norm at most T, a polytope $\hat{P}$ that contains all of them as extreme points, and a supporting plane $e(v)$ for every vector $v \in V''$ with normal $a(v)$ of norm $O(T^{1/2})$. It is not hard to ensure also that the norms will be $\Omega(T^{1/2})$ by eliminating only a constant fraction of vertices of V''.

6.2. Constructing the instance. We next use the construction of the dense three-dimensional CIS for improving the lower bound on the size of sourcewise preservers.

Let N and T, $N > 10 \cdot T^3$, be two sufficiently large positive integer parameters. We start with enlarging the polytope $\hat{P}$ by a factor of $N^{1/3}/T$.

Similarly to the two-dimensional construction, for each vertex $v \in V''$ we now build a box B_v of real dimensions $N^{1/3} \times N^{1/3} \times T$ with the square facet of dimensions $N^{1/3} \times N^{1/3}$ parallel to the plane $e(v)$ and with the rectangular facet of dimensions $N^{1/3} \times T$ perpendicular to $e(v)$. The box is placed at a distance of roughly $N^{1/3}/2$ from the origin (the exact location of the box is not crucial for the analysis; one possible way of placing those boxes is analogous to the way it is done in section 5.3 for the two-dimensional boxes). Define $F = \frac{N^{1/3}}{T}$, and let B'_v be the box B_v displaced by $F \cdot v$, that is, $B'_v = \{x' = x + F \cdot v : x \in B_v\}$.

Now, for each vertex x, x and x' are connected by the paths of length F, exactly as it was done in the two-dimensional construction. The construction of auxiliary trees is also similar, and the main difference is the ordering of points on integer aligned

segments. (An aligned segment I_v is defined analogously to Definition 5.4. It is a two-dimensional square with both edges of length $N^{1/3}$, parallel to the plane $e(v)$. An integer aligned segment is the set of integer points of the aligned segment I_v. Let $m = m(v)$ denote the cardinality of I_v.)

We next specify the way the subsets $R_i(Z_v)$, $C_j(Z_v)$, $i, j \in [\sqrt{m(v)}]$ are formed. Let $n = m^{1/4}$. Partition the square to $\sqrt{m}$ strips of real dimensions $m^{1/2} \times n$ by $n-1$ lines parallel to one of the edges of the square. Order the n strips $H_1, H_2, \ldots, H_n$ in one of the two natural orders (in other words, one of the long edges of the strip H_1 is one of the edges of the square; the strip H_2 shares the long edge with the strip H_1, etc.). Do the same with respect to the other edge of the square, and order the strips $J_1, J_2, \ldots, J_n$. This way we partitioned the aligned segment I_v into $\sqrt{m}$ subsegments $I_v(i,j) = H_i \cap J_j$, $i, j \in [n]$, of real dimensions $N^{1/6} \times N^{1/6}$ each.

Let $Z_v(i,j)$ denote the set of integer points of $I_v(i,j)$; the set $Z_v(i,j)$ will be referred to as the *integer aligned* (i,j)*-subsegment* of Z_v. For an index $i \in [\sqrt{m}]$, let $j(i) = ((i-1) \bmod n) + 1$ and $\ell(i) = ((i-j)/n) + 1$. For each index $i \in [\sqrt{m}]$, we define $R_i(Z_v) = Z_v(\ell(i), j(i))$. (In other words, each set $Z_v(\ell, j)$ becomes $R_i(Z_v)$ for $i - 1 = (j-1) + (\ell - 1) \cdot n$, $j - 1 \in [(n)]$.) To define $C_j(Z_v)$ we do the following. Insert into $C_1(Z_v)$ one arbitrary element from each subset $Z_v(i,j)$. Insert into $C_2(Z_v)$ one arbitrary element from each $Z_v(i,j)$ that was not chosen for insertion into $C_1(Z_v)$. Keep forming disjoint subsets $C_1(Z_v), C_2(Z_v), \ldots, C_{\sqrt{m}}(Z_v)$ this way. Note that $Z_v = \bigcup_{i=1}^{\sqrt{m}} R_i(Z_v) = \bigcup_{j=1}^{\sqrt{m}} C_j(Z_v)$. Note also that $|R_i(Z_v) \cap C_j(Z_v)| = 1$ for every pair of indices $i, j \in [\sqrt{m}]$.

To form the auxiliary trees $\tau_i(Z_v)$ rooted in a source $s_i(Z_v)$, connect each vertex of $R_i(Z_v)$ to $s_i(Z_v)$ via paths of length $\frac{2n}{T^{1/2}}$ that use new auxiliary vertices. (See Figure 8.) To form the auxiliary trees $\tau'_j(Z'_v)$ rooted in the vertices $s'_j(Z'_v)$ we divide the set $C_j(Z'_v)$ into $\sqrt{m}$ subsets $H_1(C_j(Z'_v)), H_2(C_j(Z'_v)), \ldots, H_n(C_j(Z'_v))$, with $H_i(C_j(Z'_v)) = H_i \cap C_j(Z'_v)$ for each index $i \in [n]$. Note that each set $H_i(C_j(Z'_v))$ contains precisely one element that belongs to the strip J_ℓ for each index $\ell \in [n]$. Order the elements of $H_i(C_j(Z'_v))$ such that its first element will belong to the strip J_1, its second element will belong to the strip J_2, etc.; its nth element will belong to the strip J_n.

Observe that the Euclidean distance between two consecutive elements of $H_i(C_j(Z'_v))$ in this ordering is at most the diameter of the parallelogram with one edge of length $2n$ and the other of length n, that is, at most $3n$.

Now for each pair of indices $i, j \in [n]$, the tree $\tau'_{ij}(Z'_v)$ is formed in almost the same way as the tree $\tau'_j(Z'_v)$ is formed in the two-dimensional construction. There are the following differences:

1. The root $s'_{ij}(Z'_v)$ is not a source, but rather an "ordinary" auxiliary vertex.
2. Each edge of the skeleton tree $\hat{\tau}$ is replaced with a path of length $20 \cdot 2^{\ell(e)} \frac{n}{T^{1/2}}$, and not $20 \cdot 2^{\ell(e)} \frac{\sqrt{m}}{T^{2/3}}$.
3. Its set of leaves is $H_i(C_j(Z'_v))$.
4. The source $s'_j(Z'_v)$ is now connected via paths of length $\frac{20n}{T^{1/2}}$ to each of the vertices $s'_{ij}(Z'_v)$, $i \in [n]$.

These paths are formed with new auxiliary vertices. This completes the construction of the instance. (See Figure 9.)

Observe that, since for every pair of indices $i, j \in [n]$, we have $|H_i(C_j(Z'_v))| = n$ and $\ell(e) \leq \log n$, each tree $\tau'_{ij}(Z'_v)$ contains $O(\frac{n^2 \cdot \log n}{T^{1/2}}) = O(\frac{m \cdot \log m}{T^{1/2}})$ vertices. The star rooted in the source $s'_j(Z'_v)$ contributes additional $O(\sqrt{m}/T^{1/2})$ edges and vertices.

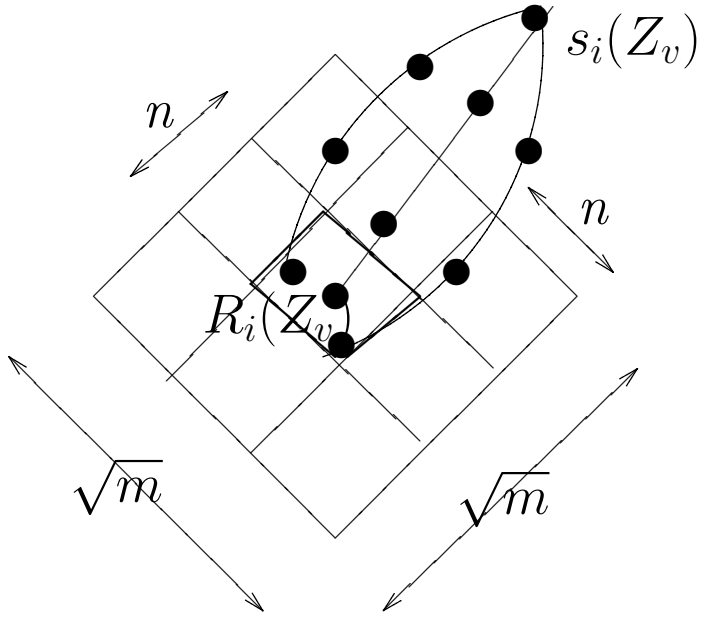


FIG. 8. *The paths between the vertices of $R_i(Z_v)$ and the source $s_i(Z_v)$ are of length $\frac{2 \cdot n}{\sqrt{T}}$. $R_i(Z_v)$ is the set of integer points of the square of real dimensions $(\Theta(n \cdot \sqrt{T})) \times (\Theta(n \cdot \sqrt{T}))$, depicted in the middle. Hence, it has $\Theta(n)$ integer points lying along each of its edges.*

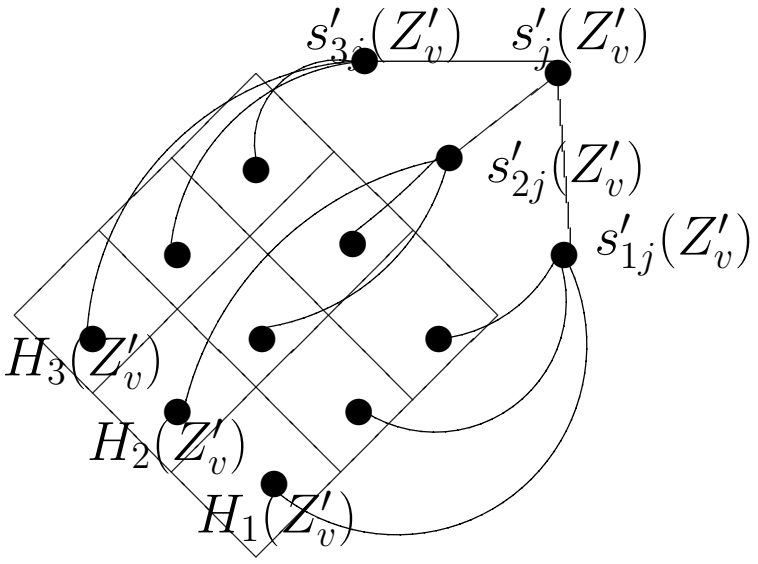


FIG. 9. *The points of the set $C_j(Z'_v)$ are depicted by large black circles inside the square. Each tree $\tau'_{ij}(Z'_v)$ spans the set $H_i(Z'_v)$; specifically, the set $H_i(Z'_v)$ is the set of leaves of the tree $\tau'_{ij}(Z'_v)$. The roots of those trees, that is, the vertices $s'_{ij}(Z'_v)$ for different indices $i \in [n]$, are connected to the source $s'_j(Z'_v)$.*

Hence, for every index $j \in [n]$,

$$|V(\tau'_j(Z'_v))| = \left(\sum_{i=1}^{n} |V(\tau'_j(Z'_v))|\right) + O\left(\frac{\sqrt{m}}{T^{1/2}}\right)$$
$$= O\left(\frac{n\sqrt{m}\cdot\log m}{T^{1/2}}\right) \;=\; O\left(\frac{m^{3/4}\cdot\log m}{T^{1/2}}\right),$$

where for a tree τ, $V(\tau)$ stands for its vertex set. It is easy to see that for every index $i \in [n]$, $|V(\tau_i(Z_v))| = O(\frac{m^{3/4}}{T^{1/2}})$.

6.3. The analysis of the three-dimensional construction. We start by analyzing the parameters of the instance.

LEMMA 6.3. *The number N' of the vertices in the graph $G = (V, E)$ is $O(N + N^{5/6}\cdot\log N\cdot T^{15/8})$, the number of edges is $|E| = O(N\cdot T^{3/2} + N^{5/6}\cdot\log N\cdot T^{15/8})$, the number of sources is $S = N^{1/3}\cdot T^{11/4}$.*

Proof. By the same argument as in the proof of Lemma 5.5, the number of vertices that are contained in $\bigcup_{v\in V''}(B_v \cup B'_v)$ is $O(N)$ (because they are all contained in a three-dimensional ball of radius $O(N^{1/3})$).

Also, since by Lemma 6.1 each aligned segment has normal a of norm at most $||a|| = O(T^{1/2})$, it follows that the set of integer points of each box B_v (or B'_v) decomposes into a disjoint union of $\Theta(T^{3/2})$ integer aligned segments, each of size $m = \Theta(\frac{N^{2/3}}{T^{1/2}})$ (as in the two-dimensional construction, this size depends on the vector v, but for every $v \in V''$, $m(v) = \Theta(\frac{N^{2/3}}{T^{1/2}})$).

For each integer aligned segment, $O(\sqrt{m})$ auxiliary trees with $O(\sqrt{m}\cdot\frac{m^{1/4}}{T^{1/2}}\cdot\log m) = O(\frac{m^{3/4}}{T^{1/2}}\cdot\log m)$ vertices each are formed, summing up to $O(\frac{m^{5/4}}{T^{1/2}}\cdot\log m)$ vertices in each aligned segment. Since each box contains $\Theta(T^{3/2})$ aligned segments, it follows that there are $O(m^{5/4}\cdot T\cdot\log m)$ auxiliary vertices per box. Also, recall that $|V''| = O(T^{3/2})$, and so there are $O(T^{3/2})$ different boxes. Hence, the overall number of auxiliary vertices is $O(m^{5/4}\cdot T^{5/2}\cdot\log m)$. Substituting $m = \Theta(\frac{N^{2/3}}{T^{1/2}})$ implies $N' = O(N + N^{5/6}\cdot T^{15/8}\cdot\log N)$.

A similar argument shows that the number of sources is $S = O(N^{1/3}\cdot T^{11/4})$ and the number of edges is $|E| = O(N\cdot T^{3/2})$. ∎

Observe that Lemma 6.3 implies the lower bound $|E| = \Omega(N^{9/11}\cdot S^{6/11})$, assuming $N' = O(N)$. The latter condition holds whenever $T = O(\frac{N^{4/45}}{(\log N)^{8/15}})$. This lower bound is superlinear in $N + S^2$ whenever $\omega(N^{1/4}) = S = o(N^{9/16})$. This happens for $T = o(N^{1/12}) = O(\frac{N^{4/45}}{(\log N)^{8/15}})$. Hence, we obtain the lower bound of $|E| = \Omega(N^{9/11}\cdot S^{6/11})$ which provides a nontrivial lower bound on the number of edges for $\omega(N^{1/4}) = S = o(N^{9/16})$ (extending the range from $S = o(N^{5/9})$ of Theorem 5.10). The new lower bound also improves the one of Theorem 5.10 in the range $\omega(N^{1/2}) = S = o(N^{9/16})$.

To complete the proof, we need to argue that no edge can be removed from the graph G without increasing the distance between at least one pair of sources from $\mathcal{S}$. To this end, observe that the auxiliary trees $\tau_i(Z_v)$, $\tau'_j(Z'_v)$ that were formed as part of the construction of the graph G do not create viable shortcuts between the vertices that correspond to points of the Euclidean space. Given this observation, the rest of the proof is completely analogous to the proof of Lemma 5.7, and is, therefore,

omitted. (Similarly to the proof of Lemma 5.7, this proof uses Lemma 4.1 to argue that short-cutting is impossible.)

6.4. Lower bound for weighted graphs. In this section we show a slightly stronger lower bound on the size of the sourcewise preserver, that, however, applies only to *weighted* graphs.

For a fixed sufficiently large positive integer parameter T, let $\hat{\Gamma}$ be the set of all pairs (ℓ, k) of relatively prime integer numbers with absolute values between $T/2$ and T.

For another fixed integer parameter $N > 10 \cdot T^2$, consider the set of integer points of the disc of radius $\sqrt{N}$ centered at the origin. For each vector $v \in \hat{\Gamma}$ form a rectangular box B_v of real dimensions $(\sqrt{N}/2) \times T$ and place it similarly to the way it was done in the two-dimensional unweighted construction (see section 5.3). Specifically, let $A_v = -\frac{v}{||v||}\sqrt{N}$, and let I_v be the segment of length $\sqrt{N}/2$ that is perpendicular to the vector v and has its center in A_v. Let $\tilde{I}_v$ be the segment I_v displaced by the vector $T \cdot \frac{v}{||v||}$. The endpoints of these two segments I_v and I'_v form the corners of the box B_v. (Note that the main difference between this location and its location in the unweighted construction is that there, the long edge of the box was parallel to the supporting line $e(v)$, while here it is perpendicular to v.) Let $F = (\sqrt{N}/T)$ and $B'_v = B_v + F \cdot v$ (as in the unweighted construction).

Notice that unlike the unweighted construction, here it is obvious that the box B'_v is contained in the disc of radius $\sqrt{N}$ centered at the origin (while there we used triangle inequality for a slightly weaker upper bound of $2\sqrt{N}$ on the radius).

Next, each vertex $x \in B_v$ is connected to $x' \in B'_v$ by a path $\{x, x+v\}, \{x+v, x+2v\}, \ldots, \{x+(F-1)\cdot v, x+F\cdot v\}$, and each edge of this path is assigned weight $||v||$.

The box B_v decomposes into $\Theta(T^2)$ integer aligned segments with $\Theta(\frac{\sqrt{N}}{T})$ integer points on each. For each integer aligned segment Z_v, $|Z_v| = m(v)$, the sets $R_i(Z_v)$, $C_j(Z_v)$, $i, j \in [\sqrt{m(v)}]$, are formed exactly in the same way as in section 5.3. For each index $i \in [\sqrt{m(v)}]$ we form a source $s_i(Z_v)$, which is a new vertex, and connect $s_i(Z_v)$ to each edge of $R_i(Z_v)$ via edges of weight N. Analogously, for each index $j \in [\sqrt{m(v)}]$ we form a source $s'_j(Z'_v)$, and connect it to each vertex of the set $C_j(Z'_v)$ via edges of weight N. Notice that these weighted stars replace the somewhat more complicated construction of auxiliary trees $\tau_i(Z_v)$, $\tau'_j(Z'_v)$ from section 5.3. Here the freedom in setting the weights of the edges enables us to guarantee that these edges will not enable shortcuts between pairs of vertices in the plane in this simple way.

Next, we analyze this construction. Observe that the number of vertices is $O(N)$, and the number of edges is $\Theta(N \cdot T^2)$ (since $|\hat{\Gamma}| = \Theta(T^2)$). For the number of sources note that for each integer aligned segment of length $m(v)$, $O(\sqrt{m(v)})$ sources are formed. Note that as in the unweighted construction, the number of integer points $m = m(v)$ in an integer aligned segment Z_v depends on v. However, for every v, $m(v) = \Theta(\frac{\sqrt{N}}{T})$. We denote the right-hand side of this equality by m.

Since there are $\Theta(T^2)$ integer aligned segments in each box, and $\Theta(T^2)$ different boxes, it follows that $S = \Theta(\sqrt{m} \cdot T^4) = \Theta(N^{1/4} \cdot T^{7/2})$. (As $S \leq N$, this implies the constraint of $T \leq N^{3/14}$ on our construction.) The lower bound of $|E| = \Omega(N^{6/7} \cdot S^{4/7})$ follows. This lower bound is superlinear in $N + S^2$ for a wider range of values of S than the lower bound of section 6.3, specifically, $\omega(N^{1/4}) = S = o(N^{3/5})$. This lower bound is also stronger than the aforementioned lower bound for every value of S in which either of the lower bounds is superlinear in $N + S^2$. (However, of course, the new lower bound does not apply to *unweighted* graphs, while the lower bound of section 6.3 does.)

To complete the proof we need only argue that no edge of the graph G can be removed without increasing the distance between some pair of vertices.

LEMMA 6.4. *For every edge $e \in E$, there exists a pair of sources $s, s' \in \mathcal{S}$ such that $dist_{G\setminus e}(s, s') > dist_G(s, s')$.*

Proof. Let E_1 be the set of edges that are not incident to any source vertex, and $E_2 = E \setminus E_1$. We will prove the lemma for the case $e \in E_1$. For the case $e \in E_2$ the proof is actually simpler and uses a similar argument (see also the proof of Lemma 5.7).

The edge $e \in E_1$ belongs to some path $\{x, x+v\}, \{x+v, x+2v\}, \ldots, \{x+(F-1)v, x+F\cdot v\}$. Let $x \in V$, $v \in \hat{\Gamma}$ be the pair of vectors as above. Let $i, j \in [\sqrt{m(v)}]$ be the unique pair of indices such that $x \in R_i(Z_v) \cap C_j(Z_v)$. We will show that $dist_{G\setminus e}(s_i(Z_v), s'_j(Z'_v)) > dist_G(s_i(Z_v), s'_j(Z'_v))$. First note that $dist_G(s_i(Z_v), s'_j(Z'_v)) = 2N + dist_G(x, x') = 2N + F \cdot ||v||$.

Observe that

$$dist_{G\setminus e}(s_i(Z_v), s'_j(Z'_v)) = 2N + \min\{dist_{G\setminus e}(y, z') : y \in R_i(Z_v),\ z \in C_j(Z_v)\}.$$

By construction, for any pair points $y, z \in Z_v$, $y \neq z$, the Euclidean distance between the points y and z', $||y-z'||$ is strictly greater than $F\cdot||v||$ (because the integer aligned segments Z_v and Z'_v are parallel to one another and perpendicular to the vector v).

Note also that for any two nonsource vertices of the graph G, the distance between them in G is greater than or equal to the Euclidean distance between them (each such vertex corresponds to a point in the plane). It follows that $dist_{G\setminus e}(y, z') \geq dist_G(y, z') \geq ||y - z'|| > F \cdot ||v||$ for every pair of vertices $y, z \in Z_v$, $y \neq z$.

Since $\{x\} = R_i(Z_v) \cap C_j(Z_v)$, it remains to argue that $dist_{G\setminus e}(x, x') > F \cdot ||v||$. The proof of this fact is analogous to the proof of Lemma 3.3. ∎

Finally, we remark that all our lower bounds hold for every sufficiently large N, even though they were proven only for all values of N that belong to certain monotone increasing infinite sequences of positive integers (such as the sequence of squares of integers in the case of Corollary 3.4). To see it note that those sequences are all sufficiently dense, and in particular, for every sufficiently large N not in the sequence there is a number N' in the sequence such that $N/2 \leq N' < N$. To derive the lower bound for N, we use the construction for N' and add to it $N - N'$ dummy isolated vertices. Since $N' \geq N/2$, the lower bound deteriorates only by a constant factor.

7. Upper bounds. In this section we present two upper bounds on the size of pairwise distance preservers. We start with few definitions, an algorithm, and a few lemmas that are used in both upper bounds.

Both our upper bounds are achieved by the same construction. The construction is very simple. Assuming that the shortest paths in the graph G are unique, for every pair $\{u, w\} \in \mathcal{P}$ we insert the shortest path between u and w. The uniqueness of shortest paths can be assumed without loss of generality since one can always slightly change the weights.

7.1. The construction. First, we need a few definitions.

Given an N-vertex weighted directed or undirected graph $G = ((V, E), wt)$ and a set $\mathcal{P}$ of ordered pairs of vertices, $\mathcal{P} \subseteq \binom{V}{2}$, form the subgraph $G' = ((V, H), wt)$, $H \subseteq E$ as follows. Order the edges and the pairs arbitrarily. (Generally, we will refer to the elements of E as "edges" if the graph G is undirected and as "arcs" if it is directed. Whenever the argument is applicable to both undirected and directed graphs, we will use the term "edge" meaning either "edge" or "arc" depending on the context.) Write

$\mathcal{P} = (p^{(1)}, p^{(2)}, \ldots, p^{(P)})$, where $P = |\mathcal{P}|$, and $E = (e^{(1)}, e^{(2)}, \ldots, e^{(m)})$, $m = |E|$. Assign auxiliary weights $\overline{wt}(e^{(i)}) = 1 + \epsilon_i$ for the edges $e^{(i)} \in E$. Choose the numbers $0 < \epsilon_1, \ldots, \epsilon_m < 1/N$ in a way that guarantees that the set $\{\epsilon_1, \ldots, \epsilon_m\}$ is linearly independent over the set $\mathbb{Q}$ of rational numbers [11]. (These numbers will be used to perturb the weights in order to ensure a consistent ordering of shortest paths.) Denote $p^{(i)} = (u^{(i)}, w^{(i)})$. Initialize the edgeset H of the subgraph G' as an empty set. Consider the following algorithm.

ALGORITHM (*Paths*).

For $i = 1, 2, \ldots, P$ **do**

Compute the path Π_i that satisfies the following conditions, and add this path to H:

1. Π_i is one of the shortest paths from $u^{(i)}$ to $w^{(i)}$ with respect to the weight function wt.
2. Among all such shortest paths Π_i is the one with the smallest auxiliary weight $\sum_{e \in \Pi_i} \overline{wt}(e_i)$. (Observe that a rule for breaking the ties is unnecessary, as no ties can possibly happen.)

Let E_i, $i \in [P]$, be the set of *new edges* that were added to the edgeset H on the ith iteration of Algorithm *Paths* (that is, they were not present in H before the ith iteration). Let e_i denote the cardinality of the set E_i. Obviously, $|H| = \sum_{i=1}^{P} e_i$. Let H_i denote the set H at the beginning of iteration i. Let $\hat{\Pi}$ denote the sequence of paths $\{\Pi_1, \ldots, \Pi_P\}$ that were computed by Algorithm *Paths*.

DEFINITION 7.1. *For a path $\Pi \in \hat{\Pi}$ from u to w, $\Pi = (u = x_0, x_1, \ldots, x_\ell = w)$, and a pair of vertices $x, y \in V(\Pi)$, we say that $x <_\Pi y$ if x is closer to u than y. Also, for an index $j \in \{0, \ldots, \ell - 1\}$ (respectively, $j \in \{1, \ldots, \ell\}$), the vertex x_{j+1} (resp., x_{j-1}) is called the* successor (*resp.,* predecessor) *of the vertex x_j in the path Π and is denoted $x_{j+1} = succ_\Pi(x_j)$ (resp., $x_{j-1} = pred_\Pi(x_j)$). The successor of x_ℓ and the predecessor of x_0 are formally defined as $\perp$; for the sake of our arguments, $\perp$ is not a vertex and does not belong to $V(\Pi)$. The vertex $u = x_0$ (resp., $w = x_\ell$) is called the* starting (*resp.,* ending) vertex of the path Π, *and both are called the* endpoints of the path Π.

LEMMA 7.2. *Every two distinct paths Π, Π' in the graph G have different auxiliary weights.*

Proof. If $|\Pi| = |\Pi'|$, then $\sum\{\epsilon_i : e^{(i)} \in \Pi\} \neq \sum\{\epsilon_i : e^{(i)} \in \Pi'\}$, since the set $\{\epsilon_1, \ldots, \epsilon_m\}$ is linearly independent over $\mathbb{Q}$, and the paths Π and Π' are distinct. ∎

We remark that one possible way to construct the linearly independent over $\mathbb{Q}$ set $\{\epsilon_1, \ldots, \epsilon_m\}$, $0 < \epsilon_i < 1/N$, is by picking m distinct primes $q_1, \ldots, q_m$, and for every $i \in [m]$, setting ϵ_i to be a rational multiple of $\sqrt{q_i}$ that satisfies $0 < \epsilon_i < 1/N$.

LEMMA 7.3. *For a pair of paths $\Pi, \Pi' \in \hat{\Pi}$, if $x, y \in V(\Pi) \cap V(\Pi')$, and $x <_\Pi y$, $x <_{\Pi'} y$, then the paths Π and Π' share the subpath that connects the vertex x to y.*

Proof. For a path Π, and a pair of vertices $z_1, z_2 \in V(\Pi)$ such that $z_1 <_\Pi z_2$, let $\Pi(z_1, z_2)$ denote the subpath of Π that starts in z_1 and ends in z_2.

Let (u, w) (resp., (u', w')) be the pair of vertices for which the path Π (resp., Π') was computed. Note that $\Pi = \Pi(u, x) \cdot \Pi(x, y) \cdot \Pi(y, w)$, and $\Pi' = \Pi'(u, x) \cdot \Pi'(x, y) \cdot \Pi'(y, w)$, where $\cdot$ stands for concatenation. If $\Pi(x, y) \neq \Pi'(x, y)$, then by definition of the auxiliary weight function $\overline{wt}$, $\overline{wt}(\Pi(x, y)) \neq \overline{wt}(\Pi'(x, y))$. Note that since both Π and Π' are the shortest paths between their endpoints (with respect to the weight function wt), it follows that both $\Pi(x, y)$ and $\Pi'(x, y)$ are the shortest paths from x to y, and, in particular, $wt(\Pi(x, y)) = wt(\Pi'(x, y))$.

Suppose without loss of generality that $\overline{wt}(\Pi(x,y)) < \overline{wt}(\Pi'(x,y))$. Consider the path $\Pi'' = \Pi'(u',x) \cdot \Pi(x,y) \cdot \Pi'(y,w')$. Note that $wt(\Pi'') = wt(\Pi')$, and $\overline{wt}(\Pi'') < \overline{wt}(\Pi')$. This is a contradiction to the assumption that Π' is the shortest path from u' to w' with the smallest (among the shortest paths) auxiliary weight. ∎

7.2. The first upper bound. In this section we prove an upper bound of $H = O(N + \sqrt{N} \cdot P)$.

DEFINITION 7.4. *A pair of distinct paths $\Pi, \Pi' \in \hat{\Pi}$ is said to* branch *in a vertex $v \in V$ if $v \in V(\Pi) \cap V(\Pi')$, and either $succ_{\Pi}(v) \notin V(\Pi')$ or $pred_{\Pi}(v) \notin V(\Pi')$. The vertex v as above is called a* branching vertex *of the paths Π and Π'.*

LEMMA 7.5. *For an* undirected *(possibly weighted) graph $G = ((V,E), wt)$, a collection $\hat{\Pi}$ of paths constructed by Algorithm Paths, and a pair $\{\Pi, \Pi'\} \in \binom{\hat{\Pi}}{2}$ of paths, there are at most two branching vertices of the paths Π and Π'.*

Proof. Consider the set Q of branching vertices of the paths Π and Π', and suppose that it contains two or more vertices (otherwise we are done). Order them with respect to the order $<_{\Pi}$ induced by the path Π. Let v_1 (resp., v_2) be the smallest (resp., the largest) vertex in Q with respect to $<_{\Pi}$. If $v_1 <_{\Pi'} v_2$, then by Lemma 7.3, the paths Π and Π' share the entire subpath between v_1 and v_2.

Consider the case that $v_1 >_{\Pi'} v_2$. Let $p = (u,w)$ (resp., $p' = (u',w')$) be the pair corresponding to the path Π (resp., Π'). Consider the collection of pairs $\mathcal{P}'$ formed by $\mathcal{P}' = \mathcal{P} \cup \{(w',u')\} \setminus \{p'\}$. It is easy to see that on an undirected graph, Algorithm *Paths* behaves identically when it accepts as input the collection $\mathcal{P}$ and when it accepts as input the collection $\mathcal{P}'$. Hence, by Lemma 7.3 the paths Π and Π' share the entire subpath between the vertices v_1 and v_2.

Consider a vertex $v \in V(\Pi)$ such that $v_1 <_{\Pi} v <_{\Pi} v_2$. It follows that $v \in V(\Pi')$, and furthermore, $succ_{\Pi}(v), pred_{\Pi}(v), succ_{\Pi'}(v), pred_{\Pi'}(v) \in V(\Pi) \cap V(\Pi')$. Hence, v is not a branching vertex of the paths Π and Π'. Hence, the paths Π and Π' branch in at most two vertices, proving the lemma. ∎

DEFINITION 7.6. *For a pair of paths $\{\Pi, \Pi'\} \in \binom{\hat{\Pi}}{2}$ and a branching vertex $v \in V$ of this pair, the pair $(v, \{\Pi, \Pi'\})$ is called a* branching event.

Let B denote the subset of vertices v with $\deg_H(v) \geq 3$.

LEMMA 7.7. *The number β of different branching events satisfies*

$$\sum_{v \in B} \left(\frac{1}{2}\right) \left\lceil \frac{\deg_H(v)}{2} \right\rceil \left(\left\lceil \frac{\deg_H(v)}{2} \right\rceil - 1 \right) \;\leq\; \beta \;\leq\; P(P-1).$$

Proof. To prove the upper bound note that by Lemma 7.5 for every pair of paths $\{\Pi, \Pi'\} \in \binom{\hat{\Pi}}{2}$, these two paths may participate together in at most two branching events.

To prove the lower bound consider a vertex $v \in V$. Suppose first $\deg_H(v) \geq 3$. Observe that each path Π that passes through v contains at most two edges that are adjacent to v. Since for every edge e of the subgraph H, there exists a path $\Pi \in \hat{\Pi}$ that contains e, it follows that at least $\lceil \frac{\deg_H(v)}{2} \rceil$ different paths $\Pi \in \hat{\Pi}$ pass through the vertex v, and, furthermore, for every pair of such different paths Π and Π', there is a branching event $(v, \{\Pi, \Pi'\})$. Hence, for every vertex $v \in V$ there are at least $(\frac{1}{2}) \lceil \frac{\deg_H(v)}{2} \rceil (\lceil \frac{\deg_H(v)}{2} \rceil - 1)$ different branching events of the form $(v, \{\Pi, \Pi'\})$ for different pairs of paths $\{\Pi, \Pi'\} \in \binom{\hat{\Pi}}{2}$. Hence, the overall number of branching

events β is at least

$$\sum_{v\in B}\left(\frac{1}{2}\right)\left\lceil\frac{\deg_H(v)}{2}\right\rceil\left(\left\lceil\frac{\deg_H(v)}{2}\right\rceil-1\right),$$

proving the lemma. ∎

For vertices $v\in V$ of degree $\deg_H(v)\le 2$, $\lceil\frac{\deg_H(v)}{2}\rceil-1=0$. Note that those vertices contribute at most $O(N)$ edges that are adjacent to them in the subgraph H. For vertices v of degree at least 3, $\lceil\frac{\deg_H(v)}{2}\rceil-1\ge\frac{\deg_H(v)}{4}$, and hence it follows that $(1/16)\sum_{v\in V}(\deg_H(v))^2\le P^2-P$. By the Cauchy–Schwarz inequality, it follows that $\sum_{v\in B}\deg_H(v)=O(\sqrt{N}\cdot P)$.

COROLLARY 7.8. *For a weighted undirected N-vertex graph $G=((V,E),wt)$, and a set $\mathcal{P}\subseteq\binom{V}{2}$ of P pairs of vertices $\mathcal{P}=\{p^{(i)}=(u^{(i)},w^{(i)}):i\in[P]\}$, there exists a subgraph $G'=((V,H),wt)$, $H\subseteq E$, that preserves the distances from u_i to w_i for every index $i\in[P]$, and such that $|H|=O(\sqrt{N}\cdot P+N)$.*

7.3. The second upper bound. The upper bound that we show in this section applies to *weighted directed* graphs.

For each index $i\in[P]$, we form the set $\mathcal{P}_i$ of ordered pairs of vertices in the following way:

$$\mathcal{P}_i=\{(x,y):(x,z),(v,y)\in E_i,x<_{\Pi_i}z<_{\Pi_i}v<_{\Pi_i}y\},$$

where the sets E_i are as defined in section 7.1.

LEMMA 7.9. *For a pair of indices $i,j\in[P]$, $\mathcal{P}_i\cap\mathcal{P}_j=\emptyset$.*

Proof. Suppose for contradiction that there exists a pair of indices $i<j$, $i,j\in[P]$, and a pair of vertices $x,y\in V$ such that $(x,y)\in\mathcal{P}_i\cap\mathcal{P}_j$. It follows that $x,y\in V(\Pi_i)\cap V(\Pi_j)$ and $x<_{\Pi_i}y$ and $x<_{\Pi_j}y$. Hence, by Lemma 7.3,

$$\Pi_i(x,y)\;=\;\Pi_j(x,y). \tag{12}$$

Let $x,v\in V$ be the vertices such that $(x,z),(v,y)\in\Pi_i(x,y)$. Since $j>i$, $E(\Pi_i(x,y))\subseteq H_j$. Hence, $(x,z),(v,y)\in H_j$. Note that $(x,y)\in\mathcal{P}_j$ implies that there exist vertices $z',v'\in V$ such that $(x,z'),(v',y)\in E_j$ that satisfy $x<_{\Pi_j}z'<_{\Pi_j}v'<_{\Pi_j}y$. Hence, by (12) $z',v'\in V(\Pi_j(x,y))=V(\Pi_i(x,y))$. Since $E_j\subseteq E(\Pi_j(u_j,w_j))$, it follows that $(x,z'),(v',y)\in E(\Pi_j(u_j,w_j))$. Since $x,y,z',v'\in V(\Pi_j(x,y))$, it follows that $(x,z'),(v',y)\in E(\Pi_j(x,y))=E(\Pi_i(x,y))$. It follows that $z=z'$, $v=v'$. Hence, $(x,z')=(x,z)$, $(v'y)=(v,y)\in E_j$. Recall that $(x,z),(v,y)\in H_j$. This is a contradiction because edges of H_j are never added to E_j. ∎

Note that by the definition of the set $\mathcal{P}_i$, $|\mathcal{P}_i|=e_i(e_i-1)/2$. By Lemma 7.9, $\sum_{i=1}^{P}|\mathcal{P}_i|\le N(N-1)$, where $N=|V|$ (because each ordered pair of vertices may appear in at most one set $\mathcal{P}_i$). Hence, $\frac{1}{2}\sum_{i=1}^{P}(e_i^2-e_i)\le N(N-1)$. By the Cauchy–Schwarz inequality, we conclude $|H|=\sum_{i=1}^{P}e_i=O(N\cdot\sqrt{P})$, as required. (Those indices for which $e_i\le 1$ contribute together at most P, and for other indices i, $e_i^2-e_i\ge\frac{3}{4}e_i^2$, and so the inequality is applicable. Since $N\ge\sqrt{P}$, $P+O(N\cdot\sqrt{P})=O(N\cdot\sqrt{P})$).

COROLLARY 7.10. *For a weighted directed N-vertex graph $G=((V,E),wt)$, and a set $\mathcal{P}\subseteq\binom{V}{2}$ of P pairs of vertices $\mathcal{P}=\{p^{(i)}=(u^{(i)},w^{(i)}):i\in[P]\}$, there exists a subgraph $G'=((V,H),wt)$, $H\subseteq E$, that preserves the distances from $u^{(i)}$ to $w^{(i)}$ for every index $i\in[P]$, and such that $|H|=O(N\cdot\sqrt{P})$.*

We remark that by a simple probabilistic argument it is possible to show an upper bound which is only by a factor of $O(\sqrt{\log n})$ weaker than the bound $|H| = O(\sqrt{P} \cdot N)$ of Corollary 7.10. (However, we are not aware of a simpler argument than the one described in section 7.2 for proving the upper bound of $|H| = O(\sqrt{N} \cdot P + N)$, or anything close to it.) Specifically, set $T = c \cdot \frac{N}{\sqrt{P}} \sqrt{\log N}$ for some constant $c > 2$, and initialize $H = \emptyset$. For every pair $(u, w) \in \mathcal{P}$, compute one of the shortest paths from u to w. Let $\mathcal{P}_1$ denote the subset of pairs $p = (u, w)$ of $\mathcal{P}$ for which the shortest path from u to w contains $T - 1$ or less internal (different from u and w) vertices. For each pair $p = (u, w) \in \mathcal{P}_1$, insert the computed shortest path from u to w into the edgeset H. Form a subset $Q \subseteq V$ of vertices by inserting into it each vertex with probability $\Theta(\frac{\log N}{T})$ independently at random. For each vertex $v \in Q$, insert the shortest-path in- and out-arborescences rooted in v into the subgraph H. Observe that altogether the expected number of inserted edges is $O(N \cdot \sqrt{P} \cdot \sqrt{\log N})$. It is not hard to see that with high probability (at least $1 - \frac{1}{n^{c-2}}$) the subgraph $G' = ((V, H), wt)$ preserves all the distances from u to w for every pair $(u, w) \in \mathcal{P}$. Hence, there exists a choice of subset Q of size $O(\frac{N}{T} \sqrt{\log N})$ for which the subgraph G' will satisfy these conditions.

Acknowledgments. The authors are grateful to the anonymous referees for multiple remarks that helped to improve the presentation of this paper.

REFERENCES

[1] N. Alon, R. M. Karp, D. Peleg, and D. West, *A graph-theoretic game and its application to the k-server problem*, SIAM J. Comput., 24 (1995), pp. 78–100.

[2] T. M. Apostol, *Introduction to Analytic Number Theory*, Springer-Verlag, Berlin, 1998.

[3] B. Awerbuch, B. Berger, L. Cowen, and D. Peleg, *Near-linear time construction of sparse neighborhood covers*, SIAM J. Comput., 28 (1998), pp. 263–277.

[4] B. Awerbuch and D. Peleg, *Network synchronization with polylogarithmic overhead*, in Proceedings of the 31st Annual IEEE Symposium on Foundations of Computer Science, IEEE, Piscataway, NJ, 1990, pp. 514–522.

[5] B. Awerbuch and D. Peleg, *Routing with polynomial communication-space trade-off*, SIAM J. Discrete Math., 5 (1992), pp. 151–162.

[6] B. Awerbuch, D. Peleg, B. Patt-Shamir, and M. Saks, *Adapting to asynchronous dynamic networks with polylogarithmic overhead*, in Proceedings of the 24th Annual ACM Symposium on Theory of Computing, ACM, New York, 1992, pp. 557–570.

[7] A. Balog and I. Barany, *On the convex hull of the integer points in a disc*, Discrete Comput. Geom., 6 (1991), pp. 39–44.

[8] I. Barany and D. G. Larman, *The convex hull of the integer points in a large ball*, Math. Ann., 312 (1998), pp. 167–181.

[9] S. Baswana, T. Kavitha, K. Mehlhorn, and S. Pettie, *New constructions of (a, β)-spanners and purely additive spanners*, in Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, ACM, New York, 2005, pp. 672–681.

[10] S. Baswana and S. Sen, *A simple linear time algorithm for computing a $(2k-1)$-spanner of $O(n^{1+1/k})$ size in weighted graphs*, in Proceedings of the 30th International Colloquium on Automata, Languages and Computing (ICALP), Springer, Berlin, 2003, pp. 384–396.

[11] A. S. Besicovitch, *On the linear independence of fractional powers of integers*, J. London Math. Soc., 15 (1940), pp. 3–6.

[12] B. Bollobás, D. Coppersmith, and M. Elkin, *Sparse distance preservers and additive spanners*, in Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, ACM, New York, 2003, pp. 414–423.

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

[14] E. Cohen, *Fast algorithms for constructing t-spanners and paths of stretch t*, in Proceedings of the 34th Annual IEEE Symposium on Foundations of Computer Science, IEEE, Piscataway, NJ, 1993, pp. 648–658.

[15] D. Coppersmith and M. Elkin, *Sparse source-wise and pair-wise distance preservers*, in Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, ACM, New York, 2005, pp. 660–669.

[16] D. Dor, D. Halperin, and U. Zwick, *All pairs almost shortest paths*, in Proceedings of the 37th Annual IEEE Symposium on Foundations of Computer Science, IEEE, Piscataway, NJ, 1997, pp. 452–461.

[17] M. Elkin, *Computing almost shortest paths*, in Proceedings of the 20th Annual ACM Symposium on Principles of Distributed Computing, ACM, New York, 2001, pp. 53–63.

[18] M. Elkin and D. Peleg, $(1+\epsilon,\beta)$*-spanner constructions for general graphs*, in Proceedings of the 33rd Annual ACM Symposium on Theory of Computing, ACM, New York, 2001, pp. 173–182.

[19] C. Gavoille, D. Peleg, S. Perennes, and R. Raz, *Distance labeling in graphs*, in Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, ACM, New York, 2001, pp. 210–219.

[20] P. Indyk and J. Matousek, *Low distortion embeddings of finite metric spaces*, in Handbook of Discrete and Computational Geometry, 2nd ed., J. E. Goodman and J. O'Rourke, eds., Chapman & Hall/CRC, Boca Raton, FL, 2004, pp. 177–196.

[21] V. Jarnik, *Uber Gitterpunkte und konvex Kurven*, Math. Z., 24 (1925), pp. 500–518.

[22] W. B. Johnson and J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, Contemp. Math., 26 (1984), pp. 189–206.

[23] Y. Mansour and D. Peleg, *An Approximation Algorithm for Minimum-Cost Network Design*, Technical Report CS94-22, The Weizmann Institute of Science, Rehovot, Israel, 1994. Available online at http://wisdomarchive.wisdom.weizmann.ac.il:81/archive/00000021.

[24] J. Matousek, *Lectures in Discrete Geometry*, Graduate Texts in Math. 212, Springer, New York, 2002.

[25] D. Peleg, *Proximity-preserving labeling schemes*, J. Graph Theory, 33 (2000), pp. 167–176.

[26] D. Peleg and A. Schäffer, *Graph spanners*, J. Graph Theory, 13 (1989), pp. 99–116.

[27] D. Peleg and E. Upfal, *A trade-off between space and efficiency for routing tables*, J. ACM, 36 (1989), pp. 510–530.

[28] L. Roditty, M. Thorup, and U. Zwick, *Roundtrip spanners and roundtrip routing in directed graphs*, in Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, ACM, New York, 2002, pp. 844–851.

[29] M. Thorup and U. Zwick, *Approximate distance oracles*, in Proceedings of the 33rd Annual ACM Symposium on Theory of Computing, ACM, New York, 2001, pp. 183–192.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 502–522

ON THE GREEDY SUPERSTRING CONJECTURE*

MAIK WEINARD† AND GEORG SCHNITGER†

Abstract. We investigate the greedy algorithm for the shortest common superstring problem. We show that the length of the greedy superstring is upper-bounded by the sum of the lengths of an optimal superstring and an optimal cycle cover, *provided* the greedy algorithm happens to merge the strings in a particular way. Thus, when restricting inputs correspondingly, we verify the well-known greedy conjecture, namely, that the approximation ratio of the greedy algorithm is within a factor of two of the optimum, and actually extend the conjecture considerably. We achieve this bound by systematically combining known conditional inequalities about overlaps and period- and string-lengths with a new familiy of string inequalities. We show that conventional systems of conditional inequalities, including the Monge inequalities, are insufficient to obtain our result.

Key words. superstring, DNA fragment assembly, approximation, performance analysis

AMS subject classifications. 68W05, 68W25, 68W40

DOI. 10.1137/040619144

1. Introduction.

We investigate the problem of finding a shortest common superstring:

> Given a set $S = \{s_1, \ldots, s_n\}$ of strings, determine a string of minimum length which contains each s_i as a substring.

(Obviously, we may assume that no string in S contains another string in S.) Apart from being an interesting problem in itself the shortest common superstring problem models the sequence assembly problem in shotgun sequencing, a fundamental problem in bioinformatics. Each string in S represents a sequenced DNA fragment created by shotgun sequencing, and the assembly problem is to deduce the original DNA string from its set S of sequenced fragments.

Blum et al. [3] show that the shortest common superstring problem is APX-complete; hence polynomial time approximation schemes are not to be expected.

A simple greedy algorithm is the basis of the best known approximation algorithms. The greedy algorithm repeatedly merges two strings with maximum *overlap* until only one string remains. The overlap of two strings a and b is the length of the largest suffix of a that is also a prefix of b.

The length of the greedy superstring in relation to the length of the optimal superstring—the approximation ratio—has been the subject of a large body of research. The following example [12] shows that the ratio is at least 2. In particular we consider the three strings $x = c(ab)^n, y = (ab)^n d$, and $z = (ba)^n$.

Overlap	$x = c(ab)^n$	$y = (ab)^n d$	$z = (ba)^n$
$c(ab)^n$	–	$2n$	$2n-1$
$(ab)^n d$	0	–	0
$(ba)^n$	0	$2n-1$	–

*Received by the editors November 18, 2004; accepted for publication (in revised form) December 15, 2005; published electronically June 9, 2006. This research was partially supported by DFG grant SCHN 503/2-1. A preliminary version of this paper appeared as [14].
http://www.siam.org/journals/sidma/20-2/61914.html
†Institut für Informatik, Johann Wolfgang Goethe–Universität Frankfurt am Main, Robert-Mayer-Straße 11-15, 60054 Frankfurt am Main, Germany (weinard@thi.informatik.uni-frankfurt.de).

The greedy algorithm first joins x and y and obtains $c(ab)^n d$ as a new string that has zero overlap with z in both directions. Hence its second iteration delivers the superstring $c(ab)^n d(ba)^n$ or $(ba)^n c(ab)^n d$ of length $4n+2$. Obviously the solution $c(ab)^{n+1}d$ of length $2n+4$ is shorter, and the length ratio approaches 2.

Blum et al. [3] have shown that the greedy algorithm provides a 4-approximation which was the first constant factor approximation achieved for the problem. In 2005 Kaplan and Shafrir [7] improved the upper bound to 3.5. Better upper bounds are not known. However, there are no lower bounds larger than 2 for the greedy heuristic. Therefore it is widely conjectured that the approximation ratio of the greedy algorithm is in fact 2 [6, 3, 11]. This *greedy conjecture* is the main subject of our paper.

Also in [3] a modified version of the greedy algorithm achieving a 3-approximation is introduced. Further improvements [13, 5, 8, 1, 2, 4] have led to the best known result of a 2.5 approximation [11].

The greedy algorithm when allowed to close cycles (i.e., merge a string with itself) determines a minimum length cycle cover [3]. We verify a stronger version of the greedy conjecture, provided the greedy algorithm processes strings according to a restricted class of orders. In particular we show that the length of the superstring obtained by the greedy heuristic is upper-bounded by the length of an optimal superstring plus the length of an optimal cycle cover whenever the input strings cause the greedy heuristic to merge the strings in a manner that we will call a *linear order*.

The rest of the paper is organized as follows. The greedy heuristic and our stronger version of the greedy conjecture are described in section 2. In section 3 we introduce conditional inequalities on string-lengths and overlaps and define the *mice game*, which allows us to analyze inequalities in a combinatorial fashion. In particular we introduce the new *Triple inequality*, which is crucial for our result. Section 4 completes the analysis. We show in section 5 that the established set of string properties is insufficient to prove even the weaker classical greedy conjecture. Our conclusion and open problems are given in section 6.

2. Superstrings and cycle covers. Crucial for the greedy heuristic is the concept of overlaps defined as follows.

DEFINITION 1. 1. *Let a and b be strings. $\langle a,b\rangle$ denotes the longest proper suffix of a which is also a proper prefix of b. The* overlap *of a and b, denoted by (a,b), is the length of $\langle a,b\rangle$.*

2. *Let a and p be strings. We say that a is p-*periodic *or* has period p *iff a is a substring of p^k for some $k \in \mathbb{N}$. If p_a is a shortest string such that a is p_a-periodic, then p_a is a* minimum period *of a.*

Note that the *self-overlap* (a,a) is different from the length of the string since we require proper prefixes and suffixes: the length of a is the sum of its self-overlap and its period length.

GREEDY ALGORITHM.

1. INPUT: A set S of n strings
2. while $|S| > 1$
   (a) choose $a, b \in S$ such that $a \neq b$ and (a,b) is maximal,
   (b) let c be the string of minimal length with a as prefix and b as suffix,
   (c) set $S = (S \cup \{c\}) \setminus \{a,b\}$.
3. OUTPUT: The greedy superstring, i.e., the one string left in S. /* since we obtain a superstring in every merge in step 2(b) we output a superstring of S. */

If we choose identical strings a and b in step 2(a) and insert the period of a into a cycle cover rather than inserting c into S, then we obtain the *cyclic greedy algorithm* which determines a cycle cover of minimum length [3].

Tarhio and Ukkonen [12] showed that the greedy heuristic achieves an approximation factor of 2 with respect to string compression: if the total length of the input strings is M and the optimum superstring has a length of $M - x$, then the greedy heuristic finds a superstring of length at most $M - \frac{x}{2}$. This, however, does not imply a bound on the ratio of the superstrings' lengths.

Let S be a superstring of $\{s_1, \ldots, s_n\}$, and let π be the permutation of $\{1, \ldots, n\}$ such that in S the string $s_{\pi(i)}$ starts before $s_{\pi(i+1)}$. As we assume that no string is a substring of another string, we know that $s_{\pi(i)}$ also ends before $s_{\pi(i+1)}$. Hence the length of S is at least

$$l = \sum_{i=1}^{n} |s_i| - \sum_{i=1}^{n-1} (s_{\pi(i)}, s_{\pi(i+1)}),$$

and, moreover, a superstring with order π and length l exists. Hence we may focus on the $n!$ superstrings that maximally exploit the overlaps of consecutive strings given by their superstring order π.

Throughout this paper we assume without loss of generality that the greedy superstring contains $s_1, \ldots, s_n$ in this order. Hence its length of the greedy superstring is

$$L(GREEDY) = \sum_{i=1}^{n} |s_i| - \sum_{i=1}^{n-1} (s_i, s_{i+1}).$$

Depending on the strings $s_1, \ldots, s_n$ there are $(n-1)!$ *greedy orders* in which GREEDY can merge the strings to produce $(s_1, \ldots, s_n)$, namely, all permutations of the pairs $\{(s_1, s_2), \ldots, (s_{n-1}, s_n)\}$. We call an order *linear* if GREEDY starts with an arbitrary connection (s_i, s_{i+1}) and later on, when $s_j, s_{j+1}, \ldots, s_{k-1}, s_k$ are already connected, picks either (s_{j-1}, s_j) or (s_k, s_{k+1}). Observe that GREEDY can be compared to Kruskals' algorithm for determining a minimum cost spanning tree: GREEDY starts from a collection of strings and merges adjacent "intervals of strings" according to a highest profit criterion until a superstring is obtained. If GREEDY merges strings in a linear order, then it always grows the same "component," and hence it behaves like Prims's algorithm.

DEFINITION 2. *Let $s_1, \ldots, s_n$ be strings. The length $L^*(c)$ of a given cycle $c = (s_{i_1}, \ldots, s_{i_k})$ is defined as $L^*(c) = \sum_{j=1}^{k} |s_{i_j}| - \sum_{j=1}^{k-1} (s_{i_j}, s_{i_{j+1}}) - (s_{i_k}, s_{i_1})$. A cycle cover $\mathcal{C}$ of $s_1, \ldots, s_n$ decomposes the set of strings into disjoint cycles $c_1, \ldots, c_r$. The length of $\mathcal{C}$ is defined as $L^*(\mathcal{C}) = \sum_{i=1}^{r} L^*(c_i)$. $L^*(GREEDY)$ is the length of the greedy superstring after merging the last string s_n with the first string s_1; i.e., $L^*(GREEDY) = L(GREEDY) - (s_n, s_1)$.*

We compare the length $L^*(GREEDY)$ of the cyclic superstring delivered by GREEDY with the lengths $L^*(\mathcal{C}_1), L^*(\mathcal{C}_2)$ of any two cycle covers. However, we require that $\mathcal{C}_1$ and $\mathcal{C}_2$ *expand* in the following sense: for any proper subset S' of S, $\mathcal{C}_1$ and $\mathcal{C}_2$ do not both contain a cycle cover of S'. This property is equivalent to the fact that the directed graph induced by the two cycle covers is strongly connected.

Our main result is the following theorem.

THEOREM 1. *If GREEDY determines a linear greedy order, then for any expanding cycle covers $\mathcal{C}_1$ and $\mathcal{C}_2$*

$$L^*(GREEDY) \leq L^*(\mathcal{C}_1) + L^*(\mathcal{C}_2)$$

holds.

The remainder of this section as well as sections 3 and 4.1 are devoted to the proof of Theorem 1. We conclude that for linear greedy orders

$$L^*(GREEDY) \leq opt^*_{\text{string}} + opt_{\text{cyclic}} \tag{1}$$

holds, where opt^*_{string} is the length of the cycle obtained by closing the shortest superstring and opt_{cyclic} is the length of the shortest cycle cover. This follows since the expansion property of Theorem 1 is fulfilled if $\mathcal{C}_1$ is a single cycle. In section 4.2 we show the following corollary.

COROLLARY 1. *If GREEDY determines a linear greedy order, then*

$$L(GREEDY) \leq opt_{\text{string}} + opt_{\text{cyclic}},$$

where opt_{string} is the length of the shortest superstring and opt_{cyclic} is the length of the shortest cycle cover.

Corollary 1 shows that the lengths of the greedy superstring and the optimal superstring differ by at most the length of an optimal cycle cover. We hence obtain a far-reaching extension of the original greedy conjecture, since a shortest cycle cover will in general be considerably smaller than a shortest superstring. Although we are not aware of any counterexample for the general case, our proof applies only to the restricted case of linear greedy orders.

We may represent a cycle cover $\mathcal{C}$ as a union of disjoint cycles on the vertices $s_1, \ldots, s_n$. If E denotes the corresponding set of edges, then $L^*(\mathcal{C}) = \sum_{i=1}^{n} |s_i| - \sum_{(s,t)\in E}(s,t)$, where we abuse notation as (s,t) denotes the overlap as well as the edge. To show Theorem 1 we have to verify $L^*(GREEDY) \leq L^*(\mathcal{C}_1) + L^*(\mathcal{C}_2)$, which is equivalent to

$$\sum_{i=1}^{n} |s_i| - \sum_{i=1}^{n-1}(s_i, s_{i+1}) - (s_n, s_1) \leq 2 \cdot \sum_{i=1}^{n} |s_i| - \sum_{(s,t)\in E_1} (s,t) - \sum_{(s,t)\in E_2} (s,t),$$

where E_i is the set of edges of $\mathcal{C}_i$. Hence it suffices to verify the inequality

$$\sum_{(s,t)\in E_1} (s,t) + \sum_{(s,t)\in E_2} (s,t) \leq \sum_{i=1}^{n} |s_i| + \sum_{i=1}^{n-1}(s_i, s_{i+1}) + (s_n, s_1). \tag{2}$$

The terms appearing in our equation are string-lengths and overlaps; hence they are not arbitrary numbers but are constrained by properties of strings. An obvious example is the fact that an overlap cannot be longer than any of its strings, i.e., $|a| \geq (a,b)$. Another example is the inequality $|a| + (b,c) \geq (a,c) + (b,a)$ (as we will see in Lemma 2). Of course we also need to exploit the fact that GREEDY proceeds greedily: GREEDY connects strings a and b at a time when it could also connect c and d, we conclude that $(a,b) \geq (c,d)$. Hence we obtain a partial order on the a priori unknown overlaps—the greedy order.

We obtain a large class of *conditional inequalities* if we take the greedy order into account. A first example is the famous class of Monge inequalities (cf. Lemma 3).

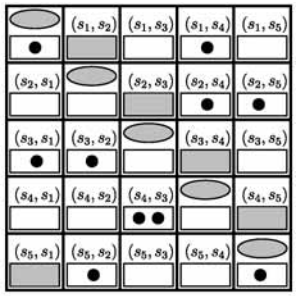


FIG. 1. *An example setup.*

We introduce a second important class of inequalities with the Triple inequality of Lemma 5. Thus we are led to a problem of linear programming conditioned on the given greedy order, namely, to show that inequality (2) is enforced by a set of linear inequalities relative to a given greedy order. A first principle analysis of the corresponding family of linear programs seems impossible, and hence we introduce a combinatorial perspective, namely, the mice game that we introduce in the next section. Our results therefore suggest the following program to analyze the greedy algorithm:

1. Formulate crucial string properties as (conditional) linear inequalities.
2. Analyze the resulting systems of inequalities via linear programming, respectively, via combinatorial methods derived from linear programming.

Corollary 1 shows that this program can be carried out in the nontrivial case of linear greedy orders. Subsequently, using the above program, our analysis has been refined in [9] to cover the larger class of bilinear greedy orders.

3. The mice game. The mice game for n strings is played on a board of $n \times n$ cells, as illustrated in Figure 1. The rectangle within the cell in row i and column j represents the overlap (s_i, s_j). In the case $i = j$ the corresponding diagonal cell has two components, namely, an ellipse representing the length of string s_i and the rectangle representing the self-overlap (s_i, s_i). The overlaps (s_i, s_{i+1}) as well as the string-lengths are shown shaded, and their locations are called holes. Observe that the corresponding terms appear on the right-hand side of inequality (2).

The terms that appear on the left-hand side of inequality (2), namely, overlaps between adjacent strings of the cycle covers $\mathcal{C}_1$ and $\mathcal{C}_2$, are the starting positions of the *mice*. We therefore have $2n$ mice on the board, with exactly two mice in each row and column. They are shown as black bullets. In Figure 1 we show the mice placement for the two expanding cycle covers $\mathcal{C}_1 = \{(s_1), (s_2, s_4, s_3), (s_5)\}$ and $\mathcal{C}_2 = \{(s_1, s_4, s_3), (s_2, s_5)\}$ that serve as a running example throughout this section.

The objective of the game is to move the mice from their starting positions into their holes by a sequence of *legal moves*. We require that a hole can accomodate only one mouse. A move is legal if the total value of terms represented by the mice does not decrease. Hence a sequence of legal moves corresponds to a proof of inequality (2) for the given instance.

3.1. Unconditional inequalities. We introduce a class of obvious inequalities and see how they translate into legal moves in our game.

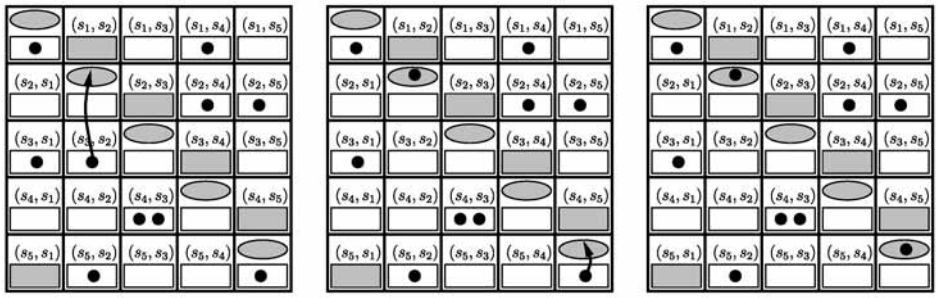


FIG. 2. *The application of one Diagonal Insertion and one Discarding of a Period. The moves are justified by the inequalities* $(s_3, s_2) \leq |s_2|$ *(resp.,* $(s_5, s_5) \leq |s_5|$*).*

LEMMA 1. *Let* a, b *be strings and let* p_a *be a minimum period of* a*. Then*

$$(a,b) < |a|, \quad (a,b) < |b|, \quad \text{and} \quad (a,a) = |a| - |p_a| < |a|.$$

As an immediate consequence we may move a mouse from cell (i, j) to the elliptic diagonal hole of row i or column j. We call this move a *Diagonal Insertion*. In the case of $i = j$ we may move a mouse within a diagonal cell from the rectangle representing the self-overlap to the elliptic hole. We call this special case *Discarding a Period*. Figure 2 shows one application of each move.

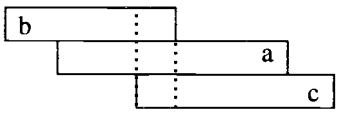
Next we describe an easy version of the Monge inequality [10].

LEMMA 2. *Let* a*,* b*, and* c *be strings. Then the inequality*

$$(b,a) + (a,c) \leq |a| + (b,c) \tag{3}$$

holds.

Proof. If $|a| \geq (b, a) + (a, c)$ holds, the claim is obvious. Otherwise we have the following diagram:



The sum $(b, a) + (a, c)$ exceeds $|a|$ by the space between the dashed vertical lines. This segment, however, is a suffix of b, which is also a prefix of c. Hence (b, c) is at least as large. □

The inequality of Lemma 2 justifies a move that involves two mice: of two mice, located in cells (i, j) and (k, i) with $k \neq i$ and $j \neq i$, one moves to the elliptic diagonal hole in (i, i) and the other at the same time to the rectangle in (k, j). Figure 3 shows two consecutive applications of this *Diagonal Monge* in our game.

Note that the moves introduced up to this point are independent of the greedy order. Since we cannot analyze GREEDY without utilizing its properties, we need to incorporate the greedy order into our argument.

3.2. Exploiting the greedy order. We fix an arbitrary greedy order. When GREEDY chooses to merge two strings a and b by picking the pair (a, b), then the options to pick (a, x), (x, b), or (b, a) are thereby eliminated for any third string x. As GREEDY picks the maximum overlap over all possible choices, we know that the off-diagonal cell (a, b) represents a value at least as large as the value of every cell whose pair was still available at the time. We assign a rank of $n - i$ to the cell that

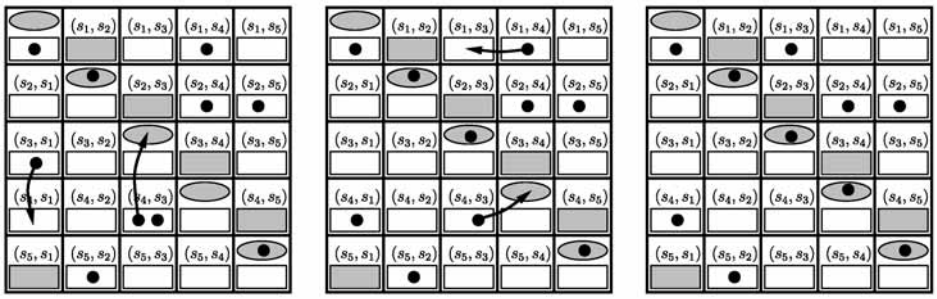


FIG. 3. *The application of two Diagonal Monges. The moves are justified by the inequalities* $(s_4, s_3) + (s_3, s_1) \leq |s_3| + (s_4, s_1)$ *(resp.,* $(s_4, s_3) + (s_1, s_4) \leq |s_4| + (s_1, s_3)$*).*

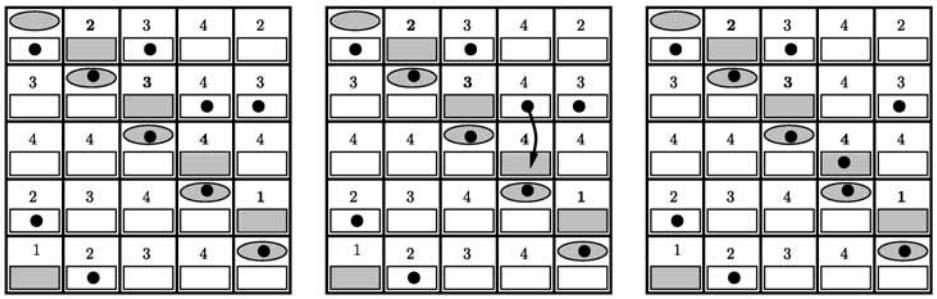


FIG. 4. *The ranks corresponding to the assumed greedy order and a Greedy Insertion. As the greedy cell* (s_3, s_4) *has rank* 4*, we have* $(s_3, s_4) \geq (s_2, s_4)$*.*

was picked in the ith iteration of GREEDY and to every cell thereby eliminated. We refer to the cells on the off-diagonal as greedy cells, as they represent overlaps that GREEDY picks.

Let us assume for our example that GREEDY chooses (s_3, s_4), (s_2, s_3), (s_4, s_5), and (s_1, s_2) in that order. In Figure 4 the coordinates of the cells have been replaced by their rank. Note that diagonal cells do not have a rank, since GREEDY may not pick them at any time. We further show a *Greedy Insertion*, a simple move of just one mouse to a greedy cell of not smaller rank.

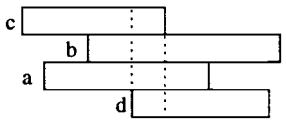
The next inequality—like Lemma 2—goes back to Gaspard Monge [10].

LEMMA 3. *Let* a, b, c, d *be strings with* $(a, b) \geq (a, d)$ *and* $(a, b) \geq (c, b)$*. Then*

$$(a, d) + (c, b) \leq (a, b) + (c, d) \tag{4}$$

holds.

Proof. The proof is rather similar to that of Lemma 2: if $(a, b) \geq (a, d) + (c, b)$ holds, then the claim is obvious. Otherwise we get the following diagram reflecting the overlaps $(c, b), (a, b), (a, d)$ from top to bottom:



The sum of (c, b) and (a, d) exceeds the length of (a, b) by the section between the dashed vertical lines. This segment, however, is a suffix of c which is also a prefix of d. Hence (c, d) is at least as large. □

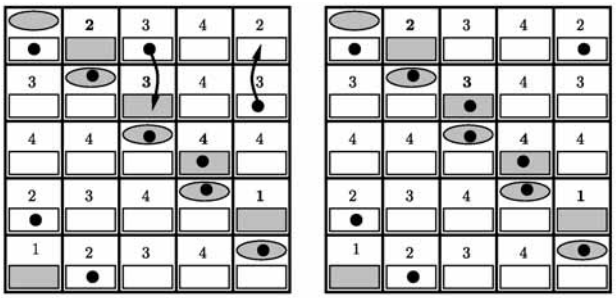


FIG. 5. *A Greedy Monge in* (s_2, s_3). *Ranks indicate that* (s_2, s_3) *is at least as large as* (s_1, s_3) *and* (s_2, s_5). *Thus the inequality* $(s_2, s_3) + (s_1, s_5) \geq (s_1, s_3) + (s_2, s_5)$ *holds.*

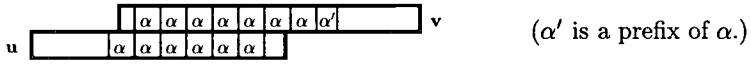
We call the corresponding move a *Greedy Monge* since we mainly apply this move if (a, b) corresponds to a greedy cell. Hence we can easily verify the legality of Greedy Monges by consulting ranks. Figure 5 shows a Greedy Monge.

We conclude this section with further inequalities that we use in section 4.2.

LEMMA 4. *Let* a, b, *and* c *be strings.*

1. *If* $(a, b) \geq (a, c)$, *then* $(a, c) \leq (b, c) + |p_b|$.
2. *If* $(a, b) \geq (b, c)$, *then* $(b, c) \leq (a, c) + |p_b|$.
3. *If* $(a, b) \geq (c, b)$, *then* $(c, b) \leq (c, a) + |p_a|$.
4. *If* $(a, b) \geq (c, a)$, *then* $(c, a) \leq (c, b) + |p_a|$.

Proof. We prove only the first two statements, since we obtain the other two by simply interchanging rows and columns. Observe that substrings inherit the periods of their superstrings: if α is a period of x, then α is also a period of every substring of x. Moreover, given a string u with an α-periodic suffix of length r and a string v with an α-periodic prefix of length s, we may conclude $\min\{r, s\} \leq (u, v) + |\alpha|$.



For the first claim of the lemma the assumption $(a, b) \geq (a, c)$ guarantees that c has a p_b-periodic prefix of length (a, c). String b, of course, is p_b-periodic. Hence $\min\{|b|, (a, c)\} \leq (b, c) + |p_b|$. The claim follows since $|b| \geq (a, b) \geq (a, c)$.

For the second statement the assumption $(a, b) \geq (b, c)$ guarantees that a has a p_b-periodic suffix of length (b, c). String c clearly has a p_b-periodic prefix of length (b, c), and we get $\min\{(b, c), (b, c)\} \leq (a, c) + |p_b|$. The claim follows. ∎

Figure 6 shows two possible applications of these inequalities to our game, but we do not execute any one of them since this would result in a dead end.

Using the moves we have established so far, we cannot proceed to win the game. In fact we cannot win this game at all from the initial configuration with our current set of moves. This can be checked by linear programming. The next section introduces the Triple inequality. Its move wins this game and is crucial for proving that every game can be won for linear greedy orders.

3.3. The Triple.

LEMMA 5. *Let* a, b, c, d, *and* x *be strings with*

$$\max\{(a, x), (x, b)\} \geq (a, b), (x, d), (c, x).$$

Then

$$(a, b) + (c, x) + (x, d) \leq (a, x) + (x, b) + (c, d) + |p_x|.$$

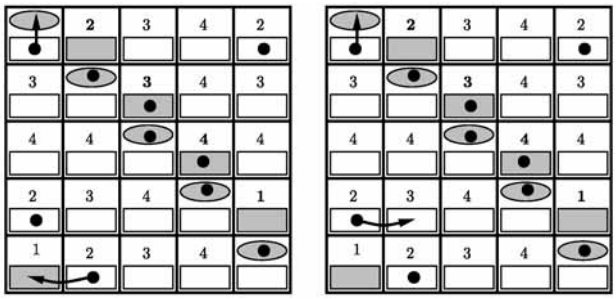


FIG. 6. *Two legal applications of Lemma* 4 *(3rd (resp., 4th) inequality). The ranks indicate that the conditions are met. By moving a mouse from* (s_1, s_1) *to* $|s_1|$*, the mouse gains the weight* $|p_{s_1}|$ *(compare with Lemma* 1*).*

Proof. We proceed by induction on (x, x). For $(x, x) = 0$ we have $p_x = x$. The Monge inequality (3) gives us

$$(c, x) + (x, d) \leq |x| + (c, d).$$

As we know that (a, b) is smaller than the maximum of (x, b) and (a, x), we conclude that

$$(a, b) \leq (a, x) + (x, b),$$

and we are done.

For the induction step assume that the claim is shown for all x with $(x, x) \leq k$. Let $(x, x) = k + 1$ and assume that the premise of the lemma is fulfilled. First we eliminate the case $\max\{(a, x), (x, b)\} \geq (x, x)$. We assume without loss of generality that $(a, x) \geq (x, b)$. As (a, x) dominates (a, b), we obtain

$$(x, x) + (a, b) \leq (a, x) + (x, b) \tag{5}$$

with the Monge inequality (4). The Monge inequality also gives us

$$(c, x) + (x, d) \leq |x| + (c, d). \tag{6}$$

Adding inequalities (5) and (6) and subtracting (x, x) on both sides yields our claim. Observe that we did not use the induction hypothesis.

Now we may assume $(x, x) > (a, x), (x, b)$ and, since every overlap $(a, b), (x, d)$, and (c, x) is dominated by (a, x) or (x, b), we know that (x, x) also dominates these overlaps. With this knowledge we may conclude $(a, x) = (a, \langle x, x\rangle)$, $(x, b) = (\langle x, x\rangle, b)$, $(c, x) = (c, \langle x, x\rangle)$, and $(x, d) = (\langle x, x\rangle, d)$.

Assuming the premise of the lemma, we can infer that the larger of $(a, \langle x, x\rangle)$ and $(\langle x, x\rangle, b)$ dominates $(a, b), (\langle x, x\rangle, d)$, and $(c, \langle x, x\rangle)$. Hence we can use the induction hypothesis since $(\langle x, x\rangle, \langle x, x\rangle) < (x, x)$ to obtain

$$(a, b) + (\langle x, x\rangle, d) + (c, \langle x, x\rangle) \leq (a, \langle x, x\rangle) + (\langle x, x\rangle, b) + (c, d) + |p_{\langle x,x\rangle}|,$$

which is equivalent to

$$(a, b) + (x, d) + (c, x) \leq (a, x) + (x, b) + (c, d) + |p_{\langle x,x\rangle}|.$$

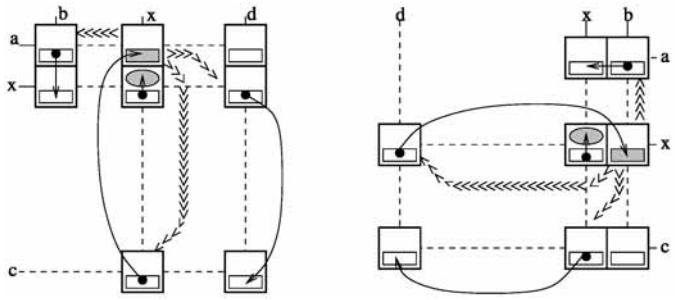


FIG. 7. *The Triple move. We distinguish the vertical and the horizontal Triple moves depending on which of the overlaps* (a, x) *or* (x, b) *is maximal. The required dominations are indicated by the sequence of arrow heads.*

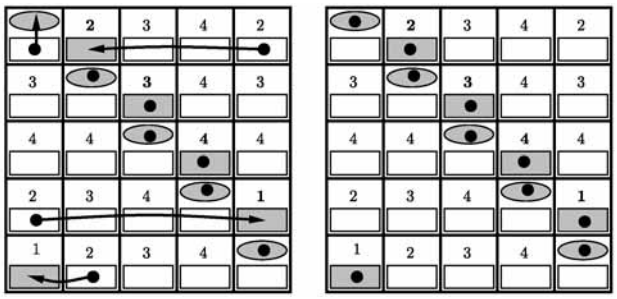


FIG. 8. *A horizontal Triple move. The correspondence to Figure* 7 *(right diagram) is given by* $x = s_1$, $a = d = s_5$, $b = s_2$, *and* $c = s_4$.

Now we need only to observe that $|p_{\langle x,x\rangle}| \leq |p_x|$, which clearly holds as $\langle x, x\rangle$ is a substring of x. □

This inequality is clearly not immediately appealing. Before using the Triple to win our example game, we visualize the corresponding moves in Figure 7.

Finally, Figure 8 shows how to win the game by applying the Triple. As a consequence we have shown inequality (2) for this particular greedy order and for these two particular expanding cycle covers.

The next section describes a winning strategy for linear greedy orders which is given in the form of a *how-to-move-the-mice algorithm.*

4. Winning strategies. Before defining our algorithm we introduce notation for the game, in particular, for the board of the game.

DEFINITION 3. 1. *For* $I \subseteq \{1, \ldots, n\}$ *we define* $Board(I)$ *as the set of cells with coordinates in* I*:* $Board(I) = \{(x, y)|x, y \in I\}$. *We say that* $Board(I)$ *is a subboard.*

2. *For* $B = Board(I)$ *we define the set of cells* $\{(x, y)|x \in I, y \notin I\}$ *as the* horizontal frame *of* B *and the set* $\{(x, y)|x \notin I, y \in I\}$ *as the* vertical frame *of* B. *The union* $Frame(B)$ *of the vertical and horizontal frames of* B *is called the* frame *of* B.

3. $Frame(B)$ *is called* deserted *iff no cell in* $Frame(B)$ *contains a mouse.*

DEFINITION 4. *We call a move of two mice rectangular if they are initially on two opposite corners* (a, b) *and* (c, d) *of a rectangle and move to the other two corners* (c, b) *and* (a, d).

A diagonal or Greedy Monge move is rectangular. A Triple move can be represented as two rectangular moves even though the moves are not individually valid.

LEMMA 6. *Assume that every row and column of the board contains two mice. Moreover, let B be a subboard. Then the following two statements hold:*

1. *The frame of B is nonempty iff the horizontal and the vertical frames are nonempty.*
2. *A rectangular move can empty only the frame of B; if one of the participating mice leaves the horizontal frame, the other leaves the vertical frame, and exactly one of them enters B.*

Proof. The proof is obvious. □

4.1. The rank descending algorithm. We describe the rank descending algorithm (RDA) which is defined on a board for a linear greedy order. RDA systematically grows a subboard B beginning with a single diagonal until the subboard coincides with the entire board. The name RDA is assigned, since we move mice of higher rank into the holes first. The algorithm is defined as follows.

THE RANK DESCENDING ALGORITHM.

(1) The **input** consists of two cycle covers $\mathcal{C}_1, \mathcal{C}_2$ and a linear greedy sequence.
(2) **Preprocessing**
  (2a) Place a mouse in the rectangle of position (i, j), if string s_j immediately follows s_i in $\mathcal{C}_1$ or in $\mathcal{C}_2$. If s_j is the immediate successor of s_i in both $\mathcal{C}_1$ and $\mathcal{C}_2$, then the rectangle of (i, j) receives two mice.
  (2b) Let i be the row of the highest ranked greedy cell. Set $I = \{i\}$.
  (2c) If (i, i) contains a mouse, then discard its period. Otherwise execute a *legal* Diagonal Monge in (i, i).
(3) **Main Loop.**

RDA will establish Theorem 1. Crucial for our proof of correctness are the following four invariants:

I1. Every row and every column contains 2 mice.
I2. Every hole in $B = Board(I)$ is filled.
I3. No elliptic diagonal hole outside of $B = Board(I)$ holds a mouse.
I4. For all subboards $B' = Board(I')$ with $\emptyset \subset I' \subset \{1, \ldots, n\}$ the frame of B' is not deserted.

Observe that if I1 holds, I4 is equivalent to stating that all $B' = Board(I')$ contain less than $2|I'|$ mice.

We will call a move legal if it does not violate any of these four invariants. Before going into detail we start with some motivating remarks about our invariants. I2 guarantees progress, since B will eventually be the entire board, and hence inequality (2) is verified. Observe that none of our moves is able to move a mouse out of an elliptic diagonal hole, and hence I3 is necessary for mice outside of B to still be *mobile*. Finally, if the frame of some Board $B' = Board(I')$ with $\emptyset \subset I' \subset \{1, \ldots, n\}$ is empty, it turns out that at least one mouse within B' is stranded.

It turns out that the preservation of invariants I1, I2, and I3 is rather straightforward. Invariant I4, however, needs special attention. We start with two observations about avoiding deserted frames.

LEMMA 7. *Assume* I4 *holds. If one of the starting locations for a rectangular move contains more than one mouse, then* I4 *holds afterward as well.*

Proof. The proof is obvious by Lemma 6. □

LEMMA 8. *Let invariants* I1 *and* I4 *hold in a given game configuration.*

1. *Assume that the cells* (x_1, y_1), (x_2, y_2) *as well as* (x_2, y_3) *contain one mouse each. Then at least one of the two rectangular moves*

$$(x_1, y_1), (x_2, y_2) \to (x_1, y_2), (x_2, y_1),$$
$$(x_1, y_1), (x_2, y_3) \to (x_1, y_3), (x_2, y_1)$$

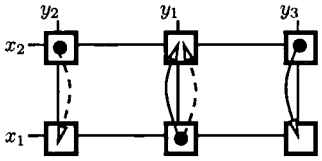
does not violate I4.

2. *Assume that the cells* (x_1, y_1), (x_2, y_2) *as well as* (x_3, y_2) *contain one mouse each. Then at least one of the two rectangular moves*

$$(x_1, y_1), (x_2, y_2) \to (x_1, y_2), (x_2, y_1),$$
$$(x_1, y_1), (x_3, y_2) \to (x_1, y_2), (x_3, y_1)$$

does not violate I4.

Proof. We prove only the first statement, as the second is obtained by simply exchanging the roles of rows and columns.



We assume that the given game configuration has no deserted frames. Hence, according to Lemma 6, I4 is in danger only if each of the two moves removes the last two mice out of the frame of some subboard. Assume the first move causes *Frame*(*Board*(I)) to be deserted and *Frame*(*Board*(J)) is deserted if the second move is executed. Observe that for $\overline{X} = \{1, \ldots, n\} \setminus X$

$$\begin{aligned} \mathit{frame}(X) &= \{(a,b) | a \in X, b \notin X\} \cup \{(a,b) | a \notin X, b \in X\} \\ &= \{(a,b) | a \notin \overline{X}, b \in \overline{X}\} \cup \{(a,b) | a \in \overline{X}, b \notin \overline{X}\} \\ &= \mathit{frame}(\overline{X}). \end{aligned}$$

We may hence assume $x_1 \in I$ and $x_1 \in J$ without loss of generality. Lemma 6.2 ensures that a rectangular move can move only the last two mice out of the frame of a board B if exactly one of the final positions is in B. Hence (x_1, y_2) is in *Board*(I) and (x_1, y_3) is in *Board*(J). Consequently $y_2 \in I$ and $y_3 \in J$ hold.

As the starting positions of the first move need to be in the horizontal, respectively vertical, frame of *Board*(I), we conclude that (x_1, y_1) is in the horizontal frame and (x_2, y_2) is in the vertical frame. We thus have $x_2 \notin I$ and $y_1 \notin I$. Similar arguments for the second move yield $x_2 \notin J$ and $y_1 \notin J$.

Assume $y3 \in I$. Since $x_2 \notin I$, the mouse on (x_2, y_3) is in the vertical frame of *Board*(I) and *Frame*(*Board*(I)) is not deserted by the first move. Hence $y_3 \notin I$. A similar argument for *Board*(J) yields $y_2 \notin J$.

In Figure 9 the board is partitioned into 16 segments. We may assume that the frames of I and J do not hold any mice beyond the three mice explicitly mentioned in the lemma. The frames of I and J are the shaded areas in the diagram. (The Frame of *Board*(I) is shaded with horizontal lines, and the Frame of *Board*(J) is shaded with vertical lines.) We now see that the vertical frame of *Board*($I \cap J$) is deserted, while the horizontal frame is not. This in conjunction with Lemma 6.1 contradicts the assumption that I1 is valid before the move. □

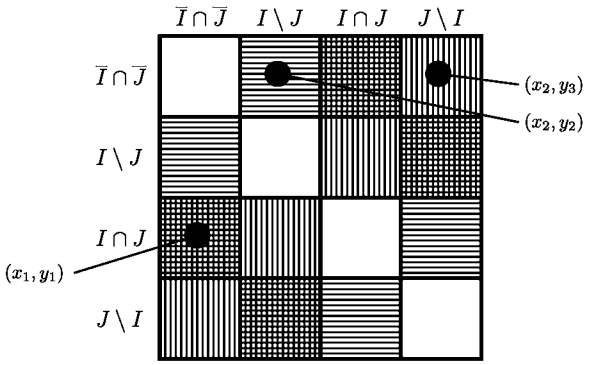


FIG. 9. *The board broken down into* 16 *nonempty segments, reflecting the sets* I *and* J *and the location of the three starting positions for the mice. The shaded areas belong to the frame of either* $Board(I)$ *or* $Board(J)$.

LEMMA 9. *Invariants* I1, ..., I4 *hold before the main loop starts.*

Proof. As to I1, observe that as $\mathcal{C}_1$ and $\mathcal{C}_2$ are cycle covers, every string gets two successors assigned and is assigned twice as the successor of another string. Hence in the initial placement every row and column contains two mice. Neither a Discard of Period nor a Diagonal Monge changes the number of mice per row and column.

$B = Board(\{i\})$ contains only one hole, the elliptic diagonal hole in cell (i, i), which is filled either by a Discarding of a Period or by a Diagonal Monge. Hence I2 holds.

As to I3, if a cell (k, k) initially holds a mouse, this mouse is located in the rectangle representing the self-overlap and not in the elliptic hole. If the Diagonal Monge executed to fill the hole in (i, i) happens to be of the form $(i, k), (k, i) \to (i, i)(k, k)$, then the mouse reaching cell (k, k) enters the rectangle for the self-overlap. Hence I3 holds.

I4 holds in the initial configuration: assume $Frame(Board(I))$ is deserted; then in $\mathcal{C}_1$ and $\mathcal{C}_2$ every string s_i with $i \in I$ has a successor s_j with $j \in I$. So both cycle covers contain a cycle cover for the set $\{s_i | i \in I\}$, in violation of the expansion property.

If the move executed to fill the hole in (i, i) consists of Discarding of a Period, I4 clearly remains valid. If cell (i, i) does not hold a mouse, the two mice in row i and the two mice in column i are four distinct mice. If the two mice in row i are in the same cell, Lemma 7 guarantees that I4 is preserved. The same holds if the two mice in column i are in the same cell. If the mice are on four different cells, Lemma 8 guarantees that one of the possible Diagonal Monges does not violate I4 and is hence legal. □

We are now ready to introduce the main loop of our algorithm. The set I is always a set of consecutive numbers. Therefore $Frame(Board(I))$ contains at most two greedy cells:

$$\begin{array}{c} (\min(I)-1,\ \min(I)) \ \text{ for } \min(I) > 1 \\ \text{and} \quad (\max(I),\ \max(I)+1) \ \text{ for } \max(I) < n. \end{array}$$

Furthermore $Frame(Board(I))$ contains cell $(n, 1)$ if either $\min(I) = 1$ or $\max(I) = n$ but not both.

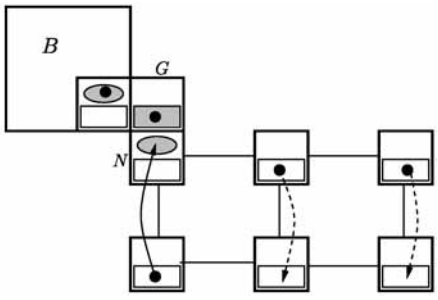
<table>
<tr><td colspan="5">while $I \neq \{1, \ldots, n\}$</td></tr>
<tr><td colspan="5">Let G be the highest ranked greedy cell in $Frame(Board(I))$
Let $N = (k, k)$ be the diagonal cell adjacent to G and not in $Board(I)$</td></tr>
<tr><td colspan="5">Does G contain a mouse?</td></tr>
<tr><td colspan="2">yes</td><td colspan="3">no</td></tr>
<tr><td colspan="2">Does N contain a mouse?</td><td colspan="3">Does N contain a mouse?</td></tr>
<tr><td>yes</td><td>no</td><td>no</td><td colspan="2">yes</td></tr>
<tr><td rowspan="3">(#1)
Discard a Period in N</td><td rowspan="3"></td><td rowspan="3">(#3)
Greedy Monge in G</td><td colspan="2">(#4)
Is the Triple in G, N legal?</td></tr>
<tr><td>no</td><td>yes</td></tr>
<tr><td>Greedy Monge in G</td><td rowspan="2">Triple</td></tr>
<tr><td colspan="2">(#2)
Diagonal Monge in N</td><td>Discard a Period in N</td></tr>
<tr><td colspan="5">Set $I = I \cup \{k\}$</td></tr>
</table>

We assume in the above description that only legal Diagonal and Greedy Monges are executed. The subsequent analysis has to show that legal moves exist when required.

We first show that RDA maintains invariants I1, I2, and I3, provided that the described moves exist. As we discard only periods and apply Monges or Triple moves, all of which leave the number of mice per row and column unchanged, invariant I1 is automatically satisfied. Since we extend the subboard $Board(I)$ only after filling the two new holes, namely G and N, invariant I2 follows. I3 is valid since, if a mouse moves into the diagonal cell (i, i) indirectly, that is, not by a Monge or Triple in (i, i), then the mouse is to be placed in the rectangle representing the self-overlap.

The existence of the moves described and their accordance with I4 will now be checked in a case-by-case study. The labels in the algorithm refer to the corresponding observations. We assume without loss of generality that $G = (\max(I), \max(I) + 1)$ and hence $k = \max(I) + 1$.

The **Discard of a Period** (#1) is executable as there is a mouse in cell N by case assumption and by I3 this mouse is in the rectangle for the self-overlap. Furthermore, the move cannot remove the last mouse from any frame since no mouse leaves its cell. The **Diagonal Monge** (#2): Cell $N = (k, k)$ does not hold a mouse, and hence the two mice in row k and the two mice in column k are four distinct mice. One of the mice in column k is in G and stays there.



The other mouse in column k can execute a Diagonal Monge with either of the two mice in row k. If the two mice of row k are located in the same cell, Lemma 7 guarantees that I4 is maintained. Otherwise, if the two mice of row k are in two distinct cells, Lemma 8.1 implies that one of the two Diagonal Monges does not violate I4.

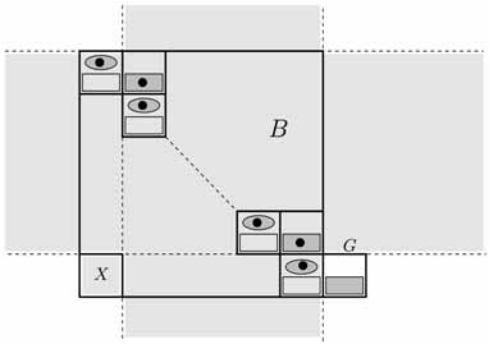


FIG. 10. *Location of cells higher ranked than G during the RDA.*

The observations for scenario (#3) and (#4) are more involved as the issue of dominance arises. For Greedy Monges as well as for Triple moves the greedy cell has to dominate the starting positions of the mice involved; i.e., its rank must be at least as high as that of the starting positions.

Let $G = (k, k+1)$. Figure 10 shows the location of cells with higher rank. Why? Let $I = \{r, r+1, \ldots, k-1, k\}$. A cell (x, y) has a rank higher than G if the overlap (s_x, s_y) is no longer a legal choice for GREEDY. We have

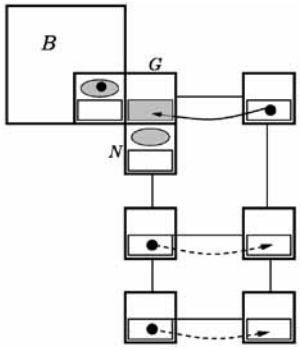
$$\text{rank}(s_x, s_y) \geq \text{rank}(s_{y-1}, s_y), \text{rank}(s, s_{x+1}),$$

since at the latest, when s_y receives a predecessor or s_x receives a successor, the choice of (s_x, s_y) is illegal. Hence, if GREEDY has already merged the strings $s_r, s_{r+1}, \ldots, s_{k-1}, s_k$, then cell (x, y) has a higher rank than G iff $r+1 \leq y \leq k$, $r \leq x \leq k-1$, or $(x = k$ and $y = r)$, since in the latter case a cycle is closed.

Therefore, the cells with rank higher than G are all those between the two horizontal dashed lines in Figure 10 as well as those between the two vertical dashed lines and the cell marked X. As we assume that all holes in B hold a mouse, the two mice in all of these rows and columns are accounted for. By I4 we also know that no mouse can be in the cell marked X, since otherwise the frame of $B = Board(I)$ is deserted.

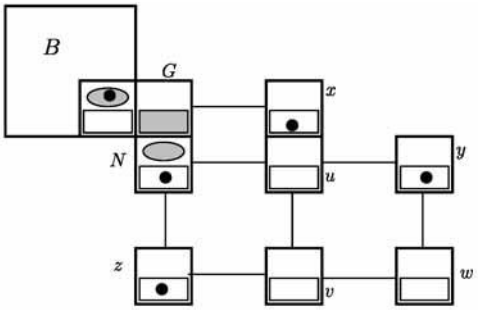
Hence the greedy cell G has a sufficiently high rank to move every mouse that is not yet in a hole and not in a diagonal cell. We can now address the two scenarios (#3) and (#4).

The arguments for the **Greedy Monge** (#3) are now quite similar to those of the Diagonal Monge in (#2): G itself does not hold a mouse, and hence the two mice of its row and column are four distinct mice.



One of the mice in the row of G is in a diagonal hole in B and stays there. The other mouse may perform a Greedy Monge with either one of the two mice in the column of G. As we have just seen, G is strong enough to move these mice and Lemma 8.2 or Lemma 7 guarantees that at least one of the two moves does not violate I4.

We conclude with the analysis of scenario (#4).



N holds a mouse, and by I3 this mouse is located in the rectangle for the self-overlap. We also know that the rank of G is sufficiently high, so that G dominates cells u, v, w, x, y, z.

Just for the sake of argument let us execute the unjustified move

$$N, x \to G, u.$$

This move is not justified by an inequality; however, it does not violate I4, since one participating mouse leaves a diagonal cell and a diagonal cell never belongs to a frame.

After this unjustified move, G is filled and we have mice located in z, u, and y. By Lemma 8.1 (or Lemma 7 if x and y are in the same column) one of the two Diagonal Monges

$$z, u \to N, v,$$
$$z, y \to N, w$$

must comply with I4. However, our unjustified first move followed by the first of these Diagonal Monges corresponds to the Greedy Monge $z, x \to G, v$ followed by a Discard of Period in N. The unjustified move combined with the second Diagonal Monge corresponds to the Triple $x, y, z \to G, u, w$. Hence, if the Triple in (#4) is illegal, we know that the Greedy Monge followed by the Discard of Period is legal.

LEMMA 10. *Invariants* I1 *thru* I4 *hold after the main loop. At the end of the main loop every hole is filled and one mouse resides in position* $(n, 1)$.

Proof. We have shown in Lemma 9 that the invariants hold before the main loop. Our case-by-case analysis of the different branches within the main loop has verified that the invariants stay intact. At the end of the main loop $I = \{1, \ldots, n\}$ and $Board(I)$ is the entire board. Each of the $2n - 1$ holes is filled, since I2 is valid, and because of I1 the only possible position for the last mouse is $(n, 1)$. □

Subsequent to [14] a refined extension of the RDA was introduced in [9]. It grows two subboards in parallel and wins the game for as many as $\Theta(4^n)$ greedy orders.

For the sake of clarity, let us cross-check how RDA wins our game of section 3. Figures 11 and 12 show one of three possible sequences of moves.

4.2. Proof of Corollary 1. RDA provides a sequence of moves, transferring the $2n$ mice representing the two cycle covers to the holes on the diagonal, the off-diagonal, and cell $(n, 1)$. If the input consists of a superstring $s_{\sigma(1)}, \ldots, s_{\sigma(n)}$ and a cycle cover, we have only $2n - 1$ mice. In this case we insert an artificial extra mouse

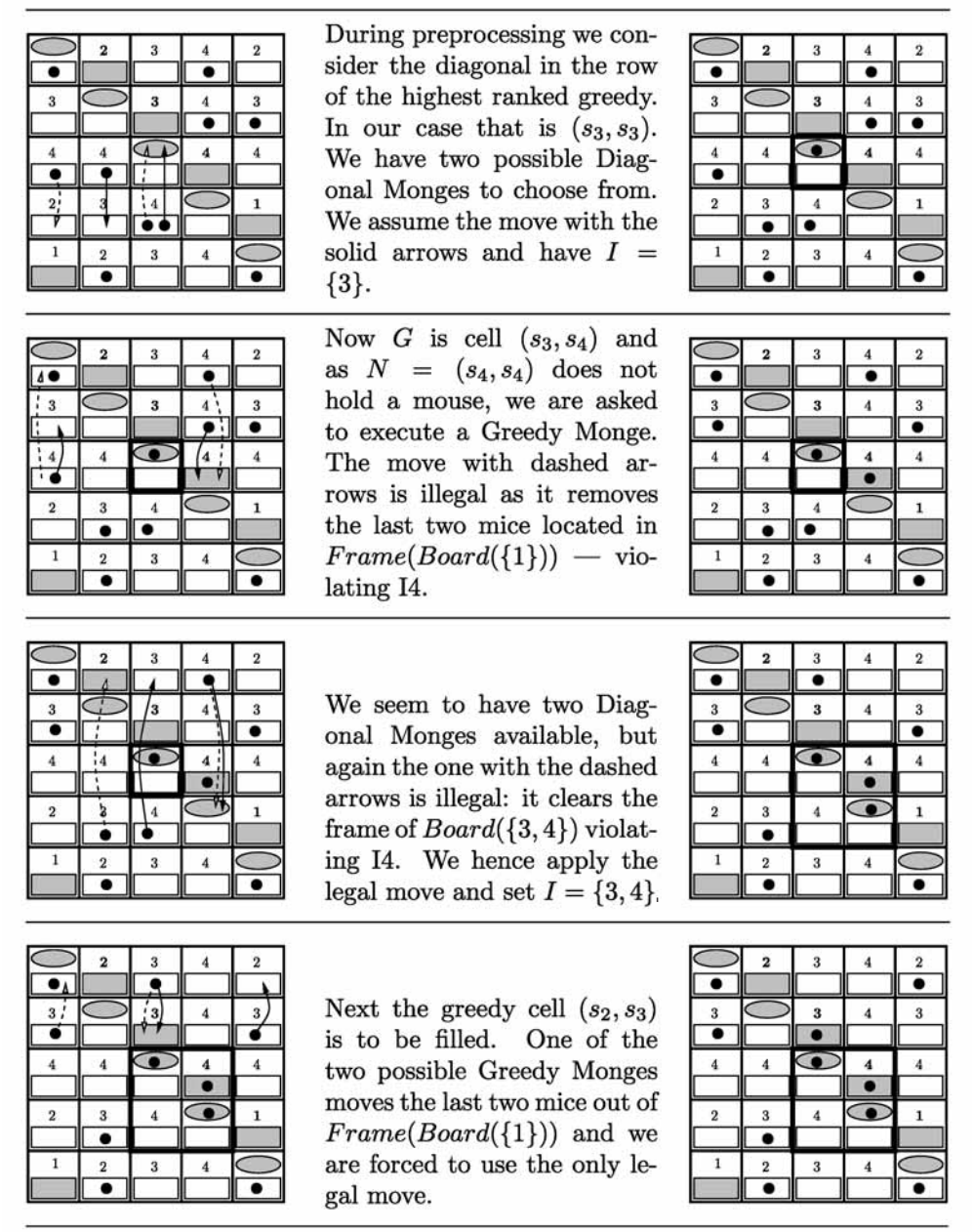


FIG. 11. *The RDA at work.*

into cell $(\sigma(n), \sigma(1))$ during preprocessing. Now, every row and every column contains two mice again.

Correspondingly the overlap $(n, 1)$ turning the greedy superstring into a cycle is not to be used in Corollary 1.

What we need to show is that, given a sequence winning the game eventually, occupying cell $(n, 1)$, and with the artificial mouse starting in $(\sigma(n), \sigma(1))$, we can obtain a sequence not involving the artificial mouse and not occupying cell $(n, 1)$.

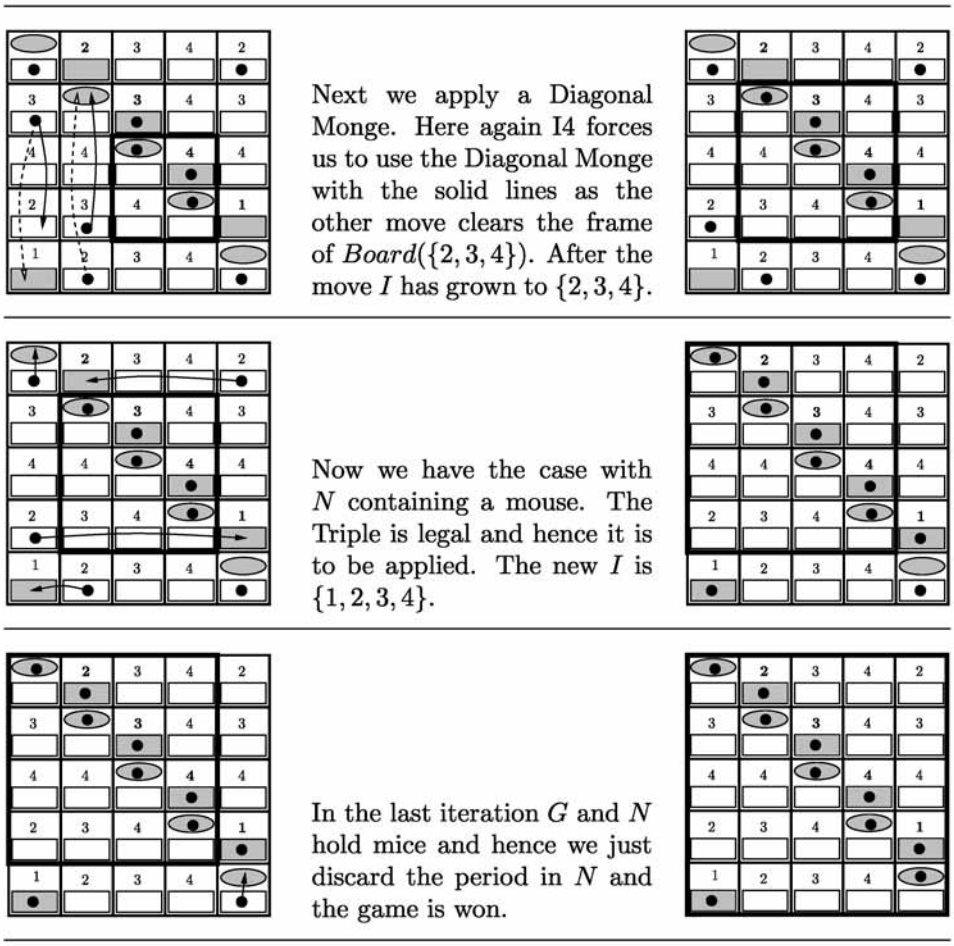


FIG. 12. *The RDA at work (continued).*

LEMMA 11. *Assume that there is an initial configuration A of mice placed on the board, as well as a final configuration F, where F can be reached from A by a sequence of Greedy and Diagonal Monges, Triple moves, and Discards of Periods. Then for any configuration A' obtained by deleting one mouse from A there is a configuration F' obtained by deleting one mouse from F such that F' is reachable from A'.*

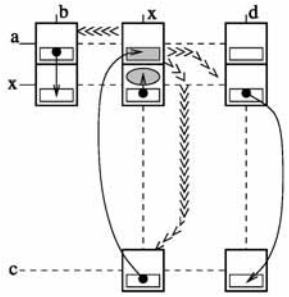
Proof. It is sufficient to show that the lemma holds for one-step transitions.

If the artificial mouse is not participating in the move, then remove it from F to obtain F', and the claim is obvious.

If the period of the artificial mouse is discarded, then remove the artificial mouse from F to obtain F'.

If the artificial mouse is part of a Diagonal or Greedy Monge, then insert the other participating mouse into the diagonal or greedy cell and remove the artificial mouse from F to obtain F'.

For the Triple move we have to differentiate three cases. We concentrate on the vertical version of the Triple move, i.e., $(a,x) \geq (x,b)$. Remember that the artificial mouse participates in the move since we are done otherwise.



Case 1. The artificial mouse sits on position (x, x). Remove (x, b) from F to obtain F' and observe that F' can be reached by a Diagonal Monge in (x, x) and a Greedy Insertion into (a, x).

Case 2. The artificial mouse sits on position (x, d) (resp., (c, x)). Remove (c, d) from F to obtain F'. To obtain F' perform a Greedy Insertion of the mouse from (c, x) (resp., (x, d)) into (a, x). Then use Lemma 4, statement 1, to move the mouse from (a, b) to (x, b), sacrificing the period in (x, x).

Case 3. The artificial mouse sits on cell (a, b). We remove cell (x, b) from F to obtain F'. Without loss of generality assume (x, d) is at least as big as (c, x). Hence according to Lemma 4, statement 4, we may move (c, x) to (c, d), sacrificing the period in (x, x). The mouse from (x, d) simply moves to (a, x) by a Greedy Insertion. □

To complete the proof of Corollary 1 we need only to observe that cell $(n, 1)$ has rank 1. If the artificial mouse did not end up in $(n, 1)$, we can just move the mouse from $(n, 1)$ into the vacant hole that the artificial mouse did end up in, either with a Greedy Insertion or with a Greedy Insertion followed by a Diagonal Insertion.

5. Monge inequalities alone are insufficient. We say that a board of overlaps is a "Monge Board" if the Monge inequalities hold throughout. For the linear greedy order $(1, 2), (2, 3), \ldots, (n-1, n)$ we found by extensive enumeration that the length of the greedy superstring is within a factor of 2 of the minimum superstring length, provided $n \leq 8$.

However, the example of a 9×9 Monge Board in Figure 13 implies that the introduction of the Triple move is necessary. In particular, we show that the length of the superstring, defined as $\sum_{i=1}^{n} |s_i| - \sum_{i=1}^{n-1} (s_i, s_{i-1})$, exceeds the minimum length by a factor of $2\frac{1}{3}$. Observe that the Triple inequality, as a consequence of Corollary 1, is necessarily violated.

We compare the linear greedy order $(1, 2), (2, 3), \ldots, (8, 9)$ with the adversarial order $(1, 9, 2, 6, 3, 8, 4, 7, 5)$. We show that the above board is a Monge Board for sufficiently large m. In particular, we have to verify that the inequalities for Diagonal and Greedy Insertions as well as for Greedy and Diagonal Monges hold.

The values on the diagonal are the highest in their row and column, and hence the inequalities for Diagonal Insertions hold. Observe that $(1, 2), \ldots, (8, 9)$ is indeed the greedy order, and therefore the inequalities for Greedy Insertions hold as well. A complete case analysis shows that the inequalities for Diagonal and Greedy Monges are also satisfied.

The length of the strings sums to $15m + 44$, and the overlaps picked by GREEDY sum up to $8m + 20$. Hence GREEDY delivers a solution of value $7m + 24$. The corresponding adversarial overlaps sum to $12m + 12$. Subtracting this from the length of the strings, we obtain a solution of value $3m + 32$. Hence the approximation ratio is $\frac{7m+24}{3m+32}$, which approaches $2\frac{1}{3}$ as m tends to infinity.

	1	2	3	4	5	6	7	8	9
1	**2m+5**	$2m$+4	m	0	0	m	0	0	2m
2	0	**3m+8**	$2m$+4	m	m	$2m$+3	m	m	0
3	0	0	**2m+8**	m+4	m+1	m	m+1	m+1	0
4	0	0	m	**m+7**	m+4	m	m+3	0	0
5	0	0	m	0	**m+6**	m+2	0	0	0
6	0	0	$2m$+2	m	m	**2m+5**	m+2	m	0
7	0	0	m+1	0	m+2	m+1	**m+4**	0	0
8	0	0	m	m+1	m+1	m	m+1	**m+1**	0
9	0	2m	m	0	0	m	0	0	**2m**

FIG. 13. *A Monge board. Cell* (i, j) *contains the overlap* (s_i, s_j). *Moreover, we set* $|s_i| = (s_i, s_i)$, *and hence all period lengths are zero.*

Hence the Monge inequality—no matter in which sequence it is applied—is insufficient to prove the bound of 2 for the superstring problem. Subsequent to our work, it was shown in [9] that the the basic inequalities, insertions and Monges, together with the Triple inequality, are still insufficient to prove the general result. This was done by providing a 10×10 Monge/Triple board, a greedy order, and an adversarial order that achieves an approximation factor of $2\frac{1}{6}$.

6. Conclusions and open problems. We have introduced the Triple inequality and have shown that conditional string inequalities are sufficient to settle the greedy superstring conjecture for linear greedy orders. The conjecture is implied by the stronger statement of Theorem 1, and it may be that this stronger version eventually turns out to be easier to verify. Moreover, we have shown that the Triple move is indeed crucial.

Of course the verification of the superstring conjecture for all greedy orders is the major open problem. In sharp contrast to the cyclic greedy algorithm that finds optimal cycle covers for strings, the performance of the greedy heuristic for short superstrings remains poorly understood. Intuitively speaking, previous research approaches have tried to *transfer* the optimality of the cyclic greedy heuristic to the case of superstrings: the upper bound of four for the performance ratio is obtained by carefully decomposing the cycles of a cycle cover and merging the pieces. Further approximation algorithms have used the cyclic greedy heuristic as a basis.

One should expect that, according to this reasoning, results are harder to obtain the more differently the cyclic greedy heuristic and the greedy heuristic for strings behave on a given input: if the optimal cycle cover is one cycle, then the two algorithms make exactly the same choices with the string algorithm stopping one iteration earlier. Here the found superstring is optimal. Observe that our result also covers inputs for which the algorithms behave completely differently right from the beginning. (If the cyclic greedy heuristic picks (s_i, s_i) first, the two algorithms may not share a pair of consecutive strings.)

Our results suggest the following program to analyze the greedy algorithm.

1. Formulate crucial string properties as (conditional) linear inequalities.
2. Analyze the resulting systems of inequalities via linear programming or via combinatorial methods derived from linear programming.

We have used linear programming extensively to isolate board configurations that could not be solved with given sets of moves. The Triple move was found as a consequence of this approach.

Acknowledgments. Many thanks to Uli Laube for carefully implementing the mice game, allowing us to recognize several dead ends ahead of time. A description of his software tool SINDBAD and how it supported our and further research can be found in [9].

We furthermore thank Stefan Kirchner as well as the unknown referees for pointing out a gap in a proof in the original version and for many more helpful remarks and suggestions.

REFERENCES

[1] C. Armen and C. Stein, *Improved length bounds for the shortest superstring problem*, in Proceedings of the 4th International Workshop on Algorithms and Data Structures (WADS), S. G. Akl, F. Dehne, J.-R. Sack, and N. Santoro, eds., Lecture Notes in Comput. Sci. 955, Springer-Verlag, New York, 1995, pp. 494–505.

[2] C. Armen and C. Stein, *A* $2\frac{2}{3}$ *superstring approximation algorithm*, Discrete Appl. Math., 88 (1998), pp. 29–57.

[3] A. Blum, T. Jiang, M. Li, J. Tromp, and M. Yannakakis, *Linear approximation of shortest superstrings*, J. ACM, 41 (1994), pp. 630–647.

[4] D. Breslauer, D. Jiang, and Z. Jiang, *Rotations of periodic strings and short superstrings*, J. Algorithms, 24 (1997), pp. 340–353.

[5] A. Czumaj, L. Gasieniec, W. Piotrow, and W. Rytter, *Parallel and sequential approximations of shortest superstrings*, J. Algorithms, 32 (2003), pp. 71–385.

[6] D. Gusfield, *Algorithms on Strings, Trees and Sequences*, Cambridge University Press, Cambridge, UK, 1997.

[7] K. Kaplan and N. Shafrir, *The greedy algorithm for shortest superstrings*, Inform. Process. Lett., 93 (2005), pp. 13–17.

[8] R. Kosaraju, J. Park, and C. Stein, *Long tours and short superstrings*, in Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science (FOCS), Santa Fe, NM, 1994, IEEE Computer Society Press, Los Alamitos, CA, 1994, pp. 166–177.

[9] U. Laube and M. Weinard, *Conditional inequalities and the shortest common superstring problem*, in Proceedings of the 4th Prague Stringology Conference, M. Šimánek and J. Holub, eds., Vydavatelstí ČVUT, Prague, 2004, pp. 124–138.

[10] G. Monge, *Mémoire sur la théorie des déblais et des remblais*, in Histoire de l'Academie Royale des Sciences, Annés MDCCLXXXI, avec les Mémoires de Mathématique et de Physique, pour la même Année, Tirés des Registres de cette Académie, 1781, pp. 666–704.

[11] Z. Sweedyk, *A* $2\frac{1}{2}$*-approximation algorithm for shortest superstring*, SIAM J. Comput., 29 (1999), pp. 954–986.

[12] J. Tarhio and E. Ukkonen, *A greedy approximation algorithm for constructing shortest common superstrings*, Theoret. Comput. Sci., 57 (1988), pp. 131–134.

[13] S. Teng and F. Yao, *Approximating a shortest superstring*, in Proceedings of the 34th Annual IEEE Symposium on Foundations of Computer Science (FOCS), Palo Alto, CA, 1993, IEEE Computer Society Press, Los Alamitos, CA, 1993, pp. 158–165.

[14] M. Weinard and G. Schnitger, *On the greedy superstring conjecture*, in Proceedings of the 23rd Conference on Foundations of Software Technology and Theoretical Computer Science (FST TCS), Mumbai, India, 2003, P. K. Pandya and J. Radhakrishnan, eds., pp. 387–398.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 523–528

A NOTE ON UNSATISFIABLE k-CNF FORMULAS WITH FEW OCCURRENCES PER VARIABLE*

SHLOMO HOORY† AND STEFAN SZEIDER‡

Abstract. The (k, s)-SAT problem is the satisfiability problem restricted to instances where each clause has exactly k literals and every variable occurs at most s times. It is known that there exists a function f such that for $s \leq f(k)$ all (k, s)-SAT instances are satisfiable, but $(k, f(k)+1)$-SAT is already NP-complete ($k \geq 3$). We prove that $f(k) = O(2^k \cdot \log k/k)$, improving upon the best known upper bound $O(2^k/k^\alpha)$, where $\alpha = \log_3 4 - 1 \approx 0.26$. The new upper bound is tight up to a $\log k$ factor with the best known lower bound $\Omega(2^k/k)$.

Key words. unsatisfiable CNF formulas, NP-completeness, occurrence of variables

AMS subject classifications. 68R05, 68Q25, 05B99

DOI. 10.1137/S0895480104445745

1. Introduction.

We consider propositional formulas in conjunctive normal form (CNF) represented as sets of clauses, where each clause is a set of literals. A literal is either a variable or a negated variable. Let k, s be fixed positive integers. We denote by (k, s)-CNF the set of formulas F where every clause of F has *exactly* k distinct literals and each variable occurs in *at most* s clauses of F. We denote the set of satisfiable formulas by SAT.

It was observed by Tovey [7] that all formulas in $(3, 3)$-CNF are satisfiable, and that the satisfiability problem restricted to $(3, 4)$-CNF is already NP-complete. This was generalized in Kratochvíl, Savický, and Tuza [4], where it is shown that for every $k \geq 3$ there is some integer $s = f(k)$ such that

1. all formulas in (k, s)-CNF are satisfiable, and
2. the satisfiability problem restricted to formulas in $(k, s + 1)$-CNF is already NP-complete.

The function f can be defined for $k \geq 1$ by the equation

$$f(k) := \max\{\, s : (k, s)\text{-CNF} \subseteq \text{SAT} \,\}.$$

Exact values of $f(k)$ are known only for $k \leq 4$. It is easy to verify that $f(1) = 1$ and $f(2) = 2$. It follows from [7] that $f(3) = 3$ and $f(k) \geq k$ in general. Also, by [6], we know that $f(4) = 4$.

Upper and lower bounds for $f(k)$, $k = 5, \ldots, 9$, have been obtained in [2, 6, 1, 3]. For larger values of k, the best known lower bound, a consequence of the Lovász local lemma, is due to Kratochvíl, Savický, and Tuza [4]:

$$f(k) \geq \left\lfloor \frac{2^k}{ek} \right\rfloor. \tag{1}$$

*Received by the editors August 25, 2004; accepted for publication (in revised form) January 28, 2006; published electronically June 9, 2006.
http://www.siam.org/journals/sidma/20-2/44574.html

†Department of Computer Science, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (shlomoh@cs.ubc.ca). This author's research was supported in part by an NSERC grant and by a PIMS postdoctoral fellowship.

‡Department of Computer Science, University of Durham, Durham DH1 3LE, UK (stefan.szeider@durham.ac.uk).

Prior to this work, the best known upper bound has been by Savický and Sgall [5]. They constructed a family of unsatisfiable k-CNF formulas with 2^k clauses and a small number of occurrences per variable. Their construction yields

$$f(k) = O\left(\frac{2^k}{k^\alpha}\right), \tag{2}$$

where $\alpha = \log_3 4 - 1 \approx 0.26$.

In this paper we asymptotically improve upon (2) and show that

$$f(k) = O\left(\frac{2^k \log k}{k}\right). \tag{3}$$

Our result reduces the gap between the upper and lower bounds to a $\log k$ factor. It turns out that the construction yielding the upper bound (3) can be generalized. We present a class of k-CNF formulas that is amenable to an exhaustive search using dynamic programming. This enables us to calculate upper bounds on $f(k)$ for values up to $k = 20000$, improving upon the bounds provided by the constructions underlying (2) and (3).

The remainder of the paper is organized as follows. In section 2 we start with a simple construction that already provides an $O(2^k \log^2 k/k)$ upper bound on $f(k)$. In section 3 we refine our construction and obtain the upper bound (3). In the last section we describe the more general construction and the results obtained using computerized searching.

2. The first construction. We denote by $\mathcal{K}(x_1, \ldots, x_k)$ the complete unsatisfiable k-CNF formula on the variables $x_1, \ldots, x_k$. This formula consists of all 2^k possible clauses. Let $\mathcal{K}^-(x_1, \ldots, x_k) = \mathcal{K}(x_1, \ldots, x_k) \setminus \{\{x_1, \ldots, x_k\}\}$. The only satisfying assignment for $\mathcal{K}^-(x_1, \ldots, x_k)$ is the all-False assignment. Also, for two CNF formulas F_1 and F_2 on disjoint sets of variables, their product $F_1 \times F_2$ is defined as $\{c_1 \cup c_2 : c_1 \in F_1 \text{ and } c_2 \in F_2\}$. Note that the satisfying assignments for $F_1 \times F_2$ are assignments that satisfy F_1 or F_2. In what follows, log and ln denote logarithms to the bases of 2 and e, respectively.

LEMMA 2.1. $f(k) < 2^k \cdot \min_{1 \le l \le k} \left((1 - 2^{-l})^{\lfloor k/l \rfloor} + 2^{-l}\right)$.

Proof. We prove the lemma by constructing, for every l, an unsatisfiable (k, s)-CNF formula F, where $s = 2^k \cdot ((1 - 2^{-l})^{\lfloor k/l \rfloor} + 2^{-l})$. Let k, l be two integers such that $1 \le l \le k$, and let $u = \lfloor k/l \rfloor$ and $v = k - l \cdot u$. Define the formula F as the union $F = F_0 \cup F_1 \cup \cdots \cup F_u$, where

$$F_0 = \mathcal{K}(z_1, \ldots, z_v) \times \prod_{i=1}^{u} \mathcal{K}^-(x_1^{(i)}, \ldots, x_l^{(i)}),$$

$$F_i = \mathcal{K}(y_1^{(i)}, \ldots, y_{k-l}^{(i)}) \times \{\{x_1^{(i)}, \ldots, x_l^{(i)}\}\} \quad \text{for } i = 1, \ldots, u.$$

Therefore, F is a k-CNF formula with n variables and m clauses, where

$$n = k + u \cdot (k - l) \le k^2/l, \tag{4}$$

$$m = 2^v \cdot (2^l - 1)^u + u \cdot 2^{k-l} = 2^k \cdot \left((1 - 2^{-l})^{\lfloor k/l \rfloor} + \lfloor k/l \rfloor \cdot 2^{-l}\right). \tag{5}$$

To see that F is unsatisfiable observe that any assignment satisfying F_0 must set all the variables $x_1^{(i)}, \ldots, x_l^{(i)}$ to False for some i. On the other hand, any satisfying assignment to F_i must set at least one of the variables $x_1^{(i)}, \ldots, x_l^{(i)}$ to True.

To bound the number of occurrences of a variable note that the variables $z_j, y_j^{(i)}$, and $x_j^{(i)}$ occur $|F_0|, |F_i|$, and $|F_0|+|F_i|$ times, respectively. Since $|F_0| = 2^v \cdot (2^l-1)^u = 2^k \cdot (1-2^{-l})^{\lfloor k/l \rfloor}$ and $|F_i| = 2^{k-l}$, we get the required result. □

For $k \geq 4$, let l be the largest integer satisfying $2^l \leq k \cdot \log e / \log^2 k$. It follows that

$$(1-2^{-l})^{\lfloor k/l \rfloor} \leq \exp(-2^{-l} \cdot \lfloor k/l \rfloor) \leq \exp\left(-\frac{\log^2 k}{k \log e} \cdot \left(\frac{k}{l} - 1\right)\right)$$
$$\leq e \cdot \exp\left(-\frac{\log^2 k}{l \log e}\right) \leq e \cdot \exp\left(-\frac{\log k}{\log e}\right) = \frac{e}{k},$$

where the last two inequalities follow from the fact that for $k \geq 4$ we have $\log^2 k < k \log e$ and $l \leq \log k$. Therefore, by Lemma 2.1 there exists an unsatisfiable k-CNF formula F where the number of occurrences of variables is bounded by

$$2^k \cdot \left(\frac{e}{k} + \frac{2 \log^2 k}{k \log e}\right).$$

It may be of interest that, by (4) and (5), the number of clauses in F is $O(2^k \cdot \log k)$ and the number of variables is $O(k^2/\log k)$. Thus, in comparison to the construction in [5], we pay for the better bound on $f(k)$ by an $O(\log k)$ factor in the number of clauses.

COROLLARY 2.2. $f(k) = O(2^k \cdot \log^2 k/k)$.

3. A better upper bound. To simplify the subsequent discussion, let us fix a value of k. We will be concerned only with CNF formulas F that have clauses of size at most k. We call a clause of size less that k an *incomplete* clause and denote $F' = \{c \in F : |c| < k\}$. A clause of size k is a *complete* clause, and we denote $F'' = \{c \in F : |c| = k\}$.

LEMMA 3.1. $f(k) < \min\{2^{k-l+1} : l \in \{0, \ldots, k\}$ *and* $l \cdot 2^l \leq \log e \cdot (k-2l)\}$.

Proof. Let l be in $\{0, \ldots, k\}$, satisfying $l \cdot 2^l \leq \log e \cdot (k-2l)$, and set $s = 2^{k-l+1}$. We will define a sequence of CNF formulas, $F_0, \ldots, F_l$. We require that (i) F_j is unsatisfiable, (ii) F_j' is a $(k-l+j)$-CNF formula, (iii) $|F_j'| \leq 2^{k-l}$, and (iv) the maximal number of occurrences of a variable in F_j is bounded by s. It follows that F_l is an unsatisfiable (k,s)-CNF formula, implying the claimed upper bound.

Set $d_j = k-l+j$ and $u_j = \lfloor (k-l+j)/(l-j+1) \rfloor$. We proceed by induction on j. For $j=0$, we define $F_0 = \mathcal{K}(x_1, \ldots, x_{k-l})$. It can be easily verified that F_0 satisfies the above four requirements. For $j > 0$, assume there is a formula F_{j-1} on the variables $y_1, \ldots, y_n$ satisfying the requirements. We define the formula $F_j = \bigcup_{i=0}^{u_j} F_{j,i}$ as follows:

$$F_{j,0} = \mathcal{K}(z_1, \ldots, z_{d_j - u_j \cdot (l-j+1)}) \times \prod_{i=1}^{u_j} \mathcal{K}^-(x_1^{(i)}, \ldots, x_{l-j+1}^{(i)}), \tag{6}$$

$$F_{j,i} = F'_{j-1}(y_1^{(i)}, \ldots, y_n^{(i)}) \times \{\{x_1^{(i)}, \ldots, x_{l-j+1}^{(i)}\}\} \cup F''_{j-1}(y_1^{(i)}, \ldots, y_n^{(i)}) \tag{7}$$
$$\text{for } i = 1, \ldots, u_j.$$

It is easy to verify that F_j' is a $(k-l+j)$-CNF formula. To see that F_j is unsatisfiable, observe that any assignment satisfying $F_{j,0}$ must set all the variables $x_1^{(i)}, \ldots, x_{l-j+1}^{(i)}$ to False for some i. On the other hand, for any satisfying assignment to $F_{j,i}$, at least one of the variables $x_1^{(i)}, \ldots, x_{l-j+1}^{(i)}$ must be set to True.

Let us consider the number of occurrences of a variable in F_j. Consider first the y-variables. These variables occur only in the u_j duplicates of F_{j-1} and therefore occur the same number of times as in F_{j-1}, which is bounded by s by induction. The number of occurrences of an x- or z-variable is $|F'_{j-1}| + |F_{j,0}|$ or $|F_{j,0}|$, respectively. By induction, $|F'_{j-1}| \leq 2^{k-l}$. Also,

$$\begin{aligned}|F'_j| = |F_{j,0}| &= 2^{d_j - u_j \cdot (l-j+1)} \cdot (2^{l-j+1} - 1)^{u_j} = 2^{d_j} \cdot (1 - 2^{-l+j-1})^{u_j} \\ &\leq 2^{k-l+j} \cdot \exp(-2^{-l+j-1} \cdot u_j) \leq 2^{k-l+j} \cdot \exp(-2^{-l+j-1} \cdot (k-2l)/l).\end{aligned}$$

Taking logarithms, we get

$$\begin{aligned}\log |F_{j,0}| &\leq k - l + j - \log e \cdot 2^{-l+j-1} \cdot (k-2l)/l \\ &\leq k - l + j - 2^{j-1} \leq k - l.\end{aligned}$$

Therefore, F_j satisfies the induction hypothesis. For $j = l$ this implies that F_l is an unsatisfiable (k, s)-CNF formula for $s = 2^{k-l+1}$, as long as

$$l \cdot 2^l \leq \log e \cdot (k - 2l). \qquad \square \tag{8}$$

Let l be the largest integer satisfying $2^l \leq \log e \cdot k/(2 \log k)$. Then (8) holds for $k \geq 2$, and we get the following corollary.

COROLLARY 3.2. *$f(k) < 2^k \cdot 8 \ln k/k$ for $k \geq 2$.*

4. Further generalization and experimental results. One way to derive better upper bounds on $f(k)$ is to generalize the constructions of sections 2 and 3. To this end, we first define a special way to compose CNF formulas capturing the essence of these constructions.

DEFINITION 4.1. *Let G_1, G_2 be unsatisfiable CNF formulas that have clauses of size at most k such that G'_i is a k_i-*CNF *formula for $i = 1, 2$. Also, assume that $k_1 \leq k_2 < k$. Then the formula $G_1 \circ G_2$ is defined as*

$$\left(\bigcup_{c \in \mathcal{K}^-(x_1, \ldots, x_{k-k_2})} G'_{1,c} \times c \cup G''_{1,c} \right) \cup G'_2 \times \{\{x_1, \ldots, x_{k-k_2}\}\} \cup G''_2,$$

where the formulas $G_{1,c}$ are copies of G_1 on distinct sets of variables. We say that $G_1 \circ G_2$ is obtained by applying $\circ G_2$ to G_1, and we let $G_1 \circ_q G_2$ denote the formula obtained by applying $\circ G_2$ to G_1 q times.

It is not difficult to verify the following.

LEMMA 4.2. *Let G_1, G_2 be formulas as above, where the number of occurrences of each variable is bounded by some number s satisfying $s \geq (2^{k-k_2} - 1) \cdot |G'_1| + |G'_2|$. Then $G = G_1 \circ G_2$ is an unsatisfiable CNF formula where each variable occurs at most s times. Furthermore, G' is a $(k_1 + k - k_2)$-*CNF *formula, and $|G'| = (2^{k-k_2} - 1) \cdot |G'_1|$.*

Given k, s, we ask whether one can obtain a k-CNF formula using the following derivation rules. We start with the unsatisfiable formula $\{\emptyset\}$ as an axiom (this formula consists of one empty clause). For a set of derivable formulas, one can apply one of the following rules:

1. If G is a derived formula such that $s \geq 2 \cdot |G'|$, then we can derive $G'_x \times \{\{x\}\} \cup G'_{\overline{x}} \times \{\{\overline{x}\}\} \cup G''_x \cup G''_{\overline{x}}$, where x is a new variable and G_x, $G_{\overline{x}}$ are two disjoint copies of G.
2. If G_1, G_2 are two derived formulas satisfying the conditions of Lemma 4.2, then we can derive the formula $G_1 \circ G_2$.

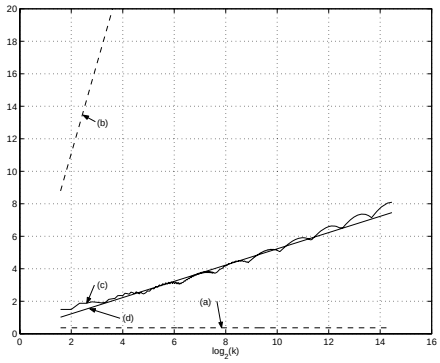


FIG. 1. *The bounds on* $f(k) \cdot k/2^k$. (a) *Lower bound of Kratochvíl, Savický, and Tuza* [4], $1/e$. (b) *Upper bound* (3) *obtained in section* 3 *of the present paper,* $8 \ln k$. (c) *Upper bound* $f_2(k) \cdot k/2^k$, *calculated by a computer program.* (d) *The line* $0.5 \log(k) + 0.23$.

One can sometimes replace $G_1 \circ G_2$ in the second rule by a more compact formula $G_1 \circ' G_2$ that avoids duplicating G_1—namely, the formula $G_1' \times \mathcal{K}^-(x_1, \ldots, x_{k-k_2}) \cup G_1'' \cup G_2' \times \{\{x_1, \ldots, x_{k-k_2}\}\} \cup G_2''$. Although this can never reduce the number of occurrences of variables, this modification reduces the number of clauses and variables. The constructions presented in sections 2 and 3 are special cases of the above derivation rule. Indeed, $\mathcal{K}(x_1, \ldots, x_v)$ can be obtained by applying the first rule v times to $\{\emptyset\}$. The formula of section 2 is just

$$F = \mathcal{K}(z_1, \ldots, z_v) \circ_u' \mathcal{K}(y_1, \ldots, y_{k-l}).$$

The formula of section 3 is inductively obtained by

$$F_0 = \mathcal{K}(z_1, \ldots, z_{k-l}),$$
$$F_j = \mathcal{K}(z_1, \ldots, z_{d_j - u_j \cdot (l-j+1)}) \circ_{u_j}' F_{j-1} \qquad \text{for } j = 1, \ldots, l.$$

Since any k-CNF formula obtained using the above procedure is an unsatisfiable (k, s)-CNF, one can define $f_2(k)$ as the maximal value of s such that no k-CNF formula can be obtained using the above procedure (clearly $f(k) \le f_2(k)$). It turns out that the function $f_2(k)$ is appealing from an algorithmic point of view. Given a value for s, one can check if $f_2(k)$ is larger than s using a simple dynamic programming algorithm. The algorithm keeps an array $a_0, \ldots, a_k$, where eventually a_l contains the minimal size of F' for a derivable formula F such that F' is an l-CNF formula.

Initialize $a_0 = 1$, $a_1 = \cdots = a_k = \infty$
Repeat until no more changes are made to $a_1, \ldots, a_k$

For $l = 0, \ldots, k-1$
If $s \ge 2l$ then $a_{l+1} \leftarrow \min(\, 2a_l,\, a_{l+1}\,)$

For $k_2 = 0, \ldots, k-1$
For $k_1 = 0, \ldots, k_2$
If $s \ge (2^{k-k_2} - 1) \cdot a_{k_1} + a_{k_2}$ then $a_{k_1+k-k_2} \leftarrow \min(\, (2^{k-k_2} - 1) \cdot a_{k_1},\, a_{k_1+k-k_2}\,)$

If $a_k < \infty$ then output "$f_2(k) \le s$" else output "$f_2(k) > s$"

This algorithm works well in practice, and we were able to calculate $f_2(k)$ for values up to $k = 20000$ to get the results depicted by the graph in Figure 1.

The computed numerical values of $f_2(k)$ seem to indicate that

$$f_2(k) \cdot k/2^k = 0.5 \log(k) + o(\log(k)), \tag{9}$$

which is better than our upper bound by a constant factor of about 11. If (9) indeed holds, then a better analysis of the function f_2 may improve our upper bound by a constant factor. However, such an approach cannot improve upon the logarithmic gap left between the known upper and lower bounds on $f(k)$.

REFERENCES

[1] P. Berman, M. Karpinski, and A. D. Scott, *Approximation Hardness and Satisfiability of Bounded Occurrence Instances of SAT*, Technical report TR03-022, Electronic Colloquium on Computational Complexity (ECCC), University of Trier, Trier, Germany, 2003.

[2] O. Dubois, *On the r, s-SAT satisfiability problem and a conjecture of Tovey*, Discrete Appl. Math., 26 (1990), pp. 51–60.

[3] S. Hoory and S. Szeider, *Computing unsatisfiable k-SAT instances with few occurrences per variable*, Theoret. Comput. Sci., 337 (2005), pp. 347–359.

[4] J. Kratochvíl, P. Savický, and Z. Tuza, *One more occurrence of variables make satisfiability jump from trivial to NP-complete*, Acta Inform., 30 (1993), pp. 397–403.

[5] P. Savický and J. Sgall, *DNF tautologies with a limited number of occurrences of every variable*, Theoret. Comput. Sci., 238 (2000), pp. 495–498.

[6] J. Stříbrná, *Between Combinatorics and Formal Logic*, Master's thesis, Department of Applied Mathematics, Charles University, Prague, Czech Republic, 1994.

[7] C. A. Tovey, *A simplified NP-complete satisfiability problem*, Discrete Appl. Math., 8 (1984), pp. 85–89.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 529–535

ON GRAPH ASSOCIATIONS*

LANDON RABERN†

Abstract. We introduce a notion of vertex association and consider sequences of these associations. This allows for slick proofs of a few known theorems as well as showing that for any induced subgraph H of G, $\chi(G) \leq \chi(H) + \frac{1}{2}(\omega(G) + |G| - |H| - 1)$. As a special case of this, we have $\chi(G) \leq \lceil \frac{\omega(G)+\tau(G)}{2} \rceil$ (here $\chi(G)$ denotes the chromatic number, $\omega(G)$ the clique number, and $\tau(G)$ the vertex cover number), which is a generalization of the Nordhaus–Gaddum upper bound. In addition, this settles a conjecture of Reed that $\chi(G) \leq \lceil \frac{\omega(G)+\Delta(G)+1}{2} \rceil$ in the case when $\delta(\overline{G}) \leq \omega(\overline{G})$.

Key words. graph coloring, Reed's conjecture

AMS subject classifications. 05C15, 05C69

DOI. 10.1137/050626545

1. Definitions and basic properties.

All graphs will be assumed finite and simple. We let $|G|$ denote the order of G, $s(G)$ the size of G, $\chi(G)$ the chromatic number, $\omega(G)$ the clique number, $\tau(G)$ the vertex cover number, $\Delta(G)$ the maximum degree, $\delta(G)$ the minimum degree, $d_G(x)$ the degree of x in G, and $N_G(x)$ the set of neighbors of x in G.

DEFINITION 1.1. *Given a graph G and nonadjacent vertices a and b, we write $G/[a,b]$ for the graph obtained from G by associating (i.e., identifying) a and b into a single vertex $[a,b]$ and discarding multiple edges.*

PROPOSITION 1.2. *Let G be a graph and $a, b, x \in V(G)$ with $a \notin N_G(b)$. Then*

$$d_{G/[a,b]}(x) = \begin{cases} d_G(x) - 1 & \text{if } x \in N_G(a) \cap N_G(b), \\ d_G(a) + d_G(b) - |N_G(a) \cap N_G(b)| & \text{if } x \in \{a, b\}, \\ d_G(x) & \text{otherwise.} \end{cases}$$

Proof. The proof is immediate from the definitions. □

The content of the following proposition is that the operations of vertex removal and association commute.

PROPOSITION 1.3. *Let G be a graph. If $a, b \in V(G)$ with $a \notin N_G(b)$ and $S \subseteq V(G) \setminus \{a, b\}$, then*

$$(G \setminus S)/[a,b] = G/[a,b] \setminus S.$$

Proof. Again, this is immediate from the definitions. □

LEMMA 1.4. *Let a and b be nonadjacent vertices in a graph G. Then*

(i) $\chi(G) \leq \chi(G/[a,b]) \leq \chi(G) + 1$, *and*
(ii) $\chi(G/[a,b]) = \chi(G)$ *if and only if there exists a coloring of G with $\chi(G)$ colors in which a and b receive the same color.*

Proof. (i) Since a and b are nonadjacent, any k-coloring of $G/[a,b]$ lifts to a k-coloring of G. This gives the first inequality. The second follows by noting that any k-coloring of G induces a k-coloring of $G/[a,b] \setminus \{[a,b]\}$ and hence a $(k+1)$-coloring of $G/[a,b]$ by introducing a new color.

*Received by the editors March 10, 2005; accepted for publication (in revised form) January 25, 2006; published electronically June 21, 2006.
http://www.siam.org/journals/sidma/20-2/62654.html
†Department of Mathematics, UC Santa Barbara, Goleta, CA 93117 (landonr@math.ucsb.edu).

(ii) Assume $\chi(G/[a,b]) = \chi(G)$. Then we have a $\chi(G)$-coloring of $G/[a,b]$ and lifting this to G gives a $\chi(G)$-coloring of G in which a and b receive the same color.

For the converse, assume we have a $\chi(G)$-coloring of G in which a and b receive the same color. Then the induced $\chi(G)$-coloring of $G/[a,b] \setminus \{[a,b]\}$ extends to a $\chi(G)$-coloring of $G/[a,b]$ by coloring $[a,b]$ the color that a and b share. □

PROPOSITION 1.5. *Let a and b be nonadjacent vertices in a graph G. Then*

$$\chi(G) = \min\{\chi(G/[a,b]), \chi(G+ab)\}.$$

Proof. If $\chi(G) = \chi(G/[a,b])$, then we are done since $\chi(G+ab) \geq \chi(G)$. Otherwise, by Lemma 1.4(ii), a and b must receive different colors in every $\chi(G)$-coloring of G. Hence, any $\chi(G)$-coloring of G extends to a $\chi(G)$-coloring of $G+ab$. Thus $\chi(G) = \chi(G+ab)$, completing the proof. □

2. Sequences of associations. We consider sequences of the form

$$G = H_0 \to H_1 \to \cdots \to H_r = K_t,$$

where each term is obtained from the previous one by associating two nonadjacent vertices. The process clearly terminates at some complete graph K_t.

LEMMA 2.1. *Let G be a graph. If G is not complete, then there exist nonadjacent vertices a and b which receive the same color in some $\chi(G)$-coloring of G.*

Proof. If not, then any given vertex must be colored differently from every other vertex in any $\chi(G)$-coloring of G. Hence, $\chi(G) = |G|$ and thus G is complete. □

PROPOSITION 2.2. *The smallest t for which there exists a sequence*

$$G = H_0 \to H_1 \to \cdots \to H_r = K_t$$

is $t = \chi(G)$.

Proof. The first inequality of Lemma 1.4(i) and the fact that $\chi(K_t) = t$ yield $t \geq \chi(G)$. We just need to show that $K_{\chi(G)}$ can be attained. If G is complete, then we are done. Otherwise, by Lemma 2.1, we have two vertices a and b which receive the same color in some $\chi(G)$-coloring of G. By Lemma 1.4(ii), $\chi(G/[a,b]) = \chi(G)$. Since $|G/[a,b]| < |G|$, the result follows by induction. □

DEFINITION 2.3. *We denote by $\psi(G)$ the largest t for which there exists a sequence*

$$G = H_0 \to H_1 \to \cdots \to H_r = K_t.$$

With a little thought, one can see that this is the same thing as the achromatic number of G.

Loose upper bounds on $\psi(G)$ can be easily obtained.

PROPOSITION 2.4. *Let G be a graph. Then*

(i) $\psi(G) \leq |G|$, *and*

(ii) $\psi(G) \leq \frac{1+\sqrt{1+8s(G)}}{2}$.

Proof. Consider the sequence

$$G = H_0 \to H_1 \to \cdots \to H_r = K_{\psi(G)}.$$

As we move from left to right, both the order and the size of the graphs do not increase; hence, $|G| \geq \psi(G)$ and $s(G) \geq \binom{\psi(G)}{2}$. The results follow. □

3. Some slick proofs.

LEMMA 3.1. *If a and b are nonadjacent vertices in a graph G, then*

$$\chi(\overline{G}) - 1 \leq \chi(\overline{G/[a,b]}) \leq \chi(\overline{G}).$$

Proof. Note that the chromatic number of $\overline{G}$ is the clique cover number of G. Assume we have a partition of $V(G)$ into n disjoint sets $\{K_1, \ldots, K_n\}$, each of which induces a clique. Since a and b are nonadjacent, they are in distinct cliques, say, $a \in K_i$, $b \in K_j$ with $i \neq j$. We see that replacing K_i with $K_i \setminus \{a\}$ and K_j with $(K_j \setminus \{b\}) \cup \{[a,b]\}$ yields a covering of $G/[a,b]$ with n cliques. This gives the second inequality. To get the first, assume we have a partition of $V(G/[a,b])$ into n disjoint sets $\{K_1, \ldots, K_n\}$, each of which induces a clique. Then $[a,b]$ is in one of the sets, say, $[a,b] \in K_i$. Let $K_i' = ((K_i \setminus \{[a,b]\}) \cap N_G(a)) \cup \{a\}$ and $K_{n+1}' = ((K_i \setminus \{[a,b]\}) \setminus K_i') \cup \{b\}$. Then $\{K_1, \ldots, K_{i-1}, K_i', K_{i+1}, \ldots, K_n, K_{n+1}'\}$ is a partition of $V(G)$ into $n+1$ disjoint sets, each of which induces a clique. □

PROPOSITION 3.2 (see Harary and Hedetniemi [2]). *Let G be a graph. Then*

$$\psi(G) + \chi(\overline{G}) \leq |G| + 1.$$

Proof. Consider the sequence

$$G = H_0 \to H_1 \to \cdots \to H_r = K_{\psi(G)}, \tag{1}$$

where $r = |G| - \psi(G)$. It follows from the first inequality of Lemma 3.1 that

$$\chi(\overline{G}) - (|G| - \psi(G)) = \chi(\overline{G}) - r \leq \chi(\overline{K_{\psi(G)}}) = 1,$$

so that $\psi(G) + \chi(\overline{G}) \leq |G| + 1$ as required. □

COROLLARY 3.3 (see Nordhaus and Gaddum [3]). *Let G be a graph. Then*

$$\chi(G) + \chi(\overline{G}) \leq |G| + 1.$$

Proof. Use $\chi(G) \leq \psi(G)$ in Proposition 3.2. □

LEMMA 3.4. *Let G be a graph. Then*

$$\chi(G) \geq 2\psi(G) - |G|.$$

Proof. It follows from (1) and the second inequality of Lemma 1.4(i) that

$$\psi(G) = \chi(K_{\psi(G)}) \leq \chi(G) + r = \chi(G) + |G| - \psi(G).$$

The result follows. □

PROPOSITION 3.5. *Let G be a graph. Then*

$$2\psi(G) + \psi(\overline{G}) \leq 2|G| + 1.$$

Proof. Lemma 3.4 applied to $\overline{G}$ yields $\chi(\overline{G}) \geq 2\psi(\overline{G}) - |G|$. Substituting this into Proposition 3.2 gives $2\psi(\overline{G}) + \psi(G) \leq 2|G| + 1$. Now substituting $\overline{G}$ for G gives the result. □

COROLLARY 3.6 (see Gupta [1]). *Let G be a graph. Then*

$$\psi(G) + \psi(\overline{G}) \leq \left\lceil \frac{4}{3}|G| \right\rceil.$$

Proof. Applying Proposition 3.5 to G and $\overline{G}$ yields the inequalities

$$2\psi(G) + \psi(\overline{G}) \leq 2|G| + 1$$

and

$$\psi(G) + 2\psi(\overline{G}) \leq 2|G| + 1,$$

respectively. By adding these, we get

$$3(\psi(G) + \psi(\overline{G})) \leq 4|G| + 2,$$

which is

$$\psi(G) + \psi(\overline{G}) \leq \frac{4}{3}|G| + \frac{2}{3}.$$

The result follows. ∎

4. The main results.

DEFINITION 4.1. *Let G be a graph and I an independent set in G. We denote by $G/[I]$ the graph obtained from G by associating I down to a single vertex $[I]$.*

LEMMA 4.2. *Let f be a real-valued graph function such that, for any graph G, $f(G \setminus \{v\}) \geq f(G) - 1$ for all $v \in V(G)$. Then, for any graph G and independent set I in G,*

$$f(G/[I]) \leq f(G \setminus I) + 1.$$

Proof. Observe that $G \setminus I = G/[I] \setminus \{[I]\}$. But $[I]$ is a single vertex; hence, $f(G \setminus I) = f(G/[I] \setminus \{[I]\}) \geq f(G/[I]) - 1$. The result follows. ∎

DEFINITION 4.3. *We say that a graph G consists of* an independent set attached to a clique *if $V(G)$ can be partitioned into two disjoint sets I and K such that I is independent and K induces a clique. We say that G consists of* an independent set strongly attached to a clique *if there is such a partition in which each vertex of K is adjacent to at least one vertex of I.*

LEMMA 4.4.

(a) *If a graph G consists of an independent set I attached to a clique K, then $\overline{G}$ consists of an independent set $\overline{K}$ attached to a clique $\overline{I}$, and $\chi(G) = \omega(G) = |K|$ or $|K| + 1$ and $\chi(\overline{G}) = \omega(\overline{G}) = \alpha(G) = |I|$ or $|I| + 1$.*

(b) *If G consists of an independent set I strongly attached to a clique K, then $\chi(\overline{G}) = \omega(\overline{G}) = \alpha(G) = |I|$.*

(c) *If I is an independent set in a graph G, then $G/[I]$ is complete if and only if G consists of I strongly attached to a clique.*

Proof. (a) Since I is independent, $\chi(G) \leq |K| + 1$ and $\chi(G) = |K| + 1$ if and only if there exists $v \in I$ such that $N_G(v) = K$; in this case, $\omega(G) = |K| + 1$ as well. The statements about $\overline{G}$ follow in a similar manner.

(b) Assume each vertex of K is adjacent in G to at least one vertex of I. Then, in $\overline{G}$, each vertex of $\overline{K}$ is nonadjacent to at least one vertex of $\overline{I}$. Hence $\omega(\overline{G}) = |I|$. The other equalities follow from (a).

(c) We have $G/[I]$ complete if and only if $N_{G/[I]}([I]) = K$. This happens if and only if each vertex of K is adjacent to at least one vertex of I. ∎

LEMMA 4.5. *Let*

$$G = H_0 \to H_1 \to \cdots \to H_{r-1} \to H_r = K_t$$

be a sequence where each term is obtained from the previous one by associating two nonadjacent vertices. If $\chi(H_{r-1}) = \chi(H_r)$*, then* $\omega(H_{r-1}) = \omega(H_r)$*.*

Proof. Since H_r is complete, H_{r-1} is an independent set of size 2 strongly attached to a clique; hence, by Lemma 4.4(a), $\omega(H_{r-1}) = \chi(H_{r-1}) = \chi(H_r) = \omega(H_r)$. □

THEOREM 4.6. *Let* $I_1, \ldots, I_m$ *be disjoint independent sets in a graph* G*. Then*

$$\chi(G) \le \frac{1}{2}\left(\omega(G) + |G| - \sum_{j=1}^{m} |I_j| + 2m - 1\right). \tag{2}$$

Proof. Associate I_1 through I_m in turn to yield a sequence

$$G = H_0 \to H_1 \to \cdots \to H_{m-1} \to H_m = B, \tag{3}$$

and let $A = H_{m-1}$, so that B is obtained from A by associating I_m to a single vertex. We distinguish two cases.

Case 1. B is complete, so that $B = K_{\chi(B)}$. Then, by Lemma 4.4(c), A consists of I_m strongly attached to a clique. By Corollary 3.3 and Lemma 4.4(b),

$$\chi(A) \le |A| - \chi(\overline{A}) + 1 = |A| - |I_m| + 1,$$

so that, since $\chi(A) = \omega(A)$ by Lemma 4.4(a),

$$2\chi(A) \le \omega(A) + |A| - |I_m| + 1. \tag{4}$$

Since $\omega(G \setminus \{v\}) \ge \omega(G) - 1$ for all $v \in V(G)$, Lemma 4.2 tells us that associating an independent set to a single point increases ω by at most one. Hence

$$\omega(A) \le \omega(G) + m - 1. \tag{5}$$

Also, $|G| - |A| = \sum_{j=1}^{m-1}(|I_j| - 1) = \sum_{j=1}^{m} |I_j| - |I_m| - m + 1$, so that

$$|A| - |I_m| = |G| - \sum_{j=1}^{m} |I_j| + m - 1. \tag{6}$$

Since $\chi(G) \le \chi(A)$ by the first inequality of Lemma 1.4(i), substituting (5) and (6) into (4) gives

$$\begin{aligned} 2\chi(G) \le 2\chi(A) &\le \omega(G) + m - 1 + |G| - \sum_{j=1}^{m} |I_j| + m - 1 + 1 \\ &= \omega(G) + |G| - \sum_{j=1}^{m} |I_j| + 2m - 1, \end{aligned}$$

which is (2).

Case 2. B is not complete. Consider the sequence

$$B \to \cdots \to C \to K_{\chi(B)}, \tag{7}$$

where each term is obtained from the previous one by associating two nonadjacent vertices. Then, by the first inequality in Lemma 1.4(i),

$$\chi(B) \le \chi(C) \le \chi(K_{\chi(B)}) = \chi(B).$$

Hence $\chi(C) = \chi(B) = \chi(K_{\chi(B)})$, and we may apply Lemma 4.5 to conclude

$$\omega(C) = \omega(K_{\chi(B)}) = \chi(B). \tag{8}$$

In addition, it is clear that

$$|C| = \chi(B) + 1. \tag{9}$$

Applying Lemma 4.2 as in (5), but this time to a combination of sequences (3) and (7) between G and C, gives

$$\omega(C) \leq \omega(G) + m + |B| - |C|, \tag{10}$$

and $|G| - |B| = \sum_{j=1}^{m} |I_j| - m$, so that, by (8), (9), and (10),

$$\begin{aligned} 2\chi(B) = \omega(C) + |C| - 1 &\leq \omega(G) + m + |B| - 1 \\ &= \omega(G) + m + |G| - \sum_{j=1}^{m} |I_j| + m - 1. \end{aligned}$$

Since $\chi(G) \leq \chi(B)$, by the first inequality of Lemma 1.4(i), the theorem then follows. □

Since the vertex-set of an induced subgraph H of G can be partitioned into $\chi(H)$ independent sets, the following is an equivalent formulation of Theorem 4.6.

THEOREM 4.7. *Let G be a graph. Then, for any induced subgraph H of G,*

$$\chi(G) \leq \chi(H) + \frac{1}{2}(\omega(G) + |G| - |H| - 1).$$

COROLLARY 4.8. *Let G be a graph. Then*

$$\chi(G) \leq \left\lceil \frac{\omega(G) + \tau(G)}{2} \right\rceil.$$

Proof. Apply Theorem 4.6 to a single independent set with $\omega(\overline{G})$ elements to get

$$\chi(G) \leq \frac{1}{2}(\omega(G) + |G| - \omega(\overline{G}) + 1). \tag{11}$$

Since $S \subseteq V(G)$ is a vertex cover if and only if $V(G) \setminus S$ is an independent set,

$$\tau(G) + \omega(\overline{G}) = |G|.$$

The result follows. □

Note that this is a generalization of the Nordhaus–Gaddum upper bound since replacing G by $\overline{G}$ in (11) and adding the two inequalities yields $\chi(G)+\chi(\overline{G}) \leq |G|+1$.

CONJECTURE 4.9 (see Reed [4]). *Let G be a graph. Then*

$$\chi(G) \leq \left\lceil \frac{\omega(G) + \Delta(G) + 1}{2} \right\rceil.$$

Corollary 4.8 establishes this for all graphs G with $\tau(G) \leq \Delta(G) + 1$ and, equivalently, for all graphs with $\delta(\overline{G}) \leq \omega(\overline{G})$. In particular, if $\delta(\overline{G}) \leq 2$, then either

$\delta(\overline{G}) \leq 2 \leq \omega(\overline{G})$ or $\omega(\overline{G}) = 1$, and hence G is complete. Thus Reed's conjecture holds for any graph G with $\Delta(G) \geq |G| - 3$.

COROLLARY 4.10. *Let G be a triangle-free graph. Then*

$$\chi(G) \leq 2 + \frac{1}{2}\delta(\overline{G}).$$

Proof. Since G is triangle-free, $\omega(\overline{G}) \geq \Delta(G)$. It follows from (11) that

$$\chi(G) \leq \frac{1}{2}(\omega(G) + |G| - \Delta(G) + 1) = \frac{1}{2}(\omega(G) + \delta(\overline{G}) + 2) \leq \frac{1}{2}(4 + \delta(\overline{G})),$$

which is the required result. □

REFERENCES

[1] R. P. GUPTA, *Bounds on the chromatic and achromatic numbers of complementary graphs*, in Recent Progress in Combinatorics, Proceedings of the 3rd Waterloo Conference in Combinatorics, Waterloo, 1968, W. T. Tutte, ed., Academic Press, New York, London, 1969, pp. 229–235.

[2] F. HARARY AND S. HEDETNIEMI, *The achromatic number of a graph*, J. Combin. Theory, 8 (1970), pp. 154–161.

[3] E. A. NORDHAUS AND J. W. GADDUM, *On complementary graphs*, Amer. Math. Monthly, 63 (1956), pp. 175–177.

[4] B. REED, *ω, Δ, and χ*, J. Graph Theory, 27 (1997), pp. 177–212.

SIAM J. DISCRETE MATH.
Vol. 20, No. 2, pp. 536–543

CONSTRUCTION OF LARGE GRAPHS WITH NO OPTIMAL SURJECTIVE $L(2,1)$-LABELINGS*

DANIEL KRÁL'†, RISTE ŠKREKOVSKI‡, AND MARTIN TANCER§

Abstract. An $L(2,1)$-labeling of a graph G is a mapping $c : V(G) \to \{0,\dots,K\}$ such that the labels of two adjacent vertices differ by at least two and the labels of vertices at distance two differ by at least one. A hole of c is an integer $h \in \{0,\dots,K\}$ that is not used as a label for any vertex of G. The smallest integer K for which an $L(2,1)$-labeling of G exists is denoted by $\lambda(G)$. The minimum number of holes in an optimal labeling, i.e., a labeling with $K = \lambda(G)$, is denoted by $\rho(G)$. Georges and Mauro [*SIAM J. Discrete Math.*, 19 (2005), pp. 208–223] showed that $\rho(G) \le \Delta$, where Δ is the maximum degree of G, and conjectured that if $\rho(G) = \Delta$ and G is connected, then the order of G is at most $\Delta(\Delta+1)$. We disprove this conjecture by constructing graphs G with $\rho(G) = \Delta$ and order $\lfloor \frac{(\Delta+1)^2}{4} \rfloor (\Delta+1) \approx \Delta^3/4$.

Key words. channel assignment problem, graph labeling with distance conditions

AMS subject classification. 05C15

DOI. 10.1137/050623061

1. Introduction.

$L(2,1)$-labelings of graphs form an important model for the frequency assignment problem [9]. An $L(2,1)$-*labeling* of a graph G is a labeling $c : V(G) \to \{0,\dots,K\}$ of the vertices of G such that the labels of any two adjacent vertices differ by at least two and the labels of any two vertices at distance two are different. The smallest K for which there exists a proper labeling of G is denoted by $\lambda(G)$.

One of the most studied problems on $L(2,1)$-labelings is the famous Delta Square Conjecture of Griggs and Yeh [8]: They conjectured that $\lambda(G) \le \Delta(G)^2$ for every graph G with maximum degree $\Delta(G) \ge 2$. Though the conjecture has been verified for several special classes of graphs, including graphs of maximum degree two, chordal graphs [15] (see also [1, 12]), and Hamiltonian cubic graphs [10, 11], it remains widely open. The original upper bound, $\lambda(G) \le \Delta(G)^2 + 2\Delta(G)$ by Griggs and Yeh [8], has been improved to $\lambda(G) \le \Delta(G)^2 + \Delta(G)$ in [2] (an analogous bound in a more general setting of the channel assignment problem was proved by McDiarmid [14]). A more general result of the first two authors [13] yields $\lambda(G) \le \Delta(G)^2 + \Delta(G) - 1$, and the present record $\lambda(G) \le \Delta(G)^2 + \Delta(G) - 2$ has been recently proven by Gonçalves [7].

*Received by the editors January 21, 2005; accepted for publication (in revised form) January 30, 2006; published electronically June 21, 2006. This research was conducted as a part of the Czech-Slovenian bilateral project MŠMT-07-0405 (Czech side) and SLO-CZ/04-05-002 (Slovenian side).
http://www.siam.org/journals/sidma/20-2/62306.html

†Institute for Mathematics, Technical University Berlin, Strasse des 17. Juni 136, D-10623 Berlin, Germany. The author was a postdoctoral fellow at TU Berlin within the framework of the European training network COMBSTRU from October 2004 to July 2005. 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 (kral@kam.mff.cuni.cz). Institute for Theoretical Computer Science is supported by Ministry of Education of Czech Republic as projects LN00A056 and 1M0545. At the present, the author is a Fulbright scholar at School of Mathematics, Georgia Institute of Technology, 686 Cherry St., Atlanta, GA 30332 (kral@math.gatech.edu).

‡Department of Mathematics, University of Ljubljana, Jadranska 19, 1111 Ljubljana, Slovenia (riste.skrekovski@fmf.uni-lj.si).

§Department of Applied Mathematics, Faculty of Mathematics and Physics, Charles University, Malostranské náměstí 25, 118 00 Prague, Czech Republic (martin@atrey.karlin.mff.cuni.cz). The work of this author was partially supported by Institute for Theoretical Computer Science (ITI).

In this paper, we focus on surjective $L(2,1)$-labelings that were first studied by Fishburn and Roberts [3] under the name of *full colorings*, and further investigated in [4, 5, 6]. If c is an $L(2,1)$-labeling of G, then a number h, $0 \leq h \leq K$, is a *hole* if there is no vertex v of G with $c(v) = h$. The minimum number of holes in an $L(2,1)$-labeling of G with $K = \lambda(G)$ is denoted by $\rho(G)$. Georges and Mauro [5] established that $\rho(G)$ never exceeds the maximum degree $\Delta(G)$ of G. In [5], Georges and Mauro also posed (among others) the following conjectures on $L(2,1)$-labelings and holes.

CONJECTURE 1. *If G is an r-regular graph and $\rho(G) \geq 1$, then $\rho(G)$ divides r.*

CONJECTURE 2. *If G is a connected graph with maximum degree $\Delta(G)$ and $\rho(G) = \Delta(G)$, then the order of G does not exceed $\Delta(\Delta+1)$.*

CONJECTURE 3. *If G is a graph with $\lambda(G) > 2\Delta(G)$, then $\rho(G) = 0$. In other words, if G is a graph with $\rho(G) > 0$, then $\lambda(G) \leq 2\Delta(G)$.*

In this paper, we focus on Conjecture 2. We provide a construction of connected r-regular graphs G of order $(r+1)\lfloor\frac{(r+1)^2}{4}\rfloor \approx r^3/4$ with $\rho(G) = r$ (Corollary 3.4). This shows that Conjecture 2 does not hold for $\Delta \geq 3$. Note that Conjecture 2 trivially holds for $\Delta = 1$ since the only graph G satisfying the assumptions of the conjecture is K_2. In [5], it was shown that Conjecture 2 also holds for $\Delta = 2$.

2. Previous results. In this section, we survey results obtained by Georges and Mauro [5] on the structure of graphs G with $\rho(G) = \Delta(G)$. The following theorem shows that the structure of such graphs is very restricted.

THEOREM 2.1. *If G is a graph with $\rho(G) = \Delta(G)$, then G is a Δ-regular graph with $\lambda(G) = 2\Delta$. Moreover, for every optimum $L(2,1)$-labeling c, i.e., a labeling using the labels $0, \ldots, \lambda(G)$, the following hold:*

- *every odd integer between 0 and $\lambda(G)$ is a hole of c;*
- *the cardinality of the preimage in c of every even number between 0 and $\lambda(G)$ is the same; and*
- *the subgraph of G induced by the preimages of any two even numbers is a perfect matching (union of disjoint edges).*

In particular, there exists an integer $t > 0$ such that the order of G is $(\Delta+1)t$.

In [5], Georges and Mauro constructed connected Δ-regular graphs G with $\rho(G) = \Delta$ of order $(\Delta+1)t$ for every $t = 1, \ldots, \Delta$. They conjectured that the number t (under the assumption that G is connected) cannot exceed Δ (this is equivalent to Conjecture 2 stated in section 1).

We now recall a construction of an r-regular graph Ω_r of order $r(r+1)$ from [5]. Consider a union of r vertex disjoint cliques of order r and number the vertices of each clique from 1 to r. Add to the graph r new vertices and join the i-th of them to the vertices of the cliques numbered with i. The resulting graph is Ω_r. Clearly, Ω_r is a connected r-regular graph. It can be shown that $\lambda(\Omega_r) = 2r$ and $\rho(\Omega_r) = r$.

In order to show that $\lambda(\Omega_r) = 2r$ and $\rho(\Omega_r) = r$ for the graph Ω_r, Georges and Mauro [5] showed that Ω_r has a special property which we call the neighborhood property in this paper. Assume that G is a connected r-regular graph of order $(r+1)t$. We say that G has the *t-neighborhood property* if the following holds for any two (disjoint) sets V and W of vertices of G: If neither V nor W contains two vertices at distance at most two and no vertex of V is adjacent to a vertex of W, then $|V| + |W| \leq t$.

We finish this section with the following proposition whose proof is implicitly contained in [5]. We include its short proof for the sake of completeness.

PROPOSITION 2.2. *If G is a connected r-regular graph of order $(r+1)t$ with $\lambda(G) \leq 2r$ that has the t-neighborhood property, then $\lambda(G) = 2r$ and $\rho(G) = r$.*

Proof. Let us consider an $L(2,1)$-labeling of G with $0,\ldots,2r$ and let V_i, $i = 0,\ldots,2r$, be the set of the vertices labeled with i. Since G has the t-neighborhood property, it holds that

$$|V_i| + |V_{i+1}| \leq t \tag{2.1}$$

for every $i = 0,\ldots,2r-1$.

First, we show that $V_i = \emptyset$ for all odd i's. Let i_0 be an odd integer between 0 and $2r$ and let $\mu = |V_{i_0}|$. We now bound the sum $|V_0| + \cdots + |V_{2r}|$ using (2.1):

$$\sum_{i=0}^{2r} |V_i| = \sum_{i=0}^{(i_0-1)/2} (|V_{2i}| + |V_{2i+1}|) + \sum_{i=(i_0+1)/2}^{r} (|V_{2i-1}| + |V_{2i}|) - |V_{i_0}| \leq (r+1)t - \mu\,.$$

Since the sets $V_0,\ldots,V_{2r}$ partition the vertex set of G, the sum of their sizes is $(r+1)t$. Therefore, $\mu = 0$. Since the choice of i_0 was arbitrary, $V_i = \emptyset$ for all odd i's as claimed.

Note that $|V_i| \leq t$ for every $i = 0,\ldots,2r$ by (2.1). Since the sum $|V_0| + \cdots + |V_{2r}|$ is equal to $(r+1)t$ and the set V_i is empty for every odd i, it must hold that $|V_i| = t$ for every $i = 0,2,4,\ldots,2r$. The statement of the proposition now readily follows. ∎

3. Construction. In this section, we present our construction of graphs of order $\Theta(\Delta^3)$ with $\rho = \Delta$ (the exact parameters of the constructed graphs can be found in Theorem 3.3). First, we describe the considered graphs in subsection 3.1. In subsection 3.2, we analyze their properties. Finally, we slightly generalize our construction to obtain additional graphs with similar properties in subsection 3.3.

3.1. The graph. In this subsection, we construct an $(\alpha+\beta-1)$-regular connected graph $\Gamma_{\alpha,\beta}$ of order $(\alpha+\beta)\alpha\beta$ with $\rho(\Gamma_{\alpha,\beta}) = \Delta(\Gamma_{\alpha,\beta}) = \alpha+\beta-1$. The vertex set of $\Gamma_{\alpha,\beta}$ is composed of two sets V_g and V_r:

$$V_g = \{[a,b,\overline{a}] \mid 1 \leq a \leq \alpha, 1 \leq b \leq \beta \text{ and } 1 \leq \overline{a} \leq \alpha\} \text{ and}$$
$$V_r = \{[a,b,\overline{b}] \mid 1 \leq a \leq \alpha, 1 \leq b \leq \beta \text{ and } 1 \leq \overline{b} \leq \beta\}.$$

Note that $|V_g| = \alpha^2\beta$ and $|V_r| = \alpha\beta^2$. The vertices of V_g are later referred to as *green* and those of V_r as *red*.

We now describe the edge set of $\Gamma_{\alpha,\beta}$. Two distinct green vertices $[a,b,\overline{a}]$ and $[a',b',\overline{a}']$ are joined by an edge if $b = b'$ and $\overline{a} = \overline{a}'$. Similarly, two distinct red vertices $[a,b,\overline{b}]$ and $[a',b',\overline{b}']$ are joined by an edge if $a = a'$ and $\overline{b} = \overline{b}'$. A green vertex $[a,b,\overline{a}]$ and a red vertex $[a',b',\overline{b}']$ are joined by an edge if $a = a'$ and $b = b'$.

Notice the following: The subgraph of $\Gamma_{\alpha,\beta}$ induced by the green vertices is composed of $\alpha\beta$ cliques of order α, the subgraph induced by the red vertices of $\alpha\beta$ cliques of order β, and the spanning subgraph containing edges between the red and green vertices is composed of $\alpha\beta$ complete bipartite graphs isomorphic to $K_{\alpha,\beta}$. It is not hard to verify that the graph $\Gamma_{\alpha,\beta}$ is connected, that its order is $(\alpha+\beta)\alpha\beta$, and that it is $(\alpha+\beta-1)$-regular. Examples of graphs $\Gamma_{\alpha,\beta}$ for some (small) values of α and β are given in Figure 3.1. Note that the graph $\Gamma_{\alpha,1}$ is isomorphic to the graph Ω_α. Also note that the graphs $\Gamma_{\alpha,\beta}$ and $\Gamma_{\beta,\alpha}$ are isomorphic for all $\alpha,\beta \geq 1$.

3.2. Analysis. In this subsection, we analyze properties of the graphs $\Gamma_{\alpha,\beta}$. First, we show an upper bound on $\lambda(\Gamma_{\alpha,\beta})$.

PROPOSITION 3.1. *For every $\alpha,\beta \geq 1$, the number $\lambda(\Gamma_{\alpha,\beta})$ does not exceed $2\alpha + 2\beta - 2$.*

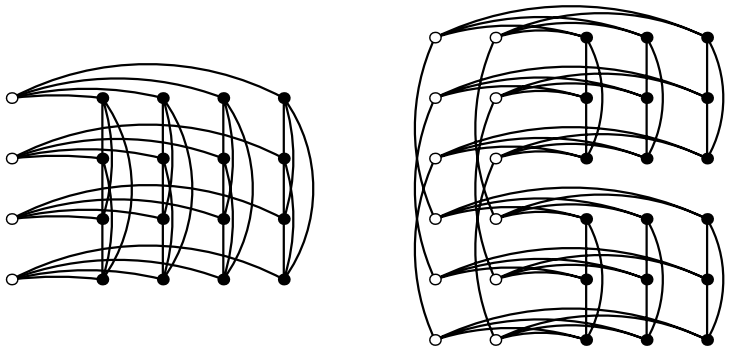


FIG. 3.1. *The graphs* $\Gamma_{1,4}$ *and* $\Gamma_{2,3}$. *Green vertices are depicted by empty circles and red vertices by full ones.*

Proof. We partition green vertices into α independent sets $V_1, \ldots, V_\alpha$ and red vertices into β independent sets $W_1, \ldots, W_\beta$. A green vertex $[a, b, \overline{a}]$ is contained in the set V_i, where i is congruent to $a + \overline{a}$ modulo α. A red vertex $[a, b, \overline{b}]$ is contained in the set W_i, where i is congruent to $b + \overline{b}$ modulo β.

Clearly, the sets $V_1, \ldots, V_\alpha$ and $W_1, \ldots, W_\beta$ are independent. We claim that the distance between any two vertices contained in the same set is at least three. Assume the opposite. By symmetry, it is enough to consider the case when V_1 contains two distinct green vertices $[a, b, \overline{a}]$ and $[a', b', \overline{a}']$ at distance two. If the common neighbor of $[a, b, \overline{a}]$ and $[a', b', \overline{a}']$ is a green vertex, then $b = b'$ and $\overline{a} = \overline{a}'$. By the definition of V_1, it follows that $a = a'$ and the vertices $[a, b, \overline{a}]$ and $[a', b', \overline{a}']$ are not distinct. On the other hand, if their common neighbor is a red vertex, then $a = a'$ and $b = b'$. The definition of V_1 now yields that $\overline{a} = \overline{a}'$, and the vertices $[a, b, \overline{a}]$ and $[a', b', \overline{a}']$ are not distinct as supposed.

We now construct an $L(2,1)$-labeling of $\Gamma_{\alpha,\beta}$. Label the vertices of V_i, $i = 1, \ldots, \alpha$, by the number $2i - 2$ and the vertices of W_i, $i = 1, \ldots, \beta$, by the number $2\alpha + 2i - 2$. The obtained labeling is a proper $L(2,1)$-labeling of $\Gamma_{\alpha,\beta}$. In particular, $\lambda(\Gamma_{\alpha,\beta}) \leq 2\alpha + 2\beta - 2$. □

In the next lemma, we present the key property of graphs $\Gamma_{\alpha,\beta}$.

LEMMA 3.2. *For every* $\alpha, \beta \geq 1$, *the graph* $\Gamma_{\alpha,\beta}$ *has the* $\alpha\beta$*-neighborhood property.*

Proof. Fix $\alpha \geq 1$ and $\beta \geq 1$ and let us consider two sets V_1 and V_2 of vertices of $\Gamma_{\alpha,\beta}$. Assume that V_1 contains no two vertices at distance at most two, V_2 contains no two vertices at distance at most two, and no two vertices of V_1 and V_2 are adjacent. We show that $|V_1| + |V_2| \leq \alpha\beta$. The statement of the lemma would then follow.

Let us construct two auxiliary matrices M_g and M_r of type $\alpha \times \beta$. For a, $1 \leq a \leq \alpha$, and b, $1 \leq b \leq \beta$, the entry $M_g[a, b]$ is the number of green vertices of the form $[a, b, \overline{a}]$, $1 \leq \overline{a} \leq \alpha$, contained in $V_1 \cup V_2$. Similarly, the entry $M_r[a, b]$ is the number of red vertices of the form $[a, b, \overline{b}]$, $1 \leq \overline{b} \leq \beta$, contained in $V_1 \cup V_2$. Next, several properties of the matrices M_g and M_r are established. We formulate the properties as a series of claims.

CLAIM 3.2.1. *All the entries of the matrices* M_g *and* M_r *are integers* 0, 1, *or* 2.

For fixed numbers a, $1 \leq a \leq \alpha$, and b, $1 \leq b \leq \beta$, all the green vertices $[a, b, \overline{a}]$, $1 \leq \overline{a} \leq \alpha$, of $\Gamma_{\alpha,\beta}$ are at distance two. In particular, at most one of them is contained in the set V_1 and at most one of them in V_2. Hence, $M_g[a, b] \leq 2$. A

symmetric argument applies to M_r.

CLAIM 3.2.2. *For every a, $1 \leq a \leq \alpha$, and b, $1 \leq b \leq \beta$, at most one of the entries $M_g[a,b]$ and $M_r[a,b]$ is nonzero.*

If $M_g[a,b] > 0$, then there is a green vertex $[a,b,\overline{a}]$, $1 \leq \overline{a} \leq \alpha$, that is contained in $V_1 \cup V_2$. Since every red vertex $[a,b,\overline{b}]$, $1 \leq \overline{b} \leq \beta$, is adjacent to the green vertex $[a,b,\overline{a}]$ contained in $V_1 \cup V_2$, no red vertex of the form $[a,b,\overline{b}]$, $1 \leq \overline{b} \leq \beta$, is contained in $V_1 \cup V_2$. Hence, $M_r[a,b]$ is equal to zero. An analogous argument yields that if $M_r[a,b] > 0$, then $M_g[a,b] = 0$.

CLAIM 3.2.3. *If $M_g[a,b] = 2$ for a, $1 \leq a \leq \alpha$, and b, $1 \leq b \leq \beta$, then $M_r[a',b] = M_r[a,b'] = 0$ for every a', $1 \leq a' \leq \alpha$, and b', $1 \leq b' \leq \beta$.*

Let $\overline{a}_1$ and $\overline{a}_2$, $1 \leq \overline{a}_1, \overline{a}_2 \leq \alpha$, be two distinct integers such that both green vertices $[a,b,\overline{a}_1]$ and $[a,b,\overline{a}_2]$ are contained in $V_1 \cup V_2$. Since the distance between the vertices $[a,b,\overline{a}_1]$ and $[a,b,\overline{a}_2]$ is two, one of them is contained in V_1 and the other in V_2. By symmetry, we can assume that $[a,b,\overline{a}_1] \in V_1$ and $[a,b,\overline{a}_2] \in V_2$.

Let us consider integers a', $1 \leq a' \leq \alpha$, and $\overline{b}$, $1 \leq \overline{b} \leq \beta$. If $a = a'$, then $M_r[a',b] = 0$ by Claim 3.2.2. In the rest, we consider the case $a \neq a'$. The green vertices $[a',b,\overline{a}_1]$ and $[a',b,\overline{a}_2]$ are neighbors of the green vertices $[a,b,\overline{a}_1]$ and $[a,b,\overline{a}_2]$, respectively. Since the red vertex $[a',b,\overline{b}]$ is a neighbor of both the green vertices $[a',b,\overline{a}_1]$ and $[a',b,\overline{a}_2]$, the vertex $[a',b,\overline{b}]$ can be included in neither V_1 nor V_2. Since the choice of $\overline{b}$ was arbitrary, $M_r[a',b]$ must be equal to zero for every $a' \neq a$, $1 \leq a \leq \alpha$. A symmetric argument yields that $M_r[a,b'] = 0$ for every b', $1 \leq b' \leq \beta$.

CLAIM 3.2.4. *If $M_r[a,b] = 2$ for a, $1 \leq a \leq \alpha$, and b, $1 \leq b \leq \beta$, then $M_g[a',b] = M_g[a,b'] = 0$ for every a', $1 \leq a' \leq \alpha$, and b', $1 \leq b' \leq \beta$.*

The proof is analogous to the proof of Claim 3.2.3.

CLAIM 3.2.5. *For every a, $1 \leq a \leq \alpha$, the sum of the entries of M_r on the ath row is at most β.*

For every $\overline{b}$, $1 \leq \overline{b} \leq \beta$, the vertices $[a,b,\overline{b}]$, $1 \leq b \leq \beta$, form a clique in $\Gamma_{\alpha,\beta}$. Hence, at most one of them can be contained in $V_1 \cup V_2$. Since there are β possible choices of $\overline{b}$, there are at most β red vertices with the first coordinate equal to a in $V_1 \cup V_2$.

CLAIM 3.2.6. *For every b, $1 \leq b \leq \beta$, the sum of the entries of M_g on the bth column is at most α.*

The proof is analogous to the proof of Claim 3.2.5.

We now continue the main part of the proof of Lemma 3.2. Let A_g be the set of all integers a such that $M_g[a,b] = 2$ for some b. Similarly, B_g is the set of all b's such that $M_g[a,b] = 2$ for some a. Analogously, A_r and B_r are sets of all integers a and b, respectively, such that $M_r[a,b] = 2$. In addition, let M be the matrix that is the sum of the matrices M_r and M_g, i.e., $M = M_r + M_g$. Note that the sum of all the entries of M is $|V_1| + |V_2|$. By Claims 3.2.3 and 3.2.4, it holds that $M_g[a,b] = M_r[a,b] = M[a,b] = 0$ for all $[a,b] \in A_r \times B_g$ and $[a,b] \in A_g \times B_r$. On the other hand, by the definitions of the sets A_g, B_g, A_r and B_r, if $[a,b] \notin (A_g \cup A_r) \times (B_g \cup B_r)$, $M_g[a,b] \leq 1$, $M_r[a,b] \leq 1$, and at most one of $M_g[a,b]$ and $M_r[a,b]$ is nonzero by Claim 3.2.2. We conclude that $M[a,b] = M_g[a,b] + M_r[a,b] \leq 1$ for every $[a,b] \notin (A_g \cup A_r) \times (B_g \cup B_r)$.

In order to finish the proof, we distinguish three cases according to the cardinalities of the sets A_g, B_g, A_r and B_r:

- $|B_g| \leq |B_r|$
  For $a \in A_r$, all the entries of M_g on the ath row are zero by Claim 3.2.4. In particular, the entries of M and M_r on the ath row coincide. Hence, the sum of the entries of the matrix M in the rows $a \in A_r$ is at most $|A_r|\beta$ by Claim 3.2.5. The sum of the entries $M[a,b]$ with $[a,b] \in A_g \times (B_g \cup B_r)$ is at most

$2|A_g||B_g| \leq |A_g|(|B_g| + |B_r|)$ by Claims 3.2.2, 3.2.3, and 3.2.4. Finally, all the remaining entries of M are at most one. We infer that the sum of all the entries of M does not exceed $\alpha\beta$.

- $|A_g| \geq |A_r|$
  An argument that is similar to that in the previous case and that involves Claim 3.2.6 applies.
- $|A_g| \leq |A_r|$ and $|B_g| \geq |B_r|$
  Observe first that the following holds:

$$\begin{aligned} |B_r| &\leq |B_g|, \\ |B_r|(|A_r| - |A_g|) &\leq |B_g|(|A_r| - |A_g|), \\ |A_g||B_g| + |A_r||B_r| &\leq |A_g||B_r| + |A_r||B_g|. \end{aligned} \tag{3.1}$$

  By Claim 3.2.1, it holds that $M[a,b] \leq 2$ for every $[a,b] \in (A_g \times B_g) \cup (A_r \times B_r)$. Since $M[a,b] = 0$ for all $[a,b] \in (A_g \times B_r) \cup (A_r \times B_g)$, the sum of the entries $M[a,b]$ for $[a,b] \in (A_g \cup A_r) \times (B_g \cup B_r)$ is at most the following (the first inequality follows from (3.1)):

$$\begin{aligned} 2(|A_g||B_g| + |A_r||B_r|) &\leq |A_g||B_g| + |A_r||B_r| + |A_g||B_r| + |A_g||B_r| \\ &\leq (|A_g| + |A_r|)(|B_g| + |B_r|). \end{aligned}$$

  Since $M[a,b] \leq 1$ for every $[a,b] \notin (A_g \cup A_r) \times (B_g \cup B_r)$, the sum of the entries of the matrix M is at most $\alpha\beta$ as desired.

Since the sum of the entries of M is equal to $|V_1| + |V_2|$, we conclude that $|V_1| + |V_2| \leq \alpha\beta$. □

The following theorem now readily follows from Propositions 2.2 and 3.1 and Lemma 3.2.

THEOREM 3.3. *For every $\alpha, \beta \geq 1$, the graph $\Gamma_{\alpha,\beta}$ has the following properties:*

- *the order of $\Gamma_{\alpha,\beta}$ is $(\alpha+\beta)\alpha\beta$;*
- *the graph $\Gamma_{\alpha,\beta}$ is connected;*
- *the graph $\Gamma_{\alpha,\beta}$ is $(\alpha+\beta-1)$-regular; and*
- *its hole number $\rho(\Gamma_{\alpha,\beta})$ is $\alpha+\beta-1$.*

An immediate corollary of Theorem 3.3 follows.

COROLLARY 3.4. *For every $r \geq 1$, there exists an r-regular connected graph G of order $(r+1)\lfloor \frac{(r+1)^2}{4} \rfloor \approx r^3/4$ with $\rho(G) = r$.*

Proof. Set $\alpha = \lfloor r/2 \rfloor + 1$ and $\beta = \lceil r/2 \rceil$, and consider the graph $\Gamma_{\alpha,\beta}$. Note that $\alpha\beta(\alpha+\beta) = (r+1)\lfloor \frac{(r+1)^2}{4} \rfloor$. □

3.3. Generalization. In this subsection, we slightly generalize our construction. If G is a graph with the vertex set $V(G)$, then $G^{[s]}$ is the graph whose vertex set is $V(G) \times \{1, \ldots, s\}$ and two distinct vertices $[v,i]$ and $[v',i']$ of $G^{[s]}$ are joined by an edge if $v = v'$ or $v \neq v'$ and vv' is an edge of G. Clearly, if G is an r-regular graph of order n, then $G^{[s]}$ is an $(rs+s-1)$-regular graph of order ns. Note that $G^{[s]}$ is the lexicographic product of G and the complete graph of order s.

We now formulate the following lemma.

LEMMA 3.5. *Let G be a connected r-regular graph of order $(r+1)t$. If G has the t-neighborhood property, then $G^{[s]}$ has also the t-neighborhood property for every $s \geq 1$.*

Proof. Let V and W be two disjoint sets of vertices of $G^{[s]}$ such that the distance between any two vertices in each of the sets is at least two and no vertex of V is

adjacent to a vertex of W. Let V' be the set of vertices v of G such that $[v,i] \in V$ for some i, $1 \leq i \leq s$. Similarly, W' is the set of vertices w such that $[w,i] \in W$. Note that the sets V' and W' are disjoint, $|V| = |V'|$, and $|W| = |W'|$. Moreover, V' and W' do not contain any two vertices at distance two and no vertex of V' is adjacent to a vertex of W'. Since G has the t-neighborhood property, $|V'| + |W'| \leq t$. Hence, $|V| + |W| \leq t$. Because the choice of V and W was arbitrary, $G^{[s]}$ has the t-neighborhood property. □

Fix $\alpha, \beta \geq 1$ and $s \geq 2$. Consider the labeling of $\Gamma_{\alpha,\beta}$ with the labels $0, 2, \ldots, 2\alpha + 2\beta - 2$ constructed in Proposition 3.1. We now construct an $L(2,1)$-labeling of $\Gamma^{[s]}_{\alpha,\beta}$. If v is a vertex of $\Gamma_{\alpha,\beta}$ that is labeled with γ, then a vertex $[v,i]$, $i = 1, \ldots, s$, of $\Gamma^{[s]}_{\alpha,\beta}$ is labeled with $\gamma + 2(i-1)(\alpha+\beta)$. The obtained labeling is a proper $L(2,1)$-labeling of $\Gamma^{[s]}_{\alpha,\beta}$. Hence, $\lambda(\Gamma^{[s]}_{\alpha,\beta}) \leq 2s(\alpha+\beta) - 2$.

The following theorem readily follows from Lemmas 3.2 and 3.5.

THEOREM 3.6. *For every $\alpha, \beta, s \geq 1$, the graph $\Gamma^{[s]}_{\alpha,\beta}$ has the following properties:*

- *the order of $\Gamma^{[s]}_{\alpha,\beta}$ is $(\alpha+\beta)\alpha\beta s$;*
- *the graph $\Gamma^{[s]}_{\alpha,\beta}$ is connected;*
- *the graph $\Gamma^{[s]}_{\alpha,\beta}$ is $((\alpha+\beta)s-1)$-regular; and*
- *its hole number $\rho(\Gamma^{[s]}_{\alpha,\beta})$ is $(\alpha+\beta)s-1$.*

Note that Theorem 3.6 yields a construction of connected r-regular graphs G of order $(r+1)t$ for some (but not all) numbers t between r and $\lfloor \frac{(r+1)^2}{4} \rfloor$.

4. Conclusion. We conclude the paper with several problems in the spirit of Conjecture 2. The first problem that comes to mind is the following.

PROBLEM 1. *Is it true that there exists a function $f(r)$ with the following property: If G is a connected r-regular graph of order $(r+1)t$ with $\rho(G) = r$, then $t \leq f(r)$? Does there exist a polynomial $f(r)$ with this property?*

Georges and Mauro [5] constructed connected r-regular graphs of order $(r+1)t$ for every $t = 1, \ldots, r$. We constructed such graphs for some numbers t larger than r, but we were not able to construct such graphs for all $t = 1, \ldots, \lfloor \frac{(r+1)^2}{4} \rfloor$. This leads us to the following problem.

PROBLEM 2. *Assume that G is a connected r-regular graph of order $(r+1)t_0$ with $\rho(G) = r$. Is it true that for every $t = 1, \ldots, t_0$, there exists a connected r-regular graph of order $(r+1)t$ with $\rho(G) = r$? In particular, is this true for $t_0 = \lfloor \frac{(r+1)^2}{4} \rfloor$?*

In the case of cubic graphs, we are aware of constructions of connected cubic graphs G of orders 4, 8, 12, and 16 with $\rho(G) = 3$. We have a computer-assisted proof that there is no such cubic graph of order 20. If the answer to Problem 2 were positive, then the answer to the following problem would also be positive.

PROBLEM 3. *Is it true that there is no connected cubic graph G with $\rho(G) = 3$ whose order is at least 20?*

REFERENCES

[1] G. J. CHANG, W.-T. KE, D. D.-F. LIU, AND R. K. YEH, *On $L(d,1)$-labellings of graphs*, Discrete Math., 3 (2000), pp. 57–66.

[2] G. J. CHANG AND D. KUO, *The $L(2,1)$-labeling problem on graphs*, SIAM J. Discrete Math., 9 (1996), pp. 309–316.

[3] P. C. FISHBURN AND F. S. ROBERTS, *Full Color Theorems for $L(2,1)$-Colorings*, DIMACS Technical Report 2000-08, 2000.

[4] P. C. Fishburn and F. S. Roberts, *No-hole $L(2,1)$-colorings*, Discrete Appl. Math., 130 (2003), pp. 513–519.
[5] J. P. Georges and D. W. Mauro, *On the structure of graphs with non-surjective $L(2,1)$-labelings*, SIAM J. Discrete Math., 19 (2005), pp. 208–223.
[6] J. P. Georges and D. W. Mauro, *A note on collections of graphs with non-surjective lambda labelings*, Discrete Appl. Math., 146 (2005), pp. 92–98.
[7] D. Gonçalves, *On the $L(p,1)$-labelling of graphs*, Discrete Math. Theoret. Comput. Sci., AE (2005), pp. 81–86.
[8] J. R. Griggs and R. K. Yeh, *Labeling graphs with a condition at distance* 2, SIAM J. Discrete Math., 5 (1992), pp. 586–595.
[9] W. K. Hale, *Frequency assignment: Theory and applications*, Proc. IEEE, 68 (1980), pp. 1497–1514.
[10] J.-H. Kang, *$L(2,1)$-labeling of* 3*-regular Hamiltonian graphs*, SIAM J. Discrete Math., submitted.
[11] J.-H. Kang, *$L(2,1)$-Labelling of* 3*-Regular Hamiltonian Graphs*, Ph.D. thesis, University of Illinois, Urbana-Champaign, IL, 2004.
[12] D. Král', *Coloring powers of chordal graphs*, SIAM J. Discrete Math., 18 (2004), pp. 451–461.
[13] D. Král' and R. Škrekovski, *A theorem about the channel assignment problem*, SIAM J. Discrete Math., 16 (2003), pp. 426–437.
[14] C. McDiarmid, *On the span in channel assignment problems: Bounds, computing and counting*, Discrete Math., 266 (2003), pp. 387–397.
[15] D. Sakai, *Labeling chordal graphs: Distance two condition*, SIAM J. Discrete Math., 7 (1994), pp. 133–140.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 545–563

ON PREEMPTIVE RESOURCE CONSTRAINED SCHEDULING: POLYNOMIAL-TIME APPROXIMATION SCHEMES*

KLAUS JANSEN† AND LORANT PORKOLAB‡

Abstract. We study resource constrained scheduling problems where the objective is to compute feasible preemptive schedules minimizing the makespan and using no more resources than what are available. We present approximation schemes along with some inapproximability results showing how the approximability of the problem changes in terms of the number of resources. The results are based on linear programming formulations (though with exponentially many variables) and some interesting connections between resource constrained scheduling and (multidimensional, multiple-choice, and cardinality constrained) variants of the classical knapsack problem. In order to prove the results we generalize a method by Grigoriadis et al. for the max-min resource sharing problem to the case with weak approximate block solvers (i.e., with only constant, logarithmic, or even worse approximation ratios). Finally, we present applications of the above results in fractional graph coloring and multiprocessor task scheduling.

Key words. scheduling, linear programming, approximation algorithms

AMS subject classifications. 68W25, 68W40, 90C05, 90C27

DOI. 10.1137/S0895480101396949

1. Introduction.

In this paper we consider the general preemptive resource constrained scheduling problem denoted by $P|res\ldots,pmtn|C_{max}$: There are given n tasks $\mathcal{T} = \{T_1, \ldots, T_n\}$, m identical machines, and s resources such that each task $T_j \in \mathcal{T}$ is processed by one machine requiring p_j units of time and r_{ij} units of resource i, $i = 1, \ldots, s$, from which only c_i units are available at each time. One may assume w.l.o.g. that $r_{ij} \in [0, 1]$ and $c_i \geq 1$. The objective is to compute a preemptive schedule of the tasks minimizing the maximum completion time C_{max}. The three dots in the notation indicate that there are no restrictions on the number of resources s, the largest possible capacity o, and resource requirement r values, respectively. If any of these is limited, the corresponding fixed limit replaces the corresponding dot in the notation (e.g., if $s \leq 1$, then $P|res1..,pmtn|C_{max}$ is used, or if $r_{ij} \leq r$, then $P|res\,..r,pmtn|C_{max}$ is used).

Let $A(I)$ denote the schedule length produced by algorithm A on I, and let $OPT(I)$ denote the minimum schedule length. We say that A is a ρ-approximation algorithm for the scheduling problem (where $\rho \geq 1$) if it generates in polynomial time a schedule with length $A(I) \leq \rho OPT(I)$ for any instance I. A polynomial-time approximation scheme (PTAS) is a family of approximation algorithms $\{A_\epsilon | \epsilon > 0\}$ such that $A_\epsilon(I) \leq (1 + \epsilon)OPT(I)$. A fully polynomial-time approximation scheme (FPTAS) is an approximation scheme $\{A_\epsilon | \epsilon > 0\}$ where each algorithm A_ϵ runs in time polynomial in the length of I and $1/\epsilon$. We will study different variants of the

*Received by the editors October 25, 2001; accepted for publication (in revised form) November 18, 2005; published electronically June 30, 2006. This research was supported in part by EU-Project APPOL, Approximation and Online Algorithms, IST-1999-14084 and IST-2001-30012, and by EU-Project CRESCCO, Critical Resource Sharing for Cooperation in Complex Systems, IST-2001-33135. An extended abstract of this paper appeared in *Integer Programming and Combinatorial Optimization*, Lecture Notes in Comput. Sci. 2337, Springer-Verlag, Berlin, 2002, pp. 329–349.
http://www.siam.org/journals/sidma/20-3/39694.html

†Institut für Informatik und Praktische Mathematik, Universität zu Kiel, Kiel, Germany (kj@informatik.uni-kiel.de).

‡Department of Computing, Imperial College, London, UK (lorant.porkolab@doc.ic.ac.uk).

problem and their applications in multiprocessor task scheduling and fractional graph coloring.

1.1. Related results. Resource constrained scheduling is one of the classical scheduling problems. Garey and Graham [20] proposed approximation algorithms for the nonpreemptive variant $P|res\ldots|C_{max}$ with approximation ratios $s+1$ (when the number of machines is unbounded, $m \geq n$) and $\min(\frac{m+1}{2}, s+2-\frac{2s+1}{m})$ (when $m \geq 2$).

Further results are known for some special cases: Garey et al. [21] proved that when $m \geq n$ and each task T_j has unit-execution time, i.e., $p_j = 1$, the problem (denoted by $P\infty|res\ldots, p_j = 1|C_{max}$) can be solved by First Fit and First Fit Decreasing heuristics providing asymptotic approximation ratio $s + \frac{7}{10}$ and a ratio between $s + ((s-1)/s(s+1))$ and $s + \frac{1}{3}$, respectively. Fernandez de la Vega and Lueker [18] gave a linear-time algorithm with asymptotic approximation ratio $s + \epsilon$ for each fixed $\epsilon > 0$. Further results and improvements for nonpreemptive variant are given in [3, 5, 10, 54, 55].

For the preemptive variant substantially fewer results are known: Blazewicz, Lenstra, and Rinnooy Kan [6] proved that when m is fixed, the scheduling problem $Pm|res\ldots, pmtn|C_{max}$ (with identical machines) and even the variant $Rm|res\ldots, pmtn|C_{max}$ (with unrelated machines) can be solved in polynomial time. Krause, Shen, and Schwetman [38, 39] studied $P|res1.., pmtn|C_{max}$, i.e., where there is only one resource ($s = 1$) and proved that both First Fit and First Fit Decreasing heuristics can guarantee a $3 - 3/n$ asymptotic approximation ratio.

A related problem is multiprocessor task scheduling, where a set $\mathcal{T}$ of n tasks has to be executed by m processors such that each processor can execute at most one task at a time and each task must be processed by several processors in parallel. In the parallel (nonmalleable) model $P|size_j|C_{max}$, there is a value $size_j \in M = \{1, \ldots, m\}$ given for each task T_j indicating that T_j can be processed on any subset of processors of cardinality $size_j$ [1, 2, 4, 13, 14, 29, 35, 50, 56]. In the malleable variant $P|fctn|C_{max}$, each task can be executed on an arbitrary subset of processors, and the execution time $p_j(\ell)$ depends on the number ℓ of processors assigned to it [41, 45, 57]. Regarding the complexity, it is known [13, 14] that the preemptive variant $P|size_j, pmtn|C_{max}$ is NP-hard. In [30], focusing on computing optimal solutions, we presented an algorithm for solving the problem $P|size_j, pmtn|C_{max}$ and showed that this algorithm runs in $O(n) + poly(m)$ time, where $poly(.)$ is a univariate polynomial. Furthermore, we extended this algorithm also to malleable tasks with running time polynomial in m and n. These results are based on methods by Grötschel, Lovász, and Schrijver [26] and use the ellipsoid method.

Another related problem is fractional graph coloring; see, e.g., [15, 17, 25, 36, 42, 44, 46, 51, 52]. Grötschel, Lovász, and Schrijver [25] proved that the weighted fractional coloring problem is NP-hard for general graphs but can be solved in polynomial time for perfect graphs. They have proved the following interesting result: For any graph class $\mathcal{G}$, if the problem of computing $\alpha(G, w)$ (the weight of the largest weighted independent set in G) for graphs $G \in \mathcal{G}$ is NP-hard, then the problem of determining the weighted fractional chromatic number $\chi_f(G, w)$ is also NP-hard. This gives a negative result of the weighted fractional coloring problem even for planar cubic graphs. Furthermore, if the weighted independent set problem for graphs in $\mathcal{G}$ is polynomial-time solvable, then the weighted fractional coloring problem for $\mathcal{G}$ can also be solved in polynomial time. The first inapproximability result for the unweighted version of the problem (i.e., where $w_v = 1$ for each vertex $v \in V$) was obtained by Lund and Yannakakis [42] who proved that there exists a $\delta > 0$ such that there is no

polynomial-time approximation algorithm for the problem with approximation ratio n^{δ}, unless P = NP. Feige and Kilian [17] showed that the fractional chromatic number $\chi_f(G)$ cannot be approximated within $\Omega(|V|^{1-\epsilon})$ for any $\epsilon > 0$, unless ZPP = NP. Recently, the authors [31] proved that fractional coloring is NP-hard even for graphs with $\chi_f(G) = 3$ and constant degree 4. Similarly, as was shown by Gerke and McDiarmid [22], the problem remains NP-hard even for triangle-free graphs. Regarding the approximability of the fractional chromatic number, Matsui [44] gave a polynomial-time 2-approximation algorithm for unit disk graphs.

1.2. New results. The results presented in this paper are based on linear programming formulations. They are typically of the following form:

$$\begin{array}{lll} \min & \sum_{h\in H} x_h & \\ \text{s.t.} & \sum_{h\in H: a\in h} x_h \;\geq\; w_a & \forall a \in A, \\ & x_h \;\geq\; 0 & \forall h \in H, \end{array} \tag{1.1}$$

where A is a finite set (usually the set of all tasks), and $H \subseteq 2^A$ is a set consisting of subsets of A satisfying some combinatorial property (usually each contains tasks that can be scheduled together at the same time). These linear programs will, in general, have exponentially many variables but special underlying structures allowing efficient approximations. A linear program of form (1.1) can be solved (approximately) by using binary search on its optimum and computing at each stage an approximate solution of a special max-min resource sharing problem of the following type:

$$\begin{array}{lll} \lambda^* \;=\; \max & \lambda & \\ \text{s.t.} & f_i(x_1,\ldots,x_N) \;\geq\; \lambda, & i = 1,\ldots,M, \\ & (x_1,\ldots,x_N) \in P, & \end{array} \tag{1.2}$$

where $f_i : P \to \mathbb{R}^+$, $i = 1,\ldots,M$, are, in general, nonnegative concave functions defined on a nonempty convex set $P \subseteq \mathbb{R}^N$. Furthermore, approximate solutions for problem (1.2) can be computed by an iterative procedure that requires in each iteration for a given M-vector $(p_1,\ldots,p_M)$ the approximate maximization of $\sum_{i=1}^{M} p_i f_i(x)$ over all $x = (x_1,\ldots,x_N) \in P$. Interestingly, these subproblems for different variants of resource constrained scheduling turn out to be knapsack-type problems (multiple-choice, multidimensional, and cardinality constrained knapsack) with efficient approximation algorithms. For fractional graph coloring the subproblem is the well-known maximum weighted independent set problem.

In section 2 we describe the methodology used for solving the max-min resource sharing problem. Let $f(x) = (f_1(x),\ldots,f_M(x))$ and $\Lambda(p) = \max_{x\in P} p^T f(x)$. Based on the paper of Grigoriadis et al. [24], we derive the following result extending some of the previous works on computing approximate solutions for fractional covering problems [24, 47, 58]: If there exists a polynomial-time approximation algorithm with approximation ratio c for the subproblem, i.e., for finding a vector $\hat{x}(p) \in P$ such that $p^T f(\hat{x}) \geq \frac{1}{c}\Lambda(p)$, then there is also a polynomial-time approximation algorithm for the max-min resource sharing problem that computes a feasible solution with objective function value $\frac{1-\epsilon}{c}\lambda^*$. Interestingly, the number of iterations (hence also the number of calls to the solver for the subproblem) is bounded by $O(M(\ln M + \ln c\epsilon^{-3} + \epsilon^{-2}))$, independently of the width [47] of P and the number of variables. If, in particular, there is an (F)PTAS for the subproblem, one also gets an (F)PTAS for the original problem and the number of iterations is at most $O(M(\ln M + \epsilon^{-2}))$ [24]. This fact can be particularly useful for models with exponentially many variables.

In section 3 we describe a linear programming approach for the preemptive resource constrained scheduling problem, where there are no assumptions on s or $m \geq n$; they are arbitrary numbers and parts of the input. We show by using the linear programming formulation that there is an approximation algorithm for our scheduling problem with approximation ratio $O(s^{\frac{1}{c_{min}}})$, where the minimum resource capacity $c_{min} = \min_i c_i \geq 1$. Furthermore, we argue that if for each resource i, capacity $c_i \geq \frac{12}{\epsilon^2}\log(2s)$, then there is a polynomial-time $(1+\epsilon)$-approximation algorithm.

Then with the aim of obtaining stronger approximation results we study restricted variants, where s is fixed. In particular, we show that for any constant $s \geq 2$, there exists a PTAS computing an ϵ-approximate preemptive schedule satisfying the s resource constraints. In fact this is the best one can expect regarding approximation, since as we show it, this variant even with $s = 2$ cannot have an FPTAS unless P = NP. However, if we assume that $s = 1$ (i.e., the number of resources is pushed to its lower extreme), the problem possesses an FPTAS. Next, we apply our approach to the case where there is only a limited number of processors. We give an FPTAS for the variant of the problem with one resource improving the previously known best $(3-\frac{1}{n})$-approximation algorithm by Krause, Shen, and Schwetman [38]. The method can be used to obtain a PTAS for a more general variant with a fixed number of resources, where the input also contains release and delivery times for each task. In section 5 we study the preemptive multiprocessor task scheduling problem $P|size_j, pmtn|C_{max}$ and its generalization $P|fctn_j, pmtn|C_{max}$ to malleable tasks. We show the existence of FPTASs for both problems.

Finally, we apply our linear programming based approach, initially introduced for preemptive resource constrained scheduling, to the problem of computing the fractional weighted chromatic number. We prove an approximation analogue of the above-mentioned classical result of Grötschel, Lovász, and Schrijver [25] on the equivalence between polynomial-time (exact) computations of $\alpha(G,w)$ and $\chi_f(G,w)$: If for a graph class $\mathcal{G}$ there exists a polynomial-time $\frac{1}{c}$-approximation algorithm for computing $\alpha(G,w)$, then there is also a polynomial-time $c(1+\epsilon)$-approximation algorithm for computing $\chi_f(G,w)$ for graphs in $\mathcal{G}$. By applying this general result for intersection graphs of disks in the plane, we also obtain a PTAS for the fractional coloring problem providing a substantial improvement on Matsui's result [44].

2. Approximate max-min resource sharing. In this section we will follow the presentation of [24] and use the notation introduced there. Let $f : B \to \mathbb{R}_+^M$ be a vector with M nonnegative, continuous, concave functions f_m, block B a nonempty, convex, compact set, and $e^T = (1,\ldots,1) \in \mathbb{R}_+^M$. Consider the optimization problem

$$(P) \qquad \lambda^* \;=\; \max\,\{\,\lambda \;:\; f(x) \;\geq\; \lambda e,\; x \in B\,\},$$

and assume w.l.o.g. that $\lambda^* > 0$. Let $\lambda(f) = \min_{1\leq m\leq M} f_m$ for any given function f. Here we are interested in computing a (c,ϵ)-approximate solution for (P); i.e., for an approximation guarantee $c = c(M) > 1$ and an additional error tolerance $\epsilon \in (0,1)$ we want to solve the following problem:

$$(P_{c,\epsilon}) \qquad \text{compute}\;\; x \in B \;\;\text{such that}\;\; f(x) \;\geq\; \left[\frac{1}{c}(1-\epsilon)\lambda^*\right] e.$$

In order to solve this resource sharing problem we study the subproblem

$$\Lambda(p) = \max\,\{\,p^T f(x) \;:\; x \in B\,\}$$

for $p \in P = \{p \in \mathbb{R}^M \ : \ \sum_{i=1}^M p_i = 1,\ p_i \geq 0\}$. Here we use an approximate block solver (ABS) that solves the following subproblem:

$$ABS(p,c) \qquad \text{compute } \ \hat{x} = x(p) \in B \ \text{ such that } \ p^T f(\hat{x}) \ \geq \ \frac{1}{c}\Lambda(p).$$

By duality we have $\lambda^* = \max_{x\in B} \min_{p\in P} p^T f(x) = \min_{p\in P} \max_{x\in B} p^T f(x)$. This implies that $\lambda^* = \min\{\Lambda(p) \ : \ p \in P\}$. Based on this equality, one can naturally define the problem of finding a (c,ϵ)-approximate dual solution:

$$(D_{c,\epsilon}) \qquad \text{compute } \ p \in P \ \text{ such that } \ \Lambda(p) \ \leq \ c(1+\epsilon)\lambda^*.$$

Then the following result holds.

THEOREM 2.1. *If there exists a polynomial-time block solver $ABS(p,c)$ for some $c \geq 1$ and any $p \in P$, then there is an approximation algorithm for the resource sharing problem that computes a solution whose objective function value is at least $\frac{1}{c}(1-\epsilon)\lambda^*$.*

The running time of the approximation algorithm depends only on c, M, and $\frac{1}{\epsilon}$. In particular, if there is an (F)PTAS for the block problem computing an $\hat{x} \in B$ such that $p^T f(\hat{x}) \geq (1-\epsilon)\Lambda(p)$ for any constant $\epsilon > 0$, then there is an (F)PTAS for the resource sharing problem [24].

The algorithm uses the logarithmic potential function

$$\Phi_t(\theta, f) = \ln\theta + \frac{t}{M}\sum_{m=1}^{M} \ln(f_m - \theta),$$

where $\theta \in \mathbb{R}$, $f = (f_1, \ldots, f_M)$ are variables associated with the coupling constraints $f_m \geq \lambda$, $1 \leq m \leq M$ and $t > 0$ is a tolerance (that depends on ϵ). For $\theta \in (0, \lambda(f))$, the function Φ_t is well defined. The maximizer $\theta(f)$ of function $\Phi_t(\theta, f)$ is given by the first order optimality condition

$$\frac{t\theta}{M} \ \sum_{m=1}^{M} \frac{1}{f_m - \theta} \ = \ 1. \tag{2.1}$$

This has a unique root since $g(\theta) = \frac{t\theta}{M}\sum_{m=1}^{M} \frac{1}{f_m - \theta}$ is a strictly increasing function of θ. The logarithmic dual vector $p = p(f)$ for a fixed f is defined by

$$p_m(f) = \frac{t}{M}\frac{\theta(f)}{f_m - \theta(f)}. \tag{2.2}$$

By (2.1), we have $p(f) \in P$. We will also use the following properties [24].

PROPOSITION 2.2.

(a) $p(f)^T f \ = \ (1+t)\theta(f)$.

(b) $\frac{\lambda(f)}{1+t} \ \leq \ \theta(f) \ \leq \ \frac{\lambda(f)}{1+t/M}$.

Now define parameter $v = v(x, \hat{x})$ by

$$v(x,\hat{x}) = \frac{p^T\hat{f} - p^T f}{p^T\hat{f} + p^T f}, \tag{2.3}$$

where $p \in P$, $f = f(x)$, $\hat{f} = f(\hat{x})$, and $\hat{x} \in B$ is an approximate block solution produced by $ABS(p,c)$. The following lemma provides a generalization of a useful result in [24].

LEMMA 2.3. *Suppose $\epsilon \in (0,1)$ and $t = \epsilon/5$. For a given $x \in B$, let $p \in P$ be computed from (2.2) and $\hat{x}$ be computed by $ABS(p,c)$. If $v(x,\hat{x}) \le t$, then the pair (x,p) solves $(P_{c,\epsilon})$ and $(D_{c,\epsilon})$, respectively.*

Proof. First rewrite condition $v \le t$ by using (2.3): $p^T\hat{f}(1-t) \le p^T f(1+t)$. Then use that $p^T\hat{f} \ge \frac{1}{c}\Lambda(p)$, $p^T f = (1+t)\theta$, and $\theta(f) < \lambda(f)$ by Proposition 2.2. This gives

$$\Lambda(p) \;\le\; cp^T\hat{f} \;\le\; c\frac{(1+t)}{(1-t)}p^T f \;=\; c\frac{(1+t)^2}{(1-t)}\theta(f) \;<\; c\frac{(1+t)^2}{(1-t)}\lambda(f) \;\le\; c(1+\epsilon)\lambda(f).$$

Using $\lambda^* \le \Lambda(p) \le c(1+\epsilon)\lambda(f)$, one has $\lambda(f) \ge \frac{1}{c}\frac{1}{1+\epsilon}\lambda^* > \frac{1}{c}(1-\epsilon)\lambda^*$ for any $\epsilon > 0$, which gives $(P_{c,\epsilon})$. Using $\lambda(f) \le \lambda^*$, one gets $\Lambda(p) \le c(1+\epsilon)\lambda(f) \le c(1+\epsilon)\lambda^*$, which is $(D_{c,\epsilon})$. ∎

The main algorithm works as follows.

ALGORITHM. *Improve*(f, B, ϵ, x).

(1) set $t := \epsilon/5$; $v := t+1$;
(2) **while** $v > t$ **do**
  (2.1) compute $\theta(f)$ and $p \in P$;
  (2.2) set $\hat{x} := ABS(p,c)$;
  (2.3) compute $v(x,\hat{x})$;
  (2.4) **if** $v > t$ **then** set $x = (1-\tau)x + \tau\hat{x}$, where $\tau \in (0,1)$ is an appropriate step length

  end
(3) return(x,p).

The step length can be defined by $\tau = \frac{t\theta v}{2M(p^T\hat{f}+p^T f)}$. Notice that $\theta < p^T\hat{f} + p^T f$ (by Proposition 2.2), and therefore $\tau \in (0,1)$. Furthermore, $v > t > 0$ implies that $\tau > 0$. For the initial solution, let $x^0 = \frac{1}{M}\sum_{m=1}^M \hat{x}^{(m)}$, where $\hat{x}^{(m)}$ is the solution given by $ABS(e_m, c)$ obtained for unit vector e_m with all zero coordinates except for its mth component which is 1. The next lemma provides a bound on $f(x^0)$.

LEMMA 2.4. *For each $p \in P$, $\lambda^* \le \Lambda(p) \le cMp^T f(x^0)$. Furthermore, $f_m(x^0) \ge \frac{1}{M}\frac{1}{c}\lambda^*$ for each $m = 1, \ldots, M$.*

Proof. The first inequality follows from duality. For the second inequality,

$$\begin{aligned}\Lambda(p) &= \max\{p^T f(x) \;:\; x \in B\} = \max\left\{\sum_{m=1}^M p_m f_m(x) \;:\; x \in B\right\}\\ &\le \sum_{m=1}^M p_m \max\{f_m(x) \;:\; x \in B\},\end{aligned}$$

where $\max\{f_m(x) \;:\; x \in B\} = \Lambda(e_m)$. Since $\hat{x}^{(m)}$ is the solution computed by $ABS(e_m, c)$, $f_m(\hat{x}^{(m)}) \ge \frac{1}{c}\Lambda(e_m)$ implying that $\Lambda(e_m) \le cf_m(x^{(m)})$. Therefore, $\Lambda(p) \le c\sum_{m=1}^M p_m f_m(\hat{x}^{(m)})$. Using the concavity of f_m we get

$$f_m(\hat{x}^{(m)}) \;\le\; \sum_{\ell=1}^M f_m(\hat{x}^{(\ell)}) \;\le\; Mf_m\left(1/M\sum_{\ell=1}^M \hat{x}^{(\ell)}\right) \;=\; Mf_m(x^0).$$

Combining the two inequalities, we obtain

$$\Lambda(p) \;\le\; c\sum_{m=1}^M p_m f_m(\hat{x}^{(m)}) \;\le\; cM\sum_{m=1}^M p_m f_m(x^0) \;=\; cMp^T f(x^0).$$

Finally, $f_m(x^0) \geq \frac{1}{M} f_m(\hat{x}^{(m)}) \geq \frac{1}{M}\frac{1}{c}\Lambda(e_m) \geq \frac{1}{M}\frac{1}{c}\lambda^*$. ∎

Let $\phi_t(f) = \Phi_t(\theta(f), f)$, which is called the reduced potential function. The following two lemmas proved in [24] are used here to bound the number of iterations.

LEMMA 2.5. *For any two consecutive iterates x and x' of Algorithm Improve, it holds that $\phi_t(f') - \phi_t(f) \geq tv^2/4M$.*

LEMMA 2.6. *For any two points $x' \in B$ and $x \in B$ with $\lambda(f) > 0$, $\phi_t(f') - \phi_t(f) \leq (1+t)\ln\frac{\Lambda(p)}{p^T f}$, where p is the vector defined by (2.2).*

THEOREM 2.7. *Algorithm Improve solves $(P_{c,\epsilon})$ and $(D_{c,\epsilon})$ in $O(\frac{M\ln M}{\epsilon} + \frac{M}{\epsilon^2} + \frac{M\ln c}{\epsilon^3})$ iterations.*

Proof. Let N_0 be the number of iterations to reach an iterate x^1 with corresponding error $v \leq 1/2$ starting from our initial solution x^0. For all iterations with $v \geq 1/2$, each iteration increases the potential by at least $tv^2/4M \geq t/16M$ (see Lemma 2.5). By Lemma 2.6, the total increase is bounded by $\phi_t(f^1) - \phi_t(f^0) \leq (1+t)\ln\frac{\Lambda(p^0)}{p^{0T}f^0}$. Since $t = \epsilon/5$ and $\Lambda(p^0) \leq cMp^{0T}f^0$ (by Lemma 2.4), we obtain that

$$N_0 \;\leq\; \frac{(1+\epsilon/5)16M\ln(cM)}{\epsilon/5} \;=\; O\left(\frac{M\ln(cM)}{\epsilon}\right).$$

Now suppose that the error $v_\ell \leq 1/2^\ell$ for iterate $x^\ell \in B$, and let N_ℓ be the number of iterations to halve this error. We get $\phi_t(f^{\ell+1}) - \phi_t(f^\ell) \geq \frac{N_\ell t v_{\ell+1}^2}{4M} = \frac{N_\ell t v_\ell^2}{16M}$. On the other hand, the definition of v_ℓ implies $p^{\ell T}\hat{f}^\ell(1-v_\ell) = p^{\ell T}f^\ell(1+v_\ell)$. Using $ABS(p^\ell, c)$, we get a solution $\hat{x}^\ell$ with $p^{\ell T}\hat{f}^\ell \geq \frac{1}{c}\Lambda(p^\ell)$. Combining the two inequalities,

$$\frac{\Lambda(p^\ell)}{p^{\ell T}f^\ell} \;\leq\; \frac{c(1+v_\ell)}{(1-v_\ell)} \;\leq\; c(1+4v_\ell).$$

The last inequality holds since $v_\ell \leq 1/2$. Since $d \geq 0$, Lemma 2.6 implies that

$$\phi_t(f^{\ell+1}) - \phi_t(f^\ell) \;\leq\; (1+t)(\ln c + \ln(1+4v_\ell)) \;\leq\; (1+t)(\ln c + 4v_\ell).$$

This gives now an upper bound

$$N_\ell \;\leq\; \frac{16M(1+t)(\ln c + 4v_\ell)}{tv_\ell^2} \;=\; O\left(\frac{M(\ln c + v_\ell)}{\epsilon v_\ell^2}\right).$$

One gets the total number of iterations by summing N_ℓ over all $\ell = 0, 1, \ldots, \lceil\ln(\frac{1}{t})\rceil$. Therefore, the total number of iterations is bounded by

$$N_0 + O\left(\frac{M\ln c}{\epsilon}\sum_{\ell=1}^{\lceil\ln(\frac{1}{t})\rceil} 2^{2\ell} + \frac{M}{\epsilon}\sum_{\ell=1}^{\lceil\ln(\frac{1}{t})\rceil} 2^\ell\right) \;\leq\; O\left(\frac{M\ln(cM)}{\epsilon} + \frac{M\ln c}{\epsilon^3} + \frac{M}{\epsilon^2}\right). \qquad ∎$$

The total number of iterations can be improved by the scaling method used in [47, 24]. The idea is to reduce the parameter t step by step to the desired accuracy. In the sth scaling phase we set $\epsilon_s = \epsilon_{s-1}/2$ and $t_s = \epsilon_s/5$ and use the current approximate point x^{s-1} as its initial solution. For phase $s = 0$, we use the initial point $x^0 \in B$. For this point we have $p^T f(x^0) \geq \frac{1}{cM}\Lambda(p)$. We set $\epsilon_0 = (1 - 1/M)$. Using Lemma 2.3, $f_m(x^0) \geq \frac{1}{M}\frac{1}{c}\lambda^*$. This implies $f_m(x^0) \geq \frac{1}{cM}\lambda^* = \frac{1}{c}(1 - 1 + \frac{1}{M})\lambda^* = \frac{1}{c}(1-\epsilon_0)\lambda^*$ for each $m = 1, \ldots, M$.

THEOREM 2.8. *For any accuracy $\epsilon > 0$, the error scaling implementation computes solutions x and p of $(P_{c,\epsilon})$ and $(D_{c,\epsilon})$, respectively, in*

$$N \;=\; O(M \ln M + M \ln c/\epsilon^3 + M/\epsilon^2)$$

iterations.

Proof. To reach the first $\epsilon_0 \in (1/2, 1)$ in the primal and dual problem we need $O(M(\ln c + \ln M))$ iterations (by Theorem 2.7). Let N_s be the number of iterations in phase s to reach ϵ_s for $s \geq 1$. By Lemma 2.5, each iteration of phase s increases the potential function by at least $t_s^3/4M = \theta(\epsilon_s^3/M)$. Lemma 2.6 implies that for $x = x^s$ and $x' = x^{s+1}$,

$$\phi_{t_s}(f^{s+1}) - \phi_{t_s}(f^s) \;\leq\; (1 + t_s) \ln \frac{\Lambda(p^s)}{p^{sT} f^s}.$$

Note that x^s is an $\epsilon_{s-1} = 2\epsilon_s$ solution of $(P_{c,\epsilon_{s-1}})$, and therefore $f(x^s) \geq (1-2\epsilon_s)\frac{1}{c}\lambda^* e$. Furthermore, since $\Lambda(p^s) \leq c(1 + 2\epsilon_s)\lambda^*$, $\Lambda(p^s) \leq c^2 \frac{1+2\epsilon_s}{1-2\epsilon_s} \lambda(f^s) \leq c^2 \frac{1+2\epsilon_s}{1-2\epsilon_s} p^{sT} f^s$, implying that $\frac{\Lambda(p^s)}{p^{sT} f^s} \leq c^2(1 + 8\epsilon_s)$. Then one can bound N_s by $O(M(\ln c + \epsilon_s)/\epsilon_s^3)$, and as before, the overall number of iterations is bounded by

$$\begin{aligned} N_0 + \sum_{s \geq 1} N_s &\leq O(M(\ln c + \ln M)) + O\left(\frac{M \ln c}{\epsilon^3}\right) + O\left(\frac{M}{\epsilon^2}\right) \\ &= O\left(M \ln M + \frac{M \ln c}{\epsilon^3} + \frac{M}{\epsilon^2}\right). \qquad \square \end{aligned}$$

Remark. The root $\theta(f)$ can often be computed only approximately, but an accuracy of $O(\epsilon^2/M)$ for $\theta(f)$ is sufficient such that the iteration bounds remain valid. With this required accuracy, the number of evaluations of the sum $\sum_{m=1}^{M} \frac{1}{f_m - \theta}$ is bounded by $O(\ln(M/\epsilon))$. This gives $O(M(\ln M/\epsilon))$ arithmetic operations to determine $\theta(f)$ approximately. The overhead can be further improved by using Newton's method to $O(M(\ln \ln(M/\epsilon)))$ [23, 24].

3. General linear programming approach. In this section we study the preemptive resource constrained scheduling problem. First we consider the case with an unlimited number of machines $m \geq n$. In fact, if $m \leq n$, the machines can be handled as the $(s+1)$st resource with requirement $r_{s+1,j} = 1$ and capacity $c_{s+1} = m$. For our scheduling problem, a *configuration* is a compatible (or feasible) subset of tasks that can be scheduled simultaneously. Let F be the set of all configurations, and for every $f \in F$, let x_f denote the length (in time) of configuration f in the schedule. Clearly, $f \in F$ iff $\sum_{j \in f} r_{ij} \leq c_i$ for $i = 1, \ldots, s$.

By using these variables, the problem of finding a preemptive schedule of the tasks with smallest makespan value (subject to the resource constraints) can be formulated as the following linear program [30]:

$$\begin{array}{lll} \min & \sum_{f \in F} x_f & \\ \text{s.t.} & \sum_{f \in F: j \in f} x_f \;\geq\; p_j, & j = 1, \ldots, n, \\ & x_f \;\geq\; 0 & \forall f \in F. \end{array} \tag{3.1}$$

One can solve (3.1) by using binary search on the optimum value and testing at each stage the feasibility of the following linear system for a given $r \in [p_{max}, np_{max}]$:

$$\sum_{f \in F: j \in f} x_f \;\geq\; p_j, \quad j = 1, \ldots, n, \quad (x_f)_{f \in F} \in P,$$

where

$$P = \left\{ (x_f)_{f\in F} \ : \ \sum_{f\in F} x_f \ = \ r, \quad x_f \geq 0, \ f \in F \right\}.$$

This can be done approximately (hence leading to an approximate decision procedure) by computing an approximate solution for the following max-min resource sharing problem:

$$\lambda^* = \max \left\{ \lambda \ : \ \sum_{f\in F: j\in f} \frac{1}{p_j} \cdot x_f \ \geq \ \lambda, \ j = 1, \ldots, n, \quad (x_f)_{f\in F} \ \in \ P \right\}. \tag{3.2}$$

The latter problem can also be viewed as a fractional covering problem with one block P, and n coupling constraints. Let the coupling (covering) constraints be represented by $Ax \geq \lambda e$. By using the approach presented in section 2, problem (3.2) can be solved approximately in $O(n(\delta^{-2} + \delta^{-3} \ln c + \ln n))$ iterations (coordination steps), each requiring for a given n-vector $y = (y_1, \ldots, y_n)$ a $\frac{1}{c}$-approximate solution of the problem

$$\Lambda(y) \ = \ \max \{ \ y^T Ax \ : \ x \in P \ \}. \tag{3.3}$$

Since P in (3.3) is just a simplex, the optimum of this linear program is also attained at a vertex $\tilde{x}$ of P corresponding to a (single) configuration $\tilde{f}$. A similar argument was used for the bin packing problem by Plotkin, Shmoys, and Tardos [47]. At this vertex $\tilde{x}_{\tilde{f}} = r$ and $\tilde{x}_f = 0$ for $f \neq \tilde{f}$. Therefore, it suffices to find a subset $\tilde{f}$ of tasks that can be executed in parallel and has the largest associated profit value $c_{\tilde{f}}$ in the profit vector $c^T = y^T A$. But for given multipliers $y_1, \ldots, y_n$, this problem can also be formulated as

$$\max \left\{ \sum_{j\in f} \frac{y_j}{p_j} \ : \ f \in F \right\},$$

or, equivalently, as a general s-dimensional knapsack problem (sD-KP) or packing integer program (PIP),

$$\begin{array}{ll} \max & \sum_{j=1}^n \frac{y_j}{p_j} x_j \\ \text{s.t.} & \sum_{j=1}^n r_{ij} x_j \ \leq \ c_i, \qquad i = 1, \ldots, s, \\ & \qquad\qquad x_j \in \{0, 1\}, \qquad j = 1, \ldots, n. \end{array} \tag{3.4}$$

Let $K(n, s, c)$ denote the time required (in the worst case) to compute a $\frac{1}{c}$-approximate solution for (3.4). At each iteration, in addition to solving (3.4) (approximately), we also need to compute the new y vector based on Ax for the current x. Though the dimension of x is exponential, the computation requires only updating the previous Ax value, since the current x is $(1 - \tau)x + \tau\hat{x}$ (for an appropriate step length $\tau \in (0, 1]$), where $\hat{x}$ is the vertex of P corresponding to the solution of (3.4) at the current iteration. Thus the number of nonzero components of x can increase by at most one at each iteration, and each update of Ax takes $O(n)$ operations.

Initially, x^0 has at most n nonzero components obtained from solving n subproblems (one for each n-dimensional unit vector as y) requiring $O(nK(n, s, c))$ time, and

computing the initial y^0 in $O(n^2)$ time. Approximating the root and determining the next price vector p can be done in $O(n \ln \ln \frac{n}{\delta}) = O(n^2)$ time (for, e.g., $\delta \geq 1/n$). Later each update of Ax^k can be done in $O(n)$ time. For any fixed r, the algorithm requires $O(n(\delta^{-2} + \delta^{-3} \ln c + \ln n)(K(n,s,c) + n \ln \ln(n\delta^{-1})))$ time.

By binary search on r one can obtain a solution $(x_f)_{f \in F}$ with $\sum_{f \in F} x_f = (1+\epsilon/4)r^*$ and $\sum_{f \in F: j \in f} x_f \geq \frac{1}{c}(1-\delta)p_j$, where r^* is the length of an optimal schedule. Now one can define $\tilde{x}_f = x_f c(1+4\delta)$ and obtain $\sum_{f \in F: j \in f} \tilde{x}_f \geq (1-\delta)(1+4\delta)p_j = (1+3\delta-4\delta^2)p_j \geq p_j$ for $\delta \leq 3/4$. In this case the length of the generated schedule is at most $cr^*(1+4\delta)(1+\epsilon/4) = cr^*(1+4\delta+\epsilon/4+\delta\epsilon) \leq cr^*(1+\epsilon)$ by choosing $\epsilon \leq 1$ and $\delta \leq 3\epsilon/20$. Since the optimum of (3.1) lies within interval $[p_{max}, np_{max}]$, the overall complexity of the algorithm can be bounded by $O(n \ln \frac{n}{\epsilon}(\epsilon^{-2} + \epsilon^{-3} \ln c + \ln n)(K(n,s,c) + n \ln \ln(n\epsilon^{-1})))$ time. For $K(n,s,c) \geq O(n \ln \ln \frac{n}{\epsilon})$ we obtain $O(n \ln(n\epsilon^{-1})(\epsilon^{-2} + \epsilon^{-3} \ln c + \ln n)K(n,s,c))$ time.

The number of iterations can be improved by computing an approximate nonpreemptive schedule with a greedy algorithm. The main idea is to use a modified list scheduling algorithm. The classical list scheduling algorithm is defined as follows. First consider the tasks in any fixed order $L = (T_{i_1}, \ldots, T_{i_n})$. At any time if there are positive quantities available from all resources, the algorithm scans L from the beginning and selects the first task T_k (if there is any) which may validly be executed and which has not been already (or is not currently) executed. If a task is finished, it will be removed from the list. Garey and Graham [20] showed that this list scheduling algorithm for nonpreemptive tasks gives an $(s+1)$-approximation ratio (comparing the length of the produced schedule and the optimum nonpreemptive schedule). To compare this with the optimal preemptive makespan C^*_{max}, we allow overpacking of the resources with one task at each time. Let $C_{max}(H)$ be the length of this (infeasible) pseudoschedule and consider a task T_k that is finished at time $C_{max}(H)$. Then for each time $t \in [0, C_{max}(H) - p_k)$ at least one resource is completely used by the tasks. Let $l(i)$ be the total length of intervals where resource i is overpacked. Clearly, we have $l(i) \leq C^*_{max}$ and $p_k \leq p_{max} \leq C^*_{max}$. The length of the pseudoschedule is at most $\sum_{i=1}^{s} l(i) + p_k \leq (s+1)C^*_{max}$. By replacing the overpacked tasks at the end we obtain a feasible schedule of length $C^{(MLS)}_{max} \leq (2s+1)C^*_{max}$. This implies that $C^*_{max} \leq C^{(MLS)}_{max} \leq (2s+1)C^*_{max}$, i.e., $1/(2s+1)C^{(MLS)}_{max} \leq C^*_{max} \leq C^{(MLS)}_{max}$. Hence the binary search for the optimum of (3.1) requires only $O(\ln \frac{s}{\epsilon})$ steps (instead of $O(\ln \frac{n}{\epsilon})$) improving the previous running time to

$$O\left(n\left(K(n,s,c) + n \ln \ln \frac{n}{\epsilon}\right) \min(\ln(s\epsilon^{-1}), \ln(n\epsilon^{-1}))(\epsilon^{-2} + \epsilon^{-3} \ln c + \ln n)\right).$$

If the block problem possesses an approximation scheme, then the factor $\epsilon^{-3} \ln c$ can be removed. As main result we obtain the following theorem.

THEOREM 3.1. *Let $\mathcal{I}$ be a set of instances of the preemptive resource constrained scheduling problem. If there is a polynomial-time approximation algorithm for the corresponding s-dimensional knapsack instance with ratio c, then for any $\epsilon > 0$ there is a polynomial-time algorithm for preemptive resource constrained scheduling restricted to $\mathcal{I}$ with approximation ratio $c(1+\epsilon)$.*

The number of configurations in the final solution can be reduced from $O(n(\epsilon^{-2} + \epsilon^{-3} \ln c + \ln n))$ to $O(n)$ within $O(n(\epsilon^{-2} + \epsilon^{-3} \ln c + \ln n)\mathcal{M}(n))$ time, where $\mathcal{M}(n)$ is the time to invert an $(n \times n)$ matrix. The main idea is to consider iteratively systems of equalities with $n+1$ variables and n equalities and to eliminate one variable in each iteration. The maximum number of tasks per configuration is bounded by

$t = \min(n, \min_{i=1}^{s} \frac{c_i}{min_j r_{ij}})$. Therefore, the number of preemptions can be bounded by $O(nt)$.

4. Approximability as a function of the number of resources. As we have seen in the previous section, the sD-KP is a key subproblem in our approach whose solution (for various inputs) is required repeatedly. It is well known that the approximability of this problem varies with the dimension s. Therefore in this section we will specialize the above general result by making different assumptions on s and using different approximation algorithms for the sD-KP. In particular, we will obtain a sequence of approximation results for our scheduling problem where the approximation will improve (constant, PTAS, and then FPTAS) as we move from an arbitrary to a fixed number of resources and eventually to the case with a single resource. To contrast these approximation algorithms, we will also present some inapproximability results for the first two variants.

4.1. Arbitrary number of resources. In this section we consider the case when s is arbitrary, i.e., when it is part of the input. First we give the presentation of our approximation algorithms, and then we briefly discuss some simple inapproximability results.

4.1.1. Approximation algorithms. It is known [48, 53] that for general s, the sD-KP or, equivalently, the PIP has an $\Omega(1/s^{1/c_{\min}})$ approximation algorithm when all $r_{ij} \in [0,1]$, $c_{\min} = \min_i c_i \geq 1$, and $\frac{y_j}{p_j} \geq 0$. This implies the following result.

THEOREM 4.1. *For any number s of resources, there is a polynomial-time approximation algorithm with performance ratio $O(s^{\frac{1}{c_{\min}}})$ for the preemptive resource constrained scheduling problem.*

This result can be further improved by using the algorithm by Srinivasan [53]. Furthermore, Srivastav and Stangier [54, 55] showed that if $c_{\min} \geq \frac{16}{\epsilon^2}\log(2s)$ and $OPT \geq 12/\epsilon^2$ (where OPT is the optimum value of the linear relaxation of (3.4)), an ϵ-approximate solution for the sD-KP can be computed in polynomial time. The running time of the algorithm is bounded by $O(K_r(n,s,1)+sn^2\ln(sn))$, where $K_r(n,s,1)$ is the time required to solve (exactly) the linear programming relaxation of (3.4). Combining these with our approach presented in the previous section and extending the algorithm to arbitrary OPT, we obtain the following.

THEOREM 4.2. *For any $\epsilon > 0$ and any number s of resources, if $c_i \geq \frac{12}{\epsilon^2}\log(2s)$ and $r_{ij} \in [0,1]$ for each i and j, there is a polynomial-time approximation algorithm that computes a $(1+\epsilon)$-approximate solution for the preemptive resource constrained scheduling problem.*

The last two results can also be generalized to $P|res\ldots,pmtn|C_{max}$ with a limited number of machines, i.e., when $m \leq n$. Note that Theorems 4.1 and 4.2 hold only under some special conditions on resource capacities and requirements. Therefore it is natural to ask whether they can be eliminated at least when s is fixed. After presenting some inapproximability results, we will show in section 4.2 that if s is fixed, though the problem remains NP-hard, approximating its optimum becomes much easier. In particular we prove that for any fixed s the general problem has a PTAS.

4.1.2. Inapproximability. For any graph $G = (V,E)$ one can construct a resource constrained scheduling problem with $n = |V|$ tasks and $s = |E|$ resources [6, 49], where vertices correspond to tasks and edges to resources in the following way: The resource capacities are all 1, i.e., $c_e = 1$ for each $e \in E$, while the resource

requirement $r_{ev} = 1$, if $v \in e$, and 0 otherwise. Independent sets of vertices correspond to sets of tasks that can be executed together at the same time; therefore the (fractional) coloring problem for graphs can be viewed as a special case of (preemptive) resource constrained scheduling. Hence the inapproximability results in [42, 17] imply the following.

THEOREM 4.3. *For any $\delta > 0$, the preemptive resource constrained scheduling problem with n tasks and s resources has no polynomial-time approximation algorithm with approximation ratio $n^{1-\delta}$, neither for some $\delta > 0$, unless $P = NP$, nor for any $\delta > 0$, unless $ZPP = NP$.*

Note that this negative result holds even for the restricted case when each processing time is of unit length, and all capacities and resource requirements are either 0 or 1. This shows that for arbitrary s not only is the problem hard, but even approximating its optimum is difficult. Using that $s \leq n^2$ in the special case above we get the following corollary.

COROLLARY 4.4. *The preemptive resource constrained scheduling problem with n tasks and s resources has no polynomial-time approximation algorithm with approximation ratio $s^{1/2-\delta}$, neither for some $\delta > 0$, unless $P = NP$, nor for any $\delta > 0$, unless $ZPP = NP$.*

4.2. Fixed number of resources—PTAS. In this section we study how the approximability of the problem changes under a restricting assumption on the number of resources. We consider here the case when $s \geq 1$ is a fixed constant larger than one. As we argue below, this restriction allows us to prove a substantially better approximation result than the one above. Namely, we will show that under the discussed assumption, the problem possesses a PTAS, and then we will also prove that, in fact, this is the best one can expect (unless P = NP).

4.2.1. Approximation algorithms. It is known [43, 19] that for any fixed s, the sD-KP has a PTAS. Let $K(n, s, \delta)$ denote the time required (in the worst case) to compute a δ-approximate solution for the sD-KP. Using that s is constant, the running time of our scheduling algorithm is bounded by $O((K(n, s, c) + n \ln\ln(n\epsilon^{-1}))n \ln(\epsilon^{-1})(\epsilon^{-2} + \ln n))$. The currently known best bound for $K(n, s, \Theta(\epsilon))$ is $O(n^{\lfloor \frac{s}{\epsilon} \rfloor - s}) = n^{O(\frac{s}{\epsilon})}$ [7]. By using this bound and the above argument, we obtain the following result.

THEOREM 4.5. *For any fixed number s of resources, there is a PTAS for the preemptive resource constrained scheduling problem with running time $n^{O(\frac{s}{\epsilon})}$.*

Notice that for fixed s, there is also an $O(n)$ time $(s+1)$-approximation algorithm for the sD-KP [7], which implies the following.

COROLLARY 4.6. *There is an $(s+1)(1+\epsilon)$-approximation algorithm for the preemptive resource constrained scheduling problem with running time $O((n^2 \ln\ln(n\epsilon^{-1})) \cdot \ln(\epsilon^{-1})(\epsilon^{-2} + \ln n))$.*

4.2.2. Inapproximability. The running time of the previously described algorithm depends exponentially on the accuracy, and as the next result shows this dependence cannot be improved to polynomial unless P = NP.

THEOREM 4.7. *For any $s \geq 2$, there is no FPTAS for the preemptive resource constrained scheduling problem with s resources unless $P = NP$.*

Proof. We use a reduction from the NP-complete partition problem: Given a set A and a size $s(a) \in \mathbb{N}$ for each $a \in A$, where $n = |A|$ is assumed to be even, decide whether there is a subset I of A such that $|I| = n/2$ and $\sum_{a \in I} s(a) = \frac{1}{2} \sum_{a \in A} s(a)$. W.l.o.g. we may assume that $s(a') \leq \frac{1}{2} \sum_{a \in A} s(a)$ for any $a' \in A$. Let $s_{max} =$

$\max_{a\in A} s(a)$. Now construct n tasks and two resources with capacities $\frac{1}{2}\sum_{a\in A} s(a)$ and $\frac{1}{2}\sum_{a\in A}(s_{max} - s(a))$, where each task $a \in A$ requires $(s(a), s_{max} - s(a))$ of the two resources and has processing time $p_a = 1$. If there is a solution I of the partition problem, then $|I| = n/2$, $\sum_{a\in I} s(a) = \frac{1}{2}\sum_{a\in A} s(a)$, and $\sum_{a\in I}(s_{max} - s(a)) = \frac{1}{2}\sum_{a\in A}(s_{max} - s(a))$. This means that set I can be executed in parallel on both resources. Furthermore, the set $A \setminus I$ is also a solution for the partition problem and can be executed also parallel on both resources. Therefore, one can schedule all tasks in two phases in a nonpreemptive way: in one phase all tasks are in I (of length 1), and in the other phase all tasks are in $A\setminus I$ (also of length 1). This gives a schedule with makespan $C_{max} = 2$ and the minimum makespan is $C^*_{max} = 2$ (by using an argument based on the required minimum area for all tasks). If there is no solution of the partition problem, then we can still split the set in three parts $I_1, \ldots, I_3$ according to resource 1 such that $\sum_{a\in I_j} s(a) \le \frac{1}{2}\sum_{a\in A} s(a)$. Now only one of these parts can have $\sum_{a\in I_j}(s_{max} - s(a)) > \frac{1}{2}\sum_{a\in A}(s_{max} - s(a))$. By splitting this set (according to resource 2) into three parts, we obtain a feasible nonpreemptive schedule of length $C_{max} \le 5$. This implies that $C^*_{max} \le 5$. Assume now that there is an FPTAS for the preemptive 2-resource constrained scheduling problem and then show that this leads to a contradiction. The FPTAS gives for each $\epsilon > 0$ a $poly(n, 1/\epsilon)$ time algorithm to obtain a schedule with length $\le C^*_{max}(1+\epsilon) \le C^*_{max} + 5\epsilon$. If we choose $\epsilon = 1/5n$, then we obtain in $poly(n)$ time a preemptive schedule with length $\le C^*_{max} + 1/n$. The length of the preemptive schedule (given by the FPTAS with $\epsilon = 1/5n$) is larger than $2+1/n$ iff the partition problem has no solution. If the length of the schedule is larger than $2+1/n$, then $C^*_{max} > 2$ implying that we have a no-instance of the partition problem. Consider the other direction: For each time step t there are at least $n/2+1$ tasks which are not executed at step t; otherwise we have a solution of the partition problem. To see this consider a set I of $n/2$ tasks executed at one time step. This implies that $\sum_{a\in I} s(a) \le 1/2\sum_{a\in A} s(a)$ and $\sum_{a\in I}(s_{max} - s(a)) \le \frac{1}{2}\sum_{a\in A}(s_{max} - s(a))$. Both of these inequalities can be transformed into

$$(n/2)s_{max} - \sum_{a\in I} s(a) \le \sum_{a\in I}(s_{max} - s(a)) \le \frac{1}{2}\sum_{a\in A}(s_{max} - s(a)) = (n/2)s_{max} - \sum_{a\in I} s(a).$$

Then, $\sum_{a\in I} s(a) = \frac{1}{2}\sum_{a\in A} s(a)$ and $\sum_{a\in I}(s_{max} - s(a)) = \frac{1}{2}\sum_{a\in A}(s_{max} - s(a))$. Therefore, I is a solution of the partition problem.

Let $ne(i)$ be the total length in interval $[0, 2]$ where task T_i is not executed. Using the property above, $\sum_{j=1}^{n} ne(i) \ge 2(n/2 + 1) = n + 2$. This implies that at least one task T_k has $ne(k) \ge 1 + 2/n$. Therefore, this task is executed at most $1 - 2/n$ in interval $[0, 2]$ and the schedule length is at least $2 + 2/n$. This argument implies that we can test (using the FPTAS) the existence of a solution for the partition problem in polynomial time, which is impossible unless P = NP. □

In the proof above we have used an idea of Korte and Schrader [37]. They proved that there is no FPTAS for the sD-KP with $s = 2$ unless P = NP. Since it was essential in the proof of Theorem 4.7 that s is (at least) 2, it is natural to ask again what happens when $s = 1$. In this case, as will be demonstrated in the next section, there is an FPTAS for the problem, and hence the negative result of Theorem 4.7 no longer holds.

4.3. Single resource—FPTAS. Clearly, the general approach presented in section 4.2 can also be used for the special case when there is only one resource. Note that the number of iterations in computing an approximate solution for (3.1)

is independent of s, so it remains $O(n(\delta^{-2} + \ln n))$, as above. The only difference is that the subproblem one has to solve (approximately) at each iteration becomes the classical (one-dimensional) knapsack problem (instead of the s-dimensional variant). This can be solved approximately with any $\Theta(\delta)$ accuracy in $O(n \min(\ln n, \ln(1/\delta)) + 1/\delta^2 \min(n, 1/\delta \ln(1/\delta))) = O(n\delta^{-2})$ time [34]. In addition, we have to count the overhead of $O(n \ln \ln(n\epsilon^{-1}))$ operations in each iteration (i.e., the computation of the root and the new price vector). Hence the previous bound can be substituted for $K(n, s, \Theta(\delta))$ in the analysis above in section 4.2, and therefore for any fixed r the procedure requires $O(n^2 \max(\delta^{-2}, \ln \ln(n\delta^{-1}))(\delta^{-2} + \ln n))$ time (including also the overheads arising from computing the initial solution).

Similarly to the discussion in section 4.2, one can use binary search on r to find a good approximation for the optimum of (3.1). The initial interval for r can be determined by a strip packing algorithm (called longest task first) [8, 57] that computes a nonpreemptive schedule of length at most three times the length of the an optimal preemptive schedule. These all imply the following.

THEOREM 4.8. *If there is only one resource, the resource constrained scheduling problem has an FPTAS which runs in* $O(n^2 \ln(\epsilon^{-1}) \max(\epsilon^{-2}, \ln \ln(n\epsilon^{-1}))(\epsilon^{-2} + \ln n))$ *time.*

So far we have assumed that m is sufficiently large (e.g., $m \geq n$), or otherwise processors can be treated as an extra resource. But having seen above the dividing line (regarding approximability) between instances with one and two resources, one may naturally ask how easy or difficult it is to compute approximate solutions for the problem when there is one resource and a limited number of machines. Krause, Shen, and Schwetman [38] gave a polynomial-time $(3 - \frac{1}{n})$-approximation algorithm for the problem. This can be substantially improved by following our approach and extending Theorem 4.8 to this variant. First formulate it as a restricted preemptive 2-resource constrained scheduling problem, where the m identical machines correspond to the second resource with $r_{2j} = 1$ for each task j and capacity $c_2 = m$. It is easy to check that the subproblem in this case is the cardinality constrained ($\sum_{j=1}^{n} x_j \leq m$) knapsack problem, which has an FPTAS with running time $O(nm^2\epsilon^{-1})$ [7]. In addition, the initial interval for the binary search on r can be bounded as for $s = 2$ resources. Hence the following holds.

THEOREM 4.9. *There is an FPTAS of running time* $O(n^2 \ln(\epsilon^{-1}) \max(m^2\epsilon^{-1}, \ln \ln(n\epsilon^{-1})(\epsilon^{-2} + \ln n)))$ *for* $P|res1.., pmtn|C_{max}$.

This result can also be extended to the variant $P|res\,1.., r_j, pmtn|L_{max}$, where the input contains release r_j dates and delivery q_j dates for each task T_j, and the objective is to find a schedule minimizing the maximum delivery completion time $L_{max} = \max_j C_j + q_j$ [27].

THEOREM 4.10. *There is an FPTAS for* $P|res\,1.., r_j, pmtn|L_{max}$ *that runs in* $poly(n, 1/\epsilon)$ *time.*

5. Multiprocessor task scheduling. In this section we address preemptive multiprocessor task scheduling problems [13], where a set $\mathcal{T} = \{T_1, \ldots, T_n\}$ of n tasks has to be executed by m processors such that each processor can execute at most one task at a time and a task must be processed simultaneously by several processors.

Since we consider here the preemptive model, each task can be interrupted any time at no cost and restarted later possibly on a different set of processors. We will focus on those preemptive schedules where migration is allowed, that is, where each task may be assigned to different processor sets during different execution phases [4, 13, 14]. The *malleable* variant of multiprocessor task scheduling, $P|fctn_j, pmtn|C_{max}$,

can be formulated as the following linear program [30], where M_j denotes the set of different cardinalities that processor sets executing task T_j can have:

$$\begin{array}{lll} \min & \sum_{f\in F} x_f & \\ \text{s.t.} & \sum_{\ell\in M_j} \frac{1}{p_j(\ell)} \sum_{f\in F:|f^{-1}(j)|=\ell}, & x_f \geq 1, \quad j=1,\ldots,n, \\ & & x_f \geq 0 \quad \forall f \in F. \end{array} \tag{5.1}$$

Here the goal is to find for a given r a vector $x \in P = \{(x_f) : \sum_{f\in F} x_f = r, x_f \geq 0, f \in F\}$ that satisfies all the other constraints in (5.1). This corresponds to a vector $x \in P$ such that $Ax \geq 1 - \delta$. Again, we get a subroutine to find a vertex $\tilde{x}$ in P such that $c^T\tilde{x} \geq c^T x$ for all $x \in P$, where $c = y^T A$. For each task T_j we have now different values in M_j and hence in the corresponding knapsack problem the profit $y_j/p_j(\ell)$ depends on the cardinality $\ell \in M_j$, while the capacity of the knapsack remains m, as before. The subroutine corresponds now to a generalized knapsack problem with different choices for tasks (items). The problem we have to solve (approximately) for a given n-vector $(y_1, \ldots, y_n)$ can be formulated as follows:

$$\begin{array}{lll} \max & \sum_{j=1}^{n} \sum_{\ell\in M_j} \frac{y_j}{p_j(\ell)} \cdot x_{j\ell} & \\ \text{s.t.} & \sum_{j=1}^{n} \sum_{\ell\in M_j} \ell \cdot x_{j\ell} \leq m, & \\ & \sum_{\ell\in M_j} x_{j\ell} \leq 1, \quad j=1,\ldots,n, & \\ & x_{j\ell} \in \{0,1\}, \quad \ell \in M_j, \quad j=1,\ldots,n. & \end{array} \tag{5.2}$$

In fact, this is the multiple-choice knapsack problem. For this problem, Lawler [40] showed that an ϵ-approximate solution can be computed in $O(\sum_j |M_j| \ln |M_j| + \sum_j |M_j| n/\epsilon) = O(nm \ln m + n^2 m/\epsilon)$ time. In order to obtain a lower bound, one can compute $d_j = \min_{1\leq\ell\leq m} p_j(\ell)$ and $d_{max} = \max_j d_j$. Then $d_{max} \leq OPT \leq n d_{max}$. In this case, the overhead $O(n \ln \ln(n/\epsilon)) = O(n \ln \ln n + n \ln \ln(1/\epsilon)) = O(n^2 + n\epsilon^{-1})$ is less than the running time required by the knapsack subroutine. Hence by using an argument similar to the one in the previous section, one can obtain the following result.

THEOREM 5.1. *There exists an FPTAS for $P|fctn_j, pmtn|C_{max}$ whose running time is bounded by $O(n(\epsilon^{-2} + \ln n) \ln(n\epsilon^{-1})(nm \ln m + n^2 m\epsilon^{-1}))$.*

Other variants of $P|fctn_j, pmtn|C_{max}$ concern preemptive scheduling on parallel processors, where the underlying interconnection network is not completely disregarded [11, 50, 59]. (Note that in the original formulation, we assumed nothing about the network architecture.) Based on the above results and a few other ideas (such as strip packing as a scheduling problem with consecutive processors and greedy packing of tasks on hypercubes) the following can be shown.

THEOREM 5.2. *If the processors are arranged in a line or hypercube network, $P|fctn_j, pmtn|C_{max}$ has an FPTAS that runs in $O(n(\epsilon^{-2} + \ln n) \ln(n\epsilon^{-1})(nm \ln m + n^2 m\epsilon^{-1}))$ time.*

6. Weighted fractional coloring. Let $G = (V, E)$ be a graph with a positive weight w_v for each vertex $v \in V$. Let $\mathcal{I}$ be the set of all independent sets of G. The weighted fractional coloring problem consists of assigning a nonnegative real value x_I to each independent set I of G such that each vertex $v \in V$ is completely covered by independent sets containing v (i.e., the sum of their values is at least w_v) and the total value $\sum_I x_I$ is minimized. This problem can also be formulated as a linear program of the form (3.1). Similarly to section 3, this linear program can be solved approximately by using binary search on the optimum value r^* and computing at each stage for the

current r an approximate solution for a fractional covering problem of form (3.2). Let $w_{\max} = \max_{v \in V} w_v$ be the maximum weight of a vertex. By binary search, one can obtain a solution $(x_I)_{I \in \mathcal{I}}$ with $\sum_{I \in \mathcal{I}} x_I = (1+\epsilon/4)r^*$ and $\sum_{I \in \mathcal{I}: v \in I} x_I \geq 1/c(1-\delta)w_v$. Now one can define $\tilde{x}_I = x_I c(1+4\delta)$ and obtain $\sum_{I \in \mathcal{I}: v \in I} \tilde{x}_I \geq (1-\delta)(1+4\delta)w_v = (1+3\delta-4\delta^2)p_j \geq p_j$ for $\delta \leq 3/4$. In this case, the length of the generated fractional coloring is at most $cr^*(1+4\delta)(1+\epsilon/4) = cr^*(1+4\delta+\epsilon/4+\delta\epsilon) \leq cr^*(1+\epsilon)$ by choosing $\epsilon \leq 1$ and $\delta \leq 3\epsilon/20$. Since the optimum lies within the interval $[w_{\max}, nw_{\max}]$, the overall complexity of the algorithm can be bounded by $O((n \ln n + n \ln c/\epsilon^3 + n/\epsilon^2)(WIS(G,n,c,d)+n \ln \ln(n/\epsilon)) \ln(n\epsilon^{-1}))$, where $WIS(n,c,d)$ is the time required to compute an approximate weighted independent set for a weighted graph (G,w). The above arguments imply the following result.

THEOREM 6.1. *Let $\mathcal{G}$ be a graph class. If there is a polynomial-time algorithm for the weighted independent set problem restricted to graphs $G \in \mathcal{G}$ with approximation ratio $1/c$ for $c \geq 1$, then for any $\epsilon > 0$ there is a polynomial-time algorithm for the fractional weighted coloring problem restricted to $\mathcal{G}$ with approximation ratio $c(1+\epsilon)$.*

COROLLARY 6.2. *Let $\mathcal{G}$ be a graph class. If there is an (F)PTAS for the computation of the weighted independent set in a graph $G \in \mathcal{G}$ and weights w, then we obtain an (F)PTAS for the fractional weighted coloring problem for graphs $G \in \mathcal{G}$.*

Using a recent result [16] for computing the maximum weighted independent set in intersection graphs of disks in the plane, we obtain the following.

COROLLARY 6.3. *There is a PTAS for the computation of the fractional weighted chromatic number for intersection graphs of disks in the plane.*

Since this graph class contains planar graphs and unit disk graphs, Corollary 6.2 implies the following result which also provides a substantial improvement on Matsui's polynomial-time 2-approximation algorithm [44] for unit disk graphs.

COROLLARY 6.4. *There is a PTAS for the computation of the fractional weighted chromatic number for planar and unit disk graphs.*

7. Conclusion. In this paper we have studied preemptive variants of resource constrained scheduling and the closely related fractional coloring problem. The approach we presented is based on linear programming formulations with exponentially many variables but with special structures allowing efficient approximations. The linear programs are solved (approximately) in an iterative way as covering problems, where at each iteration subproblems of the same type have to be solved. Interestingly, for resource constrained scheduling these subproblems turned out to be knapsack-type problems (multiple-choice, multidimensional, and cardinality constrained knapsack) with efficient approximation algorithms. For fractional coloring, it is the well-known maximum weighted independent set problem.

For some of the subproblems we have encountered, there are only relatively weak polynomial-time approximation results (i.e., with constant, logarithmic, or even worse approximation ratios). To handle these cases, too, we have extended some of the methods in [24, 47, 58] to the case where the subproblem can be solved only approximatively. The underlying algorithm is independent from the width [47] and the number of variables. We note that by using other techniques [32] (via the ellipsoid method and approximate separation) with higher running time the ratio $c(1+\epsilon)$ in Theorems 3.1 and 6.1 can be improved to ratio c. Recently, Jansen [33] proposed an improved algorithm for the max-min resource sharing problem that needs at most $O(M(\ln M + \ln(1/\epsilon)/\epsilon^2))$ iterations.

We mention in closing that by using the same approach, similar approximation results can be expected for various other preemptive scheduling and fractional graph

problems, e.g., for fractional path coloring, call scheduling, bandwidth allocation, and scheduling multiprocessor tasks on dedicated processors, as well as open, flow, and job shop scheduling.

Acknowledgments. The authors thank R. Schrader and A. Srivastav for helpful comments on the complexity of the two-dimensional knapsack problem and the approximation of sD-KP, respectively.

REFERENCES

[1] A. K. Amoura, E. Bampis, C. Kenyon, and Y. Manoussakis, *Scheduling independent multiprocessor tasks*, Algorithmica, 32 (2002), pp. 247–261.

[2] B. S. Baker, E. G. Coffman, Jr., and R. L. Rivest, *Orthogonal packings in two dimensions*, SIAM J. Comput., 9 (1980), pp. 846–855.

[3] J. Blazewicz, W. Cellary, R. Slowinski, and J. Weglarz, *Scheduling under Resource Constraints—Deterministic Models*, Ann. Oper. Res. 7, Baltzer, Basel, 1986.

[4] J. Blazewicz, M. Drabowski, and J. Weglarz, *Scheduling multiprocessor tasks to minimize schedule length*, IEEE Trans. Computers, 35 (1986), pp. 389–393.

[5] J. Blazewicz, K. H. Ecker, E. Pesch, G. Schmidt, and J. Weglarz, *Scheduling in Computer and Manufacturing Systems*, Springer-Verlag, Berlin, 1996.

[6] J. Blazewicz, J. K. Lenstra, and A. H. G. Rinnooy Kan, *Scheduling subject to resource constraints: Classification and complexity*, Discrete Appl. Math., 5 (1983), pp. 11–24.

[7] A. Caprara, H. Kellerer, U. Pferschy, and D. Pisinger, *Approximation algorithms for knapsack problems with cardinality constraints*, European J. Oper. Res., 123 (2000), pp. 333–345.

[8] E. G. Coffman, Jr., M. R. Garey, D. S. Johnson, and R. E. Tarjan, *Performance bounds for level-oriented two-dimensional packing algorithms*, SIAM J. Comput., 9 (1980), pp. 808–826.

[9] A. K. Chandra, D. S. Hirschberg, and C. K. Wong, *Approximate algorithms for some generalized knapsack problems*, Theoret. Comput. Sci., 3 (1976), pp. 293–304.

[10] C. Chekuri and S. Khanna, *On multidimensional packing problems*, SIAM J. Comput., 33 (2004), pp. 837–851.

[11] G. I. Chen and T. H. Lai, *Scheduling independent jobs on hypercubes*, in Proceedings of the 5th Symposium on Theoretical Aspects of Computer Science, Lecture Notes in Comput. Sci. 294, Springer-Verlag, New York, 1988, pp. 273–280.

[12] M. Drozdowski, *On the complexity of multiprocessor task scheduling*, Bull. Polish Acad. Sci., 43 (1995), pp. 381–392.

[13] M. Drozdowski, *Scheduling multiprocessor tasks—an overview*, European J. Oper. Res., 94 (1996), pp. 215–230.

[14] J. Du and J. Y.-T. Leung, *Complexity of scheduling parallel task systems*, SIAM J. Discrete Math., 2 (1989), pp. 473–487.

[15] T. Erlebach and K. Jansen, *Conversion of coloring algorithms into maximum weight independent set algorithms*, Discrete Appl. Math., 148 (2005), pp. 107–125.

[16] T. Erlebach, K. Jansen, and E. Seidel, *Polynomial-time approximation schemes for geometric intersection graphs*, SIAM J. Comput., 34 (2005), pp. 1302–1323.

[17] U. Feige and J. Kilian, *Zero knowledge and the chromatic number*, J. Comput. System Sci., 57 (1998), pp. 187–199.

[18] W. Fernandez de la Vega and G. S. Lueker, *Bin packing can be solved within* $1+\epsilon$ *in linear time*, Combinatorica, 1 (1981), pp. 349–355.

[19] A. M. Frieze and M. R. B. Clarke, *Approximation algorithms for the* m*-dimensional* $0-1$ *knapsack problem*, European J. Oper. Res., 15 (1984), pp. 100–109.

[20] M. R. Garey and R. L. Graham, *Bounds for multiprocessor scheduling with resource constraints*, SIAM J. Comput., 4 (1975), pp. 187–200.

[21] M. R. Garey, R. L. Graham, D. S. Johnson, and A. C.-C. Yao, *Resource constrained scheduling as generalized bin packing*, J. Combin. Theory A, 21 (1976), pp. 251–298.

[22] S. Gerke and C. McDiarmid, *Graph imperfection*, J. Combin. Theory B, 83 (2001), pp. 58–78.

[23] M. D. Grigoriadis and L. G. Khachiyan, *Coordination complexity of parallel price-directive decomposition*, Math. Oper. Res., 21 (1996), pp. 321–340.

[24] M. D. Grigoriadis, L. G. Khachiyan, L. Porkolab, and J. Villavicencio, *Approximate max-min resource sharing for structured concave optimization*, SIAM J. Optim., 11 (2001), pp. 1081–1091.

[25] M. Grötschel, L. Lovász, and A. Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 169–197.
[26] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag, Berlin, 1988.
[27] L. Hall and D. Shmoys, *Jackson's rule for single machine scheduling: Making a good heuristic better*, Math. Oper. Res., 17 (1992), pp. 22–35.
[28] O. H. Ibarra and C. E. Kim, *Fast approximation algorithms for the knapsack and sum of subset problem*, J. ACM, 22 (1975), pp. 463–468.
[29] K. Jansen and L. Porkolab, *Linear-time approximation schemes for scheduling malleable parallel tasks*, Algorithmica, 32 (2002), pp. 507–520.
[30] K. Jansen and L. Porkolab, *Computing optimal preemptive schedules for parallel tasks: Linear programming approaches*, Math. Program., 95 (2003), pp. 617–630.
[31] K. Jansen and L. Porkolab, *Preemptive scheduling on dedicated processors: Applications of fractional graph coloring*, J. Sched., 7 (2004), pp. 35–48.
[32] K. Jansen, *Approximate strong separation with application in fractional graph coloring and preemptive scheduling*, Theoret. Comput. Sci., 302 (2003), pp. 239–256.
[33] K. Jansen, *An approximation algorithm for the general max-min resource sharing problem*, Math. Program., 106 (2006), pp. 547–566.
[34] H. Kellerer and U. Pferschy, *A new fully polynomial approximation scheme for the knapsack problem*, J. Combin. Optim., 3 (1999), pp. 59–71.
[35] C. Kenyon and E. Remila, *A near-optimal solution to a two-dimensional cutting stock problem*, Math. Oper. Res., 25 (2000), pp. 645–656.
[36] K. Kilakos and O. Marcotte, *Fractional and integral colourings*, Math. Programming, 76 (1997), pp. 333–347.
[37] B. Korte and R. Schrader, *On the existence of fast approximation schemes*, in Nonlinear Programming 4, Academic Press, New York, 1981, pp. 415–437.
[38] K. L. Krause, V. Y. Shen, and H. D. Schwetman, *Analysis of several task scheduling algorithms for a model of multiprogramming computer systems*, J. ACM, 22 (1975), pp. 522–550.
[39] K. L. Krause, V. Y. Shen, and H. D. Schwetman, Errata: *"Analysis of several task scheduling algorithms for a model of multiprogramming computer systems,"* J. ACM, 24 (1977), p. 527.
[40] E. Lawler, *Fast approximation algorithms for knapsack problems*, Math. Oper. Res., 4 (1979), pp. 339–356.
[41] W. Ludwig and P. Tiwari, *Scheduling malleable and nonmalleable parallel tasks*, in Proceedings of the 5th ACM–SIAM Symposium on Discrete Algorithms, ACM, New York, SIAM, Philadelphia, 1994, pp. 167–176.
[42] C. Lund and M. Yannakakis, *On the hardness of approximating minimization problems*, J. ACM, 41 (1994), pp. 960–981.
[43] O. Oguz and M. J. Magazine, *A Polynomial Time Approximation Algorithm for the Multidimensional* $0-1$ *Knapsack Problem*, Working paper, University of Waterloo, Waterloo, Ontario, Canada, 1980.
[44] T. Matsui, *Approximation algorithms for maximum independent set problems and fractional coloring problems on unit disk graphs*, in Proceedings of the Symposium on Discrete and Computational Geometry, Lecture Notes in Comput. Sci. 1763, Springer-Verlag, New York, 2000, pp. 194–200.
[45] G. Mounie, C. Rapine, and D. Trystram, *Efficient approximation algorithms for scheduling malleable tasks*, in Proceedings of the ACM Symposium on Parallel Algorithms, ACM, New York, 1999, pp. 23–32.
[46] T. Niessen and J. Kind, *The round-up property of the fractional chromatic number for proper circular arc graphs*, J. Graph Theory, 33 (2000), pp. 256–267.
[47] S. A. Plotkin, D. B. Shmoys, and E. Tardos, *Fast approximation algorithms for fractional packing and covering problems*, Math. Oper. Res., 20 (1995), pp. 257–301.
[48] P. Raghavan and C. D. Thompson, *Randomized rounding: A technique for provably good algorithms and algorithmic proofs*, Combinatorica, 7 (1987), pp. 365–374.
[49] M. W. Schäffter, *Scheduling with forbidden sets*, Discrete Appl. Math., 72 (1997), pp. 155–166.
[50] I. Schiermeyer, *Reverse-Fit: A* 2*-optimal algorithm for packing rectangles*, in Proceedings of the 2nd European Symposium of Algorithms, Lecture Notes in Comput. Sci. 855, Springer-Verlag, New York, 1994, pp. 290–299.
[51] E. R. Schreinerman and D. H. Ullman, *Fractional Graph Theory: A Rational Approach to the Theory of Graphs*, Wiley Interscience Series in Discrete Mathematics, John Wiley, New York, 1997.

[52] P. D. Seymour, *Colouring series-parallel graphs*, Combinatorica, 10 (1990), pp. 379–392.
[53] A. Srinivasan, *Improved approximation guarantees for packing and covering integer programs*, SIAM J. Comput., 29 (1999), pp. 648–670.
[54] A. Srivastav and P. Stangier, *Algorithmic Chernoff-Hoeffding inequalities in integer programming*, Random Structures Algorithms, 8 (1996), pp. 27–58.
[55] A. Srivastav and P. Stangier, *Tight approximations for resource constrained scheduling and bin packing*, Discrete Appl. Math., 79 (1997), pp. 223–245.
[56] A. Steinberg, *A strip-packing algorithm with absolute performance bound* 2, SIAM J. Comput., 26 (1997), pp. 401–409.
[57] J. Turek, J. Wolf, and P. Yu, *Approximate algorithms for scheduling parallelizable tasks*, in Proceedings of the 4th ACM Symposium on Parallel Algorithms and Architectures, ACM, New York, 1992, pp. 323–332.
[58] N. E. Young, *Randomized rounding without solving the linear program*, in Proceedings of the 6th ACM–SIAM Symposium on Discrete Algorithms, ACM, New York, SIAM, Philadelphia, 1995, pp. 170–178.
[59] Y. Zhu and M. Ahuja, *On job scheduling on a hypercube*, IEEE Trans. Parallel Distributed Systems, 4 (1993), pp. 62–69.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 564–567

MULTICOLORED PARALLELISMS OF ISOMORPHIC SPANNING TREES*

S. AKBARI†, A. ALIPOUR†, H. L. FU‡, AND Y. H. LO‡

Abstract. A subgraph in an edge-colored graph is multicolored if all its edges receive distinct colors. In this paper, we prove that a complete graph on $2m$ ($m \neq 2$) vertices K_{2m} can be properly edge-colored with $2m-1$ colors in such a way that the edges of K_{2m} can be partitioned into m multicolored isomorphic spanning trees.

Key words. complete graph, multicolored tree, parallelism

AMS subject classifications. 05B15, 05C05, 05C15, 05C70

DOI. 10.1137/S0895480104446015

A *spanning subgraph* of a graph G is a subgraph H with $V(H) = V(G)$. A *proper k-edge coloring* of a graph G is a mapping from $E(G)$ into a set of colors $\{1, \ldots, k\}$ such that incident edges of G receive distinct colors. An *h-total-coloring* of a graph G is a mapping from $V(G) \cup E(G)$ into a set of colors $\{1, \ldots, h\}$ such that (i) adjacent vertices in G receive distinct colors, (ii) incident edges in G receive distinct colors, and (iii) any vertex and its incident edges receive distinct colors. The *edge chromatic number* of a graph G is the minimum number k for which G has a proper k-edge coloring. Throughout this paper K_m and $K_{m,n}$ denote the complete graph of order m and the complete bipartite graph with partite sets of sizes m and n, respectively. It is well known that the edge chromatic number of K_m is m if m is odd, and $m-1$ if m is even [7, p. 15]. Assume that m is a natural number. For any integer i we denote the residue of i modulo m in the set $\{1, \ldots, m\}$ by $[i]_m$. The following result is known.

LEMMA 1 (see [7, p. 16]). *If m is an odd positive integer, then K_m has an m-total coloring.*

A *Latin square* of order m is an $m \times m$ array of m symbols in which every symbol occurs exactly once in each row and column of the array. A *Room square* of side $2m-1$ is a $(2m-1) \times (2m-1)$ array whose cells are empty or contain an unordered pair of distinct integers chosen from $R = \{1, \ldots, 2m\}$, such that the entries of a given row contain every member of R precisely once, and similarly for columns, and the array contains every unordered pair of members of R precisely once. Room squares have been found for all odd $2m-1 \geq 7$ [2, p. 239]. An example of a Room square of side 7 is shown in Table 1.

A subgraph in an edge-colored graph is said to be *multicolored* if no two edges have the same color. Using a Room square of side $2m-1$ one may obtain a proper

*Received by the editors September 12, 2004; accepted for publication (in revised form) January 31, 2006; published electronically June 30, 2006.
http://www.siam.org/journals/sidma/20-3/44601.html

†Institute for Studies in Theoretical Physics and Mathematics, Tehran, Iran, and Department of Mathematical Sciences, Sharif University of Technology, P.O. Box 11365-9415, Tehran, Iran (s_akbari@sharif.edu, alipour@mehr.sharif.edu). The research of the first and second authors was supported by the Institute for Studies in Theoretical Physics and Mathematics (IPM). The research of the first author was in part supported by a grant from IPM (83050211).

‡Department of Applied Mathematics, National Chiao Tung University, Hsinchu, Taiwan 30050 (hlfu@math.nctu.edu.tw, yhlo.am93g@nctu.edu.tw). The research of the third and fourth authors was supported by NSC grant 93-2115-M-009-002.

TABLE 1

			35	17	28	46
	26	48			15	37
	13	57	68	24		
47		16		38		25
58		23	14		67	
12	78			56	34	
36	45		27			18

edge coloring of K_{2m} with $2m-1$ colors in which all edges can be partitioned into $2m-1$ multicolored perfect matchings. For example, using the rows of Table 1 we give a proper edge coloring of K_8 with 7 colors. We denote the vertices of K_8 by $1,\dots,8$. In Table 1, if rs appears in the ith row, then we color the edge rs with color i. For instance, the edges $47, 16, 38, 25$ are colored with color 4. Each column in Table 1 corresponds to a multicolored perfect matching of K_8. In a recent paper [1] the existence of the multicolored matchings in an arbitrary edge-colored complete graph has been studied. A Latin square of order m corresponds to a proper edge coloring of $K_{m,m}$ with m colors. Indeed if $L=(L_{ij})$ is a Latin square of order m and $\{u_1,\dots,u_m\}$ and $\{v_1,\dots,v_m\}$ are two parts of $K_{m,m}$, then we color the edge u_iv_j with L_{ij}. Since L has m symbols, we have an m-edge coloring of $K_{m,m}$, and since every symbol occurs exactly once in each row and each column of L, the edge coloring is proper. Also the existence of two orthogonal Latin squares of order m corresponds to a proper edge coloring of $K_{m,m}$ with m colors for which all edges can be partitioned into m multicolored perfect matchings. For example, suppose that $L=(L_{ij})$ and $R=(R_{ij})$ are two orthogonal Latin squares of order m with symbols of the set $\{1,\dots,m\}$, and $\{u_1,\dots,u_m\}$ and $\{v_1,\dots,v_m\}$ are two parts of $K_{m,m}$. As we saw before, the function c, where $c(u_iv_j)=L_{ij}$, is a proper m-edge coloring of $K_{m,m}$. For any r, $1\le r\le m$, let M_r be the set of all edges u_iv_j such that $R_{ij}=r$. Obviously $\{M_1,\dots,M_n\}$ is an edge partition of $E(K_{m,m})$. Since the symbol r occurs exactly once in each row and each column of R, M_r is a perfect matching, and since L and R are orthogonal, if $R_{ij}=r$, then the symbols L_{ij} are distinct and we conclude that M_r is multicolored. There is a classic result which says that for any natural number m, $m\neq 2,6$, there exist two orthogonal Latin squares of order m; see [3].

We say that the complete graph K_{2m} admits a *multicolored tree parallelism* (MTP) if there exists a proper edge coloring of K_{2m} with $2m-1$ colors for which all edges can be partitioned into m isomorphic multicolored spanning trees. It is clear that the complete graph K_4 does not admit an MTP. We note here that such a partition of the edges of K_{2m} can be viewed as a parallelism as defined in [5] by Cameron, with an additional property due to edge colors. In fact, finding a partition as obtained above corresponds to an arrangement of the edges of K_{2m} into an array of $2m-1$ rows and m columns such that each row contains the edges with the same color which form a perfect matching and the edges in each column form a multicolored spanning tree of K_{2m}; moreover, all the m spanning trees are isomorphic. Therefore, the partition creates a double parallelism of K_{2m}, one from the rows of the perfect matchings and the other from the columns of the edge disjoint isomorphic spanning trees. The following result has been proven in [6].

THEOREM A (see [6]). *If $m\neq 1,3$ and K_{2m} admits an MTP, then for any $r\ge 1$, K_{2^rm} admits an MTP.*

There exist three interesting conjectures on the edge partitioning of the complete graphs into multicolored spanning trees.

TABLE 2

	T_1	T_2	T_3
c_1	35	46	12
c_2	24	15	36
c_3	25	34	16
c_4	26	13	45
c_5	14	23	56

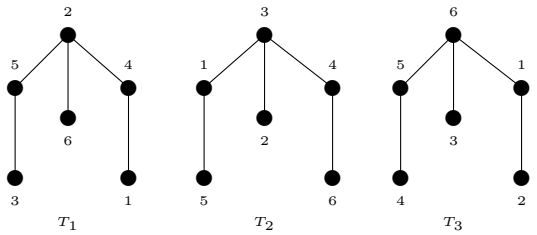


FIG. 1.

CONSTANTINE'S CONJECTURE (weak version; see [6]). *For any natural number m, $m > 2$, K_{2m} admits an MTP.*

BRUALDI–HOLLINGSWORTH CONJECTURE (see [4]). *If $m > 2$, then in any proper edge coloring of K_{2m} with $2m-1$ colors, all edges can be partitioned into m multicolored spanning trees.*

In [4] it was proved that in any proper edge coloring of K_{2m} ($m > 2$) with $2m-1$ colors there are at least two edge disjoint multicolored spanning trees.

CONSTANTINE'S CONJECTURE (strong version; see [6]). *If $m > 2$, then in any proper edge coloring of K_{2m} with $2m-1$ colors, all edges can be partitioned into m isomorphic multicolored spanning trees.*

The main goal of this paper is to prove the first conjecture.

Example 1. The complete graph K_6 admits an MTP. To see this consider the complete graph K_6 with the vertex set $\{1,\dots,6\}$. Table 2 gives a proper edge coloring of K_6 with colors $c_1,\dots,c_5$ as well as an MTP for it. The ith row of this table is the set of all edges with color c_i. Each column denotes the edges of a multicolored spanning tree. Figure 1 shows that the spanning trees T_1, T_2, T_3 are isomorphic.

In [6] it has been shown that K_8 admits an MTP.

Using the software Gap, Peter Cameron found a decomposition of $K_{6,6}$ into six isomorphic multicolored graphs $K_{1,3} \cup 3K_2 \cup 2K_1$. In the next lemma, using Cameron's decomposition we find an MTP for K_{12}.

LEMMA 2. *The complete graph K_{12} admits an MTP.*

Proof. Consider the complete graph K_{12} with the vertex set $\{u_1,\dots,u_6,v_1,\dots,v_6\}$. Table 3 gives a proper edge coloring of K_{12} with colors $c_1,\dots,c_{11}$ as well as an MTP for it. The ith row of this table is the set of all edges with color c_i. Each column denotes the edges of a multicolored spanning tree. Note that the first six rows of the table determine a decomposition of $K_{6,6}$ into six multicolored subgraphs isomorphic to $K_{1,3} \cup 3K_2 \cup 2K_1$. □

Now, we are ready to prove our main result.

THEOREM. *For $m \neq 2$, K_{2m} admits an MTP.*

Proof. First suppose that m is an odd integer. Consider the complete graph K_{2m} defined on the set $A \cup B$ where $A = \{a_1,\dots,a_m\}$ and $B = \{b_1,\dots,b_m\}$. For

TABLE 3

	T_1	T_2	T_3	T_4	T_5	T_6
c_1	u_2v_5	u_1v_6	u_6v_1	u_3v_2	u_4v_3	u_5v_4
c_2	u_2v_3	u_5v_2	u_6v_6	u_4v_5	u_3v_4	u_1v_1
c_3	u_4v_1	u_3v_3	u_6v_4	u_1v_2	u_5v_5	u_2v_6
c_4	u_1v_4	u_3v_5	u_5v_3	u_6v_2	u_2v_1	u_4v_6
c_5	u_2v_2	u_4v_4	u_1v_5	u_5v_1	u_6v_3	u_3v_6
c_6	u_5v_6	u_3v_1	u_4v_2	u_2v_4	u_1v_3	u_6v_5
c_7	u_3u_5	u_4u_6	u_1u_2	v_3v_5	v_4v_6	v_1v_2
c_8	u_2u_4	u_1u_5	u_3u_6	v_2v_4	v_1v_5	v_3v_6
c_9	u_2u_5	u_3u_4	u_1u_6	v_2v_5	v_3v_4	v_1v_6
c_{10}	u_2u_6	u_1u_3	u_4u_5	v_2v_6	v_1v_3	v_4v_5
c_{11}	u_1u_4	u_2u_3	u_5u_6	v_1v_4	v_2v_3	v_5v_6

convenience, let G and H be the complete graphs on the sets A and B, respectively. Since m is odd, G has a total coloring π which uses m colors, $1, \ldots, m$. Now, define an edge-coloring c of K_{2m} as follows:

(a) For each edge $a_ja_k \in E(G)$, let $c(a_ja_k) = \pi(a_ja_k)$.
(b) For each edge $b_jb_k \in E(H)$, let $c(b_jb_k) = \pi(a_ja_k)$.
(c) For each edge a_ib_i, $1 \le i \le m$, let $c(a_ib_i) = \pi(a_i)$.
(d) For each edge a_jb_k, $j \ne k$, let $c(a_jb_k) = [k-j]_m + m$.

Clearly, c is a proper $(2m-1)$-edge-coloring of K_{2m}. It is left to decompose K_{2m} into m multicolored isomorphic spanning trees. First, for each $i \in \{1, \ldots, m\}$, let T_i be defined on the set $A \cup B$ and $E(T_i) = \{a_ia_{[i+2t]_m}, b_ib_{[i+2t-1]_m}, b_ia_{[i+2t-1]_m}, a_{[i+1]_m}b_{[i+2t]_m} \,|\, t = 1, 2, \ldots, \frac{m-1}{2}\} \cup \{a_ib_i\}$. It is easy to check that each T_i is a multicolored spanning tree, and all the T_i's are isomorphic.

Now, if m is not an odd integer, then $2m = 2^tm'$ where $t \ge 2$ and m' is odd. In the case where $m' = 1$, t must be at least 3. Then it is a direct consequence of Theorem A. Assume $m' \ge 3$. Thus $K_{2^tm'}$ admits an MTP by Theorem A except when $m' = 3$ and $t = 2$. Since this case can be handled by Lemma 2, we conclude the proof. □

Acknowledgments. The first two authors are very grateful to professor Peter Cameron for his fruitful discussions, and we appreciate the helpful comments of the referees.

REFERENCES

[1] S. AKBARI AND A. ALIPOUR, *Transversals and multicolored matchings*, J. Combin. Des., 12 (2004), pp. 325–332.
[2] I. ANDERSON, *Combinatorial Designs: Construction Methods*, Ellis Horwood Limited, Chichester, UK, 1990.
[3] R. C. BOSE, S. S. SHRIKHANDE, AND E. T. PARKER, *Further results on the construction of mutually orthogonal Latin squares and the falsity of Euler's conjecture*, Canad. J. Math., 12 (1960), pp. 189–203.
[4] R. A. BRUALDI AND S. HOLLINGSWORTH, *Multicolored trees in complete graphs*, J. Combin. Theory Ser. B, 68 (1996), pp. 310–313.
[5] P. J. CAMERON, *Parallelisms of Complete Designs*, London Math. Soc. Lecture Notes Series 23, Cambridge University Press, Cambridge, UK, 1976.
[6] G. M. CONSTANTINE, *Multicolored parallelisms of isomorphic spanning trees*, Discrete Math. Theor. Comput. Sci., 5 (2002), pp. 121–125.
[7] H. P. YAP, *Total Colourings of Graphs*, Lecture Notes in Math. 1623, Springer-Verlag, Berlin, 1996.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 568–577

AUTOCORRELATIONS OF MAXIMUM PERIOD FCSR SEQUENCES*

HONG XU† AND WEN-FENG QI†

Abstract. Let $\underline{a}$ be a maximum period feedback with carry shift register sequence (l-sequence) with connection integer $q = p^e$ and period $T = p^{e-1}(p-1)$. It is shown that the expected value of its autocorrelations is 0, and its variance is $O(q \ln^4 q)$. Thus when q is sufficiently large, with high probability, the autocorrelations are low. Furthermore, it is shown that when $e \geq 2$, for any integer i, $1 \leq i \leq e/2$, when the shift is a multiple of $T/2p^i$, the absolute value of the autocorrelations of $\underline{a}$ is T/p^{2i-1}, and the sign relies on the parity of the multiple.

Key words. feedback with carry shift register, l-sequences, 2-adic numbers, autocorrelations, exponential sum

AMS subject classifications. 11A07, 11B50, 11L07, 11T23, 94A55, 94A60

DOI. 10.1137/050633974

1. Introduction.

Pseudorandom sequences are important in many areas of communications and computing such as cryptography, spread spectrum communications, error correcting codes, and quasi-Monte Carlo integration. In the study of pseudorandom sequences, we are often interested in the correlation properties of the sequences. These properties not only are important measures of randomness [4] but also have practical applications in spread spectrum communication systems, radar systems, cryptanalysis, and so on.

Feedback with carry shift register (FCSR) sequences, especially the l-sequences, have many fine pseudorandom properties analogous to those of m-sequences. Let $q = p^e$, with p an odd prime integer and $e \geq 1$, and let 2 be a primitive root modulo q. The class of binary sequences known as l-sequences can be described in several ways [8], [9]. An l-sequence is the output sequence from a maximum period FCSR with connection number q. It is the 2-adic expansions of a rational number r/q, where $\gcd(r, q) = 1$, and it is the sequence $a_n = (A \cdot 2^{-n} (\mathrm{mod}\, q)) (\mathrm{mod}\, 2)$, where $\gcd(A, q) = 1$.

Up to now, research has been done on the distribution properties and linear complexity of l-sequences [5], [6], [7], [8], [9], [13], [14]. The lattice test on such sequences has also been done by Couture and L'Ecuyer [1], [2], [3] and L'Ecuyer [10], [11]. Their ordinary autocorrelations have not been studied, although their arithmetic autocorrelations have been shown to be zero [5].

The autocorrelation function of a binary periodic sequence $\underline{a} = (a_0, a_1, a_2, \dots)$ with period T is defined as $C_{\underline{a}}(\tau) = \sum_{n=0}^{T-1} (-1)^{a_n + a_{n+\tau}}$ for $0 \leq \tau \leq T-1$. The sequence $\underline{a}$ is said to have "good" autocorrelation properties if, for all $\tau \neq 0$, the absolute value of $C_{\underline{a}}(\tau)$ is very small compared to T. Concerning the arithmetic autocorrelation, it can be thought of as a "with carry" analogue of the usual autocorrelation; see [5].

*Received by the editors June 19, 2005; accepted for publication (in revised form) November 30, 2005; published electronically August 7, 2006. This work was supported by the NSF of China under grant 60373092.

http://www.siam.org/journals/sidma/20-3/63397.html

†Department of Applied Mathematics, Zhengzhou Information Engineering University, P.O. Box 1001-745, Zhengzhou 450002, People's Republic of China (xuhong0504@hotmail.com, wenfeng.qi@263.net).

By use of the rational expression of l-sequences, their arithmetic autocorrelations can be easily calculated, while for the usual autocorrelations, it is far more difficult [5]. In this paper, the usual autocorrelations of l-sequences are discussed.

Let $\underline{a}$ be an l-sequence generated by an FCSR with connection integer $q = p^e$ and period $T = p^{e-1}(p-1)$. Since the l-sequences are all balanced, it is trivial that the expected autocorrelations of $\underline{a}$ are 0. In section 2, by evaluating certain exponential sums, we show that the variance of autocorrelations of $\underline{a}$ satisfies

$$Var\left(C_{\underline{a}}(\tau)\right) \le 256q \cdot \left(\frac{\ln q}{\pi} + \frac{1}{5}\right)^4 \cdot \left(\frac{1-q^{-1/2}}{1-p^{-1/2}}\right)^2.$$

This means that when q is sufficiently large, with high probability, the autocorrelations of l-sequences are low.

For $e \ge 2$, the exact autocorrelation value of $\underline{a}$ for certain shift τ is given in section 3. It is shown that for any integer i, $1 \le i \le e/2$, when the shift τ is a multiple of $T/2p^i$, the autocorrelations of $\underline{a}$ satisfy

$$C_{\underline{a}}(kT/2p^i) = \begin{cases} -T/p^{2i-1} & \text{if } 2 \mid k, \\ T/p^{2i-1} & \text{if } 2 \nmid k, \end{cases}$$

where $1 \le k \le 2p^i - 1$ and $\gcd(k,p) = 1$. The above two results also hold for the decimations of l-sequences.

2. Expectation and variance of the autocorrelations of l-sequences. For convenience, we first give the definition of l-sequences and show some of their important properties.

DEFINITION 2.1 (see [8], [9]). *The maximum period sequence generated by an FCSR with connection integer q is called an l-sequence. In this case, 2 is a primitive root modulo q. Thus q is of the form $q = p^e$, where p is an odd prime, $e \ge 1$, and the sequence has period $T = p^{e-1}(p-1)$.*

DEFINITION 2.2. *Let $\underline{a} = (a_0, a_1, a_2, \dots)$ be a binary periodic sequence of period T. The d-fold decimation of $\underline{a}$ is defined as $\underline{a}^{(d)} = (a_0, a_d, a_{2d}, \dots)$, where d is relatively prime to T.*

Let $\underline{a} = (a_0, a_1, a_2, \dots)$ be an l-sequence generated by an FCSR with connection integer $q = p^e$ and period $T = p^{e-1}(p-1)$. From [5], we know that the number of ones and the number of zeros in a period of $\underline{a}$ or its decimations are equal.

For any positive integer N, let $Z/(N) = \{0, 1, 2, \dots, N-1\}$ be the ring of integers modulo N and $(Z/(N))^*$ its multiplicative group. Throughout the article, we use the notation $(\bmod N)$ for the operator that reduces an integer modulo N to give a number between 0 and $N-1$.

With these notations, we can give the exponential representation of l-sequences as follows.

LEMMA 2.3 (see [9]). *Let $\underline{a} = (a_0, a_1, a_2, \dots)$ be an l-sequence generated by an FCSR with connection integer $q = p^e$. Then there exists $A \in (Z/(q))^*$ such that*

$$a_n = (A \cdot 2^{-n} (\bmod q)) (\bmod 2), \qquad n = 0, 1, \dots.$$

For $s \in \{0,1\}$, denote $H_s = |\{y \in Z/(q) | y(\bmod 2) = s\}|$. Obviously, $H_0 = (q+1)/2$ equals the number of evens in $Z/(q)$ and $H_1 = (q-1)/2$ equals the number of odds in $Z/(q)$.

For any integer $u \in \{0, 1, \dots, q-1\}$, there exists a unique pair (v, y) with $v \in \{0, 1\}$ and $y \in \{0, 1, \dots, H_v - 1\}$ such that $u = 2y + v$. On the other hand, for any pair (v, y) with $v \in \{0, 1\}$ and $y \in \{0, 1, \dots, H_v - 1\}$ we have $2y + v \in \{0, 1, \dots, q-1\}$.

Let $q = p^e$ as above. Suppose that g is a primitive root modulo q and $u_n = A \cdot g^n (\mathrm{mod}\, q)$. Then $\underline{u} = (u_0, u_1, u_2, \dots)$ over $Z/(q)$ is a periodic sequence with period $T = p^{e-1}(p-1)$. Set $a_n = u_n (\mathrm{mod}\, 2)$. Then $\underline{a} = (a_0, a_1, a_2, \dots)$ is an l-sequence or its decimation.

For any positive integer a, denote $e_q(a) = e^{2\pi i a/q}$. For fixed $0 \le n, \tau \le T-1$ and $s, t \in \{0, 1\}$, define

$$N(a_n = s, a_{n+\tau} = t) = \begin{cases} 1 & \text{if } a_n = s \text{ and } a_{n+\tau} = t, \\ 0 & \text{else.} \end{cases}$$

Then the autocorrelation $C_{\underline{a}}(\tau)$ of $\underline{a}$ with shift τ $(0 \le \tau \le T-1)$ can be represented as

$$\begin{aligned} C_{\underline{a}}(\tau) &= \sum_{n=0}^{T-1} (-1)^{a_n + a_{n+\tau}} = \sum_{s,t=0}^{1} (-1)^{s+t} \cdot \sum_{n=0}^{T-1} N(a_n = s, a_{n+\tau} = t) \\ &= \sum_{s,t=0}^{1} (-1)^{s+t} \cdot \left(\sum_{n=0}^{T-1} \left(\sum_{x=0}^{H_s - 1} \frac{1}{q} \sum_{b=0}^{q-1} e_q(b(u_n - 2x - s)) \right) \right. \\ &\quad \left. \cdot \left(\sum_{y=0}^{H_t - 1} \frac{1}{q} \sum_{c=0}^{q-1} e_q(c(u_{n+\tau} - 2y - t)) \right) \right) \\ &= \frac{1}{q^2} \sum_{b,c=0}^{q-1} \left(\sum_{n=0}^{T-1} e_q(bu_n + cu_{n+\tau}) \right) \cdot \left(\sum_{s=0}^{1} (-1)^s e_q(-bs) \sum_{x=0}^{H_s - 1} e_q(-2bx) \right) \\ &\quad \cdot \left(\sum_{t=0}^{1} (-1)^t e_q(-ct) \sum_{y=0}^{H_t - 1} e_q(-2cy) \right) \\ &= \frac{1}{q^2} \sum_{b,c=0}^{q-1} S_\tau(b, c) P(b) Q(c), \end{aligned}$$

where

$$S_\tau(b, c) = \sum_{n=0}^{T-1} e_q(bu_n + cu_{n+\tau}),$$

$$P(b) = \sum_{s=0}^{1} (-1)^s e_q(-bs) \sum_{x=0}^{H_s - 1} e_q(-2bx),$$

and

$$Q(c) = \sum_{t=0}^{1} (-1)^t e_q(-ct) \sum_{y=0}^{H_t - 1} e_q(-2cy).$$

Then we have $S_\tau(0, 0) = T$, $P(0) = \sum_{s=0}^{1} (-1)^s H_s = 1$, and $Q(0) = \sum_{t=0}^{1} (-1)^t H_t =$

1. Using the property that the number of ones and the number of zeros in a period of $\underline{a}$ are equal, we can show that

$$\frac{1}{q^2}\sum_{c=0}^{q-1} S_\tau(0,c)P(0)Q(c) = 0 \quad \text{and} \quad \frac{1}{q^2}\sum_{b=0}^{q-1} S_\tau(b,0)P(b)Q(0) = 0.$$

So we get

$$C_{\underline{a}}(\tau) = \frac{1}{q^2}\sum_{b,c=0}^{q-1} S_\tau(b,c)P(b)Q(c) = -\frac{T}{q^2} + \frac{1}{q^2}\sum_{b,c=1}^{q-1} S_\tau(b,c)P(b)Q(c). \tag{2.1}$$

Using this equation, based on some evaluation on certain exponential sums (given as lemmas below), we can calculate the expectation and variance of $C_{\underline{a}}(\tau)$.

LEMMA 2.4. *Let $0 \neq b, c \in Z/(q)$ and $S_\tau(b,c)$ be defined as above. Then*

$$\sum_{\tau=0}^{T-1} |S_\tau(b,c)|^2 \leq qT \cdot \gcd(c,q).$$

Proof. As $u_n = A \cdot g^n (\mathrm{mod}\, q)$ and g is a primitive root modulo q, we have

$$\begin{aligned}
\sum_{\tau=0}^{T-1} |S_\tau(b,c)|^2 &= \sum_{\tau=0}^{T-1} \left| \sum_{n=0}^{T-1} e_q((b+cg^\tau)u_n) \right|^2 \\
&\leq \sum_{\gamma=0}^{q-1} \left| \sum_{n=0}^{T-1} e_q((b+c\gamma)u_n) \right|^2 \\
&= \sum_{m,n=0}^{T-1} e_q(b(u_n-u_m)) \cdot \sum_{\gamma=0}^{q-1} e_q(c\gamma(u_n-u_m)) \\
&\leq \sum_{m,n=0}^{T-1} \left| \sum_{\gamma=0}^{q-1} e_q(c\gamma(u_n-u_m)) \right| = q \cdot |\Omega|,
\end{aligned}$$

where $\Omega = \{(m,n) | 0 \leq m,n \leq T-1 \text{ and } q|(c(u_n-u_m))\}$ and $|\Omega|$ is the number of elements in Ω. Similar to the proof of Lemma 5 in [13], we can get $|\Omega| \leq \gcd(c,q) \cdot T$. Thus

$$\sum_{\tau=0}^{T-1} |S_\tau(b,c)|^2 \leq qT \cdot \gcd(c,q).$$

This completes the proof of Lemma 2.4. ∎

LEMMA 2.5 (see [12]). *For any positive integers m and H,*

$$\sum_{a=1}^{m-1}\left|\sum_{x=0}^{H-1}e_m(ax)\right| < 2m\left(\frac{\ln m}{\pi}+\frac{1}{5}\right)$$

holds, where ln *is the natural logarithm.*

Using Lemma 2.5, we get the following.

LEMMA 2.6. *Let p be an odd prime, $q=p^e$, and $e\geq 1$. Then for any positive integer H, we have*

$$\sum_{a=1}^{q-1}\gcd(a,q)^{1/2}\cdot\left|\sum_{x=0}^{H-1}e_q(-2ax)\right| \leq 2q\cdot\left(\frac{\ln q}{\pi}+\frac{1}{5}\right)\cdot\left(\frac{1-q^{-1/2}}{1-p^{-1/2}}\right).$$

Using these three lemmas, together with (2.1), we can reach the following conclusions.

THEOREM 2.7. *Let g be a primitive root modulo $q=p^e$ $(e\geq 1)$ and $\underline{a}=(a_0,a_1,a_2,\dots)$ be a binary periodic sequence defined by $a_n=(A\cdot g^n(\bmod q))(\bmod 2)$ with period $T=p^{e-1}(p-1)$. Then the expectation of its autocorrelations is $E[C_{\underline{a}}(\tau)]=0$, and the variance of its autocorrelations satisfies*

$$Var(C_{\underline{a}}(\tau)) \leq 256q\cdot\left(\frac{\ln q}{\pi}+\frac{1}{5}\right)^4\cdot\left(\frac{1-q^{-1/2}}{1-p^{-1/2}}\right)^2.$$

Proof. The first result follows from the fact that for any sequence $\underline{a}$, $E[C_{\underline{a}}(\tau)]=I(\underline{a})^2/T$, where $I(\underline{a})$ is the imbalance of $\underline{a}$, that is, the number of zeros minus the number of ones.

Concerning the second result, we first evaluate $E\left[\left(C_{\underline{a}}(\tau)+\frac{T}{q^2}\right)^2\right]$:

$$\begin{aligned}
&E\left[\left(C_{\underline{a}}(\tau)+\frac{T}{q^2}\right)^2\right]\\
&=\frac{1}{T}\sum_{\tau=0}^{T-1}\left(\frac{1}{q^2}\sum_{b,c=1}^{q-1}S_\tau(b,c)P(b)Q(c)\right)^2\\
&=\frac{1}{q^4T}\sum_{b_1,c_1,b_2,c_2=1}^{q-1}\left(\sum_{\tau=0}^{T-1}S_\tau(b_1,c_1)S_\tau(b_2,c_2)\right)\cdot P(b_1)P(b_2)Q(c_1)Q(c_2)\\
&\leq\frac{1}{q^4T}\sum_{b_1,c_1,b_2,c_2=1}^{q-1}\left(\sum_{\tau=0}^{T-1}|S_\tau(b_1,c_1)S_\tau(b_2,c_2)|\right)\cdot|P(b_1)P(b_2)Q(c_1)Q(c_2)|.
\end{aligned}$$

Set $g=2^{-1}(\bmod q)$ in Lemma 2.4. Then by Cauchy's inequality we can get

$$\begin{aligned}
\sum_{\tau=0}^{T-1}|S_\tau(b_1,c_1)S_\tau(b_2,c_2)| &\leq \left(\sum_{\tau=0}^{T-1}|S_\tau(b_1,c_1)|^2\right)^{1/2}\cdot\left(\sum_{\tau=0}^{T-1}|S_\tau(b_2,c_2)|^2\right)^{1/2}\\
&\leq qT\cdot\gcd(c_1,q)^{1/2}\cdot\gcd(c_2,q)^{1/2}.
\end{aligned}$$

Then from Lemmas 2.5 and 2.6, we have

$$\begin{aligned}
&E\left[\left(C_{\underline{a}}(\tau)+\frac{T}{q^2}\right)^2\right]\\
&\le \frac{1}{q^4T}\sum_{b_1,c_1,b_2,c_2=1}^{q-1} qT\cdot\gcd(c_1,q)^{1/2}\cdot\gcd(c_2,q)^{1/2}\cdot|P(b_1)P(b_2)Q(c_1)Q(c_2)|\\
&\le q^{-3}\cdot\left(\sum_{b_1=1}^{q-1}|P(b_1)|\right)\cdot\left(\sum_{b_2=1}^{q-1}|P(b_2)|\right)\cdot\left(\sum_{c_1=1}^{q-1}\gcd(c_1,q)^{1/2}\cdot|Q(c_1)|\right)\\
&\quad\cdot\left(\sum_{c_2=1}^{q-1}\gcd(c_2,q)^{1/2}\cdot|Q(c_2)|\right)\\
&\le q^{-3}\cdot\left(4q\cdot\left(\frac{\ln q}{\pi}+\frac{1}{5}\right)\right)^2\cdot\left(4q\cdot\left(\frac{\ln q}{\pi}+\frac{1}{5}\right)\left(\frac{1-q^{-1/2}}{1-p^{-1/2}}\right)\right)^2\\
&= 256q\cdot\left(\frac{\ln q}{\pi}+\frac{1}{5}\right)^4\cdot\left(\frac{1-q^{-1/2}}{1-p^{-1/2}}\right)^2.
\end{aligned}$$

Thus

$$\begin{aligned}
Var(C_{\underline{a}}(\tau)) = Var\left(C_{\underline{a}}(\tau)+\frac{T}{q^2}\right) &\le E\left[\left(C_{\underline{a}}(\tau)+\frac{T}{q^2}\right)^2\right]\\
&\le 256q\cdot\left(\frac{\ln q}{\pi}+\frac{1}{5}\right)^4\cdot\left(\frac{1-q^{-1/2}}{1-p^{-1/2}}\right)^2,
\end{aligned}$$

which completes the proof of Theorem 2.7. □

Remark 2.8. Chebyshev's inequality says that for any random variable X and $\varepsilon>0$, $\Pr(|X-E[X]|\ge\varepsilon)\le Var(X)/\varepsilon^2$, where $E[X]$ denotes the expectation of X and $Var(X)$ denotes the variance of X. So, for fixed $\delta>0$, we have

$$\Pr\left(|C_{\underline{a}}(\tau)|\ge T^{(1+\delta)/2}\right)\le 256qT^{-(1+\delta)}\cdot\left(\frac{\ln q}{\pi}+\frac{1}{5}\right)^4\cdot\left(\frac{1-q^{-1/2}}{1-p^{-1/2}}\right)^2.$$

Note that $\left(\frac{1-q^{-1/2}}{1-p^{-1/2}}\right)^2$ is less than 6 (and less than 4 if $p>3$). The probability approaches 0 as q tends to infinity, which shows that when q is sufficiently large, most autocorrelations of l-sequences are small.

COROLLARY 2.9. *Let $\underline{a}$ be an l-sequence with connection integer $q=p^e$ $(e\ge1)$ and period $T=p^{e-1}(p-1)$. Then the expectation of its autocorrelations is $E[C_{\underline{a}}(\tau)]=0$, and the variance of its autocorrelations satisfies*

$$Var(C_{\underline{a}}(\tau))\le 256q\cdot\left(\frac{\ln q}{\pi}+\frac{1}{5}\right)^4\cdot\left(\frac{1-q^{-1/2}}{1-p^{-1/2}}\right)^2.$$

3. Autocorrelations of *l*-sequences with certain shifts. In this section, we will calculate the exact autocorrelation values of l-sequences with certain shifts. Before showing the main result, we first give a lemma.

LEMMA 3.1. *Let p be an odd prime and $i\ge1$. Set*

$$\begin{aligned}
\Lambda_0 &= \{(m,n)\in(Z/(p^i))^*\times(Z/(p^i))|(m+n(\mathrm{mod}\,p^i))(\mathrm{mod}\,2)=n(\mathrm{mod}\,2)\},\\
\Lambda_1 &= \{(m,n)\in(Z/(p^i))^*\times(Z/(p^i))|(m+n(\mathrm{mod}\,p^i))(\mathrm{mod}\,2)\ne n(\mathrm{mod}\,2)\}.
\end{aligned}$$

Then

$$|\Lambda_0| - |\Lambda_1| = -(p-1).$$

Proof. As $|\Lambda_0|+|\Lambda_1| = p^{2i-1}(p-1)$, we need only to calculate $|\Lambda_0|$, the cardinality of Λ_0. For any $(m,n) \in \Lambda_0$, obviously we have $0 < m+n < 2p^i$, and the cardinality of Λ_0 can be calculated according to the parity of m.

(1) If m is odd, then

$$(m + n(\mathrm{mod}\, p^i))(\mathrm{mod}\, 2) = n(\mathrm{mod}\, 2) \text{ if and only if } m+n \geq p^i.$$

Since $0 \leq n \leq p^i - 1$, the number of n satisfying $p^i - m \leq n \leq p^i - 1$ is m.

(2) If m is even, then

$$(m + n(\mathrm{mod}\, p^i))(\mathrm{mod}\, 2) = n(\mathrm{mod}\, 2) \text{ if and only if } 0 < m+n < p^i.$$

Since $0 \leq n \leq p^i - 1$, the number of n satisfying $0 \leq n \leq p^i - 1 - m$ is $p^i - m$.

From above we know that

$$\begin{aligned}
|\Lambda_0| &= \sum_{m\in(Z/(p^i))^*,\ 2\nmid m} m + \sum_{m\in(Z/(p^i))^*,\ 2|m} (p^i - m) \\
&= \sum_{m\in(Z/(p^i))^*,\ 2|m} p^i + \sum_{m\in(Z/(p^i))^*} m - 2 \sum_{m\in(Z/(p^i))^*,\ 2|m} m \\
&= [p^{2i-1}(p-1) + p^{2i-1}(p-1) - (p^{2i-1}+1)(p-1)]/2 \\
&= [p^{2i-1}(p-1) - (p-1)]/2.
\end{aligned}$$

Thus

$$|\Lambda_0| - |\Lambda_1| = p^{2i-1}(p-1) - 2|\Lambda_0| = -(p-1).$$

So the lemma follows. □

Using the same notation as in section 2, let g be a primitive root modulo q and $a_n = (A \cdot g^n(\mathrm{mod}\, q))(\mathrm{mod}\, 2)$. Then $\underline{a} = (a_0, a_1, a_2, \dots)$ refers to an l-sequence or its decimation. The autocorrelation of $\underline{a}$ with shift τ can be represented as

$$C_{\underline{a}}(\tau) = \sum_{n=0}^{T-1} (-1)^{a_n + a_{n+\tau}} = N_0 - N_1,$$

where $N_0 = \sum_{n=0}^{T-1} N(a_n = a_{n+\tau})$ is the number of n satisfying $a_n = a_{n+\tau}$, $N_1 = \sum_{n=0}^{T-1} N(a_n \neq a_{n+\tau})$ is the number of n satisfying $a_n \neq a_{n+\tau}$, and

$$N(a_n = a_{n+\tau}) = \begin{cases} 1 & \text{if } a_n = a_{n+\tau}, \\ 0 & \text{else}, \end{cases}$$

$$N(a_n \neq a_{n+\tau}) = \begin{cases} 1 & \text{if } a_n \neq a_{n+\tau}, \\ 0 & \text{else} \end{cases}$$

for fixed n, τ, $0 \leq n, \tau \leq T-1$.

Next we will calculate the difference between N_0 and N_1. Set

$$\Omega_0 = \{A \in (Z/(p^e))^* | (A \cdot g^\tau(\mathrm{mod}\, p^e))(\mathrm{mod}\, 2) = A(\mathrm{mod}\, 2)\},$$
$$\Omega_1 = \{A \in (Z/(p^e))^* | (A \cdot g^\tau(\mathrm{mod}\, p^e))(\mathrm{mod}\, 2) \neq A(\mathrm{mod}\, 2)\}.$$

Obviously, we have $N_0 = |\Omega_0|$, $N_1 = |\Omega_1|$. Thus $C_{\underline{a}}(\tau) = |\Omega_0| - |\Omega_1|$.

When the shift τ is of the form $k \cdot T/p^i$ $(1 \le i \le e-1, \gcd(k,p) = 1)$, from the primitivity of g we know that

$$g^\tau (\mathrm{mod}\, p^{e-i}) = 1, \text{ whereas } g^\tau (\mathrm{mod}\, p^{e-i+1}) \neq 1.$$

Thus we can set $g^\tau (\mathrm{mod}\, p^e) = 1 + k_0 p^{e-i} + k_1 p^{e-i+1} + \cdots + k_{i-1} p^{e-1}$, where $0 \le k_j \le p-1$, $j = 0, 1, \ldots, i-1$, and $k_0 \neq 0$. It is easy to check that when k runs through all elements in $(Z/(p^i))^*$, k_j also runs through all elements in $Z/(p)$. That is,

$$\{1 + k_0 p^{e-i} + k_1 p^{e-i+1} + \cdots + k_{i-1} p^{e-1} | 0 \le k_j \le p-1,\ j = 0, 1, \ldots, i-1, \text{ and } k_0 \neq 0\}$$
$$= \{g^{k \cdot T/p^i} (\mathrm{mod}\, p^e) | 1 \le k \le p^i - 1 \text{ and } \gcd(k,p) = 1\}.$$

For this kind of τ, we can reach the following conclusion.

THEOREM 3.2. *Let g be a primitive root modulo $q = p^e$ $(e \ge 2)$ and $\underline{a} = (a_0, a_1, a_2, \ldots)$ be a binary periodic sequence defined by $a_n = (A \cdot g^n (\mathrm{mod}\, q))(\mathrm{mod}\, 2)$ with period $T = p^{e-1}(p-1)$. Then for any positive integer i, $1 \le i \le e/2$, the autocorrelation of $\underline{a}$ with shift τ satisfies*

$$C_{\underline{a}}(k \cdot T/p^i) = -T/p^{2i-1},$$

where $1 \le k \le p^i - 1$ and $\gcd(k,p) = 1$.

Proof. From the above analysis, we know that for any $\tau = k \cdot T/p^i$, $1 \le k \le p^i - 1$, $\gcd(k,p) = 1$, g^τ is of the form

$$g^\tau (\mathrm{mod}\, p^e) = 1 + k_0 p^{e-i} + k_1 p^{e-i+1} + \cdots + k_{i-1} p^{e-1},$$

where $0 \le k_j \le p-1$, $j = 0, 1, \ldots, i-1$, and $k_0 \neq 0$. Next we will show that, for this kind of τ, $C_{\underline{a}}(\tau) = -T/p^{2i-1}$.

Using the same notation as above, we have $C_{\underline{a}}(\tau) = |\Omega_0| - |\Omega_1|$.

For any $A \in (Z/(q))^*$, let $A = A_0 + A_1 p + \cdots + A_{e-1} p^{e-1}$ be the p-adic expansion of A, $0 \le A_j \le p-1$, $j = 0, 1, \ldots, e-1$, and $A_0 \neq 0$. Then we have

$$\begin{aligned}
&A \cdot g^\tau (\mathrm{mod}\, p^e) \\
&= (A_0 + A_1 p + \cdots + A_{e-1} p^{e-1})(1 + k_0 p^{e-i} + k_1 p^{e-i+1} + \cdots + k_{i-1} p^{e-1})(\mathrm{mod}\, p^e) \\
&= (A_0 + A_1 p + \cdots + A_{e-i-1} p^{e-i-1}) + (A_0 k_0 + A_{e-i}) p^{e-i} \\
&\quad + (A_0 k_1 + A_1 k_0 + A_{e-i+1}) p^{e-i+1} + \cdots \\
&\quad + (A_0 k_{i-1} + A_1 k_{i-2} + \cdots + A_{i-1} k_0 + A_{e-1}) p^{e-1} (\mathrm{mod}\, p^e) \\
&= (A_0 + A_1 p + \cdots + A_{e-i-1} p^{e-i-1}) + p^{e-i} \cdot (A_{e-i} + p A_{e-i+1} + \cdots + A_{e-1} p^{i-1}) \\
&\quad + p^{e-i} \cdot (A_0 k_0 + (A_0 k_1 + A_1 k_0) + \cdots \\
&\quad + (A_0 k_{i-1} + A_1 k_{i-2} + \cdots + A_{i-1} k_0) p^{i-1}) (\mathrm{mod}\, p^e).
\end{aligned}$$

Set

$$m = (A_0 k_0 + (A_0 k_1 + A_1 k_0) + \cdots + (A_0 k_{i-1} + A_1 k_{i-2} + \cdots + A_{i-1} k_0) p^{i-1}) (\mathrm{mod}\, p^i),$$
$$n = (A_{e-i} + p A_{e-i+1} + \cdots + A_{e-1} p^{i-1}) (\mathrm{mod}\, p^i).$$

Then

$$A \cdot g^\tau (\mathrm{mod}\, p^e) = ((A_0 + A_1 p + \cdots + A_{e-i-1} p^{e-i-1}) + (m+n) p^{e-i}) (\mathrm{mod}\, p^e),$$
$$A = ((A_0 + A_1 p + \cdots + A_{e-i-1} p^{e-i-1}) + n p^{e-i}) (\mathrm{mod}\, p^e).$$

Thus $(A \cdot g^{\tau} (\mathrm{mod}\, p^e))(\mathrm{mod}\, 2) = A(\mathrm{mod}\, 2)$ if and only if $(m + n(\mathrm{mod}\, p^i))(\mathrm{mod}\, 2) = n(\mathrm{mod}\, 2)$.

If $1 \le i \le e/2$, then the indices of $A_0, A_1, \ldots, A_{i-1}, A_{e-i}, A_{e-i+1}, \ldots, A_{e-1}$ are pairwise distinct and can be evaluated independently. Thus when they run through all elements in $Z/(p)$, m and n run through all elements in $(Z/(p^i))^*$ and $(Z/(p^i))$, respectively. From Lemma 3.1 we know that

$$|\Lambda_0| - |\Lambda_1| = -(p-1),$$

where Λ_0, Λ_1 are defined as in Lemma 3.1.

Since the other $e - 2i$ elements of A can be chosen arbitrarily from $Z/(p)$,

$$C_{\underline{a}}(\tau) = |\Omega_0| - |\Omega_1| = -p^{e-2i}(p-1) = -T/p^{2i-1}$$

holds. □

Remark 3.3. If $\tau = (2k-1) \cdot T/2p^i$, $1 \le 2k-1 \le 2p^i - 1$, and $\gcd(2k-1, p) = 1$, then $g^{\tau}(\mathrm{mod}\, p^{e-i}) = -1(\mathrm{mod}\, p^{e-i})$. Similarly, we can get $C_{\underline{a}}(\tau) = T/p^{2i-1}$.

Especially for l-sequences, the following result holds.

COROLLARY 3.4. *Let $q = p^e$ $(e \ge 2)$ be the connection integer of an FCSR that generates an l-sequence $\underline{a}$, and $T = p^{e-1}(p-1)$. Then for any positive integers i and k, $1 \le i \le e/2$, $1 \le k \le 2p^i - 1$, and $\gcd(k, p) = 1$, we have*

$$C_{\underline{a}}(kT/2p^i) = \begin{cases} -T/p^{2i-1} & \text{if } 2 \mid k, \\ T/p^{2i-1} & \text{if } 2 \nmid k. \end{cases}$$

4. Conclusions. Experiments show that there do exist some shifts such that the corresponding autocorrelations are high, although most autocorrelations of l-sequences and their decimations are low. How to further evaluate all the autocorrelations and pick up those shifts with high correlations is still an open problem.

Acknowledgment. The authors are grateful to the two anonymous referees for their valuable comments and suggestions.

REFERENCES

[1] R. COUTURE AND P. L'ECUYER, *On the lattice structure of certain linear congruential sequences related to AWC/SWB generators*, Math. Comp., 62 (1994), pp. 799–808.

[2] R. COUTURE AND P. L'ECUYER, *Linear recurrences with carry as uniform random number generators*, in Proceedings of the 27th Conference on Winter Simulation, ACM, New York, 1995, pp. 263–267.

[3] R. COUTURE AND P. L'ECUYER, *Distribution properties of multiply-with-carry random number generators*, Math. Comp., 66 (1997), pp. 591–607.

[4] S. GOLOMB, *Shift Register Sequences*, Aegean Park, Laguna Hills, CA, 1982.

[5] M. GORESKY AND A. KLAPPER, *Arithmetic crosscorrelations of feedback with carry shift register sequences*, IEEE Trans. Inform. Theory, 43 (1997), pp. 1342–1345.

[6] M. GORESKY, A. KLAPPER, AND R. MURTY, *On the distinctness of decimations of l-sequences*, in Sequences and Their Applications—SETA 01, T. Helleseth, P. V. Kumar, and K. Yang, eds., Springer-Verlag, New York, 2001, pp. 197–208.

[7] M. GORESKY, A. KLAPPER, R. MURTY, AND I. SHPARLINSKI, *On decimations of l-sequences*, SIAM J. Discrete Math., 18 (2004), pp. 130–140.

[8] A. KLAPPER AND M. GORESKY, *Large period nearly deBruijn FCSR sequences*, in Advances in Cryptology—Eurocrypt 1995, Lecture Notes in Comput. Sci. 921, Springer-Verlag, New York, 1995, pp. 263–273.

[9] A. KLAPPER AND M. GORESKY, *Feedback shift registers, 2-adic span, and combiners with memory*, J. Cryptology, 10 (1997), pp. 111–147.

[10] P. L'Ecuyer, *Uniform random number generators: A review*, in Proceedings of the 29th Conference on Winter Simulation, ACM, New York, 1997, pp. 127–134.
[11] P. L'Ecuyer, *Uniform random number generators*, in Proceedings of the 30th Conference on Winter Simulation, IEEE Computer Society Press, Los Alamitos, CA, 1998, pp. 97–104.
[12] R. Lidl and H. Niederreiter, *Finite Fields*, Encyclopedia Math. Appl. 20, Cambridge University Press, Cambridge, UK, 1983.
[13] W. Qi and H. Xu, *Partial period distribution of FCSR sequences*, IEEE Trans. Inform. Theory, 49 (2003), pp. 761–765.
[14] C. Seo, S. Lee, Y. Sung, K. Han, and S. Kim, *A lower bound on the linear span of an FCSR*, IEEE Trans. Inform. Theory, 46 (2000), pp. 691–693.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 578–587

A SPLITTER THEOREM FOR INTERNALLY 4-CONNECTED BINARY MATROIDS*

JIM GEELEN† AND XIANGQIAN ZHOU†

Abstract. We prove that if N is an internally 4-connected minor of an internally 4-connected binary matroid M with $E(N) \geq 4$, then there exist matroids $M_0, M_1, \ldots, M_n$ such that $M_0 \cong N$, $M_n = M$, and, for each $i \in \{1, \ldots, i\}$, M_{i-1} is a minor of M_i, $|E(M_{i-1})| \geq |E(M_i)| - 2$, and M_i is 4-connected up to separators of size 5.

Key words. binary matroids, Splitter Theorem, 4-connectivity

AMS subject classification. 05B35

DOI. 10.1137/050629124

1. Introduction.

We prove the following theorem.

THEOREM 1.1 (main theorem). *Let M be a binary matroid that is 4-connected up to separators of size 5 and let N be an internally 4-connected proper minor of M. If $|E(N)| \geq 10$, then either*

- *there exists $e \in E(M)$ such that $M \setminus e$ or M/e is 4-connected up to separators of size 5 and contains an N-minor, or*
- *M has a fan $(e_1, e_2, e_3, e_4, e_5)$ such that $M/e_3 \setminus e_4$ or $M \setminus e_3/e_4$ is 4-connected up to separators of size 5 and contains an N-minor.*

A matroid M is *4-connected up to separators of size k* if M is 3-connected and for each 3-separation (A, B) of M either $|A| \leq k$ or $|B| \leq k$. A matroid is *internally 4-connected* if it is 4-connected up to separators of size 3. A sequence $(e_1, \ldots, e_i)$ of distinct elements of a matroid M is called a *fan* if the sets $\{e_1, e_2, e_3\}, \{e_2, e_3, e_4\}, \ldots, \{e_{i-2}, e_{i-1}, e_i\}$ are alternately triangles and triads. For other notation and terminology we follow Oxley [6], except we use $\mathrm{si}(M)$ and $\mathrm{co}(M)$ to denote the simplification and cosimplification, respectively, of a matroid M. Recall that M having an *N-minor* means that M has a minor isomorphic to N.

We remark that the bound $|E(N)| \geq 10$ in Theorem 1.1 is included only to simplify the proof; the result holds under the weaker hypothesis that $|E(M)| \geq 7$. (Thus we do not require a lower bound on $|E(N)|$.)

Seymour's Splitter Theorem [7] is a well-known inductive tool for studying 3-connected matroids.

THEOREM 1.2 (the Splitter Theorem). *Let M be a 3-connected matroid with $|E(M)| \geq 4$ and let N be a 3-connected proper minor of M. If M is not a wheel or a whirl, then there exists $e \in E(M)$ such that $M \setminus e$ or M/e is 3-connected and has an N-minor.*

The Splitter Theorem allows a 3-connected matroid to be built one element at a time from a given 3-connected minor so that the intermediate matroids are all 3-connected. Theorem 1.1 provides a similar result for internally 4-connected binary matroids.

*Received by the editors April 13, 2005; accepted for publication (in revised form) December 13, 2005; published electronically August 7, 2006. This research was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Sloan Foundation.
http://www.siam.org/journals/sidma/20-3/62912.html

†Department of Combinatorics and Optimization, University of Waterloo, Waterloo, ON, N2L 3G1, Canada.

COROLLARY 1.3. *Let N be an internally 4-connected minor of an internally 4-connected binary matroid M, where $|E(N)| \geq 4$. Then there exists a sequence $M_0, M_1, \ldots, M_k$ of matroids such that $M_0 \cong N$, $M_k = M$, and, for each $i \in \{1, \ldots, k\}$, M_{i-1} is a minor of M_i, $|E(M_{i-1})| \geq |E(M_i)| - 2$, and M_i is 4-connected up to separators of size 5.*

We rely heavily on results of Hall [4], who proved the following analogue of Tutte's Wheels and Whirls Theorem.

THEOREM 1.4 (Hall [4]). *If M is 4-connected up to separators of size 5 and $|E(M)| \geq 5$, then either*

- *there exists $e \in E(M)$ such that $M \setminus e$ or M/e is 4-connected up to separators of size 5, or*
- *M has a fan $(e_1, e_2, e_3, e_4, e_5)$ such that $M/e_3 \setminus e_4$ or $M \setminus e_3/e_4$ is 4-connected up to separators of size 5.*

Note that Hall's theorem holds for all matroids, while Theorem 1.1 is only for binary matroids. The main reason is simply that this is what we could prove. There is a very useful lemma (Lemma 4.3) that is particular to binary matroids. We expect that there is a reasonable analogue of the Splitter Theorem for matroids that are 4-connected up to separators of size 5—not just for binary matroids. The applicability of Theorem 1.1 (discussed below) stems from the fact that the class of binary matroids is closed under 3-sums. As there is no reasonable analogue of a 3-sum for general matroids, the proposed generalization may be of only academic interest.

It is a shortcoming of Corollary 1.3 that the intermediate matroids are only 4-connected up to separators of size 5; it would be preferable if this could be strengthened to internally 4-connected. There are, however, numerous obstacles to obtaining such a theorem, even for graphs; see Johnson and Thomas [5]. They proved that if H is an internally 4-connected minor of an internally 4-connected graph G, then either H and G belong to a family of exceptional graphs, or G can be built from H by means of four possible constructions. Their intermediate graphs are "almost" internally 4-connected. Below we give some justification that, other than causing additional case analysis, Corollary 1.3 provides a satisfactory inductive tool for internally 4-connected binary matroids.

First we will outline how one might use Corollary 1.3 to prove Seymour's decomposition of regular matroids [7]. Seymour showed that every regular matroid can be obtained from graphic matroids, cographic matroids, and copies of R_{10} via 1-, 2-, and 3-sums. Equivalently, every internally 4-connected regular matroid is either graphic or cographic or is isomorphic to R_{10}. It would suffice to prove the following claim: *If M is a regular matroid that is 4-connected up to separators of size 5 and M has an $M^*(K_{3,3})$-minor, then either M is graphic or M is isomorphic to R_{10}.* This claim reduces easily to the case that M is internally 4-connected. Therefore, one could attempt to prove the result inductively by using Corollary 1.3. Here we see that relaxing the connectivity condition slightly (from internally 4-connected to 4-connected up to separators of size 5) facilitates the use of induction.

Let $\mathcal{M}$ be a minor-closed class of binary matroids. Recall that a matroid $N \in \mathcal{M}$ is a *splitter* for $\mathcal{M}$ if there is no 3-connected matroid in $\mathcal{M}$ that contains N as a proper minor. Determining whether a 3-connected matroid N is a splitter for $\mathcal{M}$ reduces to a finite case analysis via Seymour's Splitter Theorem. Analogously we could call N a 4-*splitter* if there is no internally 4-connected matroid in $\mathcal{M}$ that contains N as a proper minor. It is a straightforward exercise to prove that, if N is internally 4-connected with $|E(N)| \geq 9$ and N is a 4-splitter for $\mathcal{M}$, then there are only finitely

many matroids in $\mathcal{M}$ that are 4-connected up to separators of size 5 and that contain N as a minor. It follows that, using Corollary 1.3, we can test whether or not N is a 4-splitter via a finite case check.

2. Small matroids. When $|E(N)| \geq 10$ it is clear that Theorem 1.1 implies Corollary 1.3. In this section we address the problems that arise for smaller matroids. There are only a few internally 4-connected binary matroids with $|E(N)| \leq 9$. The following result can be easily verified by the reader.

LEMMA 2.1. *If N is an internally 4-connected binary matroid with $|E(N)| \leq 9$, then either N is a uniform matroid with at most three elements or N is isomorphic to one of the following matroids: $M(K_4)$, F_7, F_7^*, $M(K_{3,3})$, or $M^*(K_{3,3})$.*

It follows from Tutte's Wheels and Whirls Theorem that if M is a 3-connected binary matroid with $|E(M)| \geq 4$, then M has an $M(K_4)$-minor. Thus, when $N = M(K_4)$, Corollary 1.3 is an immediate corollary of Theorem 1.4.

Using the Splitter Theorem and "blocking sequences," Zhou [9] studied internally 4-connected binary matroids with an F_7-minor.

LEMMA 2.2 (see Zhou [9]). *If M is an internally 4-connected binary matroid with a proper F_7-minor, then M has an internally 4-connected minor N with an F_7-minor and with $10 \leq |E(N)| \leq 11$.*

Let M be an internally 4-connected matroid with F_7 as a proper minor. By Lemma 2.2, M has an internally 4-connected minor N with $10 \leq |E(N)| \leq 12$ and with an F_7-minor. By the Splitter Theorem, there exists a sequence of 3-connected matroids $M_0, M_1, \ldots, M_j$ such that $M_0 \cong F_7$, $M_j = N$, and, for each $i \in \{1, \ldots, j\}$, there exists $e \in E(M_i)$ such that $M_{i-1} = M_i \setminus e$ or $M_{i-1} = M_i/e$. Since $M_0, \ldots, M_j$ have at most 11 elements, they are 4-connected up to separators of size 5. Now, applying Theorem 1.1 to N, we can prove Corollary 1.3 in the case that $N = F_7$. By duality, Corollary 1.3 holds when $N = F_7^*$.

There are exactly three 10-element binary matroids that are internally 4-connected and that contain an $M(K_{3,3})$-minor; these matroids, named R_{10}, N_{10}, and $\tilde{K_5}^*$, are defined in [7, 9]. The same techniques used by Zhou [9] in proving Lemma 2.2 can be used to prove the following result; we omit the straightforward but lengthy details.

LEMMA 2.3. *Let M be an internally 4-connected binary matroid with a proper $M(K_{3,3})$-minor. Then M has a minor isomorphic to R_{10}, N_{10}, $\tilde{K_5}^*$, or to the cycle matroid of one of the graphs in Figure 1.*

Now, considering each of the graphs in Figure 1, we can prove Corollary 1.3 when $N = M(K_{3,3})$ and $N = M(K_{3,3})^*$.

3. Basic lemmas on separations. In this section, we present some basic lemmas on separations that will be used in later sections.

Let $M = (E, r)$ be a matroid, where r is the rank function. For $A \subseteq E$, we let $\lambda_M(A)$ denote $r(A) + r(E \setminus A) - r(M)$. Then A is k-separating if and only if $\lambda_M(A) \leq k - 1$. We refer to λ_M as the *connectivity function* of M. Tutte [8] proved that the connectivity function is *submodular;* that is, if $X, Y \subseteq E(M)$, then

$$\lambda_M(X) + \lambda_M(Y) \geq \lambda_M(X \cap Y) + \lambda_M(X \cup Y).$$

The next lemma follows easily.

LEMMA 3.1. *Let X and Y be k-separating sets of M. If $X \cap Y$ is not $(k-1)$-separating in M, then $X \cup Y$ is k-separating in M.*

The *coclosure* of a set $X \subseteq E(M)$ is the closure of X in M^*. Clearly, an element $x \in E(M) \setminus X$ belongs to the coclosure of X if and only if x does not belong to the

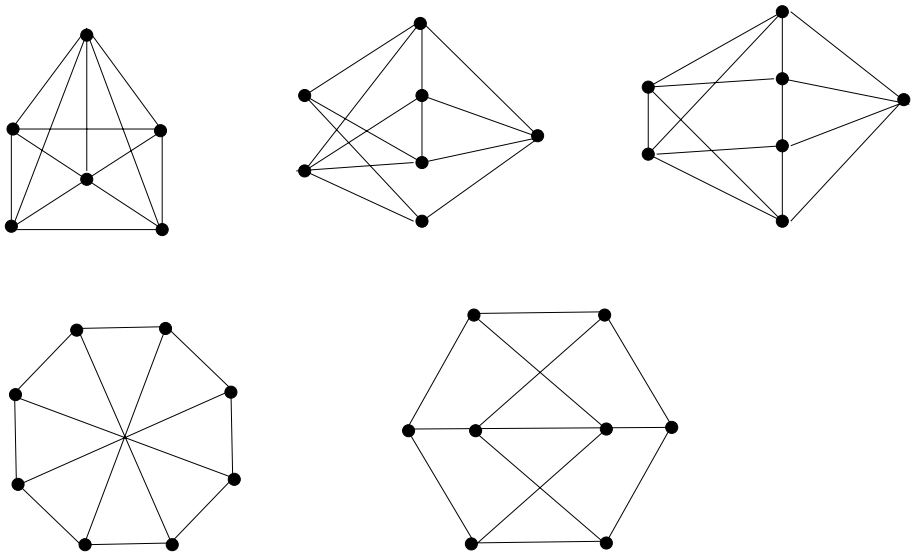


FIG. 1. *Internally 4-connected graphs.*

closure of $E(M)\backslash(X \cup \{x\})$. A set $X \subseteq E(M)$ is *coclosed* if the coclosure of X is the set X itself. We say X is *fully closed* if X is both closed and coclosed.

Let (A, B) be a k-separation of the matroid M. Following the terminology of [3], an element $x \in E(M)$ is in the *guts* of (A, B) if x belongs to the closure of both A and B. Dually, x is in the *coguts* of (A, B) if x belongs to the coclosure of both A and B. We say that (A, B) is an *exact k-separation* or A is *exactly k-separating* if $\lambda_M(A) = k - 1$. The next lemma follows easily from definitions.

LEMMA 3.2. *Let (A, B) be an exact k-separation of matroid M and let $x \in B$. Then*

- *$A \cup \{x\}$ is exactly k-separating if x belongs to either the guts or the coguts of (A, B) but not both;*
- *$A \cup \{x\}$ is exactly $(k-1)$-separating if x belongs to both the guts and the coguts of (A, B);*
- *$A \cup \{x\}$ is exactly $(k+1)$-separating if x belongs to neither the guts nor the coguts of (A, B).*

Suppose x is an element of the matroid M and let (A, B) be a k-separation of $M\backslash x$. Then x *blocks* (A, B) if neither $(A \cup \{x\}, B)$ nor $(A, B \cup \{x\})$ is a k-separation of M. Now let (A, B) be a k-separation of M/x. Then x *coblocks* (A, B) if neither $(A \cup \{x\}, B)$ nor $(A, B \cup \{x\})$ is a k-separation of M. The following lemma also follows easily from definitions.

LEMMA 3.3. *Let M be a matroid and let $\{A, B, \{x\}\}$ be a partition of $E(M)$. Then the following hold:*

- *If (A, B) is an exact k-separation of $M\backslash x$, then x blocks (A, B) if and only if x is not a coloop of M, $x \notin cl_M(A)$, and $x \notin cl_M(B)$.*
- *If (A, B) is an exact k-separation of M/x, then x coblocks (A, B) if and only if x is not a loop, $x \in cl_M(A)$, and $x \in cl_M(B)$.*

Suppose that X_1, X_2, Y_1, and Y_2 are sets. The pairs (X_1, Y_1) and (X_2, Y_2) are said to *cross* if all four sets $X_1 \cap X_2$, $X_1 \cap Y_2$, $Y_1 \cap X_2$, and $Y_1 \cap Y_2$ are nonempty. We omit the proof of the next lemma, which is a standard rank argument.

LEMMA 3.4. *Let e be an element of a 3-connected matroid M. Now, let (X_d, Y_d) be a 3-separation of $M\backslash e$ that is blocked by e, and let (X_c, Y_c) be a 3-separation of M/e that is coblocked by e. Then (X_d, Y_d) and (X_c, Y_c) cross. Moreover,*

- *one of $X_d \cap X_c$ and $Y_d \cap Y_c$ is 3-separating in M, and*
- *one of $X_d \cap Y_c$ and $Y_d \cap X_c$ is 3-separating in M.*

A matroid M is *internally 3-connected* if it is connected and, for each 2-separation (A, B) of M, either $|A| = 2$ or $|B| = 2$. The following result is due to Bixby [1].

LEMMA 3.5 (Bixby's lemma). *If e is an element of a 3-connected matroid M, then $M \setminus e$ or M/e is internally 3-connected.*

LEMMA 3.6. *Let (A, B) be a 3-separation of a 3-connected matroid M, where A is coclosed and $|A| \geq 4$. If $e \in A$ is in the guts of the separation (A, B), then $M \setminus e$ is 3-connected.*

Proof. Note that M/e is not internally 3-connected. Therefore, by Bixby's lemma, $M \setminus e$ is internally 3-connected. If $M \setminus e$ is not 3-connected, then there is a triad T of M with $e \in T$. Since $e \in \mathrm{cl}_M(B)$ and $e \in \mathrm{cl}_M(A - \{e\})$, we have $T \cap B \neq \emptyset$ and $T \cap (A - \{e\}) \neq \emptyset$. However, this contradicts the fact that A is coclosed. □

For disjoint sets $X, Y \subseteq E(M)$, we let $\kappa_M(X, Y) = \min\{\lambda_M(S) : X \subseteq S \subseteq E(M) \setminus Y\}$. It is clear that the function κ_M is minor monotone; that is, if N is a minor of M with $X \cup Y \subseteq E(N)$, then $\kappa_N(X, Y) \leq \kappa_M(X, Y)$. The following is due to Tutte [8].

THEOREM 3.7 (Tutte's Linking Theorem). *Let M be a matroid and let X and Y be disjoint subsets of $E(M)$. Then there exists a minor N of M with $E(N) = X \cup Y$ and $\lambda_N(X) = \kappa_M(X, Y)$.*

The next lemma is due to Geelen, Gerards, and Whittle [2, Lemma 4.11].

LEMMA 3.8. *Let M be a matroid and let $X, Y \subseteq E(M)$ be disjoint sets with $\kappa_M(X, Y) \geq k$. If $E(M)\backslash(X \cup Y) \neq \emptyset$, then either*

- *there exists an element $g \in E(M)\backslash(X \cup Y)$ such that $\kappa_{M/g}(X, Y) = \kappa_{M\backslash g}(X, Y) = \kappa_M(X, Y)$, or*
- *$\lambda_M(X) = k$ and there exists an ordering $b_1, b_2, \ldots, b_m$ of elements in $E(M)\backslash(X \cup Y)$ such that for $1 \leq i \leq m$, $\lambda_M(X \cup \{b_1, \ldots, b_i\}) = k$.*

4. Binary matroids and minors. We require the following lemma.

LEMMA 4.1. *Let (A, B) be a 3-separation of a matroid M, and let $C \subseteq B$ be a circuit of M with $\kappa_M(A, C) = 2$. Then there exists a minor N of M such that $A \subseteq E(N) \subseteq A \cup C$, $C \cap E(N)$ is a triangle of N, and $\lambda_N(A) = 2$.*

Proof. We start with the following claim.

4.2. *There exists a minor M' of M such that $E(M') = A \cup C$, $\lambda_{M'}(A) = 2$, and C is a circuit of M'.*

Subproof. Suppose that M' is a minor of M such that $A \cup C \subseteq E(M)$, $\kappa_{M'}(A, C) = 2$, and C is a circuit of M'. The proof is by induction on $|E(M') - (A \cup C)|$. The result is trivial if $|E(M') - (A \cup C)| = 0$; suppose otherwise, and let $e \in E(M') - (A \cup C)$. If $\kappa_{M'\backslash e}(A, C) = 2$, then the result follows inductively; we may assume otherwise. Therefore, e is in the coguts of a 3-separation (Z_1, Z_2), where $A \subseteq Z_1$ and $C \subseteq Z_2$. It follows that $e \notin \mathrm{cl}_{M'}(C)$ and, hence, that C is a circuit in M'/C. Moreover, by Tutte's Linking Theorem, $\kappa_{M'/e}(A, C) = 2$. Now, considering M'/e, the result follows inductively. □

Let M' be as given in the claim. The proof now proceeds by induction on $|C|$. If $|C| = 3$, then the result is immediate. Thus we may assume that $|C| \geq 4$. Since $\lambda_{M'}(A) = 2 < r_{M'}(C)$, there exists $e \in C - \mathrm{cl}_{M'}(A)$. Thus e is not in the guts of the 3-separation (A, C) of M'. Therefore, $\lambda_{M'/e}(A) = 2$. Moreover, $C - \{e\}$ is a circuit of M'/e; thus the result follows inductively. □

LEMMA 4.3. *Let N be an internally 4-connected minor of a binary matroid M and let (A, B) be a 3-separation of M with $|B \cap E(N)| \leq 3$. If M' is a minor of M with $A \subseteq E(N)$, $|E(M') \cap B| \geq 4$, and $\lambda_{M'}(X) \geq \min(2, |X|)$ for all $X \subseteq E(M') \cap B$, then M' has an N-minor.*

Proof. Let $B' = B \cap E(M')$. By duality, we may assume that either $|E(N) \cap B| \leq 2$ or that $E(N) \cap B$ is a triangle of N. Since M' is binary and $|B'| \geq 4$, B' cannot be a line in M'^*; thus, $r^*_{M'}(B') \geq 3$. Then $B' \not\subseteq \mathrm{cl}^*_{M'}(A)$ and, hence, B' contains a circuit C of M'. By Lemma 4.1, M' has a minor M'' such that $A \subseteq E(M'')$, $\lambda_{M'}(A) = 2$, and $B \cap E(M)$ is a triangle of M''. Evidently N is isomorphic to a minor of M'' and, hence, also of M'. □

LEMMA 4.4. *Let N be an internally 4-connected minor of a 3-connected binary matroid M and let (A, B) be a 3-separation of M with $|A|, |B| \geq 5$. If e is in the guts of (A, B), then $M \setminus e$ has an N-minor.*

Proof. By symmetry we may assume that $|E(N) \cap B| \leq 3$. Since e is in the guts of the 3-separation (A, B), M/e is not internally 3-connected. Therefore, by Bixby's lemma, $M \setminus e$ is internally 3-connected. Thus, $\mathrm{co}(M \setminus e)$ is 3-connected. Since $e \in \mathrm{cl}_M(A)$, there is no series-pair of $M \setminus e$ contained in B. Therefore, $\lambda_{M'}(X) \geq \min(2, |X|)$ for all $X \subseteq B - \{e\}$. Then, by Lemma 4.3, $M \setminus e$ has an N-minor. □

LEMMA 4.5. *Let N be an internally 4-connected minor of a 3-connected binary matroid M and let (A, B) be a 3-separation of M with $|B| \geq 5$ and $|E(N) \cap B| \leq 3$. If A is fully closed, then there exists $e \in B$ such that $M \setminus e$ and M/e both contain an N-minor.*

Proof. Assume by way of contradiction that the result is false. Let $b \in B$. By duality we may assume that M/b does not have an N-minor. Then, by Lemma 4.3, there exists a 2-separating set $Y \subseteq B - \{b\}$ of M/b with $|Y| \geq 2$. Let $X = Y \cup \{b\}$. Then $X \subseteq B$ is a 3-separating set of M.

By Lemma 3.8 and the fact that A is fully closed, there exists $e \in B - X$ such that $\kappa_{M \setminus e}(A, X) = \kappa_{M/e}(A, X) = 2$. If $|X| \geq 4$, then the result follows easily from Lemma 4.3. Thus we may assume that $|X| = 3$. Since Y is 2-separating in M/b, X is a triangle of M. Let $M' \in \{M \setminus e, M/e\}$. Thus, it suffices to prove that M' has an N-minor. Since A is fully closed in M, $X \not\subseteq \mathrm{cl}_{M'}(A)$. By Tutte's Linking Theorem there exists a partition (D, C) of $E(M) - (A \cup X)$ such that $\lambda_{M' \setminus D/C}(A) = 2$; we choose such D and C so that $|C|$ is minimal. Note that $X \subseteq \mathrm{cl}_{M' \setminus D/C}(A)$ but $X \not\subseteq \mathrm{cl}_{M'}(A)$. Thus $C \neq \emptyset$; choose $f \in C$. Now, let $M'' = M' \setminus D/(C - \{f\})$. By the minimality of C, we have $\lambda_{M'' \setminus f}(A) = 1$ and $\lambda_{M''/f}(A) = 2$. Thus $(A, X \cup \{f\})$ is a 3-separation of M'' consisting of a triangle X with a point f in the coguts. Then, by Lemma 4.3, M'' has an N-minor. Therefore, M' has an N-minor, as required. □

5. The internally 4-connected case. The goal of this section is to prove the following theorem.

THEOREM 5.1. *Let N be an internally 4-connected proper minor of an internally 4-connected binary matroid M with $|E(M)| \geq 7$. Then there exists $e \in E(M)$ such that either $M \setminus e$ or M/e is 4-connected up to separators of size 5 and has an N-minor.*

We will make use of the following lemma of Hall [4, Theorem 3.1].

LEMMA 5.2. *Let M be an internally 4-connected binary matroid and $\{a, b, c\}$ be a triangle of M. Then at least one of $M \setminus a$, $M \setminus b$, and $M \setminus c$ is 4-connected up to separators of size 5.*

Note that, by Lemma 5.2, if we find a triangle of M such that each of the three elements can be deleted to keep the N-minor, then Theorem 5.1 holds. Such a triangle will be called *an N-deletable triangle*. Similarly, an *N-contractible triad* is a triad with

the property that any one of its elements can be contracted to keep an N-minor.

Suppose M is an internally 4-connected binary matroid and M' is a minor of M. We call M' a *TT-connected minor of* M if the following hold:

- M' is internally 3-connected.
- If (X, Y) is a 3-separation of M', then either $|X| \leq 6$ or $|Y| \leq 6$.
- If (X, Y) is a 3-separation of M' with $\min(|X|, |Y|) = 6$, then one of X and Y can be partitioned into two disjoint subsets of size 3, each of which is a triangle or triad of M.

LEMMA 5.3. *Let M be an internally 4-connected binary matroid and let $e \in E(M)$. Then at least one of $M\backslash e$ and M/e is a TT-connected minor of M.*

Proof. Since M is internally 4-connected, $M\backslash e$ and M/e are both internally 3-connected. Either the lemma holds or there exist 3-separations (X_d, Y_d) and (X_c, Y_c) of $M\backslash e$ and M/e, respectively, such that the four sets X_d, Y_d, X_c, and Y_c all have size at least 6 and none of them is the union of two 3-separating sets in M. By Lemma 3.4, one of $X_d \cap X_c$ and $Y_d \cap Y_c$ is 3-separating in M, and one of $X_d \cap Y_c$ and $Y_d \cap X_c$ is 3-separating in M. By duality, we may assume that $X_d \cap X_c$ and $X_d \cap Y_c$ are 3-separating in M. Therefore X_d is the union of two 3-element 3-separating sets of M, which is a contradiction. □

LEMMA 5.4. *Let M be an internally 4-connected binary matroid and let N be an internally 4-connected minor of M with $E(N) \geq 10$. If $M\backslash e$ is a 3-connected TT-connected minor of M and has an N-minor, then there exists $f \in E(M)$ such that either $M\backslash f$ or M/f is 4-connected up to separators of size 5 and has an N-minor.*

Proof. Assume that $M\backslash e$ is not 4-connected up to separators of size 5. Then there exists a 3-separation (X, Y) of $M\backslash e$ with $|X| = 6$, $|Y| \geq 6$, and X is a disjoint union of two 3-element 3-separating sets, T_1 and T_2 of M. Since N is internally 4-connected and $E(N) \geq 10$, we must have $|E(N) \cap X| \leq 3$. Up to symmetry, we have two cases.

Case 1. T_1 is a triangle, and T_2 is a triad of M.

Since M is internally 4-connected, T_2 is closed in M and, hence, also in $M \setminus e$. Then, since M is binary, we must have $r_M(T_1 \cup T_2) = 5$. So $r^*_{M\setminus e}(T_1 \cup T_2) = 6 - (r(M) - r_M(Y)) = 6 + \lambda_{M\setminus e}(X) - r_M(X) = 3$. Now $T_1 \cup T_2$ is a rank-3 3-separating set in $(M \setminus e)^*$, and T_1 is a triad in $(M \setminus e)^*$. Therefore, $T_2 \subseteq \mathrm{cl}_{(M\setminus e)^*}(Y)$. Thus, by Lemma 4.4, T_2 is an N^*-deletable triangle in $(M \setminus e)^*$. Hence, T_2 is an N-contractible triad in M, proving the result.

Case 2. T_1 and T_2 are both triangles or both triads of M.

Choose $(M', N') \in \{(M\backslash e, N), ((M\backslash e)^*, N^*)\}$ such that T_1 and T_2 are both triads of M'. Since M'^* has no parallel pairs and since M'^* is binary, we have $r_{M'^*}(T_1 \cup T_2) = 4$. It follows that $r_{M'}(T_1 \cup T_2) = 4$. Thus, considering a geometric representation of M', T_1 and T_2 are triads spanning a common line. Now, by Lemma 4.3, we see that T_1 is an N'-contractible triad of M'. Hence, T_1 is either an N-contractible triad or an N-deletable triangle of M, proving the result. □

LEMMA 5.5. *Let M be an internally 4-connected binary matroid and let N be an internally 4-connected minor of M with $E(N) \geq 10$. Let $e \in E(M)$ such that both $M\backslash e$ and M/e have an N-minor. Then there exists $f \in E(M)$ such that either $M\backslash f$ or M/f is 4-connected up to separators of size 5 and has an N-minor.*

Proof. First assume e belongs to a triangle (or a triad) T of M. Since both $M\backslash e$ and M/e have an N-minor, T is an N-deletable triangle (or an N-contractible triad) of M. So the lemma follows from Lemma 5.2. Now we assume that e is not in a triangle or triad of M. Hence both $M\backslash e$ and M/e are 3-connected. So the result follows from Lemmas 5.3 and 5.4. □

Proof of Theorem 5.1. By the discussion in section 2, we may assume that $|E(N)| \geq 10$. By the Splitter Theorem, there exists $e \in E(M)$ and $M' \in \{M \setminus e, M/e\}$ such that M' is 3-connected and has an N-minor. Now, by Lemma 5.4, we can assume that M' is not a TT-connected minor of M. Let (A, B) be a 3-separation of M', where $|A|, |B| \geq 6$ and neither A nor B is a disjoint union of two 3-element 3-separating sets of M'. We may assume that $|E(N) \cap B| \leq 3$. Since $|E(N)| \geq 10$, $|A \cap E(N)| \geq 7$. Now, we may further assume that B is fully closed in M'.

By Lemma 5.5, we may assume that there is no element $f \in B$ such that $M' \setminus f$ and M'/f both have an N-minor. Then, by Lemma 4.5, there exists an element $f \in B$ that is in the closure or the coclosure of A in M'. By duality we may assume that $f \in \mathrm{cl}_{M'}(A)$. By Lemma 4.4, $M' \setminus f$ has a minor N' isomorphic to N.

Let $B' = B \cup \{e\} - \{f\}$. Note that $(A, B - \{f\})$ is a 2-separation of M'/f and, hence, (A, B') is a 3-separation of M/f. By Lemma 3.6, $M' \setminus f$ is 3-connected. Now it is easy to verify that e either blocks or coblocks the 3-separation $(A, B - \{f\})$ in $M' \setminus f$ and, hence, $M \setminus f$ is also 3-connected. By Lemma 5.4, we may assume that $M' \setminus f$ is not a TT-connected minor of M. Therefore, by Lemma 5.3, M'/f is a TT-connected minor of M. Now (A, B') is a 3-separation of M and $|A| \geq 7$. Therefore, $|B'| = 6$ and B' is the union of two 3-separating sets of M. Therefore there exists a triangle or triad $T \subseteq B'$ of M that contains e. First we consider the case that T is a triangle. Then, since M' is 3-connected, we have $M' = M \setminus e$. However, $T - \{e\} \subseteq B$, which contradicts the fact that e blocks the 3-separation (A, B) in M'. Now suppose that T is a triad. Then, since M' is 3-connected, we have $M' = M/e$. However, $T - \{e\} \subseteq B$, which contradicts the fact that e coblocks the 3-separation (A, B) in M'. □

6. Proof of the main theorem. In this section we complete the proof of Theorem 1.1. We break the proof into two cases depending on whether or not M is 4-connected up to separators of size 4.

LEMMA 6.1. *Let M be a binary matroid that is 4-connected up to separators of size 4 and let N be an internally 4-connected proper minor of M with $|E(N)| \geq 8$. Then there exists $e \in E(M)$ such that either $M \setminus e$ or M/e is 4-connected up to separators of size 5 and has an N-minor.*

Proof. By Theorem 5.1, we may assume that M has a 4-element 3-separating set $X = \{a, b, c, d\}$. Let $Y = E(M) - X$. By the Splitter Theorem, we may assume that $|E(M)| \geq 13$. Since M is binary, it suffices to consider the following two cases.

Case 1. (a, b, c, d) is a fan of M.

By symmetry we may assume that $\{a, b, c\}$ is a triangle. Note that N is a minor of either $M \setminus a$ or M/d. By duality we may assume that N is a minor of $M \setminus a$. Since M is 4-connected up to separators of size 4, X is fully closed in M. Then, by Lemma 3.6, $M \setminus a$ is 3-connected. Suppose that (A, B) is a 3-separation of $M \setminus a$ with $|A \cap \{b, c, d\}| \geq 2$. Then $A \cup \{b, c, d\}$ is 3-separating in $M \setminus a$ and, since $a \in \mathrm{cl}_M(\{b, c\})$, $A \cup X$ is 3-separating in M. It follows that $M \setminus a$ is 4-connected up to separators of size 5, as required.

Case 2. X is both a circuit and a cocircuit of M.

Since $|E(N)| \geq 8$ and N is internally 4-connected, we have $|E(N) \cap B| \leq 3$. By duality and symmetry, we may assume that N is a minor of $M \setminus a$. We claim that $M \setminus a$ is 4-connected up to separators of size 5. Since X is coclosed in M, $M \setminus a$ is cosimple. Suppose that (A, B) is a 2- or a 3-separation in $M \setminus a$ with $|A \cap \{b, c, d\}| \geq 2$. Then, since $a \in \mathrm{cl}_M(\{b, c, d\})$ and since $\{b, c, d\}$ is a triad in $M \setminus a$, $\lambda_M(B - X) = \lambda_M(A \cup X) = \lambda_{M \setminus a}(A \cup \{b, c, d\}) = \lambda_{M \setminus a}(A)$. Now $|B - X| \geq |B| - 1$. Thus if (A, B) is a 2-separation in $M \setminus a$, then, since M is 3-connected, $|B| \leq 2$.

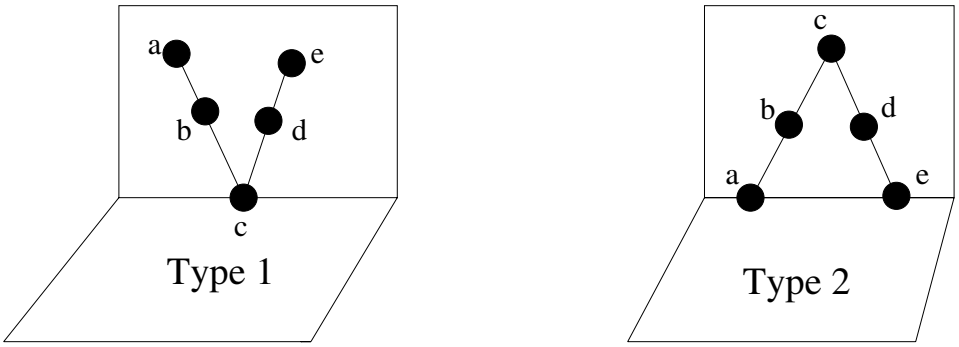


FIG. 2. 3-*separating sets of size* 5.

Since $M \setminus a$ is cosimple, $|B| \leq 1$ and, hence, $M \setminus a$ is 3-connected. Thus if (A, B) is a 3-separation in $M \setminus a$, then, since M is 4-connected up to separators of size 4, $|B| \leq 5$. Thus, $M \setminus a$ is 4-connected up to separators of size 5. □

Suppose that M is a binary matroid that is 4-connected up to separators of size 5 and that (X, Y) is a 3-separation of M with $|X| = 5$. Note that $r_M(X) + r^*_M(X) = r_M(X) + |X| - (r(M) - r_M(Y)) = |X| + \lambda_M(X) = 7$. Moreover, since M is binary, $r_M(X), r^*_M(X) \geq 3$. By duality we may assume that $r_M(X) = 3$. Now, since M is 3-connected and binary, there are either one or two elements of X in the guts of (X, Y). Thus, $X = \{a, b, c, d, e\}$ is of one of the following two types:

Type 1. $\{a, b, d, e\}$ is both a circuit and a cocircuit of M, and $\{a, b, c\}$ and $\{c, d, e\}$ are both triangles of M.

Type 2. (a, b, c, d, e) is a fan where $\{a, b, c\}$ is a triangle.

These two types of separations are depicted in Figure 2. The next lemma can be found in Hall [4].

LEMMA 6.2. *Let M be a matroid that is 4-connected up to separators of size 5 and let (X, Y) be a 3-separation of M with $X = \{a, b, c, d, e\}$.*

- *If X is a separation of Type 1, then one of $M\backslash a$, $M\backslash b$, and $M\backslash c$ is 4-connected up to separators of size 5.*
- *If X is a separation of Type 2, then one of $M \setminus a$, $M \setminus e$, and $\mathrm{co}(M \setminus c)$ is 4-connected up to separators of size 5.*

LEMMA 6.3. *Let M be a binary matroid that is 4-connected up to separators of size 5 and let N be an internally 4-connected proper minor of M with $|E(N)| \geq 8$. If $X = \{a, b, c, d, e\}$ is a 3-separating set of Type 1, then there exists $f \in X$ such that $M \setminus f$ is 4-connected up to separators of size 5 and has an N-minor.*

Proof. Since $|E(N)| \geq 8$, $|E(N) \cap X| \leq 3$. By Lemma 4.3, each of $M \setminus a$, $M \setminus c$, and $M \setminus e$ has an N-minor. So the theorem follows from Lemma 6.2. □

LEMMA 6.4. *Let M be a binary matroid that is 4-connected up to separators of size 5 and let N be an internally 4-connected proper minor of N with $|E(N)| \geq 7$. If $X = \{a, b, c, d, e\}$ is a 3-separating set of Type 2, then one of $M \setminus a$, $M \setminus e$, and $M \setminus c/d$ is 4-connected up to separators of size 5 and has an N-minor.*

Proof. Since X is a fan and N is internally 4-connected, $|E(N) \cap X| \leq 3$. By Lemma 4.3, both $M \setminus a$ and $M \setminus e$ have an N-minor. So we may assume that neither $M \setminus a$ nor $M \setminus e$ is 4-connected up to separators of size 5. So, by Lemma 6.2, $M \setminus c/d$ is 4-connected up to separators of size 5. Thus we may assume that $M \setminus c/d$ has no N-minor. It follows that $|E(N) \cap X| = 3$, that $E(N) \cap X$ is a triad of N, and that

none of M/b, M/c, and M/d has an N-minor.

6.5. *$M \setminus a$ is 3-connected and there exists a 3-separation (A, B) in $M \setminus a$ with $|A|, |B| \geq 6$ and with b or c in its coguts.*

Subproof. By Lemma 3.6, $M \setminus a$ is 3-connected. However, $M \setminus a$ is not 4-connected up to separators of size 5. So there exists a 3-separation (A, B) of $M \setminus a$ with $|A|, |B| \geq 6$. By symmetry we may assume that $|\{b, c, d\} \cap A| \geq 2$. Since $a \in \mathrm{cl}_M(\{b, c\})$ and since a blocks the separation (A, B), we have $|B \cap \{b, c\}| = 1$. Let $f \in \{b, c\} \cap B$. Since $\{b, c, d\}$ is a triad, f is in the coguts of (A, B). □

Let (A, B) be the 3-separation of $M \setminus a$ mentioned above and let $f \in \{b, c\}$ be in its coguts. By Lemma 4.4, $M \setminus a/f$ has an N-minor. But this contradicts the fact that M/f has no N-minor. □

REFERENCES

[1] R. E. Bixby, *A simple theorem on 3-connectivity*, Linear Algebra Appl., 45 (1982), pp. 123–126.

[2] J. F. Geelen, A. M. H. Gerards, and G. Whittle, *Excluding a Planar Graph from $GF(q)$-Representable Matroids*, manuscript.

[3] J. F. Geelen and G. Whittle, *Matroid 4-connectivity: A deletion-contraction theorem*, J. Combin. Theory Ser. B, 83 (2001), pp. 15–37.

[4] R. Hall, *A chain theorem for 4-connected matroids*, J. Combin. Theory Ser. B, 93 (2005), pp. 45–66.

[5] T. Johnson and R. Thomas, *Generating internally four-connected graphs*, J. Combin. Theory Ser. B, 85 (2002), pp. 21–58.

[6] J. G. Oxley, *Matroid Theory*, Oxford University Press, New York, 1992.

[7] P. D. Seymour, *Decomposition of regular matroids*, J. Combin. Theory Ser. B, 28 (1980), pp. 305–359.

[8] W. T. Tutte, *Connectivity in matroids*, Canad. J. Math., 18 (1966), pp. 1301–1324.

[9] X. Zhou, *On internally 4-connected non-regular binary matroids*, J. Combin. Theory Ser. B, 91 (2004), pp. 327–343.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 588–596

MATROID T-CONNECTIVITY*

JIM GEELEN†, BERT GERARDS‡, AND GEOFF WHITTLE§

Abstract. We introduce a new generalization of the maximum matching problem to matroids; this problem includes Gallai's T-path problem for graphs.

Key words. matroids, Gallai's T-paths theorem, paths, connectivity, matching

AMS subject classification. 05B35

DOI. 10.1137/050634190

1. Introduction. Let $G = (V, E)$ be a simple graph and let $T \subseteq V$. A *T-path* is a path in G connecting two vertices in T. Let $\nu_G(T)$ denote the maximum number of vertex disjoint T-paths in G. This parameter was introduced by Gallai [2], who showed that determining $\nu_G(T)$ is equivalent to the maximum matching problem. (Note that $\nu_G(V)$ is the size of a maximum matching in G.) As a consequence of an exact min-max theorem for $\nu_G(T)$, Gallai [2] proved the following theorem.

THEOREM 1.1 (Gallai [2]). *Let $G = (V, E)$ be a graph and $T \subseteq V$. Then there exists a set $X \subseteq V$ that hits every T-path such that $|X| \leq 2\nu_G(T)$.*

Note that if $X \subseteq V$ hits each T-path, then $\nu_G(T) \leq |X|$. Gallai's theorem shows that this natural upper bound for $\nu_G(T)$ is within a factor of 2 of being tight. We consider a matroidal generalization of $\nu_G(T)$ and prove analogous upper bounds. This problem arose naturally in proving structural results on minor-closed classes of matroids represented over finite fields. The main result presented here is needed as a lemma in that project.

Let M be a matroid. For $X \subseteq E(M)$ we let

$$\lambda_M(X) = r_M(X) + r_M(E(M) - X) - r(M).$$

For disjoint sets $S, T \subseteq E(M)$, we let

$$\kappa_M(S, T) = \min(\lambda_M(X) \,:\, S \subseteq X \subseteq E(M) - T).$$

Then, for a set $T \subseteq E(M)$, we let

$$\nu_M(T) = \max(\kappa_M(X, T - X \,:\, X \subseteq T);$$

we call $\nu_M(T)$ the *T-connectivity* of M. It is straightforward to verify that $\lambda_M(X) = \lambda_{M^*}(X)$. Therefore $\kappa_M(S, T) = \kappa_{M^*}(S, T)$ and, hence, $\nu_M(T) = \nu_{M^*}(T)$. We will consider a slightly more general parameter. Let $\mathcal{T}$ be a collection of disjoint subsets

*Received by the editors June 21, 2005; accepted for publication (in revised form) February 2, 2006; published electronically August 7, 2006. This research was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Marsden Fund of New Zealand.

http://www.siam.org/journals/sidma/20-3/63419.html

†Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Canada (jfgeelen@uwaterloo.ca).

‡CWI, Amsterdam and Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands (bert.gerards@cwi.nl).

§School of Mathematical and Computing Sciences, Victoria University, Wellington, New Zealand (geoff.whittle@mcs.vuw.ac.nz).

of $E(M)$. Then we define $\nu_M(\mathcal{T})$ to be the maximum of $\kappa_M(X,Y)$, where $X = \cup\mathcal{T}_1$ and $Y = \cup\mathcal{T}_2$ for a partition $(\mathcal{T}_1, \mathcal{T}_2)$ of $\mathcal{T}$. Thus, if $\mathcal{T}$ is a partition of a set $T \subseteq E(M)$ into singletons, then $\nu_M(T) = \nu_M(\mathcal{T})$. We also call $\nu_M(\mathcal{T})$ the *$\mathcal{T}$-connectivity* of M.

Let $G = (V, E)$ be a simple graph. We can construct a matroid M on $V \cup E$ such that V is a basis of M and, for each edge $e = uv$ of G, the element e is placed freely on the line through u and v. Note that if P is a nontrivial (u,v)-path in G, then $\{u,v\} \cup E(P)$ is a circuit of M. Now it is a straightforward application of Menger's theorem to prove that for any two disjoint subsets S and T of vertices of G, $\kappa_M(S,T)$ is equal to the maximum number of vertex disjoint (S,T)-paths in G. Now it is easy to see that, for any $T \subseteq V$, we have $\nu_M(T) = \nu_G(T)$.

Let $\mathcal{T}$ be a collection of disjoint subsets of V. Let $\nu_G(\mathcal{T})$ denote the maximum, taken over all partitions $(\mathcal{T}_1, \mathcal{T}_2)$ of $\mathcal{T}$, of the connectivity between $\cup\mathcal{T}_1$ and $\cup\mathcal{T}_2$ in G. Thus $\nu_G(\mathcal{T}) = \nu_M(\mathcal{T})$. A $\mathcal{T}$-path is a path whose ends are in distinct parts of $\mathcal{T}$. Mader [5] considered the related problem of finding the maximum number, $\mu_G(\mathcal{T})$, of vertex disjoint $\mathcal{T}$-paths. It is straightforward to show that $\nu_M(\mathcal{T}) \leq \mu_G(\mathcal{T}) \leq 2\nu_M(\mathcal{T})$. (Indeed, the first inequality is trivial and the second comes from the fact that when taking a random partition $(\mathcal{T}_1, \mathcal{T}_2)$ of $\mathcal{T}$ we expect half of Mader's $\mathcal{T}$-paths to connect $\cup\mathcal{T}_1$ and $\cup\mathcal{T}_2$.) This bound is interesting since $\mu_G(\mathcal{T})$ can be computed efficiently (see Lovász [4] or Chudnovsky, Cunningham, and Geelen [1]), while computing $\nu_G(\mathcal{T})$ is NP-hard. Indeed, suppose that G is a graph consisting of a perfect matching, $\mathcal{T}$ is a partition of $V(G)$, and G' is obtained from G by shrinking each part of $\mathcal{T}$ to a single vertex. Then $\nu_G(\mathcal{T})$ is the size of a maximum cut in G'. Therefore computing $\nu_G(\mathcal{T})$ is NP-hard, as claimed. Moreover, this implies that computing $\nu_M(\mathcal{T})$ is NP-hard.

Let M_1 and M_2 be matroids on a common ground set E. We say that M_2 is obtained by an *elementary transformation* on M_1 if there exists a matroid N on $E \cup \{e\}$ such that either $M_1 = N \setminus e$ and $M_2 = N/e$ or $M_1 = N/e$ and $M_2 = N \setminus e$. We define $\mathrm{dist}(M_1, M_2)$ to be the minimum number of elementary transformations required to transform M_1 into M_2. The following properties are straightforward to verify; the last of these properties shows that $\mathrm{dist}(M_1, M_2)$ is well defined:

- $\mathrm{dist}(M_1, M_2) = \mathrm{dist}(M_2, M_1)$.
- $\mathrm{dist}(M_1^*, M_2^*) = \mathrm{dist}(M_1, M_2)$.
- If M' is the rank-zero matroid on E, then $\mathrm{dist}(M_1, M') = r(M_1)$.
- If M_3 is a matroid on E, then $\mathrm{dist}(M_1, M_3) \leq \mathrm{dist}(M_1, M_2) + \mathrm{dist}(M_2, M_3)$.
- $\mathrm{dist}(M_1, M_2) \leq |E|$.

We use the following lemma.

LEMMA 1.2. *Let M_1 and M_2 be matroids on a common ground set E and let $\mathcal{T}$ be a collection of disjoint subsets of E. Then $\nu_{M_1}(\mathcal{T}) \leq \nu_{M_2}(\mathcal{T}) + \mathrm{dist}(M_1, M_2)$.*

Proof. By a simple inductive argument we may assume that $\mathrm{dist}(M_1, M_2) = 1$. Moreover, by duality we may assume that $M_1 = N \setminus e$ and $M_2 = N/e$. Now it is easy to check that $\nu_{M_1}(\mathcal{T}) \leq \nu_N(\mathcal{T}) \leq \nu_{M_2}(\mathcal{T}) + 1$, as required. □

Note that $\nu_M(\mathcal{T}) = 0$ if and only if no component of M contains elements from two distinct parts of $\mathcal{T}$. Let $T = \cup\mathcal{T}$ and let $\delta_M(\mathcal{T}) = \max(\kappa_M(X, T - X) : X \in \mathcal{T})$. Note that $\delta_M(\mathcal{T}) \leq \nu_M(\mathcal{T})$ and, when $\mathcal{T}$ contains only singletons, $\delta_M(\mathcal{T}) \leq 1$. The main result of this paper is the following.

THEOREM 1.3. *Let M be a matroid and let $\mathcal{T}$ be a collection of disjoint subsets of $E(M)$. Then there exists a matroid M' on the ground set $E(M)$ such that $\nu_{M'}(\mathcal{T}) = 0$ and $\mathrm{dist}(M, M') \leq 2(\delta_M(\mathcal{T}) + 1)\nu_M(\mathcal{T})$.*

The next result is an easy consequence of Theorem 1.3. We say that a partition

$\mathcal{P}$ of $E(M)$ *encloses* $\mathcal{T}$ if each set in $\mathcal{T}$ is contained in some set in $\mathcal{P}$ and no set in $\mathcal{P}$ contains two or more sets in $\mathcal{T}$. The *order* of $\mathcal{P}$, denoted by $\mathrm{ord}_M(\mathcal{P})$, is defined as $\max(\lambda_M(\cup\mathcal{Q}) : \mathcal{Q} \subseteq \mathcal{P})$. Note that if $\mathcal{P}$ encloses $\mathcal{T}$, then $\mathrm{ord}_M(\mathcal{P}) \geq \nu_M(\mathcal{T})$.

COROLLARY 1.4. *Let M be a matroid and let $\mathcal{T}$ be a collection of disjoint subsets of $E(M)$. Then there exists a partition $\mathcal{P}$ of $E(M)$ enclosing $\mathcal{T}$ where $ord_M(\mathcal{P}) \leq 2(\delta_M(\mathcal{T}) + 1)\nu_M(\mathcal{T})$.*

While Corollary 1.4 does follow from Theorem 1.3, we will not include the easy proof since Corollary 1.4 is an immediate consequence of Theorem 4.1.

We conclude the introduction by stating some open problems.

PROBLEM 1.5. *Can the bound of $2(\delta_M(\mathcal{T})+1)\nu_M(\mathcal{T})$ in Theorem 1.3 be improved to $c\nu_M(\mathcal{T})$ for some constant c?*

PROBLEM 1.6. *In the case that each element of $\mathcal{T}$ is a singleton, can the bound of $2(\delta_M(\mathcal{T}) + 1)\nu_M(\mathcal{T})$ in Theorem 1.3 be improved to $2\nu_M(\mathcal{T})$?*

We now turn to the problem of finding a tight bound on T-connectivity. If M' is a matroid on the ground set $E(M)$, then it is straightforward to prove that

$$\nu_M(T) \leq \mathrm{dist}(M, M') + \sum \left(\left\lfloor \frac{|T \cap F|}{2} \right\rfloor : F \text{ a component of } M' \right).$$

PROBLEM 1.7. *Is there always a matroid M' for which equality is attained?*

Recall that computing $\nu_M(\mathcal{T})$ is NP-hard. The final problems concern the complexity of determining $\nu_M(T)$; as usual we assume that the matroid is given by its rank oracle.

PROBLEM 1.8. *Is there a polynomial-time algorithm for computing $\nu_M(T)$?*

It is straightforward to show that $\nu_M(E(M))$ is the size of a maximum common independent set of M and M^*. So we can compute $\nu_M(E(M))$ efficiently via matroid intersection. The following special case of Problem 1.8 contains the matching problem.

PROBLEM 1.9. *Is there a polynomial-time algorithm for computing $\nu_M(B)$ where B is a basis of M?*

The above problems are all open for the class of representable matroids.

2. Submodular functions. This section contains notation, definitions, and elementary results on submodular functions.

A *set function* on a set E is an integer valued function defined on the collection of subsets of E. Let λ be a set function on E. Then

- λ is *submodular* if $\lambda(X) + \lambda(Y) \geq \lambda(X \cap Y) + \lambda(X \cup Y)$ for each $X, Y \subseteq E$;
- λ is *nonnegative* if $\lambda(X) \geq 0$ for each $X \subseteq E$;
- λ is *symmetric* if $\lambda(X) = \lambda(E - X)$ for each $X \subseteq E$.

We call $K = (E, \lambda)$ a *connectivity system* if λ is a symmetric, submodular, nonnegative set function on a finite set E. For a matroid M we define $K(M) = (E(M), \lambda_M)$; $K(M)$ is readily seen to be a connectivity system. Let $K = (E, \lambda)$ be a connectivity system and let S and T be disjoint subsets of E. Now let $\kappa_K(S, T) = \min(\lambda(X) : S \subseteq X \subseteq E - T)$. Finally, for a collection $\mathcal{T}$ of disjoint subsets of E, we let $\nu_K(\mathcal{T}) = \max \kappa_K(X, Y)$ where the maximum is taken over all partitions (X, Y) of $\cup\mathcal{T}$ where X is the union of a subcollection of $\mathcal{T}$. When $\mathcal{T}$ is a partition of a set $T \subseteq E$ into singletons, then we let $\nu_M(T) = \nu_M(\mathcal{T})$. In section 4 we provide upper bounds on $\nu_K(T)$. In the remainder of this section we consider preliminary results.

A set function r on E is *nondecreasing* if $r(X) \leq r(Y)$ whenever $X \subseteq Y$.

LEMMA 2.1. *Let $K = (E, \lambda)$ be a connectivity system, let $T \subseteq E$, and let $r(S) = \kappa_K(S, T)$ for each $S \subseteq E - T$. Then r is a nondecreasing, submodular, nonnegative set function on $E - T$.*

Proof. It is clear that r is nondecreasing and nonnegative. Let $S_1, S_2 \subseteq E - T$. Then, for $i \in \{1,2\}$, there exists a set X_i such that $S_i \subseteq X_i \subseteq E - T$ and $\lambda(X_i) = \kappa_K(S_i, T) = r(S_i)$. Note that $S_1 \cap S_2 \subseteq X_1 \cap X_2 \subseteq E - T$ and $S_1 \cup S_2 \subseteq X_1 \cup X_2 \subseteq E - T$. Therefore $\lambda(X_1 \cap X_2) \geq \kappa_K(S_1 \cap S_2, T) = r(S_1 \cap S_2)$ and $\lambda(X_1 \cup X_2) \geq \kappa_K(S_1 \cup S_2, T) = r(S_1 \cup S_2)$. Hence

$$\begin{aligned} r(S_1) + r(S_2) &= \lambda(X_1) + \lambda(X_2) \\ &\geq \lambda(X_1 \cap X_2) + \lambda(X_1 \cup X_2) \\ &\geq r(S_1 \cap S_2) + r(S_1 \cup S_2). \end{aligned}$$

Therefore r is submodular, as required. □

The following result is well known in the context of polymatroids.

LEMMA 2.2. *Let r be a nondecreasing, submodular set function on a finite set E. If $X \subseteq Y \subseteq E$ and $r(X \cup \{e\}) = r(X)$ for each $e \in Y - X$, then $r(X) = r(Y)$.*

Proof. Suppose otherwise and choose Y' minimal such that $X \subseteq Y' \subseteq Y$ and $r(Y') > r(X)$. Clearly $|Y'| \geq |X| + 2$. Let $e \in Y' - X$. By our choice of Y', $r(Y' - \{e\}) = r(X)$ and $r(X \cup \{e\}) = r(X)$. Now, by submodularity, $r(X \cup \{e\}) + r(Y' - \{e\}) \geq r(X) + r(Y')$. But then $r(Y') \leq r(Y' - \{e\}) = r(X)$; this contradiction completes the proof. □

LEMMA 2.3. *Let $K = (E, \lambda)$ be a connectivity system and let S and T be disjoint subsets of E. Then there exist sets $S' \subseteq S$ and $T' \subseteq T$ such that $\kappa_K(S', T') = \kappa_K(S, T)$ and $|S'|, |T'| \leq \kappa_K(S, T)$.*

Proof. Choose $S' \subseteq S$ maximal such that $\kappa_K(S', T) \geq |S'|$. Note that this is well defined since $\kappa_K(\emptyset, T) \geq 0$. By the definition of S' we have $\kappa_K(S' \cup \{e\}) = \kappa_K(S')$ for all $e \in S - S'$. Therefore, by Lemmas 2.1 and 2.2, $\kappa_K(S', T) = \kappa_K(S, T)$. Now choose $T' \subseteq T$ maximal such that $\kappa_K(S', T') \geq |T'|$. As above we get $\kappa_K(S', T') = \kappa_K(S', T) = \kappa_K(S, T)$, as required. □

LEMMA 2.4. *Let $K = (E, \lambda)$ be a connectivity system, let S and T be disjoint subsets of E with $\kappa_K(S, T) = k$, and let $\mathcal{S} = \{X : S \subseteq X \subseteq E - T \text{ and } \lambda(X) = k\}$. If $X, Y \in \mathcal{S}$, then $X \cap Y, X \cup Y \in \mathcal{S}$.*

Proof. Note that $S \subseteq X \cap Y \subseteq X \cup Y \subseteq E - T$. Then, since $\kappa_K(S, T) = k$ we have $\lambda(X \cap Y), \lambda(X \cup Y) \geq k$. Moreover, by submodularity, we have $2k = \lambda(X) + \lambda(Y) \geq \lambda(X \cap Y) + \lambda(X \cup Y) \geq 2k$. It follows that $\lambda(X \cap Y) = k$ and $\lambda(X \cup Y) = k$. Therefore $X \cap Y, X \cup Y \in \mathcal{S}$, as required. □

3. Homomorphisms. Let $K = (E, \lambda)$ be a connectivity system and let $X \subseteq E$. We define a set function λ' on $(E - X) \cup \{e_X\}$ such that for each $Y \subseteq E - X$, $\lambda'(Y) = \lambda(Y)$ and $\lambda'(Y \cup \{e_X\}) = \lambda(Y \cup X)$. Now let $K \circ X = ((E - X) \cup \{e_X\}, \lambda')$. It is easy to verify that $K \circ X$ is a connectivity system; we say that $K \circ X$ is obtained from K by *identifying* X. If $\mathcal{T}$ is a collection of disjoint subsets of E, then we let $K \circ \mathcal{T}$ denote the connectivity system obtained by identifying each set in $\mathcal{T}$.

Remark. If $K = (E, \lambda)$ is a connectivity system and $\mathcal{T}$ is a collection of disjoint subsets of E, and if $T = \{e_X : X \in \mathcal{T}\}$, then $\nu_K(\mathcal{T}) = \nu_{K \circ \mathcal{T}}(T)$.

By the above remark, we can reduce the problem of computing $\nu_K(\mathcal{T})$ to the apparently easier problem of computing $\nu_K(T)$.

THEOREM 3.1. *Let $K = (E, \lambda)$ be a connectivity system and let $\mathcal{T} = \{T_1, \ldots, T_l\}$ be a partition of $T \subseteq E$. Then there exists a collection $\mathcal{T}' = \{T_1', \ldots, T_l'\}$ of disjoint sets such that $\nu_K(\mathcal{T}') = \nu_K(\mathcal{T})$ and, for each $i \in \{1, \ldots, l\}$, $T_i \subseteq T_i'$ and $\lambda(T_i') = \kappa_K(T_i, T - T_i)$.*

Note that Theorem 3.1 is an immediate corollary of the following lemma.

LEMMA 3.2. *Let $K = (E, \lambda)$ be a connectivity system, let A, B, and C be disjoint subsets of E, and let X be any set satisfying $A \subseteq X \subseteq E - (B \cup C)$ and $\lambda(X) = \kappa_K(A, B \cup C)$. Then $\nu_K(\{A, B, C\}) = \nu_K(\{X, B, C\})$.*

Proof. Note that by symmetry it suffices to prove that $\kappa_K(B, A \cup C) = \kappa_K(B, X \cup C)$. Let Y be a set satisfying $B \subseteq Y \subseteq E - (A \cup C)$ and $\lambda(Y) = \kappa_K(B, A \cup C)$. Since $A \subseteq X - Y \subseteq E - (B \cup C)$ and $B \subseteq Y - X \subseteq E - (A \cup C)$, we have $\lambda(Y) \leq \lambda(Y - X)$ and $\lambda(X) \leq \lambda(X - Y)$. However, by submodularity and symmetry, we have

$$\lambda(Y) + \lambda(X) \geq \lambda(Y - X) + \lambda(X - Y).$$

Therefore $\lambda(Y) = \lambda(Y - X)$ and $\lambda(X) = \lambda(X - Y)$. Then, since $B \subseteq Y - X \subseteq E - (X \cup C)$, we have $\kappa_K(B, X \cup C) = \kappa_K(B, A \cup C)$, as required. □

4. Connectivity systems. Let $K = (E, \lambda)$ be a connectivity system and let $\mathcal{T}$ be a collection of disjoint subsets of E. Now let $\mathcal{P}$ be a partition of E. The *order* of $\mathcal{P}$, denoted $\mathrm{ord}_K(\mathcal{P})$, is $\max(\lambda(\cup\mathcal{S}) : \mathcal{S} \subseteq \mathcal{P})$. Note that if $\mathcal{P}$ encloses $\mathcal{T}$, then $\nu_K(\mathcal{T}) \leq \mathrm{ord}_K(\mathcal{P})$. Let $T = \cup\mathcal{T}$ and let $\delta_K(\mathcal{T}) = \max(\kappa_K(X, T - X) : X \in \mathcal{T})$. One of the main results of this section is the following.

THEOREM 4.1. *Let $K = (E, \lambda)$ be a connectivity system and let $\mathcal{T}$ be a collection of disjoint subsets of E. Then there exists a partition $\mathcal{P}$ of E enclosing $\mathcal{T}$ with $ord_K(\mathcal{P}) \leq 2(1 + \delta_K(\mathcal{T}))\nu_K(\mathcal{T})$.*

We conjecture that this bound can be sharpened from $2(1 + \delta_K(\mathcal{T}))\nu_K(\mathcal{T})$ to $2\nu_K(\mathcal{T})$.

The problem of computing $\mathrm{ord}_K(\mathcal{P})$ is easily seen to contain the max-cut problem and is therefore NP-hard. We will introduce another notion, a (T, k)-dissection, that also provides an upper bound on $\nu_K(T)$. However, the key properties of a (T, k)-dissection can be verified efficiently.

A triple $(A, B, \mathcal{P})$ is a (T, k)-*dissection* if it satisfies the following:

- $\mathcal{P} \cup \{A, B\}$ is a partition of E.
- $|A \cap T|, |B \cap T| \leq k$ and $|P \cap T| = 1$ for each $P \in \mathcal{P}$.
- $\kappa_K(A, B) = k$.
- $\lambda(A \cup P) = k$ for each $P \in \mathcal{P}$.

Note that the third property above is the only property that is nontrivial to verify. However, we can compute $\kappa_K(A, B)$ efficiently via submodular function minimization (see Iwata, Fleischer, and Fujishige [3] or Schrijver [7]). Therefore we can efficiently verify that a triple is a (T, k)-dissection.

THEOREM 4.2. *Let $K = (E, \lambda)$ be a connectivity system and let $T \subseteq E$ where $\nu_K(T) = k$. Then K admits a (T, k)-dissection.*

Proof. Let (T_1, T_2) be a partition of T such that $\kappa_K(T_1, T_2) = k$. By Lemma 2.3, there exists $A' \subseteq T_1$ and $B' \subseteq T_2$ such that $\kappa_K(A', B') = k$ and $|A'|, |B'| \leq k$. Let $\mathcal{A} = \{X : A' \subseteq X \subseteq E - B' \text{ and } \lambda(X) = k\}$. By Lemma 2.4, $\mathcal{A}$ is closed under intersections and unions.

For each set $Z \subseteq T$ with $A' \subseteq Z \subseteq T - B'$, we have $\kappa_K(Z, T - Z) = k$. Therefore there exists $X \in \mathcal{A}$ such that $X \cap T = Z$. Choose a set $A \in \mathcal{A}$ as large as possible such that $A \cap T = A'$. Now, for each element $e \in T - (A' \cup B')$, choose a set $A_e \in \mathcal{S}$ as large as possible such that $A_e \cap T = A' \cup \{e\}$. Note that $A \cup A_e \in \mathcal{A}$ and $(A \cup A_e) \cap T = A' \cup \{e\}$. Therefore, by the maximality of A_e, we have $A \subseteq A_e$. Now consider two distinct elements $e, f \in T - (A' \cup B')$. Note that $A \subseteq A_e \cap A_f \in \mathcal{A}$ and $(A_e \cap A_f) \cap T = A'$. Therefore, by the maximality of A, we have $A_e \cap A_f = A$. Now let $B = E - \cup(A_e : e \in T - (A' \cup B'))$ and let $\mathcal{P} = (A_e - A : e \in T - (A' \cup B'))$. Then $(A, B, \mathcal{P})$ is a (T, k)-dissection. □

For $T \subseteq E$ we let $\Delta_K(T) = \max(\lambda(\{e\} : e \in T))$.

THEOREM 4.3. *Let $K = (E, \lambda)$ be a connectivity system, let $T \subseteq E$, let $\mathcal{T}$ be the partition of T into singletons, and let $(A, B, \mathcal{P})$ be a (T, k)-dissection. Then there exist partitions $\mathcal{A}$ of A and $\mathcal{B}$ of B such that $\mathcal{A} \cup \mathcal{B} \cup \mathcal{P}$ encloses $\mathcal{T}$ and $ord_K(\mathcal{A} \cup \mathcal{B} \cup \mathcal{P}) \leq 2(1 + \Delta_K(T))k$. Hence $\nu_K(T) \leq 2(1 + \Delta_K(T))k$.*

Proof. Let $\mathcal{A} = \{A - T\} \cup \{\{e\} : e \in A \cap T\}$, $\mathcal{B} = \{B - T\} \cup \{\{e\} : e \in B \cap T\}$, and $\mathcal{C} = \mathcal{A} \cup \mathcal{B} \cup \mathcal{P}$. Note that $\mathcal{C}$ encloses $\mathcal{T}$; it remains to prove that $\text{ord}(\mathcal{C}) \leq 2(1 + \Delta_K(T))k$.

4.3.1. *$ord_K(\mathcal{P} \cup \{A, B\}) \leq 2k$.*

Subproof. By definition, $\lambda(A \cup P) = k$ for each $P \in \mathcal{P}$. Therefore, by Lemma 2.4, $\lambda(A \cup (\cup\mathcal{Q})) = k$ for each $\mathcal{Q} \subseteq \mathcal{P}$. By symmetry, $\lambda(B \cup (\cup\mathcal{Q})) = k$ for each $\mathcal{Q} \subseteq \mathcal{P}$. Now, by submodularity, $\lambda(\cup\mathcal{Q}) + \lambda(A \cup B \cup (\cup\mathcal{Q})) \leq \lambda(A \cup (\cup\mathcal{Q})) + \lambda(B \cup (\cup\mathcal{Q})) = 2k$ for each $\mathcal{Q} \subseteq \mathcal{P}$. Therefore $\lambda(\cup\mathcal{Q}) \leq 2k$ and $\lambda(A \cup B \cup (\cup\mathcal{Q})) \leq 2k$. Thus $\text{ord}_K(\mathcal{P} \cup \{A, B\}) \leq 2k$, as required. □

Consider a set $\mathcal{Q} \subseteq \mathcal{C}$. Let $X = \cup\mathcal{Q}$ and let $Y = E - X$. Note that either $|X \cap A| \leq k$ or $|Y \cap A| \leq k$. By symmetry we may assume that $|X \cap A| \leq k$. Similarly, either $|X \cap B| \leq k$ or $|Y \cap B| \leq k$. Consider the case that $|X \cap B| \leq k$. Then, by submodularity and statement 4.3.1, $\lambda(X) \leq 2k\Delta_K(T) + \lambda(X - (A \cup B)) \leq 2k\Delta_K(T) + 2k$. Finally, consider the case that $|Y \cap B| \leq k$. By submodularity and statement 4.3.1, $\lambda(X) \leq 2k\Delta_K(T) + \lambda((X - A) \cup B) \leq 2k\Delta_K(T) + 2k$. Therefore $\text{ord}_K(\mathcal{A} \cup \mathcal{B} \cup \mathcal{P}) \leq 2(1 + \Delta_K(T))k$, as required. □

We can now put these results together to prove Theorem 4.1. By Theorem 3.1 we may assume that $\lambda(X) \leq \delta_K(\mathcal{T})$ for each $X \in \mathcal{T}$. Then, by possibly applying a homomorphism, we may assume that each part of $\mathcal{T}$ is a singleton. Now Theorem 4.1 is an immediate consequence of Theorems 4.2 and 4.3.

5. Back to matroids.

LEMMA 5.1. *Let (S, A_1, A_2, T) be a partition of the elements of a matroid M such that $\lambda_M(S \cup A_1) + \lambda_M(S \cup A_2) = \lambda_M(S) + \lambda_M(S \cup A_1 \cup A_2)$. Then $\lambda_{M/S\backslash T}(A_1) = 0$.*

Proof. We have

$$\begin{aligned}
0 &= \lambda_M(S \cup A_1) + \lambda_M(S \cup A_2) - \lambda_M(S) - \lambda_M(S \cup A_1 \cup A_2) \\
&= (r_M(S \cup A_1) + r_M(T \cup A_2) - r_M(E)) \\
&\quad +(r_M(S \cup A_2) + r_M(T \cup A_1) - r_M(E)) \\
&\quad -(r_M(S) + r_M(T \cup A_1 \cup A_2) - r_M(E)) \\
&\quad -(r_M(S \cup A_1 \cup A_2) + r_M(T) - r_M(E)) \\
&= (r_M(S \cup A_1) + r_M(S \cup A_2) - r_M(S) - r_M(S \cup A_1 \cup A_2)) \\
&\quad +(r_M(T \cup A_1) + r_M(T \cup A_2) - r_M(T) - r_M(T \cup A_1 \cup A_2)) \\
&= (r_{M/S}(A_1) + r_{M/S}(A_2) - r_{M/S}(A_1 \cup A_2)) \\
&\quad +(r_{M/T}(A_1) + r_{M/T}(A_2) - r_{M/T}(A_1 \cup A_2)) \\
&= \lambda_{M/S\backslash T}(A_1) + \lambda_{M\backslash S/T}(A_1).
\end{aligned}$$

Therefore, since the last expression is the sum of two nonnegative values, we get $\lambda_{M/S\backslash T}(A_1) = 0$ and $\lambda_{M\backslash S/T}(A_1) = 0$, as required. □

LEMMA 5.2. *Let $(S, A_1, \ldots, A_l, T)$ be a partition of the elements of a matroid M such that $\kappa_M(S, T) = k$ and, for each $i \in \{1, \ldots, l\}$, $\lambda_M(S \cup A_i) = k$. Then $\lambda_{M/S\backslash T}(A_i) = 0$ for all $i \in \{1, \ldots, l\}$.*

Proof. By Lemma 2.4, $\lambda_M(S \cup (\cup_{i\in X} A_i)) = k$ for all $X \subseteq \{1,\dots,l\}$. Let $A'_2 = A_2 \cup \cdots \cup A_l$. Applying Lemma 5.1 to (S, A_1, A'_2, T) we see that $\lambda_{M/S\setminus T}(A_1) = 0$. Then, by symmetry, $\lambda_{M/S\setminus T}(A_i) = 0$ for all $i \in \{1,\dots,l\}$. □

The following result is an immediate corollary of Lemma 5.2 and Theorem 4.2.

LEMMA 5.3. *Let $M=(E,r)$ be a matroid and let $\mathcal{T}$ be a collection of disjoint subsets of $E(M)$ with $\nu_M(\mathcal{T})=k$. Then there exist disjoint sets $A, B \subseteq E(M)$ such that*

- *each set $T \in \mathcal{T}$ is contained in A, B, or $E-(A\cup B)$;*
- *A and B each contain at most k sets from $\mathcal{T}$;*
- *$\lambda_M(A) \le k$, $\lambda_M(B) \le k$;*
- *if $\mathcal{T}'$ is the collection of sets in $\mathcal{T}$ disjoint from $A\cup B$, then $\nu_{M/A\setminus B}(\mathcal{T}') = 0$.*

We need the following lemma. (Note that the proof is not self-contained; we use Theorem 6.1 from the next section.)

LEMMA 5.4. *Let M be a matroid and let (A,B) be a partition of $E(M)$. Then there exists a matroid M' on $E(M)$ such that dist$(M,M') = \lambda_M(A)$, $\lambda_{M'}(A) = 0$, $M/B = M'/B$, and $M/A = M'/A$.*

Proof. The result is vacuous when $\lambda_M(A) = 0$, so suppose that $\lambda_M(A) > 0$. By Theorem 6.1, there exists a matroid N on ground set $E(M)\cup\{e\}$ such that $M = N\setminus e$, $e \in \mathrm{cl}_N(A)$, $e \in \mathrm{cl}_N(B)$, and e is not a loop of N. Let $M'' = N/e$. Note that e is a loop in both N/A and N/B. Therefore $M''/A = (N/e)/A = (N/A)/e = (N/A)\setminus e = M/A$ and, similarly, $M''/B = M/B$. Also note that $\lambda_{M''}(A) = \lambda_M(A) - 1$ and that $\mathrm{dist}(M,M'') = 1$. The result now follows by an easy inductive argument. □

We are now ready to prove our main result, which we restate here for convenience.

THEOREM 5.5. *Let $M = (E,r)$ be a matroid and let $\mathcal{T}$ be a collection of disjoint subsets of $E(M)$. Then there exists a matroid M' on ground set $E(M)$ such that $\nu_{M'}(\mathcal{T}) = 0$ and dist$(M,M') \le 2(\delta_M(\mathcal{T})+1)\nu_M(\mathcal{T})$.*

Proof. Suppose that $\mathcal{T} = \{T_1,\dots,T_l\}$ and let $k = \nu_M(\mathcal{T})$. By Theorem 3.1, there exists a collection $\mathcal{S} = \{S_1,\dots,S_l\}$ of disjoint subsets of $E(M)$ such that $\nu_M(\mathcal{S}) = k$ and, for each $i \in \{1,\dots,l\}$, $T_i \subseteq S_i$ and $\lambda_M(S_i) \le \delta_M(\mathcal{T})$. Then, by Lemma 5.3, there exist disjoint subsets A and B of $E(M)$ such that

- each set $S \in \mathcal{B}$ is contained in A, B, or $E(M)-(A\cup B)$;
- A and B each contain at most k sets from $\mathcal{S}$;
- $\lambda_M(A) \le k$, $\lambda_M(B) \le k$;
- if $\mathcal{S}'$ is the collection of sets in $\mathcal{S}$ disjoint from $A\cup B$, then $\nu_{M/A\setminus B}(\mathcal{S}') = 0$.

By Lemma 5.4 and duality, there exists a matroid M' on ground set $E(M)$ such that $\mathrm{dist}(M,M') \le 2k$, $\lambda'_M(A) = \lambda'_M(B) = 0$, and $M'/A\setminus B = M/A\setminus B$. Note that, for each $S \in \mathcal{S}-\mathcal{S}'$, we have $\lambda_{M'}(S) \le \delta_M(\mathcal{T})$. Therefore, by Lemma 5.4, there exists a matroid M'' such that $\mathrm{dist}(M',M'') \le 2k\delta_M(\mathcal{T})$, and $\nu_{M''}(\mathcal{S}) = 0$. Then, since $T_i \subseteq S_i$ for each $i \in \{1,\dots,l\}$, we have $\nu_{M''}(\mathcal{T}) = 0$, as required. □

6. Modular cuts. In this section we prove the following theorem.

THEOREM 6.1. *Let M be a matroid and let (A,B) be a partition of $E(M)$. If $\lambda_M(A) > 0$, then there exists a matroid M' on ground set $E(M)\cup\{e\}$ such that $M = M'\setminus e$, $e \in \mathrm{cl}_{M'}(A)$, $e \in \mathrm{cl}_{M'}(B)$, and e is not a loop of M'.*

Note that Theorem 6.1 is trivial for representable matroids.

Let $X, Y \subseteq E(M)$. We call (X,Y) a *modular pair* if $r_M(X) + r_M(Y) = r_M(X\cap Y) + r_M(X\cup Y)$. A collection $\mathcal{F}$ of subsets of $E(M)$ is called a *modular cut* of M if it satisfies the following three conditions:

1. If $X \subseteq Y \subseteq E(M)$ and $X \in \mathcal{F}$, then $Y \in \mathcal{F}$.
2. If $X, Y \in \mathcal{F}$ and (X,Y) is a modular pair, then $X\cap Y \in \mathcal{F}$.
3. If $Y \in \mathcal{F}$ and $X \subseteq Y$ with $r_M(X) = r_M(Y)$, then $X \in \mathcal{F}$.

The following theorem is well known; see, for example, Oxley [6, Theorem 7.2.2].

THEOREM 6.2. *Let $\mathcal{F}$ be a modular cut in a matroid M. Then there exists a matroid N on ground set $E(M) \cup \{e\}$ such that $N \setminus e = M$ and, for each $X \subseteq E(M)$, $r_N(X \cup \{e\}) = r_M(X)$ if and only if $X \in \mathcal{F}$.*

LEMMA 6.3. *Let M be a matroid, let (A, B) be a partition of $E(M)$, and let $\mathcal{F}$ be the collection of all sets $X \subseteq E(M)$ such that $\lambda_{M/X}(A - X) = 0$. Then $\mathcal{F}$ is a modular cut of M.*

Proof. Note that $\mathcal{F}$ clearly satisfies the first condition.

6.3.1. *For any $X \subseteq E(M)$, $X \in \mathcal{F}$ if and only if $(A \cup X, B \cup X)$ is a modular pair in M.*

Subproof. Note that $\lambda_{M/X}(A-X) = r_{M/X}(A-X) + r_{M/X}(B-X) - r(M/X) = r_M(A \cup X) + r_M(B \cup X) - r(M) - r_M(X)$. Thus $\lambda_{M/X}(A - X) = 0$ if and only if $(A \cup X, B \cup X)$ is a modular pair. □

Now consider the third condition. Suppose that $Y \in \mathcal{F}$ and $X \subseteq Y$ with $r_M(X) = r_M(Y)$. By the claim, $(A \cup Y, B \cup Y)$ is a modular pair. Moreover, since $X \subseteq Y$ with $r_M(X) = r_M(Y)$, we have $r_M(A \cup Y) = r_M(A \cup X)$, $r_M(B \cup Y) = r_M(B \cup X)$, $r_M((A \cup Y) \cap (B \cup Y)) = r_M((A \cup X) \cap (B \cup X))$, and $r_M((A \cup Y) \cup (B \cup Y)) = r_M((A \cup X) \cup (B \cup X))$. Therefore $(A \cup X, B \cup X)$ is a modular pair and hence, by the claim, $X \in \mathcal{F}$. This verifies the third condition.

Finally consider the second condition. Let $X_1, X_2 \in \mathcal{F}$ such that (X_1, X_2) is a modular pair. By the definition of $\mathcal{F}$, $X_1 \cup X_2 \in \mathcal{F}$. Then, by statement 6.3.1, each of $(A \cup X_1, B \cup X_1)$, $(A \cup X_2, B \cup X_2)$, $(A \cup (X_1 \cup X_2), B \cup (X_1 \cup X_2))$ is a modular pair. Now

$$\begin{aligned}
&r_M(A \cup (X_1 \cap X_2)) + r_M(B \cup (X_1 \cap X_2)) \\
&\qquad = r_M((A \cup X_1) \cap (A \cup X_2)) + r_M((B \cup X_1) \cap (B \cup X_2)) \\
&\qquad \le (r_M(A \cup X_1) + r_M(A \cup X_2) - r_M(A \cup X_1 \cup X_2)) \\
&\qquad\quad + (r_M(B \cup X_1) + r_M(B \cup X_2) - r_M(B \cup X_1 \cup X_2)) \\
&\qquad = (r_M(A \cup X_1) + r_M(B \cup X_1)) \\
&\qquad\quad + (r_M(A \cup X_2) + r_M(B \cup X_2)) \\
&\qquad\quad - (r_M(A \cup X_1 \cup X_2) + r_M(B \cup X_1 \cup X_2)) \\
&\qquad = (r_M(X_1) + r(M)) + (r_M(X_2) + r(M)) \\
&\qquad\quad - (r_M(X_1 \cup X_2) + r(M)) \\
&\qquad = (r_M(X_1) + r_M(X_2) - r_M(X_1 \cup X_2)) + r(M) \\
&\qquad = r_M(X_1 \cap X_2)) + r(M).
\end{aligned}$$

So $(A \cup (X_1 \cap X_2), B \cup (X_1 \cap X_2))$ is a modular pair. Then, by statement 6.3.1, $X_1 \cap X_2 \in \mathcal{F}$. Hence $\mathcal{F}$ is a modular cut, as required. □

Now Theorem 6.1 is an immediate consequence of Theorem 6.2 and Lemma 6.3.

Acknowledgment. We thank the referee for carefully reading the manuscript and for correcting a significant error in our definition of a modular cut.

REFERENCES

[1] M. CHUDNOVSKY, W. H. CUNNINGHAM, AND J. GEELEN, *An Algorithm for Packing Non-zero A-Paths in Group-Labelled Graphs*, preprint.

[2] T. Gallai, *Maximum-minimum Sätze und verallgemeinerte Faktoren von Graphen*, Acta. Math. Acad. Sci. Hung., 12 (1961), pp. 131–173.
[3] S. Iwata, L. Fleischer, and S. Fujishige, *A combinatorial strongly polynomial algorithm for minimizing submodular functions*, J. ACM, 48 (2001), pp. 761–777.
[4] L. Lovász, *Matroid matching and some applications*, J. Combin. Theory Ser. B, 28 (1980), pp. 208–236.
[5] W. Mader, *Über die Maximalzahl kreuzungsfreier H-Wege*, Arch. Math. (Basel), 31 (1978), pp. 382–402.
[6] J. G. Oxley, *Matroid Theory*, Oxford University Press, New York, 1992.
[7] A. Schrijver, *A combinatorial algorithm minimizing submodular functions in strongly polynomial time*, J. Combin. Theory Ser. B, 80 (2000), pp. 346–355.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 597–602

A NOTE ON QUASI-TRIANGULATED GRAPHS*

ION GORGOS†, CHÍNH T. HOÀNG‡, AND VITALY VOLOSHIN§

Abstract. A graph is quasi-triangulated if each of its induced subgraphs has a vertex which is either simplicial (its neighbors form a clique) or cosimplicial (its nonneighbors form an independent set). We prove that a graph G is quasi-triangulated if and only if each induced subgraph H of G contains a vertex that does not lie in a hole, or an antihole, where a hole is a chordless cycle with at least four vertices, and an antihole is the complement of a hole. We also present an algorithm that recognizes a quasi-triangulated graph in $O(nm)$ time.

Key words. triangulated graphs, chordal graphs, quasi-triangulated graphs

AMS subject classifications. 05C75, 05C85

DOI. 10.1137/S0895480104444399

1. Introduction.

In a graph G, a vertex x is *simplicial* if its neighborhood $N(x)$ induces a complete subgraph of G. A graph is *triangulated (chordal)* if it does not contain a chordless cycle of length at least four (a *hole*) as an induced subgraph. A famous theorem of Dirac [3] states that every triangulated graph has a simplicial vertex. Actually, Dirac proved more: every triangulated graph different from a clique contains two nonadjacent simplicial vertices. Let us say that a vertex is *cosimplicial* if its nonneighbors form an independent subset of vertices and that a graph is *cotriangulated* if it does not contain the complement of a chordless cycle on at least four vertices (an *antihole*). Dirac's theorem says equivalently that every cotriangulated graph has a cosimplicial vertex. Our purpose is to investigate the larger class of graphs which are called *quasi-triangulated* graphs (QT for short), defined as follows: a graph G is in class QT if and only if every induced subgraph H of G has a vertex which is either simplicial or cosimplicial in H. Quasi-triangulated graphs have been introduced by the third author in [9, 11] as a generalization of chordal graphs. The problem of characterizing the class QT was raised in [9] and, independently, in [7] (where they are called *good*). The reader is referred to [1] for more information on the class QT.

Following [7], we say that an order $v_1 < v_2 < \cdots < v_n$ on a graph G is *good* if, for any induced subgraph H of G, either the largest vertex of $(H, <)$ is simplicial or the smallest vertex of $(H, <)$ is cosimplicial. Good orders are perfect in the sense of [2].

Simplicial vertices cannot lie in a hole; and cosimplicial vertices cannot lie in an antihole. A graph with each vertex belonging to some hole and some antihole is called *latticed.*

The third author conjectured (unpublished) and the first author proved in [4, 5] the following.

*Received by the editors June 10, 2004; accepted for publication (in revised form) November 9, 2005; published electronically August 25, 2006.
http://www.siam.org/journals/sidma/20-3/44439.html

†Academy of Economic Studies of Moldova, 61 Banulescu-Bodoni str. MD-2005, Chisinau, Moldova.

‡Department of Physics and Computer Science, Wilfrid Laurier University, 75 University Ave. W., Waterloo, ON N2L 3C5, Canada (choang@wlu.ca). This author's research was supported by the NSERC.

§Department of Mathematics and Physics, Troy University, Troy, AL 36082 (vvoloshin@troy.edu). This author's research was partially supported by a Troy University research grant.

THEOREM 1. *For a graph G, the following three conditions are equivalent:*

(i) *G is quasi-triangulated.*

(ii) *G does not contain a latticed subgraph as an induced subgraph.*

(iii) *G admits a good order.*

As usual, n (respectively, m) denote the number of vertices (respectively, edges) of the input graph. For the quasi-triangulated graph recognition problem, the third author [10] proposed an $O(n^4)$ algorithm, Spinrad [12] proposed an $O(n^{2.77})$ algorithm, and the second author [6] independently proposed an $O(nm)$ algorithm.

THEOREM 2. *There is an $O(nm)$-time $O(n^2)$-space algorithm to recognize a quasi-triangulated graph.*

Theorems 1 and 2 are known by researchers in the field and have been referred to in the literature, but their proofs have never been published. The purpose of this paper is to provide the proofs of these two theorems.

2. Proof of Theorem 1. To prove Theorem 1, we will need the following lemma, which was included in the original proof in [4] and was rediscovered independently in [8].

LEMMA 1. *Let G be a graph and x be a vertex of G that does not lie in a hole. Then any minimal cutset C of G which is contained in the neighborhood $N(x)$ of x is a clique.*

Proof of Lemma 1. Define G, x, C as in the lemma. Let Y be a component of $G - C$ that does not contain x. We may assume that there are nonadjacent vertices u, v in C, for otherwise we are done. Since C is a minimal cutset, each of u and v has a neighbor in Y. It follows that there is a chordless path of length at least two joining u to v whose interior vertices lie in Y. This path together with x forms a hole, a contradiction to our assumption on x. □

Proof of Theorem 1. It is easy to see that (i) and (iii) are equivalent, and (i) implies (ii). So, we need only to prove that (ii) implies (i). We shall prove this by induction on the number of vertices. Let G be a graph satisfying (ii). We may assume G contains no simplicial vertex and no cosimplicial vertex, for otherwise we are done by the induction hypothesis. If G is disconnected, then each component of G contains a hole (for otherwise, it is triangulated and contains a simplicial vertex that remains simplicial in G); thus, G contains the union of two disjoint holes, a contradiction to (ii). So, G must be connected.

By replacing G by its complement $\overline{G}$ if necessary, we may assume that G contains a vertex that does not lie in a hole.

Define $X = \{x \mid x \text{ does not lie in a hole of } G\}$.

Our assumption on G implies that $X \neq \emptyset$. Let $G' = G - X$. G' is nonempty, for otherwise G is triangulated and thus contains a simplicial vertex by Dirac's theorem. By the induction hypothesis, G' contains a simplicial or cosimplicial vertex y. Since every vertex of G' lies in a hole, y is cosimplicial. We shall prove that y is adjacent to all vertices of X (this will imply y is cosimplicial in G, a contradiction).

Let x be a vertex in X. Since G is connected and x is not cosimplicial (by assumption), there is a nonempty set C of vertices in $N(x)$ that is a minimal cutset of G. By Lemma 1, C is a clique. Let G_1, G_2 be induced subgraphs of G such that $G = G_1 \cup G_2, G_1 \cap G_2 = C$, and there is no edge between $G_1 - C$ and $G_2 - C$.

Suppose G_1 is triangulated. We claim that there is a simplicial vertex s in $G_1 - C$. If G_1 is a clique, then the claim obviously holds; otherwise, by Dirac's theorem, G_1 contains two nonadjacent simplicial vertices, one of which must lie in $G_1 - C$ since C

is a clique. But s remains a simplicial vertex of G, a contradiction to our assumption on G. Thus G_1, and similarly G_2, cannot be triangulated.

Therefore, G_1 contains a hole. Since C is a clique, one edge, say e_1, of this hole lies completely in $G_1 - C$. Similarly, there is an edge, say e_2, that lies completely in $G_2 - C$ and belongs to a hole. Since y is cosimplicial in G' and all endpoints of e_1, e_2 are in G', y must be in C, and therefore adjacent to x, as desired. □

3. A recognition algorithm for quasi-triangulated graphs. In this section, we prove Theorem 2 by describing an algorithm that recognizes a quasi-triangulated graph in $O(nm)$ time using $O(n^2)$ space.

For a vertex x, an S-*obstruction* is a triple (a, b, x) that induces a P_3 with x being the interior vertex of the path; a C-*obstruction* is a triple (a, b, x) that induces an S-obstruction (a, b, x) in the complement.

A straightforward algorithm to recognize quasi-triangulated graphs proceeds as follows. First, for all vertices x, list all S- and C-obstructions. Then find a vertex y with no S- or C-obstructions; if no such vertex exists, then the graph is not quasi-triangulated. Remove y and update the lists of obstructions for the remaining vertices. Repeat this process to eliminate all vertices. If all vertices can be eliminated in this way, then the graph is quasi-triangulated; otherwise, it is not.

Since a vertex has $O(n^2)$ obstructions, we will need a data structure to store $O(n^3)$ obstructions of the graph. Thus the algorithm runs in $O(n^3)$ time using $O(n^3)$ space. We are going to show that the algorithm can be refined to run in time $O(nm)$ using $O(n^2)$ space.

Proof of Theorem 2. We may suppose there is a total order $<$ on the vertices of a given graph G. We say that (a, b, x) is less than (c, d, x), denoted by $(a, b, x) < (c, d, x)$, if $a < c$, or $a = c$ and $b < d$. To achieve the $O(nm)$ time bound, we will list only the smallest S-obstruction and C-obstruction for each vertex. When removing a vertex y, if a vertex x loses an obstruction, then we will find a smallest obstruction for x in the remaining graph. We shall show that, over the life of the algorithm, the time needed to list the (currently) smallest obstructions for a vertex x is $O(m)$. The outline of our algorithm is as follows.

Outline of algorithm. We begin with the input graph G and proceed to eliminate vertices one by one using the following steps.

1. For each vertex x of graph G, list a smallest S-obstruction (a, b, x) and a smallest C-obstruction (g, d, x).

2. If every remaining vertex has an S-obstruction and a C-obstruction, then G is not quasi-triangulated.

3. If a vertex z has no S-obstruction or no C-obstruction, eliminate z from G, and for each remaining vertex x that loses an S-obstruction (respectively, C-obstruction), generate a new smallest S-obstruction (respectively, C-obstruction). Replace G by $G - z$, and repeat step 2.

The graph G is quasi-triangulated if and only if recursive applications of step 3 eliminate all vertices. To anticipate, our algorithm lists the S-obstructions in $O(nm)$ time using $O(n+m)$ space and the C-obstructions in $O(nm)$ time using $O(n^2)$ space.

Let $N(x)$ be the adjacency list of vertex x. Without loss of generality, we may assume for all x that the lists $N(x)$ are sorted in increasing order.

Listing the smallest S-obstruction for a vertex *x*. For each vertex x, we use two pointers, $\alpha(x)$ and $\beta(x)$. Initially $\alpha(x)$ points to the first vertex α in $N(x)$ and $\beta(x)$ points to the immediate successor β of α in $N(x)$ (for simplicity, we let α (respectively, β) denote the name of the vertex pointed to by the pointer $\alpha(x)$

(respectively, $\beta(x)$). If $\alpha(x)$ or $\beta(x)$ cannot be initialized, then x has no S-obstruction. We simply advance $\beta(x)$ on $N(x)$ until we find that α and β are nonadjacent. When $\beta(x)$ reaches the end of $N(x)$ (i.e., it has value *null*), we advance $\alpha(x)$ in $N(x)$ and initialize $\beta(x)$ (making $\beta(x)$ point to the immediate successor of $\alpha(x)$ in $N(x)$). If $\alpha(x) = null$, then x has no S-obstruction, and a message "No S-obstruction" is produced. We can summarize this process as follows. (In the following procedure, the function IsEdge(a, b) returns true if and only if ab is an edge.)

PROCEDURE LISTSMALLEST-S-OBSTRUCTION(x).

```
while true
{   if α(x) = null
        then return "no S-obstruction for x"
    if IsEdge(α, β) = true
        then advance β(x) in N(x)
    else
        return (α, β, x)
    while ((α(x) ≠ null) and β(x) = null))
    {   advance α(x) in N(x)
        initialize β(x)
    }
}
```

Suppose we eliminate a vertex z and x loses its S-obstruction (a, b, x) (because $a = z$ or $b = z$). If b (respectively, a) is eliminated, then we advance $\beta(x)$ (respectively, $\alpha(x)$) and call Procedure ListSmallest-S-Obstruction(x) to get the smallest S-obstruction for x. The number of movements of the pointers $\alpha(x), \beta(x)$ in $N(x)$ is proportional to $O(n + m)$ since we advance $\beta(x)$ only in the presence of an edge, and $\alpha(x)$ is reset at most the degree of x times. If we have the incidence matrix of G at our disposal, then each call to Procedure IsEdge takes only constant time, but this method needs $O(n^2)$ space. We are going to show that for each vertex x we can implement Procedure IsEdge in $O(n + m)$ time using only the adjacency lists of G ($O(n + m)$ space).

Now we describe Procedure IsEdge(α, β), which returns true if and only if $\alpha\beta$ is an edge. For a vertex x, there is a pointer $p(x)$ which initially points to the first vertex in $N(\alpha)$ (recall that $\alpha(x)$ is the pointer associated with vertex x). If $p(x)$ cannot be initialized, then $\alpha\beta$ is not an edge. Pointer $p(x)$ is advanced in $N(\alpha)$ until it points to either (i) β ($\alpha\beta$ is an edge) or (ii) the smallest vertex in $N(\alpha)$ that is greater than β ($\alpha\beta$ is not an edge). The vertex pointed to by $p(x)$ is denoted by p.

PROCEDURE ISEDGE(α, β).

```
while true
{   if p(x) = null
        return false
    if p < β
        advance p(x) in N(α)
    else if p = β
        return true
    else if p > β
        return false
}
```

For each vertex x and each $\alpha(x)$, $p(x)$ scans $N(\alpha)$ only once. Thus, for each x, the cost of testing for all edges $\alpha\beta$ is $O(n + m)$.

Listing the smallest C-obstruction for a vertex x. Assume that the vertices are numbered $1, 2, \ldots, n$. For each vertex x, we maintain an integer variable counter $\gamma(x)$ that refers to the smallest nonneighbor of x. We need to generate the smallest C-obstruction of the form (γ, δ, x) for some vertex δ (that must be adjacent to γ and nonadjacent to x). This can be done as follows.

For each vertex x, we maintain a 0-1 characteristic vector $I(x)$ of size n to represent the neighborhood of x (the jth entry of $I(x)$ is 1 if and only if vertex j is a neighbor of x; in other words, $I(x)$ is the xth row of the incidence matrix of G). This is necessary so that testing of an edge of the form xy can be done in constant time. Given $\gamma(x)$, we find the smallest neighbor δ of γ (the vertex referred to by $\gamma(x)$) such that $\gamma < \delta$ and δ is nonadjacent to x by scanning the list $N(\gamma)$ and, for each vertex y in this list, testing whether yx is an edge. We use a pointer $\delta(x)$ to point to the location of δ in $N(\gamma)$. If $\delta(x)$ cannot be initialized, then there is no C-obstruction of the form (γ, δ, x); in this case, we increase $\gamma(x)$ by one and repeat the process (the initial value of $\gamma(x)$ is one). This is summarized in the following procedure (we leave the pointer initialization problem to the reader).

Procedure ListSmallest-C-Obstruction(x).

```
while true
{   if δ(x) = null
        repeat
            increase γ(x) by one
            if (γ > n)
                return "No C-obstruction for x"
            let δ(x) point to the first vertex in N(γ)
        until (xγ is not an edge) and δ(x) is not null
    if (xδ is not an edge) and (γ < δ)
        return the C-obstruction (γ, δ, x)
    advance δ(x) in N(γ)
}
```

Suppose we eliminate a vertex z and x loses its C-obstruction (g, d, x) (because $g = z$ or $d = z$). If d is eliminated, then we advance $\delta(x)$ in $N(\gamma(x))$ and call Procedure ListSmallest-C-Obstruction. If g is eliminated, then we repeatedly increase $\gamma(x)$ by one until we get the next smallest nonneighbor of x and call Procedure ListSmallest-C-Obstruction.

For each vertex x and each $\gamma(x)$, the list $N(\gamma(x))$ is scanned at most once. Thus, for each x, we can list the smallest C-obstruction in $O(n + m)$ time over the life of the algorithm. This method requires $O(n^2)$ space. □

In the case of listing the C-obstructions, we do not know how to implement our algorithm in $O(nm)$ time using linear space. We leave this as an open problem. We note Spinrad's algorithm [12] uses $O(n^2)$ space since it relies on matrix multiplications.

REFERENCES

[1] A. Brandstädt, V. B. Le, J. P. Spinrad, *Graph Classes: A Survey*, SIAM Monogr. Discrete Math. Appl. 3, SIAM, Philadelphia, 1999.

[2] V. Chvátal, *Perfectly ordered graphs*, in Topics on Perfect Graphs, C. Berge and V. Chvátal, eds., Ann. Discrete Math. 21, North–Holland, Amsterdam, 1984, pp. 63–65.

[3] G. A. Dirac, *On rigid circuit graphs*, Abh. Math. Sem. Univ. Hamburg, 25 (1961), pp. 71–76.

[4] I. M. Gorgos, *A Characterization of Quasi-triangulated Graphs*, Preprint 11B494, Kishinev State University, Kishinev, Moldova, 1984 (in Russian).

[5] I. M. Gorgos, *Method of Alternating Chains and Its Applications*, Ph.D. Thesis, Kishinev State University, Kishinev, Moldova, 1985 (in Russian).
[6] C. T. Hoàng, *Recognizing Quasi-triangulated Graphs in $O(nm)$ Time*, manuscript.
[7] C. T. Hoàng and N. V. R. Mahadev, *A note on perfect orders*, Discrete Math., 74 (1989), pp. 77–84.
[8] C. T. Hoàng, S. Hougardy, F. Maffray, and N. V. R. Mahadev, *On simplicial and co-simplicial vertices in graphs*, Discrete Appl. Math., 138 (2004), pp. 117–132.
[9] V. I. Voloshin, *Quasi-triangulated Graphs*, Preprint 5569-81, Kishinev State University, Kishinev, Moldova, 1981 (in Russian).
[10] V. I. Voloshin, *Quasi-triangulated Graphs Recognition Program*, Algorithms and Programs P006124, Moscow, Russia, 1983 (in Russian).
[11] V. I. Voloshin, *Triangulated Graphs and Their Generalizations*, Ph.D. Thesis, Kishinev State University, Kishinev, Moldova, 1983 (in Russian).
[12] J. Spinrad, *Recognizing quasi-triangulated graphs*, Discrete Appl. Math., 138 (2004), pp. 203–213.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 603–609

CYCLE DECOMPOSITIONS OF $K_{n,n} - I^*$

JUN MA†, LIQUN PU†, AND HAO SHEN†

Abstract. Let $K_{n,n}$ denote the complete bipartite graph with n vertices in each bipartition set and $K_{n,n} - I$ denote $K_{n,n}$ with a 1-factor removed. An *m-cycle system* of $K_{n,n} - I$ is a collection T of m-cycles such that each edge of $K_{n,n} - I$ is contained in a unique m-cycle of T. In this paper, it is proved that the necessary and sufficient conditions for the existence of an m-cycle system of $K_{n,n} - I$ are $n \equiv 1 \pmod 2$, $m \equiv 0 \pmod 2$, $4 \le m \le 2n$, and $n(n-1) \equiv 0 \pmod m$.

Key words. decomposition, cycle, complete bipartite graph, 1-factor

AMS subject classification. 05C38

DOI. 10.1137/050626363

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 natural to ask when there exists an m-cycle system of G.

It is not difficult to verify that the following conditions are necessary for the existence of an m-cycle system of G:

$$\begin{cases} 3 \le m \le |V(G)|, \\ |E(G)| \equiv 0 \pmod m, \\ d(u) \equiv 0 \pmod 2 \text{ for each } u \in V(G), \end{cases}$$

where $d(u)$ denotes the number of edges incident with u in G.

Let K_v denote a complete graph of order v, $K_v - I$ denote K_v with a 1-factor removed, and $K_{x,y}$ denote a complete bipartite graph with partite sets of sizes x and y. When G is K_v, $K_v - I$, or $K_{x,y}$, the existence problem of m-cycle systems of G has been completely settled [2, 3, 4].

THEOREM 1 (see [3, 4]). *Let m and v be positive integers. Then there exists an m-cycle system of K_v if and only if $v \equiv 1 \pmod 2$, $3 \le m \le v$, and $v(v-1) \equiv 0 \pmod{2m}$.*

THEOREM 2 (see [3, 4]). *Let m and v be positive integers. Then there exists an m-cycle system of $K_v - I$ if and only if $v \equiv 0 \pmod 2$, $3 \le m \le v$, and $v(v-2) \equiv 0 \pmod{2m}$.*

THEOREM 3 (see [2]). *Let $m \equiv 0 \pmod 2$ and $m \ge 4$. Then there exists an m-cycle system of $K_{x,y}$ if and only if x, $y \ge \frac{1}{2}m$, $x \equiv y \equiv 0 \pmod 2$, and $xy \equiv 0 \pmod m$.*

In this paper, we consider the case when G is $K_{x,y} - I$, where $K_{x,y} - I$ denotes $K_{x,y}$ with a 1-factor removed.

Obviously if $K_{x,y}$ has a 1-factor, then necessarily we have $x = y$. Let $x = y = n$. Simple counting gives the following necessary conditions.

*Received by the editors March 9, 2005; accepted for publication (in revised form) February 23, 2006; published electronically August 25, 2006. This project was supported by National Natural Science Foundation of China under grant 10471093.

http://www.siam.org/journals/sidma/20-3/62636.html

†Department of Mathematics, Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China (mj904@sjtu.edu.cn, sarah_009@sjtu.edu.cn, haoshen@sjtu.edu.cn).

LEMMA 4. *If there exists an m-cycle system of $K_{n,n}-I$, then $n \equiv 1 \pmod 2$, $m \equiv 0 \pmod 2$, $4 \le m \le 2n$, and $n(n-1) \equiv 0 \pmod m$.*

It was proved [5] that m-cycle systems of $K_{n,n}-I$ exist in the special case $m = 2n$. The following theorem was obtained by Archdeacon et al. [1].

THEOREM 5. *Let $m \equiv 0 \pmod 2$, $m \ge 4$, and $n \equiv 1 \pmod 2$. If $m \equiv 0 \pmod 4$ and $m < n$, or if $m \equiv 2 \pmod 4$ and $m < 2n$, then there exists an m-cycle system of $K_{n,n} - I$ if and only if $n(n-1) \equiv 0 \pmod m$.*

But, it is still not known whether m-cycle systems of $K_{n,n} - I$ exist when $n \equiv 1 \pmod 2$, $n \ge 3$, $m \equiv 0 \pmod 4$, $n < m < 2n$, and $n(n-1) \equiv 0 \pmod m$.

The main purpose of this paper is to determine the existence of m-cycle systems of $K_{n,n} - I$ for the open case in Theorem 5. In fact, we will give a unified and simple proof to the following theorem.

THEOREM 6. *Let m and n be positive integers. Then there exists an m-cycle system of $K_{n,n} - I$ if and only if*

$$\begin{cases} n \equiv 1 \pmod 2, \\ m \equiv 0 \pmod 2 \text{ and } 4 \le m \le 2n, \\ n(n-1) \equiv 0 \pmod m. \end{cases}$$

2. Cycle decomposition of $K_{n,n} - I$ with $\frac{1}{2}m \le n \le \frac{3}{2}m$. A cycle on m vertices is denoted by C_m. If a graph G is the edge-disjoint union of m-cycles, then we say that G is C_m-decomposable. We shall also write $C_m \mid G$.

In [1], Archdeacon et al. skillfully proved the following lemma.

LEMMA 7. *Let $m \equiv 2 \pmod 4$, $n \equiv 1 \pmod 2$, and $6 \le m \le 2n$. Then $C_m \mid K_{n,n} - I$ if and only if $m \mid n(n-1)$.*

By Lemma 7, we obtain the following corollary immediately.

COROLLARY 8. *Let $m \equiv 2 \pmod 4$ and $m \ge 6$. If $n \in \{\frac{1}{2}m, \frac{3}{2}m\}$, then $C_m \mid K_{n,n} - I$.*

Now, for a positive integer n, let $D \subseteq Z_n$ and $X(n; D)$ be a graph with vertex set $Z_n \times Z_2$ and edge set $\{\{i_0, (i+d)_1\} \mid d \in D, i \in Z_n\}$. Clearly, $K_{n,n}$ can be viewed as $X(n; Z_n)$. 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).

For any integer x, let

$$\varepsilon(x) = \begin{cases} 0 & \text{if} \quad x \equiv 0 \pmod 2, \\ 1 & \text{if} \quad x \equiv 1 \pmod 2. \end{cases}$$

We use the difference method to give constructions of m-cycle systems of $X(n; D)$ which we need in this paper.

LEMMA 9. *Let m be a positive integer. If $m \equiv 0 \pmod 2$ and $m \ge 4$, then $C_m \mid K_{m+1,m+1} - I$.*

Proof. We view $K_{m+1,m+1}$ as $X(m+1; Z_{m+1})$. For $r = 0, 1, \ldots, m$, define

$d_r \in Z_{m+1}$ as

$$d_r = \begin{cases} 0 & \text{if} \quad r = 0, \\ 1 - \varepsilon(\frac{1}{2}m) & \text{if} \quad r = 1, \\ r & \text{if} \quad 2 \le r \le \frac{1}{2}m - \varepsilon(\frac{1}{2}m), \\ r+1 & \text{if} \quad \frac{1}{2}m + 1 - \varepsilon(\frac{1}{2}m) \le r \le m-1, \\ \frac{1}{2}m + 1 - \varepsilon(\frac{1}{2}m) & \text{if} \quad r = m. \end{cases}$$

Let $e_k = \sum_{r=0}^{k} (-1)^r d_r$ for $0 \le k \le m$.

Let

$$\theta = \begin{cases} 0 & \text{if} \quad k < \frac{1}{2}m + 1 - \varepsilon(\frac{1}{2}m), \\ 1 & \text{if} \quad k \ge \frac{1}{2}m + 1 - \varepsilon(\frac{1}{2}m). \end{cases}$$

Then

$$e_k = \begin{cases} 0 & \text{if} \quad k = 0 \text{ or } m, \\ \frac{1}{2}k + \varepsilon(\frac{1}{2}m) & \text{if} \quad k \equiv 0 \pmod 2 \quad \text{and} \quad 2 \le k \le m-2, \\ -\frac{1}{2}(k+1) - \theta + \varepsilon(\frac{1}{2}m) & \text{if} \quad k \equiv 1 \pmod 2. \end{cases}$$

Let C be the following closed trail:

$$(e_0)_1, (e_1)_0, (e_2)_1, (e_3)_0, \ldots, (e_{m-2})_1, (e_{m-1})_0, (e_m)_1.$$

The differences used in C are $d_1, d_2, \ldots, d_m$.

Since

$$0 = e_0 < e_2 < e_4 < \cdots < e_{m-2} = \frac{1}{2}m - 1 + \varepsilon\left(\frac{1}{2}m\right)$$

and

$$\begin{aligned} m + \varepsilon\left(\frac{1}{2}m\right) &= e_1 + m + 1 > e_3 + m + 1 > e_5 + m + 1 > \cdots > e_{m-1} + m + 1 \\ &= \frac{1}{2}m + \varepsilon\left(\frac{1}{2}m\right), \end{aligned}$$

then C is an m-cycle.

Let $T = (C)$ and $I = \{\{(i)_0, (i + \varepsilon(\frac{1}{2}m))_1\} \mid i \in Z_{m+1}\}$. Then I is a 1-factor in $K_{m+1,m+1}$, T is an m-cycle system of $K_{m+1,m+1} - I$, and $C_m \mid K_{m+1,m+1} - I$. ∎

LEMMA 10. *Let u be an integer and let $m \equiv 0 \pmod 2$, $m \ge 4$, $n \equiv 1 \pmod 2$, $m < 2n$, $g = gcd(m,n) > 1$, and $1 \le h \le \frac{m}{g}$. For $r = 1, 2, \ldots, \frac{m}{g}$, define $d_r \in Z_n$ as*

$$d_r = \begin{cases} u\frac{m}{g} + r & \text{if} \quad 1 \le r \le h-1, \\ u\frac{m}{g} + r + 1 & \text{if} \quad h \le r \le \frac{m}{g} - 1, \\ u\frac{m}{g} + \frac{m}{2g} + \varepsilon(h) + \frac{n}{g} & \text{if} \quad r = \frac{m}{g}. \end{cases}$$

Let $D = \{d_1, d_2, \ldots, d_{\frac{m}{g}}\}$. Then $C_m \mid X(n; D)$.

Proof. Let $d_0 = 0$ and $e_k = \sum_{r=0}^{k} (-1)^r d_r$ for $0 \le k \le \frac{m}{g}$.

Let

$$\theta = \begin{cases} 0 & \text{if} \quad k < h, \\ 1 & \text{if} \quad k \ge h. \end{cases}$$

Then

$$e_k = \begin{cases} \frac{1}{2}k + \theta(1-\varepsilon(h)) & \text{if} \quad k \equiv 0 \pmod 2 \quad \text{and} \quad 0 \le k \le \frac{m}{g} - 2, \\ -u\frac{m}{g} - \frac{1}{2}(k+1) - \theta\varepsilon(h) & \text{if} \quad k \equiv 1 \pmod 2, \\ \frac{n}{g} & \text{if} \quad k = \frac{m}{g}. \end{cases}$$

Let P be the following trail:

$$(e_0)_1, (e_1)_0, (e_2)_1, (e_3)_0, \ldots, (e_{\frac{m}{g}-2})_1, (e_{\frac{m}{g}-1})_0, (e_{\frac{m}{g}})_1.$$

The differences used in P are $d_1, d_2, \ldots, d_{\frac{m}{g}}$.

Since

$$0 = e_0 < e_2 < e_4 < \cdots < e_{\frac{m}{g}-2} = \frac{m}{2g} - \varepsilon(h)$$

and

$$-u\frac{m}{g} - 1 - \theta\varepsilon(h) = e_1 > e_3 > e_5 > \cdots > e_{\frac{m}{g}-1} = -u\frac{m}{g} - \frac{m}{2g} - \theta\varepsilon(h),$$

P is a path. Moreover, the first and last vertices are the only ones which are congruent modulo $\frac{n}{g}$. It follows that

$$C = P \bigcup \left(P + \frac{n}{g}\right) \bigcup \left(P + \frac{2n}{g}\right) \bigcup \cdots \bigcup \left(P + \frac{(g-1)n}{g}\right)$$

is an m-cycle. In C, each difference in D occurs exactly g times, and for each $j \in Z_2$, if vertices $(i_1)_j$ and $(i_2)_j$ are both incident with edges of difference d, then $i_1 \equiv i_2 (\mathrm{mod}\ \frac{n}{g})$. Let $T = (C)$. It follows that T is an m-cycle system of $X(n; D)$ and $C_m \mid X(n; D)$. □

LEMMA 11. *Let $m \equiv 0 \pmod 2$, $m \ge 4$, $n \equiv 1 \pmod 2$, $m < 2n$, $g = gcd(m, n) > 1$, $h \equiv 1 \pmod 2$, and $2 \le h \le \frac{m}{g}$. For $r = 1, 2, \ldots, \frac{m}{g}$, define $d_r \in Z_n$ as*

$$d_r = \begin{cases} 0 & \text{if} \quad r = 1, \\ r & \text{if} \quad 2 \le r < h, \\ r+1 & \text{if} \quad h \le r \le \frac{m}{g} - 1, \\ \frac{m}{2g} + \frac{n}{g} & \text{if} \quad i = \frac{m}{g}. \end{cases}$$

Let $D = \{d_1, d_2, \ldots, d_{\frac{m}{g}}\}$. Then $C_m \mid X(n; D)$.

Proof. Let $d_0 = 0$ and $e_k = \sum_{r=0}^{k} (-1)^r d_r$ for $1 \le k \le \frac{m}{g}$. Furthermore, let

$$\theta = \begin{cases} 0 & \text{if} \quad k < h, \\ 1 & \text{if} \quad k \ge h. \end{cases}$$

Then

$$e_k = \begin{cases} 0 & \text{if} \quad k = 0, \\ \frac{1}{2}k + 1 & \text{if} \quad k \equiv 0 \pmod 2 \quad \text{and} \quad 2 \le k \le \frac{m}{g} - 2, \\ -\frac{1}{2}(k-1) - \theta & \text{if} \quad k \equiv 1 \pmod 2, \\ \frac{n}{g} & \text{if} \quad k = \frac{m}{g}. \end{cases}$$

Let P be the following trail:

$$(e_0)_1, (e_1)_0, (e_2)_1, (e_3)_0, \ldots, (e_{\frac{m}{g}-2})_1, (e_{\frac{m}{g}-1})_0, (e_{\frac{m}{g}})_1.$$

The differences used in P are $d_1, d_2, \ldots, d_{\frac{m}{g}}$.

Since

$$0 = e_0 < e_2 < e_4 < \cdots < e_{\frac{m}{g}-2} = \frac{m}{2g} \leq \frac{n}{g} - 1$$

and

$$0 = e_1 > e_3 > e_5 > \cdots > e_{\frac{m}{g}-1} = -\frac{m}{2g} + 1 - \theta \geq -\left(\frac{n}{g} - 1\right),$$

P is a path. Moreover, the first and last vertices are the only ones which are congruent modulo $\frac{n}{g}$. It follows that

$$C = P \bigcup \left(P + \frac{n}{g}\right) \bigcup \left(P + \frac{2n}{g}\right) \bigcup \cdots \bigcup \left(P + \frac{(g-1)n}{g}\right)$$

is an m-cycle. In C, each difference in D occurs exactly g times, and for each $j \in Z_2$, if vertices $(i_1)_j$ and $(i_2)_j$ are both incident with edges of difference d, then $i_1 \equiv i_2 (\text{mod } \frac{n}{g})$. Let $T = (C)$. It follows that T is an m-cycle system of $X(n; D)$ and $C_m \mid X(n; D)$. □

With the above preparations, we now prove the following theorem.

THEOREM 12. *Let m and n be positive integers. If $m \equiv 0 \pmod 2$, $m \geq 4$, $n \equiv 1 \pmod 2$, $\frac{1}{2}m < n < \frac{3}{2}m$, and $n(n-1) \equiv 0 \pmod m$, then $C_m \mid K_{n,n} - I$.*

Proof. When $g = gcd(m, n) = 1$, $n = m + 1$ since $n(n-1) \equiv 0 \pmod m$. So, $C_m \mid K_{m+1,\ m+1} - I$ by Lemma 9.

If $n \neq m+1$, then $g > 1$. Since $n(n-1) \equiv 0 \pmod m$, we have $n-1 \equiv 0 \pmod{\frac{m}{g}}$. Let $s = \frac{(n-1)g}{m}$.

We view $K_{n,n}$ as $X(n; Z_n)$. For $t = 0, 1, \ldots, s-1$, let

$$D_t = \begin{cases} \{0, 1, \ldots, \frac{m}{g}\} & \text{if} \quad t = 0, \\ \{t\frac{m}{g} + 1, t\frac{m}{g} + 2, \ldots, t\frac{m}{g} + \frac{m}{g}\} & \text{if} \quad 1 \leq t \leq s-1. \end{cases}$$

Let $\delta = 1 - \varepsilon(s)[1 - \varepsilon(\frac{1}{2}m)]$. For $t = 0, 1, \ldots, s-1$, define h_t as

$$h_t = \begin{cases} t\frac{m}{g} + \frac{m}{2g} + \varepsilon(t)[1 - \varepsilon(\frac{1}{2}m)] + \frac{n}{g} & \text{if} \quad 0 \leq t \leq s-2, \\ \frac{n}{g} - \frac{m}{2g} - 1 + \delta - \varepsilon(\frac{1}{2}m) & \text{if} \quad t = s-1. \end{cases}$$

Observe that $h_t \in D_{t+1}$ for $0 \leq t \leq s-2$, $h_{s-1} \in D_0$, and $h_{s-1} \geq \delta$.

Thus, we let

$$\hat{D}_0 = \begin{cases} (D_0 \cup \{h_0\}) \backslash \{h_{s-1}, \delta\} & \text{if} \quad h_{s-1} > \delta, \\ (D_0 \cup \{h_0\}) \backslash \{h_{s-1}, d\} & \text{if} \quad h_{s-1} = \delta, \end{cases}$$

where $d \in D_0 \backslash \{\delta\}$ and $\varepsilon(d) = \delta$.

For $1 \leq t \leq s-1$, let

$$\hat{D}_t = (D_t \backslash \{h_{t-1}\}) \cup \{h_t\}.$$

When $t = 0$, $h_0 = \frac{m}{2g} + \frac{n}{g}$. Since $\delta = 0$ or 1, there are the following two cases.

Case 1. $\delta = 0$.

Then $h_{s-1} = \frac{n}{g} - \frac{m}{2g} - 1 - \varepsilon(\frac{1}{2}m)$. It is easy to check that $\varepsilon(h_{s-1}) = 0$. We take $u = 0$ and

$$h = \begin{cases} h_{s-1} & \text{if} \quad h_{s-1} > 0, \\ d & \text{if} \quad h_{s-1} = 0. \end{cases}$$

Clearly, $h_0 = \frac{m}{2g} + \frac{n}{g} + \varepsilon(h)$. By Lemma 10, we have $C_m \mid X(n; \hat{D}_0)$.

Case 2. $\delta = 1$.

Then $h_{s-1} = \frac{n}{g} - \frac{m}{2g} - \varepsilon(\frac{1}{2}m)$. It is easy to check that $\varepsilon(h_{s-1}) = 1$. We take $u = 0$ and

$$h = \begin{cases} h_{s-1} & \text{if} \quad h_{s-1} > 1, \\ d & \text{if} \quad h_{s-1} = 1. \end{cases}$$

Clearly, $h \geq 2$. By Lemma 11, we have $C_m \mid X(n; \hat{D}_0)$.

For each $t = 1, 2, \ldots, s-1$, we take $u = t$ and $h = h_{t-1} - t\frac{m}{g}$. It is easy to check that $h_t = u\frac{m}{g} + \frac{m}{2g} + \varepsilon(h) + \frac{n}{g}$. By Lemma 10, we have $C_m \mid X(n; \hat{D}_t)$.

Clearly,

$$\bigcup_{t=0}^{s-1} \hat{D}_t = \begin{cases} Z_n \backslash \{\delta\} & \text{if} \quad h_{s-1} > \delta, \\ Z_n \backslash \{d\} & \text{if} \quad h_{s-1} = \delta, \end{cases}$$

and

$$\hat{D}_t \cap \hat{D}_r = \phi \text{ for } t \neq r.$$

Suppose that T_t is an m-cycle system of $X(n; \hat{D}_t)$ for $0 \leq t \leq s-1$. Let $T = \bigcup_{t=0}^{s-1} T_t$ and

$$I = \begin{cases} \{\{i_0, (i+\delta)_1\} \mid i \in Z_n\} & \text{if} \quad h_{s-1} > \delta, \\ \{\{i_0, (i+d)_1\} \mid i \in Z_n\} & \text{if} \quad h_{s-1} = \delta. \end{cases}$$

Then T is an m-cycle system of $K_{n,n} - I$ and $C_m \mid K_{n,n} - I$. □

3. The proof of Theorem 6. Now, we are in a position to prove the main theorem of this paper.

Proof of Theorem 6. For $\frac{1}{2}m \leq n \leq \frac{3}{2}m$, we have $C_m \mid K_{n,n} - I$ by Corollary 8 and Theorem 12.

If $n > \frac{3}{2}m$, then we may write $n = qm + r$ with $\frac{1}{2}m < r \leq \frac{3}{2}m$ and $q \geq 1$.

Since

$$n \equiv 1 \pmod 2 \text{ and } n(n-1) \equiv 0 \pmod m,$$

we have

$$r \equiv 1 \pmod 2 \text{ and } r(r-1) \equiv 0 \pmod m.$$

Suppose the vertex set of $K_{n,n}$ is $\{v_0, v_1, \ldots, v_{qm+r-1}\} \cup \{u_0, u_1, \ldots, u_{qm+r-1}\}$ and $I = \{\{u_i, v_i\} \mid 0 \leq i \leq qm + r - 1\}$ is a 1-factor in $K_{n,n}$.

For $1 \leq i \leq q$, let $V_i = \{v_{(i-1)m+j} \mid 1 \leq j \leq m\}$ and $U_i = \{u_{(i-1)m+j} \mid 1 \leq j \leq m\}$. Let $V_{q+1} = \{v_{qm+j} \mid 1 \leq j \leq r-1\}$ and $U_{q+1} = \{u_{qm+j} \mid 1 \leq j \leq r-1\}$.

For $1 \le i \le q+1$, let $H_{i,i}$ be the subgraph of $K_{n,n} - I$ induced by $(V_i \cup \{v_0\}) \cup (U_i \cup \{u_0\})$. By Corollary 8 and Theorem 12, $C_m \mid H_{i,i}$. Let $T_{i,i}$ be an m-cycle system of $H_{i,i}$.

For $1 \le i, j \le q+1$ and $i \ne j$, let $H_{i,j}$ be the subgraph of $K_{n,n} - I$ induced by $V_i \cup U_j$. By Theorem 3, $C_m \mid H_{i,j}$. Let $T_{i,j}$ be an m-cycle system of $H_{i,j}$.

Let $T = \bigcup_{1 \le i,j \le q+1} T_{i,j}$. Then T is an m-cycle system of $K_{n,n} - I$ and $C_m \mid K_{n,n} - I$. This completes the proof. □

Acknowledgment. The authors are thankful to the referees for their helpful comments to improve the paper.

REFERENCES

[1] 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.

[2] 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.

[3] B. Alspach and H. Gavlas, *Cycle decompositions of K_n and $K_n - I$*, J. Combin. Theory Ser. B, 81 (2001), pp. 77–99.

[4] M. Šajna, *Cycle decompositions,* III: *Complete graphs and fixed length cycles*, J. Combin. Des., 10 (2002), pp. 27–78.

[5] R. Laskar and B. Auerbach, *On decomposition of r-partite graphs into edge-disjoint Hamilton circuits*, Discrete Math., 14 (1976), pp. 265–268.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 610–622

SHORTEST PATHS IN THE TOWER OF HANOI GRAPH AND FINITE AUTOMATA*

DAN ROMIK†

Abstract. We present efficient algorithms for constructing a shortest path between two configurations in the Tower of Hanoi graph and for computing the length of the shortest path. The key element is a finite-state machine which decides, after examining on the average only a small number of the largest discs (asymptotically, $\frac{63}{38} \approx 1.66$), whether the largest disc will be moved once or twice. This solves a problem raised by Andreas Hinz and results in a better understanding of how the shortest path is determined. Our algorithm for computing the length of the shortest path is typically about twice as fast as the existing algorithm. We also use our results to give a new derivation of the average distance $\frac{466}{885}$ between two random points on the Sierpiński gasket of unit side.

Key words. Tower of Hanoi, finite automata, Sierpiński gasket

AMS subject classifications. 68R05, 28A80

DOI. 10.1137/050628660

1. Introduction. The *Tower of Hanoi* puzzle, invented in 1883 by the French mathematician Edouard Lucas, has become a classic example in the analysis of algorithms and discrete mathematical structures (see, e.g., [4, section 1.1]). The puzzle consists of n discs, no two of the same size, stacked on three vertical pegs, in such a way that no disc lies on top of a smaller disc. A permissible *move* is to take the top disc from one of the pegs and move it to one of the other pegs, as long as it is not placed on top of a smaller disc. The set of configurations of the puzzle, together with the permissible moves, thus forms a graph in a natural way. The number of vertices in the n-disc Hanoi graph is 3^n.

The main question of interest is to find *shortest paths* in the configuration graph, i.e., shortest sequences of moves leading from a given initial configuration to a given terminal configuration. The simplest and most well known case is that in which it is required to move all the discs from one of the pegs to another, i.e., where the initial and terminal configurations are two of the three "perfect" configurations with all the discs on the same peg. This is very easy, and can be shown to take exactly $2^n - 1$ moves. More difficult is to get from a given arbitrary initial configuration to one of the perfect configurations—Hinz [6] calls this the "p1" problem. This takes $2^n - 1$ moves in the worst case (which is, for example, when the initial configuration is another perfect configuration), and on the average $\frac{2}{3} \cdot (2^n - 1)$ moves for a randomly chosen initial configuration [3]. Moreover, there is a simple and efficient algorithm to compute the shortest path in this case.

In the most general case of arbitrary initial *and* terminal configurations, however, the question of computing the shortest path and its length (the "p2" problem [6]) in the most efficient manner has not been completely resolved so far. (The worst-case behavior is still $2^n - 1$ moves, and the average number of moves for random initial and terminal configurations has been shown [2], [5] to be asymptotically $(1 + o(1))\frac{466}{885} \cdot 2^n$.) The main obstacle in the understanding of the behavior of the shortest

*Received by the editors April 6, 2005; accepted for publication (in revised form) January 30, 2006; published electronically August 29, 2006.
http://www.siam.org/journals/sidma/20-3/62866.html
†Department of Statistics, University of California, 367 Evans Hall, Berkeley, CA 94720-3860 (romik@stat.berkeley.edu).

path has been the behavior of the largest disc that "separates" the initial and terminal configurations, i.e., the largest disc which is not on the same peg in both configurations (trivially, any larger discs may simply be ignored). It is not difficult to see [6] that in a shortest path, this disc will be moved either once (from the source peg to the target peg) or twice (from the source to the target, via the third peg). The problem is to decide which of the two alternatives is the correct one. Once this is settled, the path may be constructed by two applications of the algorithm for the p1 problem. Hinz [6] proposed an algorithm for the computation of the shortest path based on this idea. The algorithm consists essentially of computing the length of the path for both alternatives and choosing the shorter of the two.

In this paper, we propose a more thorough explanation of the process whereby it is decided which of the two paths is the shortest. We show that it is possible to keep track of the relevant information using a finite-state machine, which at each step reads the locations of the next-smaller disc in the initial and terminal configurations and changes its internal state accordingly. Eventually, the machine reaches a terminal state, whereupon it pronounces which of the two paths is the shorter. For a random input, its expected stopping time is computed to be $\frac{63}{38}$, asymptotically when the number of discs grows to infinity. In other words, after observing on the average the locations of just the ≈ 1.66 largest discs in the initial and terminal configurations, we will know which of the paths to choose, and we will be able to continue using the algorithm for the p1 problem. If one is interested just in the length of the shortest path, then our algorithm is typically about twice as fast as the algorithm proposed by Hinz [6] (with a small constant overhead due to the initial 1.66 discs), since it overrides the need to compute both the distance for the path that moves the largest disc once and the path that moves it twice.

The paper is organized as follows: In the next section, we define the *discrete Sierpiński gasket* graph, a graph which is isomorphic to the Tower of Hanoi configuration graph, but for which the labeling of the vertices is simpler to understand. In section 3, we present the main ideas for the discrete Sierpiński gasket graph, and then in section 4 show how to translate the results to the Hanoi graph by a relabeling of the vertices. In section 5 we perform a probabilistic analysis of the finite-state machine, to compute the average number $\frac{63}{38}$ of discs that need to be read in order to decide whether the largest disc will be moved once or twice, and to give a new derivation of the asymptotic value $(1 + o(1))\frac{466}{885} \cdot 2^n$ for the average distance between two random configurations in the n-disc Hanoi graph, or equivalently of the statement that the average shortest-path distance between two random points in the Sierpiński gasket of unit side is equal to $\frac{466}{885}$. In section 6 we discuss extensions and some open problems. For an extensive bibliography of papers related to the Tower of Hanoi, we refer the interested reader to [13].

2. The discrete Sierpiński gasket. We now define a family of graphs called *discrete Sierpiński gaskets*. These graphs are finite versions of the famous fractal constructed by the Polish mathematician Waclaw Sierpiński in 1915. The connection between the Tower of Hanoi problem and the Sierpiński gasket was first observed by Stewart [12] and was later used by Hinz and Schief [9] in their calculation of the average distance between points on the Sierpiński gasket. The discrete Sierpiński gasket graphs that we define are identical to the graphs $S(n, 3)$ defined by Klavzar and Milutinovic in [10], and similar (although this requires proof) to the graphs S_n defined in [9], so some of the discussion below parallels the discussion in those papers.

The nth discrete Sierpiński gasket graph, which we denote by SG_n, consists of

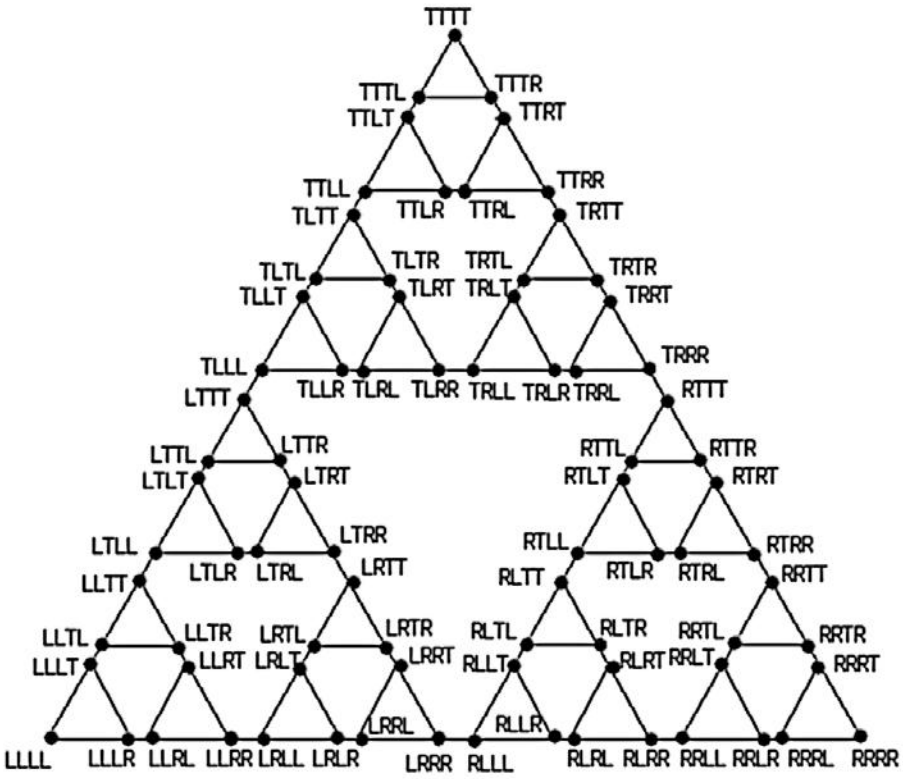


FIG. 1. *The graph SG_4.*

the vertex set $V(SG_n) = \{T, L, R\}^n$ (the symbols T, L, R indicate "top," "left," and "right," respectively), with the edges defined as follows: First, for each $x = a_{n-1}a_{n-2}\dots a_1a_0 \in V(SG_n)$ (for reasons that will become apparent below, this will be our standard indexing of the coordinates of the vertices of SG_n) we have edges connecting x to

$$a_{n-1}a_{n-2}\dots a_1\beta, \qquad \beta \in \{T, L, R\} \setminus \{a_0\}.$$

Second, define the *tail* of $x = a_{n-1}a_{n-2}\dots a_0$ as the suffix $a_ka_{k-1}\dots a_1a_0$ of x, where k is maximal such that $a_k = a_{k-1} = \cdots = a_0$. If x has a tail of length $k+1 < n$, then x is of the form $a_{n-1}a_{n-2}\dots a_{k+2}\beta\alpha\alpha\dots\alpha$, in which case connect x with an edge to the vertex

$$a_{n-1}a_{n-2}\dots a_{k+2}\alpha\beta\beta\dots\beta.$$

One possible embedding of SG_n in the plane is illustrated in Figure 1. This embedding makes clear the meaning of the labeling of the vertices: The first letter (the "most significant digit") signifies whether the vertex is in the top, left, or right triangles inside the big triangle; the next letter locates the vertex within the top, left, or right thirds of that triangle, etc.

It will be shown in section 4 that SG_n is isomorphic, in a computationally straightforward way, to the n-disc Hanoi graph. (The same was shown in [10], with less emphasis on explicit computation of the isomorphism.) Thus, the problem of shortest

paths on the Hanoi graph reduces to that of shortest paths in the discrete Sierpiński gasket. We tackle this problem in the next section.

3. Shortest paths in SG_n. For vertices $x, y \in V(SG_n)$, we define the distance $d(x, y)$ to be the length of a shortest path from x to y. Our goal is to write down a recursion equation for this distance, which is at the heart of the finite-state machine we will construct to compute $d(x, y)$. First, let us review briefly some of the known facts about $d(x, y)$ in the simple case when y is one of the "perfect" configurations $LLL\ldots L, RR\ldots R, TT\ldots T$. For concreteness, assume that $y = LLL\ldots L$, and let $x = a_{n-1}a_{n-2}\ldots a_1a_0 \in V(SG_n)$ as before. Then it is known that

$$d(x, y) = \sum_{a_k \neq L} 2^k.$$

A simple algorithm exists for computing a shortest path from x to y in this case. In the Hanoi labeling of the graph, the algorithm is described in [6]. In the current labeling, the algorithm is even simpler and is based on the binary number system: if one identifies the symbol L with 0 and the symbols R and T with 1, then traversing the edges of the graph becomes equivalent to the operations of subtraction or addition of 1 in binary notation. The number of steps to reach $LL\ldots L \equiv 00\ldots 0$ is then clearly the right-hand side in the above equation.

With these preparatory remarks, we now attack the problem of general $x = a_{n-1}a_{n-2}\ldots a_0$, $y = b_{n-1}b_{n-2}\ldots b_0$. First, observe that we may assume that $a_{n-1} \neq b_{n-1}$, since otherwise we may simply consider x and y as vertices in the graph SG_{n-1} (note the self-similar structure in the definition of the graph, also apparent in the Tower of Hanoi puzzle when one ignores the largest disc). For concreteness, we begin by analyzing in detail the case where $a_{n-1} = T$, $b_{n-1} = R$. Referring to Figure 1 for convenience, we see that

$$d(x, y) = \min\Big(1 + d(x, TRRR\ldots R) + d(y, RTT\ldots T),$$
$$1 + 2^{n-1} + d(x, TLLL\ldots L) + d(y, RLL\ldots L)\Big),$$

since in a shortest path from x to y, one must go from the top triangle to the right triangle either through the edge $\{TRR\ldots R, RTT\ldots T\}$ (we call this Alternative 1; see Theorem 1 below) or through a shortest path from $LTT\ldots T$ to $LRR\ldots R$ (Alternative 2). In the Tower of Hanoi language, this is an indication of the fact that in a shortest sequence of moves the largest disc must move either once or twice; see [6].

To simplify the next few equations, introduce the following notation: if $u = c_{n-1}c_{n-2}\ldots c_0 \in \{T, L, R\}^n$, let $u' = c_{n-2}c_{n-3}\ldots c_0$, and define for any $\alpha \in \{L, T, R\}$

$$f_\alpha(u) = \sum_{c_k \neq \alpha} 2^k.$$

Then we have

$$d(x, y) = 1 + \min\Big(f_R(x') + f_T(y'),\ \ 2^{n-1} + f_L(x') + f_L(y')\Big).$$

The recursion equations which will enable us to construct our finite-state machine and compute $d(x, y)$ are now given by the following theorem.

THEOREM 1 (the finite-state machine). *For* $u = c_{n-1}c_{n-2}\dots c_0$, $v = d_{n-1}d_{n-2}\dots d_0 \in \{T, L, R\}^n$, *define the functions*

$$p(u,v) = \min\Big(f_R(u) + f_T(v),\ 2^n + f_L(u) + f_L(v)\Big),$$
$$q(u,v) = \min\Big(2^n + f_R(u) + f_T(v),\ f_L(u) + f_L(v)\Big),$$
$$r(u,v) = \min\Big(f_R(u) + f_T(v),\ f_L(u) + f_L(v)\Big).$$

(Note that p, q, r *depend implicitly on the length* n *of the strings.) Then we have the equations*

$$p(u,v) = \begin{cases} f_R(u) + f_T(v) & \begin{array}{l} c_{n-1} = R, d_{n-1} = T \text{ or} \\ c_{n-1} = R, d_{n-1} = L \text{ or} \\ c_{n-1} = R, d_{n-1} = R \text{ or} \\ c_{n-1} = L, d_{n-1} = T \text{ or} \\ c_{n-1} = T, d_{n-1} = T \text{ or} \\ c_{n-1} = T, d_{n-1} = R, \end{array} \\ 2^n + p(u', v') & \begin{array}{l} c_{n-1} = T, d_{n-1} = L \text{ or} \\ c_{n-1} = L, d_{n-1} = R, \end{array} \\ 2^n + r(u', v') & c_{n-1} = L, d_{n-1} = L, \end{cases} \quad \textit{(Alternative 1)}$$

$$q(u,v) = \begin{cases} f_L(u) + f_L(v) & \begin{array}{l} c_{n-1} = L, d_{n-1} = L \text{ or} \\ c_{n-1} = L, d_{n-1} = T \text{ or} \\ c_{n-1} = L, d_{n-1} = R \text{ or} \\ c_{n-1} = R, d_{n-1} = L \text{ or} \\ c_{n-1} = T, d_{n-1} = L \text{ or} \\ c_{n-1} = T, d_{n-1} = R, \end{array} \\ 2^n + q(u', v') & \begin{array}{l} c_{n-1} = T, d_{n-1} = T \text{ or} \\ c_{n-1} = R, d_{n-1} = R, \end{array} \\ 2^n + r(u', v') & c_{n-1} = R, d_{n-1} = T, \end{cases} \quad \textit{(Alternative 2)}$$

$$r(u,v) = \begin{cases} f_R(u) + f_T(v) & c_{n-1} = R, d_{n-1} = T, \quad \textit{(Alternative 1)} \\ f_L(u) + f_L(v) & c_{n-1} = L, d_{n-1} = L, \quad \textit{(Alternative 2)} \\ 2^{n-1} + r(u', v') & \begin{array}{l} c_{n-1} = L, d_{n-1} = T \text{ or} \\ c_{n-1} = R, d_{n-1} = L, \end{array} \\ 2^n + r(u', v') & c_{n-1} = T, d_{n-1} = R, \\ 2^{n-1} + p(u', v') & \begin{array}{l} c_{n-1} = R, d_{n-1} = R \text{ or} \\ c_{n-1} = T, d_{n-1} = T, \end{array} \\ 2^{n-1} + q(u', v') & \begin{array}{l} c_{n-1} = T, d_{n-1} = L \text{ or} \\ c_{n-1} = L, d_{n-1} = R. \end{array} \end{cases}$$

Alternatives 1 *and* 2 *in the parentheses signify whether the minimum is attained by its first or second arguments, respectively. These equations will hold even for* $n = 1$ *if one sets trivially for* $u, v = \emptyset \in \{T, L, R\}^0 = \{\emptyset\}$*:*

$$f_\alpha(u) = 0,$$

$$p(u, v) = 0 \text{ (Alternative 1)}, \tag{1}$$

$$q(u, v) = 0 \text{ (Alternative 2)}, \tag{2}$$

$$r(u, v) = 0 \text{ (tie)}. \tag{3}$$

Proof. First, note that if $\alpha \in \{T, L, R\}$ and $w \in \{T, L, R\}^n$, then trivially $f_\alpha(w) \leq 2^n - 1$.

Here is the proof of the equation for $p(u, v)$ in several sample cases; the full proof is a slightly tedious case-by-case verification and consists of similar computations, so we omit it.

Sample case 1. Assume that $(c_{n-1}, d_{n-1}) = (R, T)$. In that case, we have

$$\begin{aligned} p(u, v) &= \min\Big(f_R(u') + f_T(v'), 2^n + 2^{n-1} + 2^{n-1} + f_L(u') + f_L(v')\Big) \\ &= \min\Big(f_R(u') + f_T(v'), 2^{n+1} + f_L(u') + f_L(v')\Big) \\ &= f_R(u') + f_T(v') = f_R(u) + f_T(v), \end{aligned}$$

since $f_R(u') + f_T(v') \leq 2^{n-1} - 1 + 2^{n-1} - 1 < 2^{n+1}$, so the minimum can only be attained by the first argument.

Sample case 2. Assume that $(c_{n-1}, d_{n-1}) = (R, L)$. Then we have

$$\begin{aligned} p(u, v) &= \min\Big(2^{n-1} + f_R(u') + f_T(v'), 2^n + 2^{n-1} + f_L(u') + f_L(v')\Big) \\ &= 2^{n-1} + \min\Big(f_R(u') + f_T(v'), 2^n + f_L(u') + f_L(v')\Big) \\ &= 2^{n-1} + f_R(u') + f_T(v') = f_R(u) + f_T(v), \end{aligned}$$

again since $f_R(u') + f_T(v') \leq 2^n - 2 < 2^n$, so again Alternative 1 must hold.

Sample case 3. Assume that $(c_{n-1}, d_{n-1}) = (T, L)$. Then

$$\begin{aligned} p(u, v) &= \min\Big(2^{n-1} + 2^{n-1} + f_R(u') + f_T(v'), 2^n + 2^{n-1} + f_L(u') + f_L(v')\Big) \\ &= 2^n + \min\Big(f_R(u') + f_T(v'), 2^{n-1} + f_L(u') + f_L(v')\Big) \\ &= 2^n + p(u', v'). \end{aligned}$$

Note that the order of the arguments in the minimum is preserved, so that once the correct alternative for $p(u', v')$ is determined, this is propagated back to $p(u, v)$.

Sample case 4. Assume that $(c_{n-1}, d_{n-1}) = (L, L)$. Then

$$\begin{aligned} p(u, v) &= \min\Big(2^{n-1} + 2^{n-1} + f_R(u') + f_T(v'), 2^n + f_L(u') + f_L(v')\Big) \\ &= 2^n + r(u', v'). \end{aligned}$$

□

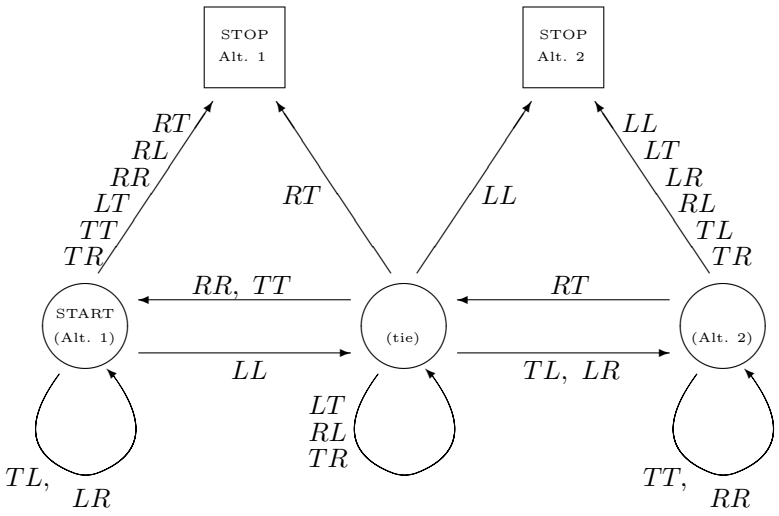


FIG. 2. *The finite-state machine: deciding between Alternatives* 1 *and* 2. *The two letters signify the two inputs from* x *and* y, *reading at each step the next-most-significant symbol. The parentheses in the nonterminal states indicate that if the input terminates without a decision, then in the START state Alternative* 1 *wins, in the rightmost state Alternative* 2 *wins, and in the middle state there is a tie, meaning that the shortest path is not unique and both alternatives are valid. (Termination of the input corresponds to the recursion equations leading to an evaluation of either* $p(u, v)$, $q(u, v)$, *or* $r(u, v)$ *with* $u = v = \emptyset$, *so the above claim follows from equations* (1), (2), (3) *together with the fact mentioned in the proof of Theorem* 1 *that the order of the arguments is propagated throughout the recursion.)*

A schematic representation of the finite-state machine is shown in Figures 2 and 3. We present two variants of the machine: the machine in Figure 2 only decides between Alternatives 1 and 2, in the case in which x begins with the symbol T and y begins with R. The machine in Figure 3, which has auxiliary counters for the distance and for the variable n (so strictly speaking it is not really a finite-state automaton), actually computes $d(x, y)$, and it is designed to treat the general case of any two configurations $x, y \in V(SG_n)$. This is done by including an initial component that discards the first few symbols which are identical for x and y, and another component that permutes the symbols T, L, R to fit the design of the basic machine in Figure 2.

4. Translating between the Hanoi graph and SG_n. We now define the graph of configurations in the n-disc Tower of Hanoi puzzle and show that it is isomorphic to SG_n. The isomorphism may be computed by reading sequentially the locations of the discs, starting with the largest one (which corresponds to the most significant digit in the Sierpiński gasket labeling), and following a diagram of permutations translating the labels of the three pegs into the symbols T, L, R (another finite-state machine!). Together with the results of the previous section, this will give an effective means of computing the length of the shortest path between any two vertices in the Hanoi graph, and of deciding whether the largest disc will be moved once or twice in a shortest path. After that, we describe briefly an algorithm for actually constructing the shortest path, based on the algorithm for getting to a perfect configuration.

Label the three pegs in the Tower of Hanoi with the symbols $0, 1, 2$. Since in a

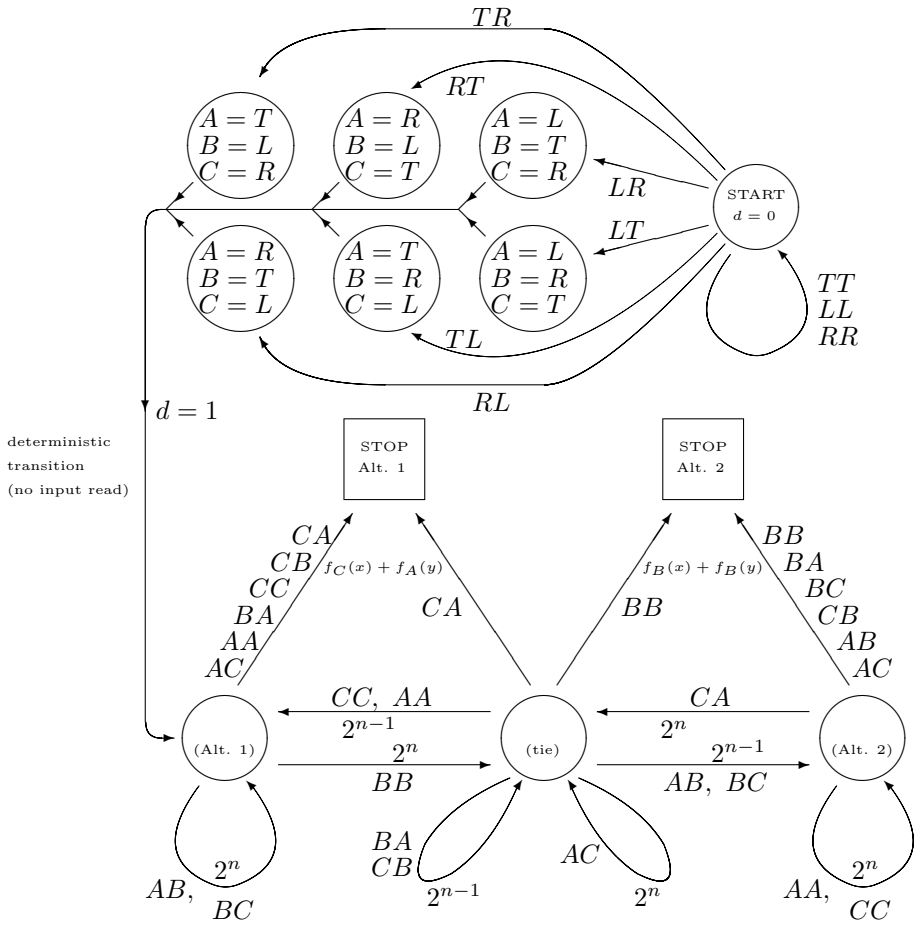


FIG. 3. *The finite-state machine: computing $d(x, y)$, the general case. Add to d the number on each edge traversed, decrease n by 1, and replace x by x' and y by y'. For the deterministic transition, do not read input or decrease n.*

legal configuration, on each of the pegs the discs are arranged in increasing size from top to bottom, a configuration is described uniquely by specifying, for any disc, the label of its peg. Thus, we define H_n, the nth *Hanoi graph*, to be the graph whose vertex set is the set $V(H_n) = \{0, 1, 2\}^n$ (with the coordinates of the vectors specifying, from left to right, the labels of the pegs of the largest disc, second-largest disc, etc.), and where edges between configurations correspond to permissible moves. Figure 4 shows the graph H_4.

The isomorphism between H_n and SG_n is now described by the following theorem.

THEOREM 2. *H_n and SG_n are isomorphic graphs. The finite-state machine shown in Figure 5 translates a Hanoi configuration $s \in \{0, 1, 2\}^n$ into a Sierpiński gasket labeling $z \in \{T, L, R\}^n$ by reading the digits from left to right and outputting the symbols T, L, R at each step according to the identifications in its internal state, then changing the internal state according to the input.*

Proof. This is Theorem 2 in [10]. There it was claimed simply that H_n and

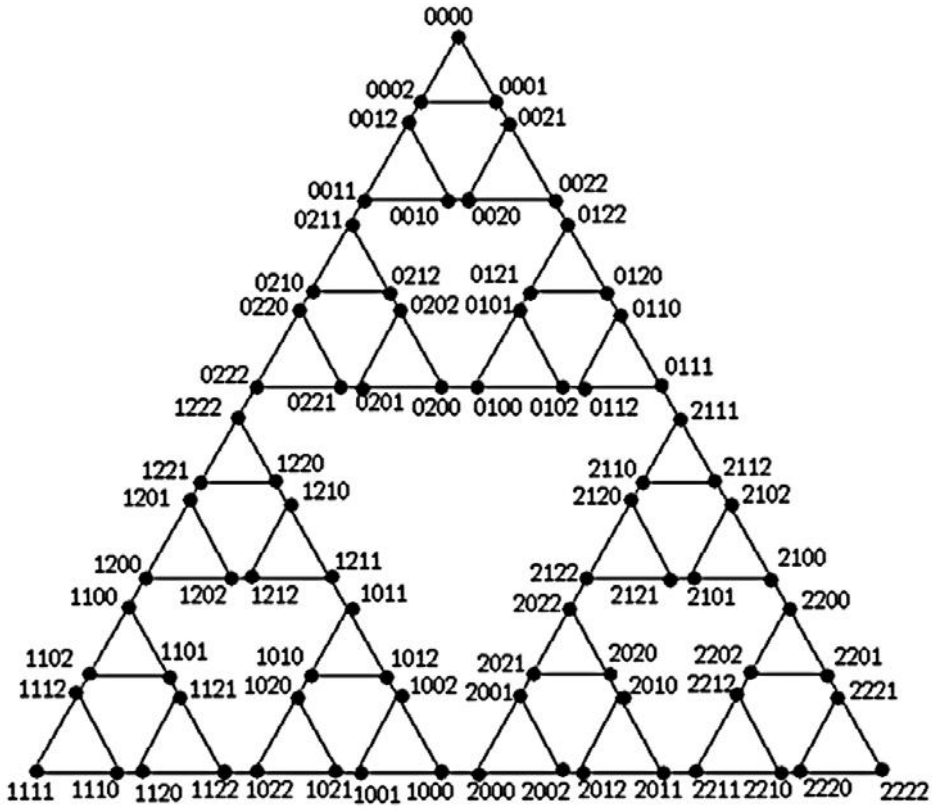


FIG. 4. *The graph H_4.*

SG_n are isomorphic, but the proof, which is by induction, actually describes how to compute the isomorphism, and this is easily seen to be equivalent to our finite-state machine formulation. (A similar argument is used in the proof of Lemma 2 in [9], which constructs an isomorphism between H_n and a different "discrete Sierpiński graph," defined in a geometrical way which is not obviously related to the current SG_n graph.) □

Summary. By running the machines of Figures 3 and 5 in parallel, we now have an algorithm for computing $d(x, y)$ for two arbitrary configurations in the Hanoi graph, and for solving the decision problem for the largest disc, i.e., to decide whether the largest disc which it is necessary to move will move once or twice. As we will show in the next section, when x and y are randomly chosen configurations, the expected stopping time of the machine is $\frac{63}{38}$. (This random variable even has an exponential tail distribution, so with very high probability only a small number of discs will need to be read to solve the decision problem.) Having solved the decision problem, the shortest path may now be computed in a straightforward manner, as described in [6], using the algorithm for getting to a perfect configuration (use the algorithm described in [6], or the algorithm for the Sierpiński gasket described in section 3 together with the machine of Figure 5—which incidentally leads to an algorithm for getting to perfect configurations which we have not found in the literature).

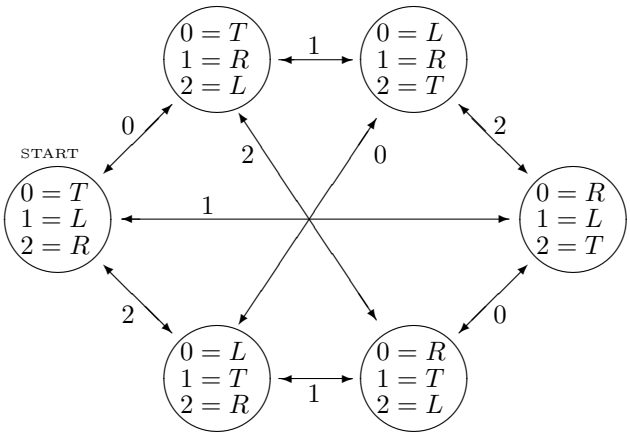


FIG. 5. *Computing the isomorphism between H_n and SG_n.*

5. The case of random inputs.

5.1. How many discs must be read to solve the decision problem? In this section, we calculate the average number of discs that must be read in order to decide whether in a shortest path the largest disc will be moved once or twice. Let $x = a_{n-1}a_{n-2}\dots a_0 \in V(H_n)$, $y = b_{n-1}b_{n-2}\dots b_0 \in V(H_n)$. Assume that we have already discarded the largest discs which for x and y were on the same peg, so that $a_{n-1} \neq b_{n-1}$. The algorithm for solving the decision problem then tells us to run the machines of Figures 3 and 5 until they reach a terminal state (or we run out of input). Since we have already initialized by discarding irrelevant discs, we will really be using the machine of Figure 2 (keeping track of the correct identification of the symbols L, T, R with the pegs $0, 1, 2$). Since we are dealing with random inputs, what we are really interested in is the absorption time of the Markov chain whose transition matrix is

$$\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} 2/9 & 1/9 & 0 & 2/3 & 0 \\ 2/9 & 1/3 & 2/9 & 1/9 & 1/9 \\ 0 & 1/9 & 2/9 & 0 & 2/3 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

into the terminal states 4 and 5. We may identify these two states to get the simpler matrix

$$\begin{array}{c} 1 \\ 2 \\ 3 \\ (45) \end{array} \begin{pmatrix} 2/9 & 1/9 & 0 & 2/3 \\ 2/9 & 1/3 & 2/9 & 2/9 \\ 0 & 1/9 & 2/9 & 2/3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

For $i = 1, 2, 3$, denote by t_i the expected time to get to state (45), *starting from state* i. Then clearly we have the equations

$$\begin{aligned} t_1 &= 1 + \frac{2}{9}t_1 + \frac{1}{9}t_2, \\ t_2 &= 1 + \frac{2}{9}t_1 + \frac{1}{3}t_2 + \frac{2}{9}t_3, \\ t_3 &= 1 \qquad\quad + \frac{1}{9}t_2 + \frac{2}{9}t_3. \end{aligned}$$

It may easily be verified that the solution to this system of equations is

$$t_1 = \frac{63}{38}, \quad t_2 = \frac{99}{38}, \quad t_3 = \frac{63}{38}.$$

The value $t_1 = \frac{63}{38}$ is our expected stopping time, since $i = 1$ corresponds to the initial state. Note that this value is the limit as $n \to \infty$ of the average number of discs that must be read; in reality, for finite n the value will be slightly smaller since after n steps we run out of input and the machine terminates even if it has not reached a terminal state. We summarize in the following theorem.

THEOREM 3. *The decision problem for shortest paths can be solved in average time $O(1)$. Specifically, the average number of disc pairs that our algorithm must read, once identical discs have been discarded, is bounded from above by, and converges as $n \to \infty$ to, $\frac{63}{38}$.*

5.2. The average distance between points on the Sierpiński gasket. Hinz and Schief [9] computed the average length $\frac{466}{885}$ of a shortest path between two random points on the *infinite* Sierpiński gasket of unit side. An equivalent result of Hinz [5] and of Chan [2], in terms of the Tower of Hanoi, is that the average number of moves in a shortest path between two random configurations in the n-disc Tower of Hanoi is asymptotically $(1 + o(1))(\frac{466}{885}) \cdot 2^n$ as $n \to \infty$.

Without going into too much detail, we show that it is possible to obtain the value of $\frac{466}{885}$ just by looking at the finite-state machine of Figure 3. Since we are dealing with the infinite gasket, we start with $n = 0$ and, as before, decrease the value of n after each step, so that n will go into the negative integers. Let d_1, d_2, d_3, d_4 be the expected accumulated values of the variable d if one starts the machine, with initial values $n = 0, d = 0$, at either of the four nonterminal states, in order of their distance from the state START (so d_1 is the total distance; d_2 is the distance after discarding identical most-significant digits of x and y, etc.). Then we have the equations

$$\begin{aligned}
d_1 &= \frac{1}{3} \cdot \frac{1}{2} d_1 + \frac{2}{3} \cdot \frac{1}{2} d_2, \\
d_2 &= \frac{2}{9} \cdot \left(1 + \frac{1}{2} d_2\right) + \frac{1}{9} \cdot \left(1 + \frac{1}{2} d_3\right) + \frac{2}{3} \cdot \left(\frac{1}{2} + \frac{2}{3}\right), \\
d_3 &= \frac{2}{9} \cdot \left(\frac{1}{2} + \frac{1}{2} d_2\right) + \frac{2}{9} \cdot \left(\frac{1}{2} + \frac{1}{2} d_3\right) + \frac{1}{9} \cdot \left(1 + \frac{1}{2} d_3\right) \\
&\quad + \frac{2}{9} \cdot \left(\frac{1}{2} + \frac{1}{2} d_4\right) + \frac{2}{9} \cdot \frac{2}{3}, \\
d_4 &= \frac{1}{9} \cdot \left(1 + \frac{1}{2} d_3\right) + \frac{2}{9} \cdot \left(1 + \frac{1}{2} d_4\right) + \frac{2}{3} \cdot \left(\frac{1}{2} + \frac{2}{3}\right).
\end{aligned}$$

The value $(\frac{1}{2} + \frac{2}{3})$ in the second and fourth equations is the expected value of $f_C(x) + f_A(y)$ (respectively, $f_B(x) + f_B(y)$), given that the first pair of inputs in the lower part of Figure 3 is one of the six values AC, AA, BA, CC, CB, CA (respectively, BB, BA, BC, CB, AB, AC).

Again, it may be verified that the solution to this system of equations is

$$d_1 = \frac{466}{885}, \quad d_2 = \frac{233}{177}, \quad d_3 = \frac{188}{177}, \quad d_4 = \frac{233}{177},$$

which gives our claimed value for d_1.

6. Extensions and open problems. We mention possible connections of our work to other questions related to the Tower of Hanoi and to the study of fractal structures similar to the Sierpiński gasket.

- *Higher-dimensional Sierpiński gaskets and other fractals.* For each $n \geq 2$, there is a fractal known as the Sierpiński gasket in $\mathbb{R}^n$ analogous to the Sierpiński gasket in $\mathbb{R}^2$. Bandt and Kuschel [1] showed that the average distance between two points in the Sierpiński gasket in $\mathbb{R}^n$ is equal to

$$\frac{n}{(2n+1)(n+1)}\left(2n - \frac{n^2-1}{n^3+7n^2+7n+9}\right).$$

  In this case, the problem is again one of determining which of several parts of the gasket a shortest path between two given points should pass through. It seems very likely that one can construct a finite-state machine to solve this problem, and that the result of Bandt and Kuschel can be re-proved using this method. More generally, one can ask similar questions for the class of postcritically finite fractals (see [1] for the definition), and it would be interesting to characterize the family of such fractals for which one can solve the shortest path problem using a finite-state machine, and to give a general method for constructing such a machine given the symmetries of the fractal. As an example, we have computed the average distance between two points in the modified Sierpiński gasket (in $\mathbb{R}^2$) which has side lengths 2, 2, and 1. It is equal to

$$\frac{147644401107013}{168923515522320} \approx 0.955.$$

  We omit the computation, which is somewhat tedious and uses basically the same ideas as the ones presented here.
- *Nonunique shortest paths in H_n.* In a recent paper [8], Hinz et al. proved the following formula for the number a_n of pairs (x, y) of vertices in the Tower of Hanoi graph H_n for which there are two shortest paths connecting x and y:

$$a_n = \frac{3}{4\sqrt{17}}\left[\left(\sqrt{17}+1\right)\left(\frac{5+\sqrt{17}}{2}\right)^n - 2\sqrt{17}\cdot 3^n + \left(\sqrt{17}-1\right)\left(\frac{5-\sqrt{17}}{2}\right)^n\right].$$

  Their proof of this formula makes use of Stern's diatomic sequence. However, as the authors point out, this formula can also be proved using our finite automaton, since basically a_n counts the number of paths in the graph of states of the automaton in Figure 3 leading from the state START to the state (tie). By writing down the adjacency matrix of the graph of states and diagonalizing, one can obtain the formula above.
  A related result, not mentioned in [8] but easily seen to follow from the same ideas, is the following: Let S be the Sierpiński gasket fractal in $\mathbb{R}^2$. Let A be the subset of $S \times S$ consisting of all pairs (x, y) of points in S for which there are two shortest paths connecting x and y in S. Then the Hausdorff dimension of A is $\log[(5+\sqrt{17})/2]/\log 2$.

REFERENCES

[1] C. Bandt and T. Kuschel, *Self-similar sets* VIII. *Average interior distance in some fractals*, Measure Theory (Oberwolfach, 1990), Rend. Circ. Mat. Palermo (2) Suppl., 28 (1992), pp. 307–317.

[2] T. Chan, *A statistical analysis of the Towers of Hanoi problem*, Internat. J. Comput. Math., 28 (1988), pp. 543–623.

[3] M. C. Er, *An analysis of the generalized Towers of Hanoi problem*, BIT, 23 (1983), pp. 429–435.

[4] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics*, 2nd ed., Addison-Wesley, Reading, MA, 1994.

[5] A. Hinz, *The Tower of Hanoi*, Enseign. Math., 35 (1989), pp. 289–321.

[6] A. Hinz, *Shortest paths between regular states of the Tower of Hanoi*, Inform. Sci., 63 (1992), pp. 173–181.

[7] A. Hinz, *The Tower of Hanoi*, in Algebras and Combinatorics (ICAC'97, Hong Kong), Springer, Singapore, 1999, pp. 277–289.

[8] A. Hinz, S. Klavzar, U. Milutinovic, D. Parisse, and C. Petr, *Metric properties of the Tower of Hanoi graphs and Stern's diatomic sequence*, European J. Combin., 26 (2005), pp. 693–708.

[9] A. Hinz and A. Schief, *The average distance on the Sierpiński gasket*, Probab. Theory Related Fields, 87 (1990), pp. 129–138.

[10] S. Klavzar and U. Milutinovic, *Graphs $S(n, k)$ and a variant of the Tower of Hanoi problem*, Czechoslovak Math. J., 47 (1997), pp. 95–104.

[11] S. Klavzar, U. Milutinovic, and C. Petr, *On the Frame-Stewart algorithm for the multi-peg Tower of Hanoi problem*, Discrete Appl. Math., 120 (2002), pp. 141–157.

[12] I. Stewart, *Le lion, le lama et la laitue*, Pour la Science, 142 (1989), pp. 102–107.

[13] P. K. Stockmeyer, *The Tower of Hanoi: A Historical Survey and Bibliography*, version 0.2, preprint, 2001. Available online from http://www.cs.wm.edu/~pkstoc/h_papers.html.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 623–627

DENSE ARRANGEMENTS ARE LOCALLY VERY DENSE. I*

JÓZSEF SOLYMOSI†

Abstract. The Szemerédi–Trotter theorem [*Combinatorica*, 3 (1983), pp. 381–392] gives a bound on the maximum number of incidences between points and lines on the Euclidean plane. In particular it says that n lines and n points determine $O(n^{4/3})$ incidences. Let us suppose that an arrangement of n lines and n points defines $cn^{4/3}$ incidences, for a given positive c. It is widely believed that such arrangements have special structure, but no results are known in this direction. Here we show that for any natural number, k, one can find k points of the arrangement in general position such that any pair of them is incident to a line from the arrangement, provided by $n \geq n_0(k)$. In a subsequent paper we will establish a similar statement for hyperplanes.

Key words. point-line incidences, Szemerédi–Trotter theorem, regularity lemma

AMS subject classifications. 52C10, 52C30, 52C45

DOI. 10.1137/05062826X

1. Introduction. The celebrated Szemerédi–Trotter theorem [21] states that for n points on the plane, the number of m-rich lines cannot exceed

$$O(n^2/m^3 + n/m), \tag{1.1}$$

and this bound is tight in the worst case. This result has numerous applications not only in geometry [11, 22] but also in number theory [4]. The Szemerédi–Trotter theorem has various proofs; the most elegant is the one by Székely [22]. However, the proofs provide very limited insight into the view of the structure of extremal arrangements. It is widely believed that a point-line arrangement which defines many incidences has a special, somehow rigid structure. For example, let us mention here a question of Elekes. Is it true that for every $c > 0$ there is a $c' > 0$ such that if a set of n points on the plane contains at least cn^2 collinear triples, then at least $n^{c'}$ points are along an algebraic curve of degree d, where d is a universal constant?

The main purpose of this paper is to show that any arrangement with close to the maximum number of incidences is locally a collection of complete geometric graphs. For the sake of simplicity we state the theorem for the balanced case, when the number of lines equals the number of points, but it is quite straightforward to see the similar statement for unbalanced cases as well.

Recent work of Gowers [6] and Nagle, Rödl, Schacht, and Skokan (see [9, 12, 13]) has established a hypergraph removal lemma, which in turn implies similar results to hyperplanes; however, a slightly different approach is needed, mainly because the higher dimensional extensions of the Szemerédi–Trotter theorem are not as well defined as in the planar case. To obtain sharp bounds one needs certain restrictions on the arrangements. Therefore the corresponding structure theorems will appear in a subsequent paper.

*Received by the editors April 1, 2005; accepted for publication (in revised form) February 27, 2006; published electronically August 29, 2006. This research was supported by OTKA and NSERC grants.

http://www.siam.org/journals/sidma/20-3/62826.html

†Department of Mathematics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2 (solymosi@math.ubc.ca).

A point set or a set of lines is in *general position* if no three of the elements are collinear or concurrent.

THEOREM 1.1. *For every natural number k and real $c > 0$ there is a threshold $n_0 = n_0(k, c)$ such that if an arrangement of $n \geq n_0$ lines and n points defines at least $cn^{4/3}$ incidences, then one can always find k points of the arrangement in general position, such that any pair of them is incident to a line from the arrangement.*

As we will see from the proof, the complete k-tuple is "local" in the sense that for any pair of points of the k-tuple, p_1 and p_2, the number of points from the arrangement, incident to the line segment (p_1, p_2), is less than k.

2. Proof of Theorem 1.1. The main tool of the proof is Szemerédi's regularity lemma [19, 20]. We will use its *counting lemma* form, because it is easier to extend to hypergraphs which we will need for the higher dimensional extensions. Let us prove first the simplest case, to show that there is always a triangle. This "simplest case" is interesting in its own right; the statement of Lemma 2.1 implies Roth's theorem [14] about arithmetic progressions on dense subsets of integers. For the details we refer to [16, 17].

LEMMA 2.1. *For every $c > 0$ there is a threshold $n_0 = n_0(c)$ and a positive $\delta = \delta(c)$ such that, for any set of $n \geq n_0$ lines L and any set of $m \geq cn^2$ points P, if every point is incident to three lines, then there are at least δn^3 triangles in the arrangement. (A triangle is a set of three distinct points from P such that any two are incident to a line from L.)*

This lemma follows the following theorem of Ruzsa and Szemerédi [15], which is also called the *triangle removal lemma* or the counting lemma for triangles.

THEOREM 2.2 (see [15]). *Let G be a graph on n vertices. If G is the union of cn^2 edge-disjoint triangles, then G contains at least δn^3 triangles, where δ depends on c only.*

The same theorem from a different angle is the following.

THEOREM 2.3. *Let G be a graph on n vertices. If G contains $o(n^3)$ triangles, then one can remove $o(n^2)$ edges to make G triangle-free.*

To prove Lemma 2.1, let us construct a graph where L is the vertex set and two vertices are adjacent if and only if the corresponding lines cross at a point of P. This graph is the union of cn^2 disjoint triangles; every point of P defines a unique triangle, so we can apply Theorem 2.2.

To determine the number of triangles in any arrangement of lines and points seems to be a hard task. A related conjecture of de Caen and Székely [1] is that n points and m lines cannot determine more than nm triangles.

One can repeat the same argument, now with k instead of 3. The corresponding counting lemma can be proven using Szemerédi's regularity lemma. The proof is analogous to the Ruzsa–Szemerédi theorem. There are slightly different ways to state the regularity lemma; for our purposes the so-called *degree form* is convenient. For the notations and proofs we refer to the survey paper of Komlós and Simonovits [7].

THEOREM 2.4 (regularity lemma). *For every $\epsilon > 0$ there is an $M = M(\epsilon)$ such that if $G = (V, E)$ is any graph and $d \in (0, 1]$ is any real number, then there is a partition of the vertex set V into $k + 1$ clusters $V_0, V_1, \ldots, V_k$, and there is a subgraph $G' \subset G$ with the following properties:*

- *$k \leq M$,*
- *$|V_0| \leq \epsilon |V|$,*
- *all clusters V_i, $i \geq 1$, are of the same size $m \leq \lceil \epsilon |V| \rceil$,*
- *$\deg_{G'}(v) > \deg_G(v) - (d + \epsilon)|V|$ for all $v \in V$,*

- $e(G'(V_i)) = 0$ *for each* $i \geq 1$,
- *all pairs* $G'(V_i, V_j)$ $(1 \leq i < j \leq k)$ *are* ϵ*-regular, each with a density either* 0 *or greater than* d.

Armed with the regularity lemma we are ready to prove the following statement, which is crucial in the proof of Theorem 1.1.

LEMMA 2.5. *For every* $c > 0$ *there is a threshold* $n_0 = n_0(c)$ *and a positive* $\delta = \delta(c)$ *such that, for any set of* $n \geq n_0$ *lines* L *and any set of* $m \geq cn^2$ *points* P*, if every point is incident to* k *lines, then there are at least* δn^k *complete* k*-tuples in the arrangement. (A complete* k*-tuple is a set of* k *distinct lines in general position from* L *such that any two intersect in a point from* P*.)*

Proof. To avoid having too many degenerate k-tuples, we remove some points from P which have many lines incident to them. Let P', which is the subset of P, consist of points incident to at most $100/c$ lines from L. We can apply (1.1) to see that P' is a large subset of P, say $2|P'| > |P|$. Let us construct a graph G where L is the vertex set and two vertices are adjacent if and only if the corresponding lines cross at a point of P'. This graph, G, is the union of at least $\frac{c}{2}n^2$ edge-disjoint K_ks. Find a subgraph, G', provided by Theorem 2.4 with $\epsilon \ll c$. In G' we still have some complete K_ks (when going from G to G' we removed $(\epsilon + d)n^2$ edges, much less than cn^2). The edges of such a complete graph are connecting V_is such that the bipartite graphs between them are dense and regular. This already implies the existence of many complete subgraphs, K_ks, as the following lemma, quoted from [7], shows. □

LEMMA 2.6. *Given* $d > \epsilon > 0$*, a graph* R *on* k *vertices, and a positive integer* m*, let us construct a graph* G *by replacing every vertex of* R *by* m *vertices, and replacing the edges of* R *with* ϵ*-regular pairs of density at least* d*. Then* G *has at least* αm^k *copies of* R*, where* α *depends on* ϵ, d*, and* k *but not on* m.

Most of the complete k-vertex subgraphs of graph G' define a complete k-tuple in the arrangement; i.e., the corresponding lines are in general position. To see this, let us count the "degenerate" k-tuples, where at least one triple is concurrent. The number of concurrent triples is at most $cn^2\binom{100/c}{3} \leq c'n^2$. For every concurrent triple one can select $k-3$ lines to get a degenerate k-tuple. The expression $c'n^{k-1}$ is clearly an upper bound on the degenerate k-tuples; therefore most of the complete graphs on k vertices in G' are complete k-tuples if n is large enough, $n \geq n_0 = n_0(c)$.

The final step of the proof of Theorem 1.1 is to show that arrangements with many incidences always have a substructure where one uses Lemma 2.1. We divide the arrangement into smaller parts where we apply the dual of Lemma 2.1. The common technique to do that is so-called *cutting*, which was introduced by Chazelle (see in [2] or in [10]) about 20 years ago. Here we use a more general result, a theorem of Matoušek [8].

LEMMA 2.7. *Let* P *be a point set,* $P \subset \boldsymbol{R}^d, |P| = n$*, and let* r *be a parameter,* $1 \ll r \ll n$*. Then the set* P *can be partitioned into* t *sets* $\Delta_1, \Delta_2, \ldots, \Delta_t$*, in such a way that* $n/r \leq |\Delta_i| \leq 2n/r$ *for all* i*, and any hyperplane crosses no more than* $O(r^{1-1/d})$ *sets, where* $t = O(r)$.

One can use the $d = 2$ case and we choose the value $r = \beta_k n^{2/3}$, where β_k is a constant that depends on k and which we will specify later. Let us count the number of incidences along the lines of L, according to the partition of P. For a given line $\xi \in L$, we count the sum $\sum_{i=1}^{t} \lfloor |\Delta_i \bigcap \xi|/k \rfloor$, which is not much smaller than the number of incidences on ξ over k if ξ is rich of incidences, say, incident to much more than $r^{1/2}k$ points of P. From the condition of Theorem 1.1 and the properties of the

partition we have the following inequality:

$$\frac{c}{k}n^{4/3} \leq \sum_{\xi\in L}\sum_{i=1}^{t}\left\lfloor \frac{|\Delta_i \bigcap \xi|}{k}\right\rfloor + |L|r^{1/2}.$$

Choosing $\beta_k = \frac{c}{2k}$, the inequality becomes

$$\frac{cn^{4/3}}{2k} = c_k n^{\frac{4}{3}} \leq \sum_{\xi\in L}\sum_{i=1}^{t}\left\lfloor \frac{|\Delta_i \bigcap \xi|}{k}\right\rfloor = \sum_{i=1}^{t}\sum_{\xi\in L}\left\lfloor \frac{|\Delta_i \bigcap \xi|}{k}\right\rfloor.$$

Therefore there is an index i, such that

$$c_k n^{2/3} \leq \sum_{\xi\in L}\left\lfloor \frac{|\Delta_i \bigcap \xi|}{k}\right\rfloor.$$

If $s = \lfloor \frac{|\Delta_i \cap \xi|}{k} \rfloor$, then we can partition the points incident to ξ into s consecutive k-tuples. We can break the line into s k-rich line segments and consider them as separate lines. Our combinatorial argument in Lemma 2.5 is robust enough to allow such modifications. Then we have some $c'n^{2/3}$ k-rich lines on $|\Delta_i| = c''n^{1/3}$ points. (Another possible way to show that there are at least $c'n^{2/3}$ k-rich lines is to apply the Szemerédi–Trotter theorem, (1.1), to show that most of the lines are not "very rich.") To complete the proof of Theorem 1.1, we apply the dual statement of Lemma 2.5.

REFERENCES

[1] D. de Caen and L. A. Székely, *On dense bipartite graphs of girth eight and upper bounds for certain configurations in planar point-line systems*, J. Combin. Theory Ser. A., 77 (1997), pp. 268–278.

[2] B. Chazelle, *The Discrepancy Method*, Cambridge University Press, Cambridge, UK, 2000.

[3] K. L. Clarkson, H. Edelsbrunner, L. J. Guibas, M. Sharir, and E. Welzl, *Combinatorial complexity bounds for arrangements of curves and spheres*, Discrete Comput. Geom., 5 (1990), pp. 99–160.

[4] Gy. Elekes, *SUMS versus PRODUCTS in number theory, algebra and Erdős geometry*, in Paul Erdos and His Mathematics II, Bolyai Soc. Math. Stud. 11, János Bolyai Math. Soc., Budapest, 2002, pp. 241–290.

[5] Gy. Elekes and Cs. D. Tóth, *Incidences of not-too-degenerate hyperplanes*, in Proceedings of the 21st ACM Symposium in Computer Geometrics (Pisa, 2005), ACM Press, New York, pp. 16–21.

[6] W. T. Gowers, *Hypergraph Regularity and the Multidimensional Szemerédi Theorem*, preprint.

[7] J. Komlós and M. Simonovits, *Szemerédi's regularity lemma and its applications in graph theory*, in Combinatorics, Paul Erdős is eighty, Vol. 2 (Keszthely, 1993), Bolyai Soc. Math. Stud. 2, János Bolyai Math. Soc., Budapest, 1996, pp. 295–352.

[8] J. Matoušek, *Efficient partition trees*, Discrete Comput. Geom., 8 (1992), pp. 315–334.

[9] B. Nagle, V. Rödl, and M. Schacht, *The counting lemma for regular k-uniform hypergraphs*, Random Structures Algorithms, 28 (2006), pp. 113–179.

[10] J. Pach and P. K. Agarwal, *Combinatorial Geometry*, John Wiley, New York, 1995.

[11] J. Pach and M. Sharir, *Geometric incidences*, in Towards a Theory of Geometric Graphs, Contemp. Math. 342, AMS, Providence, RI, 2004, pp. 185–223.

[12] V. Rödl and J. Skokan, *Regularity lemma for k-uniform hypergraphs*, Random Structures Algorithms, 25 (2004), pp. 1–42.

[13] V. Rödl and J. Skokan, *Applications of the regularity lemma for uniform hypergraphs*, Random Structures Algorithms, 28 (2006), pp. 180–194.

[14] K. F. Roth, *On certain sets of integers*, J. London Math. Soc., 28 (1953), pp. 245–252.

[15] I. Ruzsa and E. Szemerédi, *Triple systems with no six points carrying three triangles*, Colloq. Math. Soc. Janos Bolyai, 18 (1978), pp. 939–945.

[16] J. Solymosi, *Note on a generalization of Roth's theorem*, in Discrete and Computational Geometry, Algorithms Combin. 25, Springer-Verlag, Berlin, 2003, pp. 825–827.
[17] J. Solymosi, *A note on a question of Erdős and Graham*, Combin. Probab. Comput., 13 (2004), pp. 263–267.
[18] J. Solymosi and Cs. D. Tóth, *Distinct distances in the plane*, Discrete Comput. Geom., 25 (2001), pp. 629–634.
[19] E. Szemerédi, *On sets of integers containing no four elements in arithmetic progression*, Acta Math. Acad. Sci. Hungar., 20 (1969), pp. 89–104.
[20] E. Szemerédi, *Regular partitions of graphs*, in Problèmes Combinatoires et Théorie des Graphes, Proc. Colloque Inter. CNRS, J.-C. Bermond, J.-C. Fournier, M. Las Vergnas, and D. Sotteau, eds., CNRS, Paris, 1978, pp. 399–401.
[21] E. Szemerédi and W. T. Trotter, Jr., *Extremal problems in discrete geometry*, Combinatorica, 3 (1983), pp. 381–392.
[22] L. A. Székely, *Crossing numbers and hard Erdős problems in discrete geometry*, Combin. Probab. Comput., 6 (1997), pp. 353–358.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 628–648

COMPARING PARTIAL RANKINGS*

RONALD FAGIN†, RAVI KUMAR‡, MOHAMMAD MAHDIAN§, D. SIVAKUMAR¶, AND ERIK VEE†

Abstract. We provide a comprehensive picture of how to compare *partial rankings*, that is, rankings that allow ties. We propose several metrics to compare partial rankings and prove that they are within constant multiples of each other.

Key words. partial ranking, bucket order, permutation, metric

AMS subject classifications. 06A06, 68R99

DOI. 10.1137/05063088X

1. Introduction. The study of metrics on permutations (i.e., full rankings) is classical and several well-studied metrics are known [10, 22], including the Kendall tau distance and the Spearman footrule distance. The rankings encountered in practice, however, often have ties (hence the name *partial rankings*), and metrics on such rankings are much less studied.

Aside from its purely mathematical interest, the problem of defining metrics on partial rankings is valuable in a number of applications. For example the *rank aggregation* problem for partial rankings arises naturally in multiple settings, including in online commerce, where users state their preferences for products according to various criteria, and the system ranks the products in a single, cohesive way that incorporates all the stated preferences, and returns the top few items to the user. Specific instances include the following: selecting a restaurant from a database of restaurants (where the ranking criteria include culinary preference, driving distance, star rating, etc.), selecting an air-travel plan (where the ranking criteria include price, airline preference, number of hops, etc.), and searching for articles in a scientific bibliography (where the articles may be ranked by relevance of subject, year, number of citations, etc.). In all of these scenarios, it is easy to see that many of the ranking criteria lead to ties among the underlying set of items. To formulate a mathematically sound aggregation problem for such partially ranked lists (as has been done successfully for fully ranked lists [12] and "top k lists" [16]), it is sometimes necessary to have a well-defined distance measure (preferably a metric) between partial rankings.

In this paper we focus on four metrics between partial rankings. These are obtained by suitably generalizing the Kendall tau distance and the Spearman footrule distance on permutations in two different ways. In the first approach, we associate

*Received by the editors May 6, 2005; accepted for publication (in revised form) February 7, 2006; published electronically September 5, 2006. This paper is an expansion of a portion of the paper [14].

http://www.siam.org/journals/sidma/20-3/63088.html

†IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120 (fagin@almaden.ibm.com, vee@almaden.ibm.com).

‡Yahoo! Research, 701 First Ave., Sunnyvale, CA 94089 (ravikumar@yahoo-inc.com). This author's work was done at the IBM Almaden Research Center.

§Microsoft Research, Redmond, WA 98052 (mahdian@microsoft.com). Part of this author's work was supported by NSF grant CCR-0098066. Part of this work was done while the author was visiting the IBM Almaden Research Center.

¶Google, Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 (siva@google.com). This author's work was done at the IBM Almaden Research Center.

with each partial ranking a "profile vector" and we define the distance between the partial rankings to be the L_1 distance between the corresponding profile vectors. In the second approach, we associate with each partial ranking the family of all full rankings that are obtained by breaking ties in all possible ways. The distance between partial rankings is then taken to be the Hausdorff distance between the corresponding sets of full rankings.[1] In addition to the four metrics we obtain by extending the Kendall tau distance and the Spearman footrule distance using these two approaches, we consider a method obtained by generalizing the Kendall tau distance where we vary a certain parameter. For some choices of the parameter, we obtain a metric, and for one natural choice, we obtain our Kendall profile metric. All the metrics we define admit efficient computation. These metrics are defined and discussed in section 3.

Having various metrics on partial rankings is good news, but exactly which one should a practitioner use to compare partial rankings? Furthermore, which one is best suited for formulating an aggregation problem for partial rankings? Our summary answer to these questions is that the exact choice does not matter much. Namely, following the lead of [16], we define two metrics to be *equivalent* if they are within constant multiples of each other. This notion was inspired by the Diaconis–Graham inequality [11], which says that the Kendall tau distance and the Spearman footrule distance are within a factor of two of each other. Our main theorem says that all of our metrics are equivalent in this sense. The methods where we generalize the Kendall tau distance by varying a certain parameter are easily shown to be equivalent to each other, and in particular to the profile version of the Kendall tau distance (since one choice of the parameter leads to the profile version). It is also simple to show that the Hausdorff versions of the Kendall tau distance and the Spearman footrule distance are equivalent and that the Hausdorff and the profile versions of the Kendall tau metric are equivalent. Proving equivalence for the profile metrics turns out to be rather tricky and requires us to uncover considerable structure inside partial rankings. We present these equivalence results in section 4.

Related work. The Hausdorff versions of the Kendall tau distance and the Spearman footrule distance are due to Critchlow [9]. Fagin, Kumar, and Sivakumar [16] studied a variation of these for top k lists. Kendall [23] defined two versions of the Kendall tau distance for partial rankings; one of these versions is a normalized version of our Kendall tau distance through profiles. Baggerly [5] defined two versions of the Spearman footrule distance for partial rankings; one of these versions is similar to our Spearman footrule metric through profiles. However, neither Kendall nor Baggerly proceeded significantly beyond simply providing the definition. Goodman and Kruskal [20] proposed an approach for comparing partial rankings, which was recently utilized [21] for evaluating strategies for similarity search on the Web. A serious disadvantage of Goodman and Kruskal's approach is that it is not always defined (this problem did not arise in the application of [21]).

Rank aggregation and partial rankings. As alluded to earlier, rank aggregation is the problem of combining several ranked lists of objects in a robust way to produce a single consensus ranking of the objects. In computer science, rank aggregation has proved to be a useful and powerful paradigm in several applications including metasearch [4, 12, 24, 25, 26, 29], combining experts [8], synthesizing rank functions from multiple indices [15], biological databases [28], similarity search [17], and classification [17, 24].

[1] The Hausdorff distance between two point sets A and B in a metric space with metric $d(\cdot,\cdot)$ is defined as $\max\{\max_{\gamma_1 \in A} \min_{\gamma_2 \in B} d(\gamma_1, \gamma_2), \max_{\gamma_2 \in B} \min_{\gamma_1 \in A} d(\gamma_1, \gamma_2)\}$.

There has been an extensive body of work in economics and computer science on providing a mathematical basis for aggregation of rankings. In the "axiomatic approach," one formulates a set of desiderata that the aggregation function is supposed to satisfy, and characterizes various aggregation functions in terms of the "axioms" they satisfy. The classical result of Arrow [2] shows that a small set of fairly natural requirements cannot be simultaneously achieved by any nontrivial aggregation function. For a comprehensive account of specific criteria satisfied by various aggregation methods, see the survey by Fishburn [18]. In the "metric approach," one starts with a metric on the underlying set of rankings (such as permutations or top k lists) and defines the aggregation problem as that of finding a consensus ranking (permutation or top k list, respectively) whose total distance to the given rankings is minimized. It is, of course, natural to study which axioms a given metric method satisfies, and indeed several such results are known (again, see Fishburn's survey [18]).

A prime consideration in the adoption of a metric aggregation method in computer science applications is whether it admits an efficient exact or provably approximate solution. Several metric methods with excellent properties (e.g., aggregating full lists with respect to the Kendall tau distance) turn out to be NP-hard to solve exactly [6, 12]; fortunately, results like the Diaconis–Graham inequality rescue us from this despair, since if two metrics are equivalent and one of them admits an efficient algorithm, we automatically obtain an efficient approximation algorithm for the other! This is one of the main reasons for our interest in obtaining equivalences between metrics.

While the work of [12, 16] and follow-up efforts offer a fairly clear picture on how to compare and aggregate full or top k lists, the context of database-centric applications poses a new, and rather formidable, challenge. As outlined earlier through the example of online commerce systems, as a result of nonnumeric/few-valued attributes, we encounter partial rankings much more than full rankings in some contexts. While it is possible to treat this issue heuristically by arbitrarily ordering the tied elements to produce a full ranking, a mathematically well-founded treatment becomes possible once we are equipped with metrics on partial rankings. By the equivalence outlined above, it follows that every constant-factor approximation algorithm for rank aggregation with respect to one of our metrics automatically yields a constant-factor approximation algorithm with respect to all of our metrics. These facts were crucially used in [14] to obtain approximation algorithms for the problem of aggregating partial rankings.

2. Preliminaries. *Bucket orders.* A *bucket order* is, intuitively, a (strict) linear order with ties. More formally, a bucket order is a transitive binary relation $\prec$ for which there are sets $\mathcal{B}_1, \ldots, \mathcal{B}_t$ (the *buckets*) that form a partition of the domain such that $x \prec y$ if and only if there are i, j with $i < j$ such $x \in \mathcal{B}_i$ and $y \in \mathcal{B}_j$. If $x \in \mathcal{B}_i$, we may refer to $\mathcal{B}_i$ as the *bucket of* x. We may say that bucket $\mathcal{B}_i$ *precedes* bucket $\mathcal{B}_j$ if $i < j$. Thus, $x \prec y$ if and only if the bucket of x precedes the bucket of y. We think of the members of a given bucket as "tied." A linear order is a bucket order where every bucket is of size 1. We now define the *position* of bucket $\mathcal{B}$, denoted $pos(\mathcal{B})$. Let $\mathcal{B}_1, \ldots, \mathcal{B}_t$ be the buckets in order (so that bucket $\mathcal{B}_i$ precedes bucket $\mathcal{B}_j$ when $i < j$). Then $pos(\mathcal{B}_i) = (\sum_{j<i} |\mathcal{B}_j|) + (|\mathcal{B}_i| + 1)/2$. Intuitively, $pos(\mathcal{B}_i)$ is the average location within bucket $\mathcal{B}_i$.

Comment on terminology.[2] A bucket order $\prec$ is *irreflexive*, that is, there is no x for which $x \prec x$ holds. The corresponding reflexive version $\preceq$ is defined by

[2]The authors are grateful to Bernard Monjardet for providing the information in this paragraph.

saying $x \preceq y$ precisely if either $x \prec y$ or $x = y$. What we call a bucket order is sometimes called a "weak order" (or "weak ordering") [1, 19]. But unfortunately, the corresponding reflexive version $\preceq$ is also sometimes called a weak order (or weak ordering) [2, 13, 27]. A bucket order is sometimes called a "strict weak order" (or "strict weak ordering") [7, 27]. The reflexive version is sometimes called a "complete preordering" [3] or a "total preorder" [7]. We are using the terminology bucket order because it is suggestive and unambiguous.

Partial ranking. Just as we can associate a ranking with a linear order (i.e., permutation), we associate a *partial ranking* $\boldsymbol{\sigma}$ with each bucket order, by letting $\boldsymbol{\sigma}(x) = pos(\mathcal{B})$ when $\mathcal{B}$ is the bucket of x. We refer to a partial ranking associated with a linear order as a *full ranking*. When it is not otherwise specified, we assume that all partial rankings have the same domain, denoted D. We say that x *is ahead of* y *in* $\boldsymbol{\sigma}$ if $\boldsymbol{\sigma}(x) < \boldsymbol{\sigma}(y)$. We say that x *and* y *are tied in* $\boldsymbol{\sigma}$ if $\boldsymbol{\sigma}(x) = \boldsymbol{\sigma}(y)$. When we speak of the buckets of a partial ranking, we are referring to the buckets of the corresponding bucket order.

We define a *top* k *list* to be a partial ranking where the top k buckets are singletons, representing the top k elements, and the bottom bucket contains all other members of the domain. Note that in [16] there is no bottom bucket in a top k list. This is because in [16] each top k list has its own domain of size k, unlike our scenario where there is a fixed domain.

Given a partial ranking $\boldsymbol{\sigma}$ with domain D, we define its *reverse*, denoted $\boldsymbol{\sigma}^{\mathrm{R}}$, in the expected way. That is, for all $d \in D$, let $\boldsymbol{\sigma}^{\mathrm{R}}(d) = |D| + 1 - \boldsymbol{\sigma}(d)$.

We also define the notion of *swapping* in the normal way. If $a, b \in D$, then *swapping* a *and* b *in* $\boldsymbol{\sigma}$ produces a new order $\boldsymbol{\sigma}'$, where $\boldsymbol{\sigma}'(a) = \boldsymbol{\sigma}(b)$, $\boldsymbol{\sigma}'(b) = \boldsymbol{\sigma}(a)$, and $\boldsymbol{\sigma}'(d) = \boldsymbol{\sigma}(d)$ for all $d \in D \setminus \{a, b\}$.

Refinements of partial rankings. Given two partial rankings $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$, both with domain D, we say that $\boldsymbol{\sigma}$ is a *refinement* of $\boldsymbol{\tau}$ and write $\boldsymbol{\sigma} \succeq \boldsymbol{\tau}$ if the following holds: for all $i, j \in D$, we have $\boldsymbol{\sigma}(i) < \boldsymbol{\sigma}(j)$ whenever $\boldsymbol{\tau}(i) < \boldsymbol{\tau}(j)$. Notice that when $\boldsymbol{\tau}(i) = \boldsymbol{\tau}(j)$, there is no order forced on $\boldsymbol{\sigma}$. When $\boldsymbol{\sigma}$ is a full ranking, we say that $\boldsymbol{\sigma}$ is a *full refinement* of $\boldsymbol{\tau}$. Given two partial rankings $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$, both with domain D, we frequently make use of a particular refinement of $\boldsymbol{\sigma}$ in which ties are broken according to $\boldsymbol{\tau}$. Define the $\boldsymbol{\tau}$*-refinement of* $\boldsymbol{\sigma}$, denoted $\boldsymbol{\tau} * \boldsymbol{\sigma}$, to be the refinement of $\boldsymbol{\sigma}$ with the following properties. For all $i, j \in D$, if $\boldsymbol{\sigma}(i) = \boldsymbol{\sigma}(j)$ and $\boldsymbol{\tau}(i) < \boldsymbol{\tau}(j)$, then $(\boldsymbol{\tau} * \boldsymbol{\sigma})(i) < (\boldsymbol{\tau} * \boldsymbol{\sigma})(j)$. If $\boldsymbol{\sigma}(i) = \boldsymbol{\sigma}(j)$ and $\boldsymbol{\tau}(i) = \boldsymbol{\tau}(j)$, then $(\boldsymbol{\tau} * \boldsymbol{\sigma})(i) = (\boldsymbol{\tau} * \boldsymbol{\sigma})(j)$. Notice that when $\boldsymbol{\tau}$ is in fact a full ranking, then $\boldsymbol{\tau} * \boldsymbol{\sigma}$ is also a full ranking. Also note that $*$ is an associative operation, so that if $\boldsymbol{\rho}$ is another partial ranking with domain D, it makes sense to talk about $\boldsymbol{\rho} * \boldsymbol{\tau} * \boldsymbol{\sigma}$.

Notation. When f and g are functions with the same domain D, we denote the L_1 distance between f and g by $L_1(f, g)$. Thus, $L_1(f, g) = \sum_{i \in D} |f(i) - g(i)|$.

2.1. Metrics, near metrics, and equivalence. A binary function d is called *symmetric* if $d(x, y) = d(y, x)$ for all x, y in the domain, and it is called *regular* if $d(x, y) = 0$ if and only if $x = y$. A *distance measure* is a nonnegative, symmetric, regular binary function. A *metric* is a distance measure d that satisfies the *triangle inequality*: $d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, z in the domain.

The definitions and results in this section were derived in [16], in the context of comparing top k lists. Two seemingly different notions of a "near metric" were defined in [16]: their first notion of near metric is based on "relaxing" the polygonal inequality that a metric is supposed to satisfy.

DEFINITION 1 (near metric). *A distance measure on partial rankings with domain D is a* near metric *if there is a constant c, independent of the size of D, such that the distance measure satisfies the relaxed polygonal inequality: $d(x,z) \le c(d(x,x_1) + d(x_1,x_2) + \cdots + d(x_{n-1},z))$ for all $n > 1$ and $x, z, x_1, \ldots, x_{n-1} \in D$.*

It makes sense to say that the constant c is independent of the size of D when, as in [16], each of the distance measures considered is actually a family, parameterized by D. We need to make an assumption that c is independent of the size of D, since otherwise we are simply considering distance measures over finite domains, where there is always such a constant c.

The other notion of near metric given in [16] is based on bounding the distance measure above and below by positive constant multiples of a metric. It was shown that both the notions of near metrics coincide.[3] This theorem inspired a definition of what it means for a distance measure to be "almost" a metric, and a robust notion of "similar" or "equivalent" distance measures. We modify the definitions in [16] slightly to fit our scenario, where there is a fixed domain D.

DEFINITION 2 (equivalent distance measures). *Two distance measures d and d' between partial rankings with domain D are* equivalent *if there are positive constants c_1 and c_2, independent of the size of D, such that $c_1 d'(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \le d(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \le c_2 d'(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ for every pair $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2$ of partial rankings.*

It is clear that the above definition leads to an equivalence relation (i.e., reflexive, symmetric, and transitive). It follows from [16] that a distance measure is equivalent to a metric if and only if it is a near metric.

2.2. Metrics on full rankings. We now review two well-known notions of metrics on full rankings, namely the Kendall tau distance and the Spearman footrule distance.

Let $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2$ be two full rankings with domain D. The *Spearman footrule distance* is simply the L_1 distance $L_1(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$. The definition of the Kendall tau distance requires a little more work.

Let $\mathcal{P} = \{\{i,j\} \mid i \ne j \text{ and } i,j \in D\}$ be the set of unordered pairs of distinct elements. The *Kendall tau distance* between full rankings is defined as follows. For each pair $\{i,j\} \in \mathcal{P}$ of distinct members of D, if i and j are in the same order in $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$, then let the penalty $\bar{K}_{i,j}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = 0$; and if i and j are in the opposite order (such as i being ahead of j in $\boldsymbol{\sigma}_1$ and j being ahead of i in $\boldsymbol{\sigma}_2$), then let $\bar{K}_{i,j}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = 1$. The Kendall tau distance is given by $K(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = \sum_{\{i,j\} \in \mathcal{P}} \bar{K}_{i,j}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$. The Kendall tau distance turns out to be equal to the number of exchanges needed in a bubble sort to convert one full ranking to the other.

Diaconis and Graham [11] proved a classical result, which states that for every two full rankings σ_1, σ_2,

$$K(\sigma_1, \sigma_2) \le F(\sigma_1, \sigma_2) \le 2K(\sigma_1, \sigma_2). \tag{1}$$

Thus, the Kendall tau distance and the Spearman footrule distance are equivalent metrics for full rankings.

3. Metrics for comparing partial rankings. In this section we define metrics on partial rankings. The first set of metrics is based on profile vectors (section 3.1). As part of this development, we consider variations of the Kendall tau distance where

[3] This result would not hold if instead of relaxing the polygonal inequality, we simply relaxed the triangle inequality.

we vary a certain parameter. The second set of metrics is based on the Hausdorff distance (section 3.2). Section 3.3 compares these metrics (when the partial rankings are top k lists) with the distance measures for top k lists that are developed in [16].

3.1. Metrics based on profiles. Let $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2$ be two partial rankings with domain D. We now define a family of generalizations of the Kendall tau distance to partial rankings. These are based on a generalization [16] of the Kendall tau distance to top k lists.

Let p be a fixed parameter, with $0 \le p \le 1$. Similar to our definition of $\bar{K}_{i,j}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ for full rankings $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2$, we define a penalty $\bar{K}_{i,j}^{(p)}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ for partial rankings $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2$ for $\{i, j\} \in \mathcal{P}$. There are three cases.

Case 1. i and j are in different buckets in both $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$. If i and j are in the same order in $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ (such as $\boldsymbol{\sigma}_1(i) > \boldsymbol{\sigma}_1(j)$ and $\boldsymbol{\sigma}_2(i) > \boldsymbol{\sigma}_2(j)$), then let $\bar{K}_{i,j}^{(p)}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = 0$; this corresponds to "no penalty" for $\{i, j\}$. If i and j are in the opposite order in $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ (such as $\boldsymbol{\sigma}_1(i) > \boldsymbol{\sigma}_1(j)$ and $\boldsymbol{\sigma}_2(i) < \boldsymbol{\sigma}_2(j)$), then let the penalty $\bar{K}_{i,j}^{(p)}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = 1$.

Case 2. i and j are in the same bucket in both $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$. We then let the penalty $\bar{K}_{i,j}^{(p)}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = 0$. Intuitively, both partial rankings agree that i and j are tied.

Case 3. i and j are in the same bucket in one of the partial rankings $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$, but in different buckets in the other partial ranking. In this case, we let the penalty $\bar{K}_{i,j}^{(p)}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = p$.

Based on these cases, define $K^{(p)}$, the *Kendall distance with penalty parameter* p, as follows:

$$K^{(p)}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = \sum_{\{i,j\} \in \mathcal{P}} \bar{K}_{i,j}^{(p)}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2).$$

We now discuss our choice of penalty in Cases 2 and 3. In Case 2, where i and j are in the same bucket in both $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$, what if we had defined there to be a positive penalty $\bar{K}_{i,j}^{(p)}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = q > 0$? Then if $\boldsymbol{\sigma}$ were an arbitrary partial ranking that had some bucket of size at least 2, we would have had $K^{(p)}(\boldsymbol{\sigma}, \boldsymbol{\sigma}) \ge q > 0$. So $K^{(p)}$ would not have been a metric, or even a distance measure, since we would have lost the property that $K^{(p)}(\boldsymbol{\sigma}, \boldsymbol{\sigma}) = 0$. The next proposition shows the effect of the choice of p in Case 3.

PROPOSITION 3. *$K^{(p)}$ is a metric when $1/2 \le p \le 1$, is a near metric when $0 < p < 1/2$, and is not a distance measure when $p = 0$.*

Proof. Let us first consider the case $p = 0$. We now show that $K^{(0)}$ is not even a distance measure. Let the domain D have exactly two elements a and b. Let $\boldsymbol{\tau}_1$ be the full ranking where a precedes b, let $\boldsymbol{\tau}_2$ be the partial ranking where a and b are in the same bucket, and let $\boldsymbol{\tau}_3$ be the full ranking where b precedes a. Then $K^{(0)}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2) = 0$ even though $\boldsymbol{\tau}_1 \ne \boldsymbol{\tau}_2$. So indeed, $K^{(0)}$ is not a distance measure. Note also that the near triangle inequality is violated badly in this example, since $K^{(0)}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2) = 0$ and $K^{(0)}(\boldsymbol{\tau}_2, \boldsymbol{\tau}_3) = 0$, but $K^{(0)}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_3) = 1$.

It is easy to see that $K^{(p)}$ is a distance measure for every p with $0 < p \le 1$. We now show that $K^{(p)}$ does not satisfy the triangle inequality when $0 < p < 1/2$ and satisfies the triangle inequality when $1/2 \le p \le 1$. Let $\boldsymbol{\tau}_1$, $\boldsymbol{\tau}_2$, and $\boldsymbol{\tau}_3$ be as in our previous example. Then $K^{(p)}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2) = p$, $K^{(p)}(\boldsymbol{\tau}_2, \boldsymbol{\tau}_3) = p$, and $K^{(p)}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_3) = 1$. So the triangle inequality fails for $0 < p < 1/2$, since $K^{(p)}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_3) > K^{(p)}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2) + K^{(p)}(\boldsymbol{\tau}_2, \boldsymbol{\tau}_3)$. On the other hand, the triangle inequality holds for $1/2 \le p \le 1$, since

then it is easy to verify that $\bar{K}^{(p)}_{i,j}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_3) \leq \bar{K}^{(p)}_{i,j}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) + \bar{K}^{(p)}_{i,j}(\boldsymbol{\sigma}_2, \boldsymbol{\sigma}_3)$ for every i, j, and so $K^{(p)}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_3) \leq K^{(p)}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) + K^{(p)}(\boldsymbol{\sigma}_2, \boldsymbol{\sigma}_3)$.

We now show that $K^{(p)}$ is a near metric when $0 < p < 1/2$. It is easy to verify that if $0 < p < p' \leq 1$, then $K^{(p)}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \leq K^{(p')}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \leq (p'/p)K^{(p)}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$. Hence, all of the distance measures $K^{(p)}$ are equivalent whenever $0 < p$. As noted earlier, it follows from [16] that a distance measure is equivalent to a metric if and only if it is a near metric. Since $K^{(p)}$ is equivalent to the metric $K^{(1/2)}$ when $0 < p$, we conclude that in this case, $K^{(p)}$ is a near metric. □

It is worth stating formally the following simple observation from the previous proof.

PROPOSITION 4. *All of the distance measures $K^{(p)}$ are equivalent whenever $0 < p \leq 1$.*

For the rest of the paper, we focus on the natural case $p = 1/2$, which corresponds to an "average" penalty for two elements i and j that are tied in one partial ranking but not in the other partial ranking. We show that $K^{(1/2)}$ is equivalent to the other metrics we define. It thereby follows from Proposition 4 that each of the distance measures $K^{(p)}$ for $0 < p \leq 1$, and in particular the metrics $K^{(p)}$ for $1/2 \leq p \leq 1$, is equivalent to these other metrics.

We now show there is an alternative interpretation for $K^{(1/2)}$ in terms of a "profile." Let $\mathcal{O} = \{(i, j) : \ i \neq j \text{ and } i, j \in D\}$ be the set of ordered pairs of distinct elements in the domain D. Let $\boldsymbol{\sigma}$ be a partial ranking (as usual, with domain D). For $(i, j) \in \mathcal{O}$, define p_{ij} to be $1/4$ if $\boldsymbol{\sigma}(i) < \boldsymbol{\sigma}(j)$, to be 0 if $\boldsymbol{\sigma}(i) = \boldsymbol{\sigma}(j)$, and to be $-1/4$ if $\boldsymbol{\sigma}(i) > \boldsymbol{\sigma}(j)$. Define the *$K$-profile* of $\boldsymbol{\sigma}$ to be the vector $\langle p_{ij} : (i, j) \in \mathcal{O} \rangle$ and $K_{\text{prof}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ to be the L_1 distance between the K-profiles of $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$. It is easy to verify that $K_{\text{prof}} = K^{(1/2)}$.[4] It is also easy to see that the K-profile of $\boldsymbol{\sigma}$ uniquely determines $\boldsymbol{\sigma}$.

It is clear how to generalize the Spearman footrule distance to partial rankings—we simply take it to be $L_1(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$, just as before. We refer to this value as $F_{\text{prof}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$, for reasons we now explain. Let us define the *F-profile* of a partial ranking $\boldsymbol{\sigma}$ to be the vector of values $\boldsymbol{\sigma}(i)$. So the F-profile is indexed by D, whereas the K-profile is indexed by $\mathcal{O}$. Just as the K_{prof} value of two partial rankings (or of the corresponding bucket orders) is the L_1 distance between their K-profiles, the F_{prof} value of two partial rankings (or of the corresponding bucket orders) is the L_1 distance between their F-profiles. Since K_{prof} and F_{prof} are L_1 distances, and since the K-profile and the F-profile each uniquely determine the partial ranking, it follows that K_{prof} and F_{prof} are both metrics.

3.2. The Hausdorff metrics. Let A and B be finite sets of objects and let d be a metric on objects. The *Hausdorff distance* between A and B is given by

$$d_{\text{Haus}}(A, B) = \max \left\{ \max_{\gamma_1 \in A} \min_{\gamma_2 \in B} d(\gamma_1, \gamma_2), \max_{\gamma_2 \in B} \min_{\gamma_1 \in A} d(\gamma_1, \gamma_2) \right\}. \tag{2}$$

Although this looks fairly nonintuitive, it is actually quite natural, as we now explain. The quantity $\min_{\gamma_2 \in B} d(\gamma_1, \gamma_2)$ is the distance between γ_1 and the set B. Therefore, the quantity $\max_{\gamma_1 \in A} \min_{\gamma_2 \in B} d(\gamma_1, \gamma_2)$ is the maximal distance of a member of A from the set B. Similarly, the quantity $\max_{\gamma_2 \in B} \min_{\gamma_1 \in A} d(\gamma_1, \gamma_2)$ is the

[4] The reason that the values of p_{ij} in the K-profile are $1/4$, 0, and $-1/4$ rather than $1/2$, 0, and $-1/2$ is that each pair $\{i, j\}$ with $i \neq j$ is counted twice, once as (i, j) and once as (j, i).

maximal distance of a member of B from the set A. Therefore, the Hausdorff distance between A and B is the maximal distance of a member of A or B from the other set. Thus, A and B are within Hausdorff distance s of each other precisely if every member of A and B is within distance s of some member of the other set. The Hausdorff distance is well known to be a metric.

Critchlow [9] used the Hausdorff distance to define a metric, which we now define, between partial rankings. Given a metric d that gives the distance $d(\gamma_1, \gamma_2)$ between full rankings γ_1 and γ_2, define the distance d_{Haus} between partial rankings $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ to be

$$d_{\mathrm{Haus}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = \max\left\{ \max_{\gamma_1 \succeq \boldsymbol{\sigma}_1} \min_{\gamma_2 \succeq \boldsymbol{\sigma}_2} d(\gamma_1, \gamma_2), \max_{\gamma_2 \succeq \boldsymbol{\sigma}_2} \min_{\gamma_1 \succeq \boldsymbol{\sigma}_1} d(\gamma_1, \gamma_2) \right\}, \tag{3}$$

where γ_1 and γ_2 are full rankings. In particular, when d is the footrule distance, this gives us a metric between partial rankings that we call F_{Haus}, and when d is the Kendall distance, this gives us a metric between partial rankings that we call K_{Haus}. Both F_{Haus} and K_{Haus} are indeed metrics, since they are special cases of the Hausdorff distance.

The next theorem, which is due to Critchlow (but which we state using our notation), gives a complete characterization of F_{Haus} and K_{Haus}. For the sake of completeness, we prove this theorem in the appendix.[5]

THEOREM 5 (see [9]). *Let $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ be partial rankings, let $\boldsymbol{\sigma}^{\mathrm{R}}$ be the reverse of $\boldsymbol{\sigma}$, and let $\boldsymbol{\tau}^{\mathrm{R}}$ be the reverse of $\boldsymbol{\tau}$. Let ρ be any full ranking. Then*

$$\begin{aligned} F_{\mathrm{Haus}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \max\{&F(\rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}, \rho * \boldsymbol{\sigma} * \boldsymbol{\tau}), \\ &F(\rho * \boldsymbol{\tau} * \boldsymbol{\sigma}, \rho * \boldsymbol{\sigma}^{\mathrm{R}} * \boldsymbol{\tau})\}, \\ K_{\mathrm{Haus}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \max\{&K(\rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}, \rho * \boldsymbol{\sigma} * \boldsymbol{\tau}), \\ &K(\rho * \boldsymbol{\tau} * \boldsymbol{\sigma}, \rho * \boldsymbol{\sigma}^{\mathrm{R}} * \boldsymbol{\tau})\}. \end{aligned}$$

Theorem 5 gives us a simple algorithm for computing $F_{\mathrm{Haus}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$ and $K_{\mathrm{Haus}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$: we simply pick an arbitrary full ranking ρ and do the computations given in Theorem 5. Thus, let $\sigma_1 = \rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}$, let $\tau_1 = \rho * \boldsymbol{\sigma} * \boldsymbol{\tau}$, let $\sigma_2 = \rho * \boldsymbol{\tau} * \boldsymbol{\sigma}$, and let $\tau_2 = \rho * \boldsymbol{\sigma}^{\mathrm{R}} * \boldsymbol{\tau}$. Theorem 5 tells us that $F_{\mathrm{Haus}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \max\{F(\sigma_1, \tau_1), F(\sigma_2, \tau_2)\}$ and $K_{\mathrm{Haus}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \max\{K(\sigma_1, \tau_1), K(\sigma_2, \tau_2)\}$. It is interesting that the same pairs, namely (σ_1, τ_1) and (σ_2, τ_2), are the candidates for exhibiting the Hausdorff distance for both F and K. Note that the only role that the arbitrary full ranking ρ plays is to arbitrarily break ties (in the same way for $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$) for pairs (i, j) of distinct elements that are in the same bucket in both $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$. A way to describe the pair (σ_1, τ_1) intuitively is as follows: break the ties in $\boldsymbol{\sigma}$ based on the reverse of the ordering in $\boldsymbol{\tau}$, break the ties in $\boldsymbol{\tau}$ based on the ordering in $\boldsymbol{\sigma}$, and break any remaining ties arbitrarily (but in the same way in both). A similar description applies to the pair (σ_2, τ_2).

The algorithm just described for computing $F_{\mathrm{Haus}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$ and $K_{\mathrm{Haus}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$ is based on creating pairs (σ_1, τ_1) and (σ_2, τ_2), one of which must exhibit the Hausdorff distance. The next theorem gives a direct algorithm for computing $K_{\mathrm{Haus}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$ that we make use of later.

THEOREM 6. *Let $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ be partial rankings. Let S be the set of pairs $\{i, j\}$ of distinct elements such that i and j appear in the same bucket of $\boldsymbol{\sigma}$ but in different*

[5] Our proof arose when, unaware of Critchlow's result, we derived and proved this theorem.

buckets of $\boldsymbol{\tau}$, let T be the set of pairs $\{i,j\}$ of distinct elements such that i and j appear in the same bucket of $\boldsymbol{\tau}$ but in different buckets of $\boldsymbol{\sigma}$, and let U be the set of pairs $\{i,j\}$ of distinct elements that are in different buckets of both $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ and are in a different order in $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$. Then $K_{\mathrm{Haus}}(\boldsymbol{\sigma},\boldsymbol{\tau}) = |U| + \max\{|S|,|T|\}$.

Proof. As before, let $\sigma_1 = \rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}$, let $\tau_1 = \rho * \boldsymbol{\sigma} * \boldsymbol{\tau}$, let $\sigma_2 = \rho * \boldsymbol{\tau} * \boldsymbol{\sigma}$, and let $\tau_2 = \rho * \boldsymbol{\sigma}^{\mathrm{R}} * \boldsymbol{\tau}$. It is straightforward to see that the set of pairs $\{i,j\}$ of distinct elements that are in a different order in σ_1 and τ_1 is exactly the union of the disjoint sets U and S. Therefore, $K(\sigma_1,\tau_1) = |U| + |S|$. Identically, the set of pairs $\{i,j\}$ of distinct elements that are in a different order in σ_2 and τ_2 is exactly the union of the disjoint sets U and T, and hence $K(\sigma_2,\tau_2) = |U| + |T|$. But by Theorem 5, we know that $K_{\mathrm{Haus}}(\boldsymbol{\sigma},\boldsymbol{\tau}) = \max\{K(\sigma_1,\tau_1), K(\sigma_2,\tau_2)\} = \max\{|U|+|S|, |U|+|T|\}$. The result follows immediately. ∎

3.3. Metrics in this paper for top k lists vs. distance measures defined in [10]. Metrics on partial rankings naturally induce metrics on top k lists. We now compare (a) the metrics on top k lists that are induced by our metrics on partial rankings with (b) the distance measures on top k lists that were introduced in [16]. Recall that for us, a top k list is a partial ranking consisting of k singleton buckets, followed by a bottom bucket of size $|D|-k$. However, in [16], a top k list is a bijection of a domain ("the top k elements") onto $\{1,\ldots,k\}$. Let $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ be top k lists (of our form). Define the *active domain for $\boldsymbol{\sigma}$, $\boldsymbol{\tau}$* to be the union of the elements in the top k buckets of $\boldsymbol{\sigma}$ and the elements in the top k buckets of $\boldsymbol{\tau}$. In order to make our scenario compatible with the scenario of [16], we assume during our comparison that the domain D equals the active domain for $\boldsymbol{\sigma}$, $\boldsymbol{\tau}$. Our definitions of $K^{(p)}$, F_{Haus}, and K_{Haus} are then exactly the same in the two scenarios. (Unlike the situation in section 3.1, even the case $p = 0$ gives a distance measure, since the unpleasant case where $K^{(0)}(\boldsymbol{\sigma}_1,\boldsymbol{\sigma}_2) = 0$ even though $\boldsymbol{\sigma}_1 \neq \boldsymbol{\sigma}_2$ does not arise for top k lists $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$.) Nevertheless, $K^{(p)}$, F_{Haus}, and K_{Haus} are only near metrics in [16] in spite of being metrics for us. This is because, in [16], the active domain varies depending on which pair of top k lists is being compared.

Our definition of $K_{\mathrm{prof}}(\boldsymbol{\sigma},\boldsymbol{\tau})$ is equivalent to the definition of $K_{\mathrm{avg}}(\boldsymbol{\sigma},\boldsymbol{\tau})$ in [16], namely the average value of $K(\sigma,\tau)$ over all full rankings σ, τ with domain D, where $\sigma \succeq \boldsymbol{\sigma}$ and $\tau \succeq \boldsymbol{\tau}$. It is interesting to note that if $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ were not top k lists but arbitrary partial rankings, then K_{avg} would not be a distance measure, since $K_{\mathrm{avg}}(\boldsymbol{\sigma},\boldsymbol{\sigma})$ can be strictly positive if $\boldsymbol{\sigma}$ is an arbitrary partial ranking.

Let ℓ be a real number greater than k. The *footrule distance with location parameter ℓ*, denoted $F^{(\ell)}$, is defined by treating each element that is not among the top k elements as if it were in position ℓ, and then taking the L_1 distance [16]. More formally, let $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ be top k lists (of our form). Define the function $f_{\boldsymbol{\sigma}}$ with domain D by letting $f_{\boldsymbol{\sigma}}(i) = \boldsymbol{\sigma}(i)$ if $1 \le \boldsymbol{\sigma}(i) \le k$, and $f_{\boldsymbol{\sigma}}(i) = \ell$ otherwise. Similarly, define the function $f_{\boldsymbol{\tau}}$ with domain D by letting $f_{\boldsymbol{\tau}}(i) = \boldsymbol{\tau}(i)$ if $1 \le \boldsymbol{\tau}(i) \le k$, and $f_{\boldsymbol{\tau}}(i) = \ell$ otherwise. Then $F^{(\ell)}(\boldsymbol{\sigma},\boldsymbol{\tau})$ is defined to be $L_1(f_{\boldsymbol{\sigma}}, f_{\boldsymbol{\tau}})$. It is straightforward to verify that $F_{\mathrm{prof}}(\boldsymbol{\sigma},\boldsymbol{\tau}) = F^{(\ell)}(\boldsymbol{\sigma},\boldsymbol{\tau})$ for $\ell = (|D|+k+1)/2$.

4. Equivalence between the metrics. In this section we prove our main theorem, which says that our four metrics are equivalent.

THEOREM 7. *The metrics F_{prof}, K_{prof}, F_{Haus}, and K_{Haus} are all equivalent, that is, within constant multiples of each other.*

Proof. First, we show

$$K_{\mathrm{Haus}}(\boldsymbol{\sigma}_1,\boldsymbol{\sigma}_2) \le F_{\mathrm{Haus}}(\boldsymbol{\sigma}_1,\boldsymbol{\sigma}_2) \le 2K_{\mathrm{Haus}}(\boldsymbol{\sigma}_1,\boldsymbol{\sigma}_2). \tag{4}$$

The proof of this equivalence between F_{Haus} and K_{Haus} uses the robustness of the Hausdorff definition with respect to equivalent metrics. It is fairly easy, and is given in section 4.1.

Next, we show

$$K_{\mathrm{prof}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \le F_{\mathrm{prof}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \le 2K_{\mathrm{prof}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2). \tag{5}$$

We note that (5) is much more complicated to prove than (4) and is also much more complicated to prove than the Diaconis–Graham inequality (1). The proof involves two main concepts: "reflecting" each partial ranking so that every element has a mirror image and using the notion of "nesting," which means that the interval spanned by an element and its image in one partial ranking sits inside the interval spanned by the same element and its image in the other partial ranking. The proof is presented in section 4.2.

We note that the equivalences given by (4) and (5) are interesting in their own right.

Finally, we show in section 4.3 that

$$K_{\mathrm{prof}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \le K_{\mathrm{Haus}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \le 2K_{\mathrm{prof}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2). \tag{6}$$

This is proved using Theorem 6.

Using (4), (5), and (6), the proof is complete, since (4) tells us that the two Hausdorff metrics are equivalent, (5) tells us that the two profile metrics are equivalent, and (6) tells us that some Hausdorff metric is equivalent to some profile metric. □

4.1. Equivalence of F_{Haus} and K_{Haus}. In this section, we prove the simple result that the Diaconis–Graham inequalities (1) extend to F_{Haus} and K_{Haus}. We begin with a lemma. In this lemma, for metric d, we define d_{Haus} as in (2), and similarly for metric d'.

LEMMA 8. *Assume that d and d' are metrics where there is a constant c such that $d \le c \cdot d'$. Then $d_{\mathrm{Haus}} \le c \cdot d'_{\mathrm{Haus}}$.*

Proof. Let A and B be as in (2). Assume without loss of generality (by reversing A and B if necessary) that $d_{\mathrm{Haus}}(A, B) = \max_{\gamma_1 \in A} \min_{\gamma_2 \in B} d(\gamma_1, \gamma_2)$. Find γ_1 in A that maximizes $\min_{\gamma_2 \in B} d(\gamma_1, \gamma_2)$, and γ_2 in B that minimizes $d(\gamma_1, \gamma_2)$. Therefore, $d_{\mathrm{Haus}}(A, B) = d(\gamma_1, \gamma_2)$. Find γ'_2 in B that minimizes $d'(\gamma_1, \gamma'_2)$. (There is such a γ'_2 since by assumption on the definition of Hausdorff distance, A and B are finite sets.) Then $d_{\mathrm{Haus}}(A, B) = d(\gamma_1, \gamma_2) \le d(\gamma_1, \gamma'_2)$, since γ_2 minimizes $d(\gamma_1, \gamma_2)$. Also $d(\gamma_1, \gamma'_2) \le c \cdot d'(\gamma_1, \gamma'_2)$, by assumption on d and d'. Finally $c \cdot d'(\gamma_1, \gamma'_2) \le c \cdot d'_{\mathrm{Haus}}(A, B)$, by definition of d'_{Haus} and the fact that γ'_2 minimizes $d'(\gamma_1, \gamma'_2)$. Putting these inequalities together, we obtain $d_{\mathrm{Haus}}(A, B) \le c \cdot d'_{\mathrm{Haus}}(A, B)$, which completes the proof. □

We can now show that the Diaconis–Graham inequalities (1) extend to F_{Haus} and K_{Haus}.

THEOREM 9. *Let $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ be partial rankings. Then $K_{\mathrm{Haus}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \le F_{\mathrm{Haus}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \le 2K_{\mathrm{Haus}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$.*

Proof. The first inequality $K_{\mathrm{Haus}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \le F_{\mathrm{Haus}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ follows from the first Diaconis–Graham inequality $K(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \le F(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ and Lemma 8, where we let the roles of d, d', and c be played by K, F, and 1, respectively. The second inequality $F_{\mathrm{Haus}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \le 2K_{\mathrm{Haus}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ follows from the second Diaconis–Graham inequality $F(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \le 2K(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ and Lemma 8, where we let the roles of d, d', and c be played by F, K, and 2, respectively. □

4.2. Equivalence of $F_{\mathbf{prof}}$ and $K_{\mathbf{prof}}$. In order to generalize the Diaconis–Graham inequalities to F_{prof} and K_{prof}, we convert a pair of partial rankings into full rankings (on an enlarged domain) in such a way that the F_{prof} distance between the partial rankings is precisely 4 times the F distance between the full rankings, and the K_{prof} distance between the partial rankings is precisely 4 times the K distance between the full rankings. Given a domain D, produce a "duplicate set" $D^\sharp = \{i^\sharp : i \in D\}$. Given a partial ranking $\boldsymbol{\sigma}$ with domain D, produce a new partial ranking $\boldsymbol{\sigma}^\sharp$, with domain $D \cup D^\sharp$, as follows. Modify the bucket order associated with $\boldsymbol{\sigma}$ by putting $i^\sharp$ in the same bucket as i for each $i \in D$. We thereby double the size of every bucket. Let $\boldsymbol{\sigma}^\sharp$ be the partial ranking associated with this new bucket order. Since $i^\sharp$ is in the same bucket as i, we have $\boldsymbol{\sigma}^\sharp(i) = \boldsymbol{\sigma}^\sharp(i^\sharp)$. We now show that $\boldsymbol{\sigma}^\sharp(i) = 2\boldsymbol{\sigma}(i) - 1/2$ for all i in D.

Fix i in D, let p be the number of elements j in D such that $\boldsymbol{\sigma}(j) < \boldsymbol{\sigma}(i)$, and let q be the number of elements k in D such that $\boldsymbol{\sigma}(k) = \boldsymbol{\sigma}(i)$. By the definition of the ranking associated with a bucket order, we have

$$\boldsymbol{\sigma}(i) = p + (q+1)/2. \tag{7}$$

Since each bucket doubles in size for the bucket order associated with $\boldsymbol{\sigma}^\sharp$, we similarly have

$$\boldsymbol{\sigma}^\sharp(i) = 2p + (2q+1)/2. \tag{8}$$

It follows easily from (7) and (8) that $\boldsymbol{\sigma}^\sharp(i) = 2\boldsymbol{\sigma}(i) - 1/2$, as desired.

We need to obtain a full ranking from the partial ranking $\boldsymbol{\sigma}^\sharp$. First, for every full ranking π with domain D, define a full ranking $\pi^\dagger$ with domain $D \cup D^\sharp$ as follows:

$$\begin{aligned} \pi^\dagger(d) &= \pi(d) \quad \text{for all } d \in D, \\ \pi^\dagger(d^\sharp) &= 2|D| + 1 - \pi(d) \quad \text{for all } d \text{ in } D \end{aligned}$$

so that $\pi^\dagger$ ranks elements of D in the same order as π, elements of $D^\sharp$ in the reverse order of π, and all elements of D before all elements of $D^\sharp$.

We define $\boldsymbol{\sigma}_\pi = \pi^\dagger * (\boldsymbol{\sigma}^\sharp)$. For instance, suppose $\mathcal{B}$ is a bucket of $\boldsymbol{\sigma}^\sharp$ containing the items $a, b, c, a^\sharp, b^\sharp, c^\sharp$, and suppose that π orders the items $\pi(a) < \pi(b) < \pi(c)$. Then $\boldsymbol{\sigma}_\pi$ will contain the sequence $a, b, c, c^\sharp, b^\sharp, a^\sharp$. Also notice that in this example, $\frac{1}{2}(\boldsymbol{\sigma}_\pi(a) + \boldsymbol{\sigma}_\pi(a^\sharp)) = \frac{1}{2}(\boldsymbol{\sigma}_\pi(b) + \boldsymbol{\sigma}_\pi(b^\sharp)) = \frac{1}{2}(\boldsymbol{\sigma}_\pi(c) + \boldsymbol{\sigma}_\pi(c^\sharp)) = pos(\mathcal{B})$. In fact, because of this "reflected-duplicate" property, we see that in general, for every $d \in D$,

$$\frac{1}{2}(\boldsymbol{\sigma}_\pi(d) + \boldsymbol{\sigma}_\pi(d^\sharp)) = \boldsymbol{\sigma}^\sharp(d) = \boldsymbol{\sigma}^\sharp(d^\sharp) = 2\boldsymbol{\sigma}(d) - 1/2. \tag{9}$$

The following lemma shows that no matter what order π we choose, the Kendall distance between $\boldsymbol{\sigma}_\pi$ and $\boldsymbol{\tau}_\pi$ is exactly 4 times the K_{prof} distance between $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$.

LEMMA 10. *Let $\boldsymbol{\sigma}, \boldsymbol{\tau}$ be partial rankings, and let π be any full ranking on the same domain. Then $K(\boldsymbol{\sigma}_\pi, \boldsymbol{\tau}_\pi) = 4K_{\mathrm{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$.*

Proof. Assume that i and j are in D. Let us consider the cases in the definition of $K^{(p)}$ (recall that K_{prof} equals $K^{(p)}$ when $p = 1/2$).

Case 1. i and j are in different buckets in both $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$. If i and j are in the same order in $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$, then the pair $\{i, j\}$ contributes no penalty to $K_{\mathrm{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$, and no pair of members of the set $\{i, j, i^\sharp, j^\sharp\}$ contribute any penalty to $K(\boldsymbol{\sigma}_\pi, \boldsymbol{\tau}_\pi)$. If i and j are in the opposite order in $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$, then the pair $\{i, j\}$ contributes a penalty of 1 to

$K_{\text{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$, and the pairs among $\{i, j, i^\sharp, j^\sharp\}$ that contribute a penalty to $K(\boldsymbol{\sigma}_\pi, \boldsymbol{\tau}_\pi)$ are precisely $\{i, j\}$, $\{i^\sharp, j^\sharp\}$, $\{i, j^\sharp\}$, and $\{i^\sharp, j\}$, each of which contributes a penalty of 1.

Case 2. i and j are in the same bucket in both $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$. Then the pair $\{i, j\}$ contributes no penalty to $K_{\text{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$, and no pair of members of the set $\{i, j, i^\sharp, j^\sharp\}$ contribute any penalty to $K(\boldsymbol{\sigma}_\pi, \boldsymbol{\tau}_\pi)$.

Case 3. i and j are in the same bucket in one of the partial rankings $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$, but in different buckets in the other partial ranking. Then the pair $\{i, j\}$ contributes a penalty of $1/2$ to $K_{\text{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$. Assume without loss of generality that i and j are in the same bucket in $\boldsymbol{\sigma}$ and that $\boldsymbol{\tau}(i) < \boldsymbol{\tau}(j)$. There are now two subcases, depending on whether $\pi(i) < \pi(j)$ or $\pi(j) < \pi(i)$. In the first subcase, when $\pi(i) < \pi(j)$, we have

$$\boldsymbol{\sigma}_\pi(i) < \boldsymbol{\sigma}_\pi(j) < \boldsymbol{\sigma}_\pi(j^\sharp) < \boldsymbol{\sigma}_\pi(i^\sharp)$$

and

$$\boldsymbol{\tau}_\pi(i) < \boldsymbol{\tau}_\pi(i^\sharp) < \boldsymbol{\tau}_\pi(j) < \boldsymbol{\tau}_\pi(j^\sharp).$$

So the pairs among $\{i, j, i^\sharp, j^\sharp\}$ that contribute a penalty to $K(\boldsymbol{\sigma}_\pi, \boldsymbol{\tau}_\pi)$ are precisely $\{i^\sharp, j\}$ and $\{i^\sharp, j^\sharp\}$, each of which contribute a penalty of 1.

In the second subcase, when $\pi(j) < \pi(i)$, we have

$$\boldsymbol{\sigma}_\pi(j) < \boldsymbol{\sigma}_\pi(i) < \boldsymbol{\sigma}_\pi(i^\sharp) < \boldsymbol{\sigma}_\pi(j^\sharp)$$

and

$$\boldsymbol{\tau}_\pi(i) < \boldsymbol{\tau}_\pi(i^\sharp) < \boldsymbol{\tau}_\pi(j) < \boldsymbol{\tau}_\pi(j^\sharp).$$

So the pairs among $\{i, j, i^\sharp, j^\sharp\}$ that contribute a penalty to $K(\boldsymbol{\sigma}_\pi, \boldsymbol{\tau}_\pi)$ are precisely $\{i, j\}$ and $\{i^\sharp, j\}$, each of which contribute a penalty of 1.

In all cases, the amount of penalty contributed to $K(\boldsymbol{\sigma}_\pi, \boldsymbol{\tau}_\pi)$ is 4 times the amount of penalty contributed to $K_{\text{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$. The lemma then follows. □

Notice that Lemma 10 holds for every choice of π. The analogous statement is not true for F_{prof}. In that case, we need to choose π specifically for the pair of partial rankings we are given. In particular, we need to avoid a property we call "nesting."

Given fixed $\boldsymbol{\sigma}, \boldsymbol{\tau}$, we say that an element $d \in D$ is *nested* with respect to π if either

$$\begin{aligned} &[\boldsymbol{\sigma}_\pi(d), \boldsymbol{\sigma}_\pi(d^\sharp)] \sqsubset [\boldsymbol{\tau}_\pi(d), \boldsymbol{\tau}_\pi(d^\sharp)] \\ \text{or } &[\boldsymbol{\tau}_\pi(d), \boldsymbol{\tau}_\pi(d^\sharp)] \sqsubset [\boldsymbol{\sigma}_\pi(d), \boldsymbol{\sigma}_\pi(d^\sharp)], \end{aligned}$$

where the notation $[s, t] \sqsubset [u, v]$ for numbers s, t, u, v means that $[s, t] \subseteq [u, v]$ and $s \neq u$ and $t \neq v$. It is sometimes convenient to write $[u, v] \sqsupset [s, t]$ for $[s, t] \sqsubset [u, v]$.

The following lemma shows us why we want to avoid nesting.

LEMMA 11. *Given partial rankings $\boldsymbol{\sigma}, \boldsymbol{\tau}$ and full ranking π, suppose that there are no elements that are nested with respect to π. Then $F(\boldsymbol{\sigma}_\pi, \boldsymbol{\tau}_\pi) = 4F_{\text{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$.*

Proof. Assume $d \in D$. By assumption, d is not nested with respect to π. We now show that

$$\begin{aligned} &|\boldsymbol{\sigma}_\pi(d) - \boldsymbol{\tau}_\pi(d)| + |\boldsymbol{\sigma}_\pi(d^\sharp) - \boldsymbol{\tau}_\pi(d^\sharp)| \\ &= |\boldsymbol{\sigma}_\pi(d) - \boldsymbol{\tau}_\pi(d) + \boldsymbol{\sigma}_\pi(d^\sharp) - \boldsymbol{\tau}_\pi(d^\sharp)|. \end{aligned} \tag{10}$$

There are three cases, depending on whether $\boldsymbol{\sigma}_\pi(d) = \boldsymbol{\tau}_\pi(d)$, $\boldsymbol{\sigma}_\pi(d) < \boldsymbol{\tau}_\pi(d)$, or $\boldsymbol{\sigma}_\pi(d) > \boldsymbol{\tau}_\pi(d)$.

If $\boldsymbol{\sigma}_\pi(d) = \boldsymbol{\tau}_\pi(d)$, then (10) is immediate. If $\boldsymbol{\sigma}_\pi(d) < \boldsymbol{\tau}_\pi(d)$, then necessarily $\boldsymbol{\sigma}_\pi(d^\sharp) \leq \boldsymbol{\tau}_\pi(d^\sharp)$, since d is not nested. But then the left-hand side and right-hand side of (10) are each $\boldsymbol{\tau}_\pi(d) - \boldsymbol{\sigma}_\pi(d) + \boldsymbol{\tau}_\pi(d^\sharp) - \boldsymbol{\sigma}_\pi(d^\sharp)$, and so (10) holds. If $\boldsymbol{\sigma}_\pi(d) > \boldsymbol{\tau}_\pi(d)$, then necessarily $\boldsymbol{\sigma}_\pi(d^\sharp) \geq \boldsymbol{\tau}_\pi(d^\sharp)$, since d is not nested. But then the left-hand side and right-hand side of (10) are each $\boldsymbol{\sigma}_\pi(d) - \boldsymbol{\tau}_\pi(d) + \boldsymbol{\sigma}_\pi(d^\sharp) - \boldsymbol{\tau}_\pi(d^\sharp)$, and so once again, (10) holds.

From (9) we obtain $\boldsymbol{\sigma}_\pi(d) + \boldsymbol{\sigma}_\pi(d^\sharp) = 4\boldsymbol{\sigma}(d) - 1$. Similarly, we have $\boldsymbol{\tau}_\pi(d) + \boldsymbol{\tau}_\pi(d^\sharp) = 4\boldsymbol{\tau}(d) - 1$. Therefore

$$|\boldsymbol{\sigma}_\pi(d) - \boldsymbol{\tau}_\pi(d) + \boldsymbol{\sigma}_\pi(d^\sharp) - \boldsymbol{\tau}_\pi(d^\sharp)| = 4|\boldsymbol{\sigma}(d) - \boldsymbol{\tau}(d)|. \tag{11}$$

From (10) and (11) we obtain

$$|\boldsymbol{\sigma}_\pi(d) - \boldsymbol{\tau}_\pi(d)| + |\boldsymbol{\sigma}_\pi(d^\sharp) - \boldsymbol{\tau}_\pi(d^\sharp)| = 4|\boldsymbol{\sigma}(d) - \boldsymbol{\tau}(d)|.$$

Hence,

$$\begin{aligned} F(\boldsymbol{\sigma}_\pi, \boldsymbol{\tau}_\pi) &= \sum_{d \in D} (|\boldsymbol{\sigma}_\pi(d) - \boldsymbol{\tau}_\pi(d)| + |\boldsymbol{\sigma}_\pi(d^\sharp) - \boldsymbol{\tau}_\pi(d^\sharp)|) \\ &= \sum_{d \in D} 4|\boldsymbol{\sigma}(d) - \boldsymbol{\tau}(d)| \\ &= 4F_{\text{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau}). \qquad \square \end{aligned}$$

In the proof of the following lemma, we show that in fact, there is always a full ranking π with no nested elements.

LEMMA 12. *Let $\boldsymbol{\sigma}, \boldsymbol{\tau}$ be partial rankings. Then there exists a full ranking π on the same domain such that $F(\boldsymbol{\sigma}_\pi, \boldsymbol{\tau}_\pi) = 4F_{\text{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$.*

Proof. By Lemma 11, we need only show that there is some full ranking π with no nested elements. Assume that every full ranking has a nested element; we shall derive a contradiction. For a full ranking π, we say that its *first nest* is $\min_d \pi(d)$, where d is allowed to range over all nested elements of π. Choose π to be a full ranking whose first nest is as large as possible.

Let a be the element such that $\pi(a)$ is the first nest of π. By definition, a is nested. Without loss of generality, assume that

$$[\boldsymbol{\sigma}_\pi(a), \boldsymbol{\sigma}_\pi(a^\sharp)] \supseteq [\boldsymbol{\tau}_\pi(a), \boldsymbol{\tau}_\pi(a^\sharp)]. \tag{12}$$

The intuition behind the proof is the following. We find an element b such that it appears in the left-side interval but not in the right-side interval of (12). We swap a and b in the ordering π and argue that b is not nested in this new ordering. Furthermore, we also argue that no element occurring before a in π becomes nested due to the swap. Hence, we produce a full ranking whose first nest—if it has a nested element at all—is later than the first nest of π, a contradiction. We now proceed with the formal details.

Define the sets S_1 and S_2 as follows:

$$S_1 = \left\{ d \in D \setminus \{a\} \;\middle|\; [\boldsymbol{\sigma}_\pi(a), \boldsymbol{\sigma}_\pi(a^\sharp)] \supseteq [\boldsymbol{\sigma}_\pi(d), \boldsymbol{\sigma}_\pi(d^\sharp)] \right\} \text{ and}$$

$$S_2 = \left\{ d \in D \setminus \{a\} \;\middle|\; [\boldsymbol{\sigma}_\pi(a), \boldsymbol{\sigma}_\pi(a^\sharp)] \supseteq [\boldsymbol{\tau}_\pi(d), \boldsymbol{\tau}_\pi(d^\sharp)] \right\}.$$

We now show that $S_1 \setminus S_2$ is nonempty. This is because $|S_1| = \frac{1}{2}|[\boldsymbol{\sigma}_\pi(a), \boldsymbol{\sigma}_\pi(a^\sharp)]| - 1$, while $|S_2| \leq \frac{1}{2}|[\boldsymbol{\sigma}_\pi(a), \boldsymbol{\sigma}_\pi(a^\sharp)]| - 2$, since $[\boldsymbol{\sigma}_\pi(a), \boldsymbol{\sigma}_\pi(a^\sharp)] \supset [\boldsymbol{\tau}_\pi(a), \boldsymbol{\tau}_\pi(a^\sharp)]$ but a is not counted in S_2. Choose b in $S_1 \setminus S_2$. Note that the fact that $b \in S_1$ implies that a and b are in the same bucket for $\boldsymbol{\sigma}$. It further implies that $\pi(a) < \pi(b)$.

We now show that a and b are in different buckets for $\boldsymbol{\tau}$. Suppose that a and b were in the same bucket for $\boldsymbol{\tau}$. Then since $\pi(a) < \pi(b)$, we would have $\boldsymbol{\tau}_\pi(a) < \boldsymbol{\tau}_\pi(b)$ and $\boldsymbol{\tau}_\pi(a^\sharp) > \boldsymbol{\tau}_\pi(b^\sharp)$. That is, $[\boldsymbol{\tau}_\pi(a), \boldsymbol{\tau}_\pi(a^\sharp)] \supset [\boldsymbol{\tau}_\pi(b), \boldsymbol{\tau}_\pi(b^\sharp)]$. If we combine this fact with (12), we obtain $[\boldsymbol{\sigma}_\pi(a), \boldsymbol{\sigma}_\pi(a^\sharp)] \supset [\boldsymbol{\tau}_\pi(a), \boldsymbol{\tau}_\pi(a^\sharp)] \supset [\boldsymbol{\tau}_\pi(b), \boldsymbol{\tau}_\pi(b^\sharp)]$. This contradicts the fact that $b \notin S_2$. Hence, a and b must be in different buckets for $\boldsymbol{\tau}$.

Now, produce π' by swapping a and b in π. Since $\pi(a) < \pi(b)$, we see that $\pi'(b) = \pi(a) < \pi(b) = \pi'(a)$. We wish to prove that the first nest for π'—if it has a nested element at all—is larger than the first nest for π, which gives our desired contradiction. We do so by showing that b is unnested for π' and further, that d is unnested for π' for all d in D such that $\pi'(d) < \pi'(b)$. In order to prove this, we need to examine the effect of swapping a and b in π.

We first consider $\boldsymbol{\sigma}$. We know that a and b appear in the same bucket of $\boldsymbol{\sigma}$. Let $\mathcal{B}_{ab}$ be the bucket of $\boldsymbol{\sigma}$ that contains both a and b. Swapping a and b in π has the effect of swapping the positions of a and b in $\boldsymbol{\sigma}_\pi$ (so in particular $\boldsymbol{\sigma}_{\pi'}(b) = \boldsymbol{\sigma}_\pi(a)$), swapping the positions of $a^\sharp$ and $b^\sharp$ in $\boldsymbol{\sigma}_\pi$ (so in particular $\boldsymbol{\sigma}_{\pi'}(b^\sharp) = \boldsymbol{\sigma}_\pi(a^\sharp)$) and leaving all other elements d and $d^\sharp$ in $\mathcal{B}_{ab}$ in the same place (so $\boldsymbol{\sigma}_\pi(d) = \boldsymbol{\sigma}_{\pi'}(d)$ and $\boldsymbol{\sigma}_\pi(d^\sharp) = \boldsymbol{\sigma}_{\pi'}(d^\sharp)$). Since $\boldsymbol{\sigma}_{\pi'}(b) = \boldsymbol{\sigma}_\pi(a)$ and $\boldsymbol{\sigma}_{\pi'}(b^\sharp) = \boldsymbol{\sigma}_\pi(a^\sharp)$, and since two closed intervals of numbers are equal precisely if their left endpoints and their right endpoints are equal, we have

$$[\boldsymbol{\sigma}_{\pi'}(b), \boldsymbol{\sigma}_{\pi'}(b^\sharp)] = [\boldsymbol{\sigma}_\pi(a), \boldsymbol{\sigma}_\pi(a^\sharp)]. \tag{13}$$

Now, let $\mathcal{B}$ be a bucket of $\boldsymbol{\sigma}$ other than $\mathcal{B}_{ab}$. Then swapping a and b in π has no effect (as far as $\boldsymbol{\sigma}_\pi$ is concerned) on the elements in $\mathcal{B}$, since the relative order of all elements in $\mathcal{B}$ is precisely the same with or without the swap. That is, $\boldsymbol{\sigma}_\pi(d) = \boldsymbol{\sigma}_{\pi'}(d)$ and $\boldsymbol{\sigma}_\pi(d^\sharp) = \boldsymbol{\sigma}_{\pi'}(d^\sharp)$ for all d in $\mathcal{B}$. But we noted earlier that these same two equalities hold for all elements d in $\mathcal{B}_{ab}$ other than a and b. Therefore, for all elements d other than a or b (whether or not these elements are in $\mathcal{B}_{ab}$), we have

$$[\boldsymbol{\sigma}_{\pi'}(d), \boldsymbol{\sigma}_{\pi'}(d^\sharp)] = [\boldsymbol{\sigma}_\pi(d), \boldsymbol{\sigma}_\pi(d^\sharp)]. \tag{14}$$

We now consider $\boldsymbol{\tau}$. We know that a and b appear in different buckets of $\boldsymbol{\tau}$. Let $\mathcal{B}$ be a bucket of $\boldsymbol{\tau}$ containing neither a nor b (if there is such a bucket). As with $\boldsymbol{\sigma}$, we see that elements in $\mathcal{B}$ are unaffected by swapping a and b in π. That is, $\boldsymbol{\tau}_\pi(d) = \boldsymbol{\tau}_{\pi'}(d)$ and $\boldsymbol{\tau}_\pi(d^\sharp) = \boldsymbol{\tau}_{\pi'}(d^\sharp)$ for all d in $\mathcal{B}$.

Now, let $\mathcal{B}_a$ be the bucket of $\boldsymbol{\tau}$ containing a (but not b). Notice that for all d in $\mathcal{B}_a$ such that $\pi(d) < \pi(a)$, we have $\pi(d) = \pi'(d)$. Hence, the relative order among these most highly ranked elements of $\mathcal{B}_a$ remains the same. Therefore, $\boldsymbol{\tau}_\pi(d) = \boldsymbol{\tau}_{\pi'}(d)$ and $\boldsymbol{\tau}_\pi(d^\sharp) = \boldsymbol{\tau}_{\pi'}(d^\sharp)$ for all d in $\mathcal{B}_a$ such that $\pi(d) < \pi(a)$. Furthermore, $\pi'(a) > \pi(a)$, and so a is still ranked after all the aforementioned d's in $\boldsymbol{\tau}_{\pi'}$. Hence, $\boldsymbol{\tau}_\pi(a) \leq \boldsymbol{\tau}_{\pi'}(a)$ and $\boldsymbol{\tau}_\pi(a^\sharp) \geq \boldsymbol{\tau}_{\pi'}(a^\sharp)$. That is,

$$[\boldsymbol{\tau}_{\pi'}(a), \boldsymbol{\tau}_{\pi'}(a^\sharp)] \subseteq [\boldsymbol{\tau}_\pi(a), \boldsymbol{\tau}_\pi(a^\sharp)]. \tag{15}$$

Finally, let $\mathcal{B}_b$ be the bucket of $\boldsymbol{\tau}$ that contains b (but not a). As before, for all d in $\mathcal{B}_b$ such that $\pi(d) < \pi(a)$, we have $\pi(d) = \pi'(d)$. Hence, the relative order among these most highly ranked elements of $\mathcal{B}_b$ remains the same. Therefore, $\boldsymbol{\tau}_\pi(d) = \boldsymbol{\tau}_{\pi'}(d)$

and $\boldsymbol{\tau}_\pi(d^\sharp) = \boldsymbol{\tau}_{\pi'}(d^\sharp)$ for all d in $\mathcal{B}_b$ such that $\pi(d) < \pi(a)$. That is, for every d such that $\pi(d) < \pi(a)$ (i.e., every d such that $\pi'(d) < \pi'(b)$), we have

$$[\boldsymbol{\tau}_{\pi'}(d), \boldsymbol{\tau}_{\pi'}(d^\sharp)] = [\boldsymbol{\tau}_\pi(d), \boldsymbol{\tau}_\pi(d^\sharp)]. \tag{16}$$

Furthermore, $\pi'(b) < \pi(b)$, and so b is still ranked before all d' in $\mathcal{B}_b$ such that $\pi(b) < \pi(d') = \pi'(d')$. Hence, $\boldsymbol{\tau}_\pi(b) \geq \boldsymbol{\tau}_{\pi'}(b)$ and $\boldsymbol{\tau}_\pi(b^\sharp) \leq \boldsymbol{\tau}_{\pi'}(b^\sharp)$. That is,

$$[\boldsymbol{\tau}_{\pi'}(b), \boldsymbol{\tau}_{\pi'}(b^\sharp)] \supseteq [\boldsymbol{\tau}_\pi(b), \boldsymbol{\tau}_\pi(b^\sharp)]. \tag{17}$$

From (14) and (16), we see that d remains unnested for all d such that $\pi'(d) < \pi'(b)$. So we need only show that b is unnested for π' to finish the proof. If b were nested for π', then either $[\boldsymbol{\sigma}_{\pi'}(b), \boldsymbol{\sigma}_{\pi'}(b^\sharp)] \sqsupset [\boldsymbol{\tau}_{\pi'}(b), \boldsymbol{\tau}_{\pi'}(b^\sharp)]$ or $[\boldsymbol{\tau}_{\pi'}(b), \boldsymbol{\tau}_{\pi'}(b^\sharp)] \sqsupset [\boldsymbol{\sigma}_{\pi'}(b), \boldsymbol{\sigma}_{\pi'}(b^\sharp)]$. First, suppose that $[\boldsymbol{\sigma}_{\pi'}(b), \boldsymbol{\sigma}_{\pi'}(b^\sharp)] \sqsupset [\boldsymbol{\tau}_{\pi'}(b), \boldsymbol{\tau}_{\pi'}(b^\sharp)]$. Then

$$\begin{aligned}
[\boldsymbol{\sigma}_\pi(a), \boldsymbol{\sigma}_\pi(a^\sharp)] &= [\boldsymbol{\sigma}_{\pi'}(b), \boldsymbol{\sigma}_{\pi'}(b^\sharp)] \text{ by (13)} \\
&\sqsupset [\boldsymbol{\tau}_{\pi'}(b), \boldsymbol{\tau}_{\pi'}(b^\sharp)] \text{ by supposition} \\
&\supseteq [\boldsymbol{\tau}_\pi(b), \boldsymbol{\tau}_\pi(b^\sharp)] \text{ by (17)}.
\end{aligned}$$

But this contradicts the fact that $b \notin S_2$. Now, suppose that $[\boldsymbol{\tau}_{\pi'}(b), \boldsymbol{\tau}_{\pi'}(b^\sharp)] \sqsupset [\boldsymbol{\sigma}_{\pi'}(b), \boldsymbol{\sigma}_{\pi'}(b^\sharp)]$. Then

$$\begin{aligned}
[\boldsymbol{\tau}_{\pi'}(b), \boldsymbol{\tau}_{\pi'}(b^\sharp)] &\sqsupset [\boldsymbol{\sigma}_{\pi'}(b), \boldsymbol{\sigma}_{\pi'}(b^\sharp)] \text{ by supposition} \\
&= [\boldsymbol{\sigma}_\pi(a), \boldsymbol{\sigma}_\pi(a^\sharp)] \text{ by (13)} \\
&\sqsupset [\boldsymbol{\tau}_\pi(a), \boldsymbol{\tau}_\pi(a^\sharp)] \text{ by (12)} \\
&\supseteq [\boldsymbol{\tau}_{\pi'}(a), \boldsymbol{\tau}_{\pi'}(a^\sharp)] \text{ by (15)}.
\end{aligned}$$

But this implies that a and b are in the same bucket for $\boldsymbol{\tau}$, a contradiction. Hence, b must not be nested for π', which was to be shown. □

We can now prove our desired theorem that F_{prof} and K_{prof} are equivalent.

THEOREM 13. *Let $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ be partial rankings. Then $K_{\text{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) \leq F_{\text{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) \leq 2K_{\text{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$.*

Proof. Given $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$, let π be the full ranking guaranteed by Lemma 12. Then we have

$$\begin{aligned}
K_{\text{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) &= 4K(\boldsymbol{\sigma}_\pi, \boldsymbol{\tau}_\pi) \text{ by Lemma 10} \\
&\leq 4F(\boldsymbol{\sigma}_\pi, \boldsymbol{\tau}_\pi) \text{ by (1)} \\
&= F_{\text{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) \text{ by Lemma 12}.
\end{aligned}$$

And similarly,

$$\begin{aligned}
F_{\text{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) &= 4F(\boldsymbol{\sigma}_\pi, \boldsymbol{\tau}_\pi) \text{ by Lemma 12} \\
&\leq 8K(\boldsymbol{\sigma}_\pi, \boldsymbol{\tau}_\pi) \text{ by (1)} \\
&= 2K_{\text{prof}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) \text{ by Lemma 10.}
\end{aligned}$$ □

4.3. Equivalence of K_{Haus} and K_{prof}. We now prove (6), which is the final step in proving Theorem 7.

THEOREM 14. *Let $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ be partial rankings. Then $K_{\text{prof}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \leq K_{\text{Haus}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \leq 2K_{\text{prof}}(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$.*

Proof. As in Theorem 6 (where we let $\boldsymbol{\sigma}_1$ play the role of $\boldsymbol{\sigma}$, and let $\boldsymbol{\sigma}_2$ play the role of $\boldsymbol{\tau}$), let S be the set of pairs $\{i,j\}$ of distinct elements such that i and j appear in the same bucket of $\boldsymbol{\sigma}_1$ but in different buckets of $\boldsymbol{\sigma}_2$, let T be the set of pairs $\{i,j\}$ of distinct elements such that i and j appear in the same bucket of $\boldsymbol{\sigma}_2$ but in different buckets of $\boldsymbol{\sigma}_1$, and let U be the set of pairs $\{i,j\}$ of distinct elements that are in different buckets of both $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ and are in a different order in $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$. By Theorem 6, we know that $K_{\mathrm{Haus}}(\boldsymbol{\sigma}_1,\boldsymbol{\sigma}_2) = |U| + \max\{|S|,|T|\}$. It follows from the definition of K_{prof} that $K_{\mathrm{prof}}(\boldsymbol{\sigma}_1,\boldsymbol{\sigma}_2) = |U| + \frac{1}{2}|S| + \frac{1}{2}|T|$. The theorem now follows from the straightforward inequalities $|U| + \frac{1}{2}|S| + \frac{1}{2}|T| \le |U| + \max\{|S|,|T|\} \le 2(|U| + \frac{1}{2}|S| + \frac{1}{2}|T|)$. □

This concludes the proof that all our metrics are equivalent.

5. An alternative representation. Let $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ be partial rankings. Assume that the buckets of $\boldsymbol{\sigma}$ are, in order, $\mathcal{B}_1,\ldots,\mathcal{B}_t$, and the buckets of $\boldsymbol{\sigma}'$ are, in order, $\mathcal{B}'_1,\ldots,\mathcal{B}'_{t'}$. Critchlow [9] defines n_{ij} (for $1 \le i \le t$ and $1 \le j \le t'$) to be $|\mathcal{B}_i \cap \mathcal{B}'_j|$. His main theorem gives formulas for $K_{\mathrm{Haus}}(\boldsymbol{\sigma},\boldsymbol{\sigma}')$ and $F_{\mathrm{Haus}}(\boldsymbol{\sigma},\boldsymbol{\sigma}')$ (and for other Hausdorff measures) in terms of the n_{ij}'s. His formula for $K_{\mathrm{Haus}}(\boldsymbol{\sigma},\boldsymbol{\sigma}')$ is particularly simple, and is given by the following theorem.

THEOREM 15 (see [9]). *Let $\boldsymbol{\sigma}$, $\boldsymbol{\sigma}'$, and the n_{ij}'s be as above. Then*

$$K_{\mathrm{Haus}}(\boldsymbol{\sigma},\boldsymbol{\sigma}') = \max\left\{ \sum_{i<i',\ j\ge j'} n_{ij}n_{i'j'},\ \sum_{i\le i',\ j>j'} n_{ij}n_{i'j'} \right\}.$$

It is straightforward to derive Theorem 6 from Theorem 15, and to derive Theorem 15 from Theorem 6, by using the simple fact that if S,T,U are as in Theorem 6, then

$$|U| = \sum_{i<i',\ j>j'} n_{ij}n_{i'j'},$$

$$|S| = \sum_{i=i',\ j>j'} n_{ij}n_{i'j'},$$

$$|T| = \sum_{i<i',\ j=j'} n_{ij}n_{i'j'}.$$

Let us define the *Critchlow profile* of the pair $(\boldsymbol{\sigma},\boldsymbol{\sigma}')$ to be a $t \times t'$ matrix, where t is the number of buckets of $\boldsymbol{\sigma}$, t' is the number of buckets of $\boldsymbol{\sigma}'$, and the (i,j)th entry is n_{ij}. We noted that Critchlow gives formulas for $K_{\mathrm{Haus}}(\boldsymbol{\sigma},\boldsymbol{\sigma}')$ and $F_{\mathrm{Haus}}(\boldsymbol{\sigma},\boldsymbol{\sigma}')$ in terms of the Critchlow profile. The reader may find it surprising that the Critchlow profile contains enough information to compute $K_{\mathrm{Haus}}(\boldsymbol{\sigma},\boldsymbol{\sigma}')$ and $F_{\mathrm{Haus}}(\boldsymbol{\sigma},\boldsymbol{\sigma}')$. The following theorem implies that this "surprise" is true not just about K_{Haus} and F_{Haus}, but about every function d (not even necessarily a metric) whose arguments are a pair of partial rankings, as long as d is "name-independent" (that is, the answer is the same when we rename the elements). Before we state the theorem, we need some more terminology. The theorem says that the Critchlow profile "uniquely determines $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$, up to renaming of the elements." What this means is that if $(\boldsymbol{\sigma},\boldsymbol{\sigma}')$ has the same Critchlow profile as $(\boldsymbol{\tau},\boldsymbol{\tau}')$, then the pair $(\boldsymbol{\sigma},\boldsymbol{\sigma}')$ is isomorphic to the pair $(\boldsymbol{\tau},\boldsymbol{\tau}')$. That is, there is a one-to-one function f from the common domain D onto itself such that $\boldsymbol{\sigma}(i) = \boldsymbol{\tau}(f(i))$ and $\boldsymbol{\sigma}'(i) = \boldsymbol{\tau}'(f(i))$ for every i in D. Intuitively, the pair $(\boldsymbol{\tau},\boldsymbol{\tau}')$ is obtained from the pair $(\boldsymbol{\sigma},\boldsymbol{\sigma}')$ by the renaming function f.

THEOREM 16. *The Critchlow profile uniquely determines $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$, up to renaming of the elements.*

Proof. We first give an informal proof. The only relevant information about an element is which $\mathcal{B}_i$ it is in and which $\mathcal{B}'_j$ it is in. So the only information that matters about the pair $\boldsymbol{\sigma}, \boldsymbol{\sigma}'$ of partial rankings is, for each i, j, how many elements are in $\mathcal{B}_i \cap \mathcal{B}'_j$. That is, we can reconstruct $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$, up to renaming of the elements, by knowing only the Critchlow profile.

More formally, let $(\boldsymbol{\sigma}, \boldsymbol{\sigma}')$ and $(\boldsymbol{\tau}, \boldsymbol{\tau}')$ each be pairs of partial rankings with the same Critchlow profile. That is, assume that the buckets of $\boldsymbol{\sigma}$ are, in order, $\mathcal{B}_1, \ldots, \mathcal{B}_t$, the buckets of $\boldsymbol{\sigma}'$ are, in order, $\mathcal{B}'_1, \ldots, \mathcal{B}'_{t'}$, the buckets of $\boldsymbol{\tau}$ are, in order, $\mathcal{C}_1, \ldots, \mathcal{C}_t$, and the buckets of $\boldsymbol{\tau}'$ are, in order, $\mathcal{C}'_1, \ldots, \mathcal{C}_{t'}$, where $|\mathcal{B}_i \cap B'_j| = |\mathcal{C}_i \cap C'_j|$ for each i, j. (Note that the number t of buckets of $\boldsymbol{\sigma}$ is the same as the number of buckets of $\boldsymbol{\tau}$, and similarly the number t' of buckets of $\boldsymbol{\sigma}'$ is the same as the number of buckets of $\boldsymbol{\tau}'$; this follows from the assumption that $(\boldsymbol{\sigma}, \boldsymbol{\sigma}')$ and $(\boldsymbol{\tau}, \boldsymbol{\tau}')$ have the same Critchlow profile.) Let f_{ij} be a one-to-one mapping of $\mathcal{B}_i \cap \mathcal{B}'_j$ onto $\mathcal{C}_i \cap \mathcal{C}'_j$ (such an f_{ij} exists because $|\mathcal{B}_i \cap \mathcal{B}'_j| = |\mathcal{C}_i \cap \mathcal{C}'_j|$). Let f be the function obtained by taking the union of the functions f_{ij} (we think of functions as sets of ordered pairs, so it is proper to take the union). It is easy to see that $(\boldsymbol{\sigma}, \boldsymbol{\sigma}')$ and $(\boldsymbol{\tau}, \boldsymbol{\tau}')$ are isomorphic under the isomorphism f. This proves the theorem. □

The Critchlow profile differs in several ways from the K-profile and the F-profile, as defined in section 3.1. First, the K-profile and the F-profile are each profiles of a single partial ranking, whereas the Critchlow profile is a profile of a pair of partial rankings. Second, from the K-profile of $\boldsymbol{\sigma}$ we can completely reconstruct $\boldsymbol{\sigma}$ (not just up to renaming of elements, but completely), and a similar comment applies to the F-profile. On the other hand, from the Critchlow profile we can reconstruct the pair $(\boldsymbol{\sigma}, \boldsymbol{\sigma}')$ only up to a renaming of elements. Thus, the Critchlow profile "loses information," whereas the K-profile and F-profile do not.

6. Conclusions. In this paper we consider metrics between partial rankings. We define four natural metrics between partial rankings. We obtain efficient polynomial time algorithms to compute these metrics. We also show that these metrics are all within constant multiples of each other.

Appendix. Proof of Theorem 5. In this appendix, we prove Theorem 5. First, we state a fact that we use several times.

LEMMA 17. *Suppose $a \le b$ and $c \le d$. Then $|a - c| + |b - d| \le |a - d| + |b - c|$.*

Proof. To see this, first note that by symmetry, we can assume, without loss of generality, that $a \le c$. Now there are three cases: $a \le b \le c \le d$, $a \le c \le b \le d$, and $a \le c \le d \le b$. In the first case (when $a \le b \le c \le d$), it is easy to verify that both the left-hand side and the right-hand side of the inequality equal $|a-b|+2|b-c|+|c-d|$, and so the left-hand side and the right-hand side are equal. In both the second case (when $a \le c \le b \le d$) and the third case (when $a \le c \le d \le b$), it is easy to verify that the right-hand side equals $|a-c|+2|b-c|+|b-d|$, which exceeds the left-hand side by $2|b-c|$. □

We next show a simple lemma.

LEMMA 18. *Let π be a full ranking, and let $\boldsymbol{\sigma}$ be a partial ranking. Suppose that $\pi \ne \boldsymbol{\sigma}$. Then there exist i, j such that $\pi(j) = \pi(i) + 1$ while $\boldsymbol{\sigma}(j) \le \boldsymbol{\sigma}(i)$. If $\boldsymbol{\sigma}$ is in fact a full ranking, then $\boldsymbol{\sigma}(j) < \boldsymbol{\sigma}(i)$.*

Proof. For each m with $1 \le m \le |D|$, let d_m be the member of the domain D, where $\pi(d_m) = m$. Thus, $D = \{d_1, \ldots, d_{|D|}\}$ and $\pi(d_1) < \pi(d_2) < \cdots < \pi(d_{|D|})$. If $\boldsymbol{\sigma}(d_\ell) < \boldsymbol{\sigma}(d_{\ell+1})$ for all ℓ, then we would have $K_{\text{prof}}(\boldsymbol{\sigma}, \pi) = 0$, contradicting the fact

that $\pi \neq \boldsymbol{\sigma}$. Hence, there must be some ℓ for which $\boldsymbol{\sigma}(d_{\ell+1}) \leq \boldsymbol{\sigma}(d_\ell)$. Setting $i = d_\ell$ and $j = d_{\ell+1}$ gives us the lemma.

If $\boldsymbol{\sigma}$ is a full ranking, then $\boldsymbol{\sigma}(j) \neq \boldsymbol{\sigma}(i)$, showing $\boldsymbol{\sigma}(j) < \boldsymbol{\sigma}(i)$. □

The next two lemmas will be helpful in obtaining a characterization of the Hausdorff distance.

LEMMA 19. *Let σ be a full ranking, and let $\boldsymbol{\tau}$ be a partial ranking. Then the quantity $F(\sigma,\tau)$, taken over all full refinements $\tau \succeq \boldsymbol{\tau}$, is minimized for $\tau = \sigma * \boldsymbol{\tau}$. Similarly, the quantity $K(\sigma,\tau)$, taken over all full refinements $\tau \succeq \boldsymbol{\tau}$, is minimized for $\tau = \sigma * \boldsymbol{\tau}$.*

Proof. First, note that if τ is a full ranking with $\tau \succeq \boldsymbol{\tau}$, then there is a full ranking π such that $\tau = \boldsymbol{\tau} * \pi$. We show that $F(\sigma, \sigma * \boldsymbol{\tau}) \leq F(\sigma, \pi * \boldsymbol{\tau})$ and $K(\sigma, \sigma * \boldsymbol{\tau}) \leq K(\sigma, \pi * \boldsymbol{\tau})$ for every full ranking π. The lemma will then follow. Let

$$U = \{\pi \mid \pi \text{ is a full ranking and } F(\sigma, \sigma * \boldsymbol{\tau}) > F(\sigma, \pi * \boldsymbol{\tau})\},$$

$$V = \{\pi \mid \pi \text{ is a full ranking and } K(\sigma, \sigma * \boldsymbol{\tau}) > K(\sigma, \pi * \boldsymbol{\tau})\},$$

and let $S = U \cup V$. If S is empty, then we are done. So suppose not; we derive a contradiction. Over all full rankings $\pi \in S$, choose π to be a full ranking that minimizes $K(\sigma,\pi)$. In other words, choose a full ranking in S that is as close to σ as possible, according to the Kendall distance.

Clearly $\sigma \notin S$, and so $\pi \neq \sigma$ (since $\pi \in S$). Since $\pi \neq \sigma$, Lemma 18 guarantees that we can find a pair i, j such that $\pi(j) = \pi(i) + 1$, but $\sigma(j) < \sigma(i)$. Produce π' by swapping i and j in π. Clearly, π' has one fewer inversion with respect to σ than π does. Hence, $K(\sigma,\pi') < K(\sigma,\pi)$. If we can show that $\pi' \in S$, then we obtain our desired contradiction, since π is the full ranking in S that minimizes $K(\sigma,\pi)$. So we need only show that $\pi' \in S$.

If i and j are in different buckets for $\boldsymbol{\tau}$, then $\pi' * \boldsymbol{\tau} = \pi * \boldsymbol{\tau}$. Hence, $F(\sigma, \pi' * \boldsymbol{\tau}) = F(\sigma, \pi * \boldsymbol{\tau})$ and $K(\sigma, \pi' * \boldsymbol{\tau}) = K(\sigma, \pi * \boldsymbol{\tau})$. So if $\pi \in U$, then $\pi' \in U$, and if $\pi \in V$, then $\pi' \in V$. In either case, $\pi' \in S$, and we are done.

On the other hand, assume that i and j are in the same bucket for $\boldsymbol{\tau}$. Then $\pi' * \boldsymbol{\tau}(i) = \pi * \boldsymbol{\tau}(j)$ and $\pi' * \boldsymbol{\tau}(j) = \pi * \boldsymbol{\tau}(i)$. Furthermore, since $\pi(i) < \pi(j)$ and i and j are in the same bucket for $\boldsymbol{\tau}$, we have $\pi * \boldsymbol{\tau}(i) < \pi * \boldsymbol{\tau}(j)$, while $\sigma(j) < \sigma(i)$.

Either $\pi \in U$ or $\pi \in V$. First, consider the case where $\pi \in U$. We have

$$\begin{aligned} &|\pi' * \boldsymbol{\tau}(j) - \sigma(j)| + |\pi' * \boldsymbol{\tau}(i) - \sigma(i)| \\ &= |\pi * \boldsymbol{\tau}(i) - \sigma(j)| + |\pi * \boldsymbol{\tau}(j) - \sigma(i)| \\ &\leq |\pi * \boldsymbol{\tau}(i) - \sigma(i)| + |\pi * \boldsymbol{\tau}(j) - \sigma(j)|, \end{aligned} \tag{18}$$

where the inequality follows from Lemma 17 with $a = \pi * \boldsymbol{\tau}(i)$, $b = \pi * \boldsymbol{\tau}(j)$, $c = \sigma(j)$, and $d = \sigma(i)$. We also have $|\pi' * \boldsymbol{\tau}(d) - \sigma(d)| = |\pi * \boldsymbol{\tau}(d) - \sigma(d)|$ for all $d \in D \setminus \{i, j\}$, since $\pi' * \boldsymbol{\tau}$ and $\pi * \boldsymbol{\tau}$ agree everywhere but at i and j. If we sum over all d (where we make use of (18) for $d = i$ and $d = j$), we obtain $F(\sigma, \pi' * \boldsymbol{\tau}) \leq F(\sigma, \pi * \boldsymbol{\tau})$. Since $\pi \in U$, we have $F(\sigma, \pi * \boldsymbol{\tau}) < F(\sigma, \sigma * \boldsymbol{\tau})$. Combining these last two inequalities, we obtain $F(\sigma, \pi' * \boldsymbol{\tau}) < F(\sigma, \sigma * \boldsymbol{\tau})$. Therefore, $\pi' \in U$, and so $\pi' \in S$, which was to be shown.

Now consider the case where $\pi \in V$. Since $\pi(j) = \pi(i) + 1$ and since i and j are in the same bucket of $\boldsymbol{\tau}$, we have $\pi * \boldsymbol{\tau}(j) = \pi * \boldsymbol{\tau}(i) + 1$. Similarly, $\pi' * \boldsymbol{\tau}(i) = \pi' * \boldsymbol{\tau}(j) + 1$. And as we noted earlier, $\pi * \boldsymbol{\tau}$ and $\pi' * \boldsymbol{\tau}$ agree everywhere except at i and j. In other words, $\pi' * \boldsymbol{\tau}$ is just $\pi * \boldsymbol{\tau}$, with the *adjacent* elements i and j swapped. Since

$\sigma(i) > \sigma(j)$ we see that $\pi' * \boldsymbol{\tau}$ has exactly one fewer inversion with respect to σ than $\pi * \boldsymbol{\tau}$ does. Hence, $K(\sigma, \pi' * \boldsymbol{\tau}) < K(\sigma, \pi * \boldsymbol{\tau})$. Since $\pi \in V$, we have $K(\sigma, \pi * \boldsymbol{\tau}) < K(\sigma, \sigma * \boldsymbol{\tau})$. Combining these last two inequalities, we obtain $K(\sigma, \pi' * \boldsymbol{\tau}) < K(\sigma, \sigma * \boldsymbol{\tau})$. Therefore, $\pi' \in V$, and so $\pi' \in S$, which was to be shown. □

LEMMA 20. *Let $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ be partial rankings, and let ρ be any full ranking. Then the quantity $F(\sigma, \sigma * \boldsymbol{\tau})$, taken over all full refinements $\sigma \succeq \boldsymbol{\sigma}$, is maximized when $\sigma = \rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}$. Similarly, the quantity $K(\sigma, \sigma * \boldsymbol{\tau})$, taken over all full refinements $\sigma \succeq \boldsymbol{\sigma}$, is maximized when $\sigma = \rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}$.*

Proof. First, note that for any full refinement $\sigma \succeq \boldsymbol{\sigma}$, there is some full ranking π such that $\sigma = \pi * \boldsymbol{\sigma}$. We show that for all full rankings π,

$$F(\rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}, \rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma} * \boldsymbol{\tau}) \geq F(\pi * \boldsymbol{\sigma}, \pi * \boldsymbol{\sigma} * \boldsymbol{\tau})$$
$$\text{and } K(\rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}, \rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma} * \boldsymbol{\tau}) \geq K(\pi * \boldsymbol{\sigma}, \pi * \boldsymbol{\sigma} * \boldsymbol{\tau}).$$

The lemma will then follow.

Let $U = \{\text{full } \pi \mid F(\rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}, \rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma} * \boldsymbol{\tau}) < F(\pi * \boldsymbol{\sigma}, \pi * \boldsymbol{\sigma} * \boldsymbol{\tau})\}$, let $V = \{\text{full } \pi \mid K(\rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}, \rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma} * \boldsymbol{\tau}) < K(\pi * \boldsymbol{\sigma}, \pi * \boldsymbol{\sigma} * \boldsymbol{\tau})\}$, and let $S = U \cup V$. If S is empty, then we are done. So suppose not; we derive a contradiction. Over all full rankings $\pi \in S$, choose π to be the full ranking that minimizes $K(\rho * \boldsymbol{\tau}^{\mathrm{R}}, \pi)$.

Clearly $\rho * \boldsymbol{\tau}^{\mathrm{R}} \notin S$, and so $\pi \neq \rho * \boldsymbol{\tau}^{\mathrm{R}}$ (since $\pi \in S$). Since $\pi \neq \rho * \boldsymbol{\tau}^{\mathrm{R}}$, Lemma 18 guarantees that we can find a pair i, j such that $\pi(j) = \pi(i) + 1$, but $\rho * \boldsymbol{\tau}^{\mathrm{R}}(j) < \rho * \boldsymbol{\tau}^{\mathrm{R}}(i)$. Produce π' by swapping i and j. Clearly, π' has one fewer inversion with respect to $\rho * \boldsymbol{\tau}^{\mathrm{R}}$ than π does. Hence, $K(\rho * \boldsymbol{\tau}^{\mathrm{R}}, \pi') < K(\rho * \boldsymbol{\tau}^{\mathrm{R}}, \pi)$. We now show that $\pi' \in S$, producing a contradiction.

If i and j are in different buckets for $\boldsymbol{\sigma}$, then $\pi' * \boldsymbol{\sigma} = \pi * \boldsymbol{\sigma}$. Hence, $F(\pi' * \boldsymbol{\sigma}, \pi' * \boldsymbol{\sigma} * \boldsymbol{\tau}) = F(\pi * \boldsymbol{\sigma}, \pi * \boldsymbol{\sigma} * \boldsymbol{\tau})$ and $K(\pi' * \boldsymbol{\sigma}, \pi' * \boldsymbol{\sigma} * \boldsymbol{\tau}) = K(\pi * \boldsymbol{\sigma}, \pi * \boldsymbol{\sigma} * \boldsymbol{\tau})$. So if $\pi \in U$, then $\pi' \in U$, and if $\pi \in V$, then $\pi' \in V$. In either case, $\pi' \in S$, and we are done.

Likewise, if i and j are in the same bucket for both $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$, then swapping i and j in π swaps their positions in both $\pi * \boldsymbol{\sigma} * \boldsymbol{\tau}$ and $\pi * \boldsymbol{\sigma}$ and leaves all other elements in their same positions in both $\pi * \boldsymbol{\sigma} * \boldsymbol{\tau}$ and $\pi * \boldsymbol{\sigma}$. So again, we see $F(\pi' * \boldsymbol{\sigma}, \pi' * \boldsymbol{\sigma} * \boldsymbol{\tau}) = F(\pi * \boldsymbol{\sigma}, \pi * \boldsymbol{\sigma} * \boldsymbol{\tau})$ and $K(\pi' * \boldsymbol{\sigma}, \pi' * \boldsymbol{\sigma} * \boldsymbol{\tau}) = K(\pi * \boldsymbol{\sigma}, \pi * \boldsymbol{\sigma} * \boldsymbol{\tau})$. As before, $\pi' \in S$.

The only remaining situation is when i and j are in the same bucket for $\boldsymbol{\sigma}$, but in different buckets for $\boldsymbol{\tau}$. Let us consider this situation. First of all, $\pi' * \boldsymbol{\sigma}$ is just $\pi * \boldsymbol{\sigma}$ with the adjacent elements i and j swapped, since i and j are in the same bucket for $\boldsymbol{\sigma}$. Second, $\pi' * \boldsymbol{\sigma} * \boldsymbol{\tau} = \pi * \boldsymbol{\sigma} * \boldsymbol{\tau}$ since i and j are in different buckets for $\boldsymbol{\tau}$.

Since $\pi(i) < \pi(j)$, we have $\pi * \boldsymbol{\sigma}(i) < \pi * \boldsymbol{\sigma}(j)$. Further, $\boldsymbol{\tau}(i) < \boldsymbol{\tau}(j)$ since $\rho * \boldsymbol{\tau}^{\mathrm{R}}(j) < \rho * \boldsymbol{\tau}^{\mathrm{R}}(i)$ and $\rho * \boldsymbol{\tau}^{\mathrm{R}}$ is a refinement of the reverse of $\boldsymbol{\tau}$. Since $\boldsymbol{\tau}(i) < \boldsymbol{\tau}(j)$, we have $\pi * \boldsymbol{\sigma} * \boldsymbol{\tau}(i) < \pi * \boldsymbol{\sigma} * \boldsymbol{\tau}(j)$.

Either $\pi \in U$ or $\pi \in V$. Let us first examine the case that $\pi \in U$. Substituting $a = \pi * \boldsymbol{\sigma}(i)$, $b = \pi * \boldsymbol{\sigma}(j)$, $c = \pi * \boldsymbol{\sigma} * \boldsymbol{\tau}(i)$, $d = \pi * \boldsymbol{\sigma} * \boldsymbol{\tau}(j)$ in Lemma 17 gives us

$$\begin{aligned} &|\pi * \boldsymbol{\sigma}(i) - \pi * \boldsymbol{\sigma} * \boldsymbol{\tau}(i)| + |\pi * \boldsymbol{\sigma}(j) - \pi * \boldsymbol{\sigma} * \boldsymbol{\tau}(j)| \\ &\leq |\pi * \boldsymbol{\sigma}(i) - \pi * \boldsymbol{\sigma} * \boldsymbol{\tau}(j)| + |\pi * \boldsymbol{\sigma}(j) - \pi * \boldsymbol{\sigma} * \boldsymbol{\tau}(i)| \\ &= |\pi' * \boldsymbol{\sigma}(j) - \pi' * \boldsymbol{\sigma} * \boldsymbol{\tau}(j)| + |\pi' * \boldsymbol{\sigma}(i) - \pi' * \boldsymbol{\sigma} * \boldsymbol{\tau}(i)|, \end{aligned} \tag{19}$$

where the equality follows from the facts that (a) $\pi * \boldsymbol{\sigma}(i) = \pi' * \boldsymbol{\sigma}(j)$ and $\pi * \boldsymbol{\sigma}(j) = \pi' * \boldsymbol{\sigma}(i)$ since $\pi' * \boldsymbol{\sigma}$ is just $\pi * \boldsymbol{\sigma}$ with the adjacent elements i and j swapped, and (b) $\pi' * \boldsymbol{\sigma} * \boldsymbol{\tau} = \pi * \boldsymbol{\sigma} * \boldsymbol{\tau}$. Also, since $\pi' * \boldsymbol{\sigma}$ is just $\pi * \boldsymbol{\sigma}$ with the adjacent elements i and j swapped, $|\pi' * \boldsymbol{\sigma}(d) - \pi' * \boldsymbol{\sigma} * \boldsymbol{\tau}(d)| = |\pi * \boldsymbol{\sigma}(d) - \pi * \boldsymbol{\sigma} * \boldsymbol{\tau}(d)|$ for all $d \in D \setminus \{i, j\}$. If we sum over all d (where we make use of (19) for $d = i$ and $d = j$),

we obtain $F(\pi * \boldsymbol{\sigma}, \pi * \boldsymbol{\sigma} * \boldsymbol{\tau}) \leq F(\pi' * \boldsymbol{\sigma}, \pi' * \boldsymbol{\sigma} * \boldsymbol{\tau})$. Since $\pi \in U$, we have that $F(\rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}, \rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma} * \boldsymbol{\tau}) < F(\pi * \boldsymbol{\sigma}, \pi * \boldsymbol{\sigma} * \boldsymbol{\tau})$. Combining these last two inequalities, we obtain $F(\rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}, \rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma} * \boldsymbol{\tau}) < F(\pi' * \boldsymbol{\sigma}, \pi' * \boldsymbol{\sigma} * \boldsymbol{\tau})$. Therefore, $\pi' \in U$, and so $\pi' \in S$, which was to be shown.

We now examine the case that $\pi \in V$. From above, we see that $\pi' * \boldsymbol{\sigma} * \boldsymbol{\tau} = \pi * \boldsymbol{\sigma} * \boldsymbol{\tau}$, while $\pi' * \boldsymbol{\sigma}$ and $\pi * \boldsymbol{\sigma}$ differ only by swapping the adjacent elements i and j. Since, as shown above, $\pi' * \boldsymbol{\sigma}(i) > \pi' * \boldsymbol{\sigma}(j)$ while $\pi' * \boldsymbol{\sigma} * \boldsymbol{\tau}(i) < \pi' * \boldsymbol{\sigma} * \boldsymbol{\tau}(j)$, we see that there is exactly one more inversion between $\pi' * \boldsymbol{\sigma}$ and $\pi' * \boldsymbol{\sigma} * \boldsymbol{\tau}$ than between $\pi * \boldsymbol{\sigma}$ and $\pi * \boldsymbol{\sigma} * \boldsymbol{\tau}$. Hence, $K(\pi * \boldsymbol{\sigma}, \pi * \boldsymbol{\sigma} * \boldsymbol{\tau}) < K(\pi' * \boldsymbol{\sigma}, \pi' * \boldsymbol{\sigma} * \boldsymbol{\tau})$. By our assumption, $\pi \in V$, and so $K(\rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}, \rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma} * \boldsymbol{\tau}) < K(\pi * \boldsymbol{\sigma}, \pi * \boldsymbol{\sigma} * \boldsymbol{\tau})$. Combining these last two inequalities, we obtain $K(\rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}, \rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma} * \boldsymbol{\tau}) < K(\pi' * \boldsymbol{\sigma}, \pi' * \boldsymbol{\sigma} * \boldsymbol{\tau})$. Therefore, $\pi' \in V$, and so $\pi' \in S$, which was to be shown. □

We can now prove Theorem 5. We prove the theorem for F_{Haus}. The proof for K_{Haus} is analogous. Recall that

$$F_{\mathrm{Haus}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \max\left\{\max_{\sigma} \min_{\tau} F(\sigma, \tau), \max_{\tau} \min_{\sigma} F(\sigma, \tau)\right\},$$

where throughout this proof, σ and τ range through all full refinements of $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$, respectively. We show $\max_{\sigma} \min_{\tau} F(\sigma, \tau) = F(\rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}, \rho * \boldsymbol{\sigma} * \boldsymbol{\tau})$. A similar argument shows that $\max_{\tau} \min_{\sigma} F(\sigma, \tau) = F(\rho * \boldsymbol{\tau} * \boldsymbol{\sigma}, \rho * \boldsymbol{\sigma}^{\mathrm{R}} * \boldsymbol{\tau})$. The claim about F_{Haus} in the statement of the theorem follows easily.

Think for now of $\sigma \succeq \boldsymbol{\sigma}$ as fixed. Then by Lemma 19, the quantity $F(\sigma, \tau)$, where τ ranges over all full refinements of $\boldsymbol{\tau}$, is minimized when $\tau = \sigma * \boldsymbol{\tau}$. That is, $\min_{\tau} F(\sigma, \tau) = F(\sigma, \sigma * \boldsymbol{\tau})$.

By Lemma 20, the quantity $F(\sigma, \sigma * \boldsymbol{\tau})$, where σ ranges over all full refinements of $\boldsymbol{\sigma}$, is maximized when $\sigma = \rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}$. Hence, $\max_{\sigma} \min_{\tau} F(\sigma, \tau) = F(\rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}, \rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma} * \boldsymbol{\tau})$. Since $\rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma} * \boldsymbol{\tau} = \rho * \boldsymbol{\sigma} * \boldsymbol{\tau}$, we have $\max_{\sigma} \min_{\tau} F(\sigma, \tau) = F(\rho * \boldsymbol{\tau}^{\mathrm{R}} * \boldsymbol{\sigma}, \rho * \boldsymbol{\sigma} * \boldsymbol{\tau})$, as we wanted.

REFERENCES

[1] F. Aleskerov and B. Monjardet, *Utility Maximization, Choice, and Preference*, Stud. Econom. Theory 16, Springer-Verlag, Berlin, 2002.

[2] K. J. Arrow, *Social Choice and Individual Values*, John Wiley and Sons, New York, 1951.

[3] K. J. Arrow and M. D. Intriligator, *Handbook of Mathematical Economics*, North-Holland, Amsterdam, 1982.

[4] J. A. Aslam and M. Montague, *Models for metasearch*, in Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, 2001, pp. 276–284.

[5] K. A. Baggerly, *Visual Estimation of Structure in Ranked Data*, Ph.D. thesis, Rice University, Houston, TX, 1995.

[6] J. J. Bartholdi, C. A. Tovey, and M. A. Trick, *Voting schemes for which it can be difficult to tell who won the election*, Social Choice and Welfare, 6 (1989), pp. 157–165.

[7] D. S. Bridges and G. B. Mehta, *Representations of Preference Orderings*, Lecture Notes in Econom. and Math. Systems 422, Springer-Verlag, Heidelberg, Berlin, New York, 1995.

[8] W. W. Cohen, R. E. Schapire, and Y. Singer, *Learning to order things*, J. Artificial Intelligence Res., 10 (1999), pp. 243–270.

[9] D. E. Critchlow, *Metric Methods for Analyzing Partially Ranked Data*, Lecture Notes in Statist. 34, Springer-Verlag, Berlin, 1980.

[10] P. Diaconis, *Group Representation in Probability and Statistics*, IMS Lecture Series Monogr. Ser. 11, Institute of Mathematical Statistics, Hayward, CA, 1988.

[11] P. Diaconis and R. Graham, *Spearman's footrule as a measure of disarray*, J. Roy. Statist. Soc. Ser. B, 39 (1977), pp. 262–268.

[12] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, *Rank aggregation methods for the web*, in Proceedings of the 10th International World Wide Web Conference, Hong Kong, 2001, pp. 613–622.

[13] J. Eatwell, M. Milgate, and P. Newman, *The New Palgrave: A Dictionary of Economics*, MacMillan, London, 1987.

[14] R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee, *Comparing and aggregating rankings with ties*, in Proceedings of the 23rd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Denver, 2004, pp. 47–58.

[15] R. Fagin, R. Kumar, K. McCurley, J. Novak, D. Sivakumar, J. Tomlin, and D. Williamson, *Searching the workplace web*, in Proceedings of the 12th International World Wide Web Conference, Budapest, 2003, pp. 366–375.

[16] R. Fagin, R. Kumar, and D. Sivakumar, *Comparing top k lists*, SIAM J. Discrete Math., 17 (2003), pp. 134–160.

[17] R. Fagin, R. Kumar, and D. Sivakumar, *Efficient similarity search and classification via rank aggregation*, in Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, San Diego, 2003, pp. 301–312.

[18] P. Fishburn, *Condorcet social choice functions*, SIAM J. Appl. Math., 33 (1977), pp. 469–489.

[19] P. C. Fishburn, *Interval Orders and Interval Graphs: A Study of Partially Ordered Sets*, John Wiley and Sons, New York, 1985.

[20] L. A. Goodman and W. H. Kruskal, *Measures of association for cross classification*, J. Amer. Statist. Assoc., 49 (1954), pp. 732–764.

[21] T. H. Haveliwala, A. Gionis, D. Klein, and P. Indyk, *Evaluating strategies for similarity search on the web*, in Proceedings of the 11th International World Wide Web Conference, Honolulu, 2002, pp. 432–442.

[22] M. Kendall and J. D. Gibbons, *Rank Correlation Methods*, Edward Arnold, London, 1990.

[23] M. G. Kendall, *The treatment of ties in ranking problems*, Biometrika, 33 (1945), pp. 239–251.

[24] G. Lebanon and J. D. Lafferty, *Cranking: Combining rankings using conditional probability models on permutations*, in Proceedings of the 19th International Conference on Machine Learning, Sydney, Australia, 2002, pp. 363–370.

[25] M. Montague and J. A. Aslam, *Condorcet fusion for improved retrieval*, in Proceedings of the 11th International Conference on Information and Knowledge Management, McLean, VA, 2002, pp. 538–548.

[26] M. E. Renda and U. Straccia, *Web metasearch: Rank vs. score based rank aggregation methods*, in Proceedings of the 18th Annual Symposium on Applied Computing, Melbourne, FL, 2003, pp. 841–846.

[27] F. S. Roberts, *Measurement Theory, with Applications to Decisionmaking, Utility, and the Social Sciences*, Encyclopedia Math. Appl., Addison-Wesley, Reading, MA, 1979.

[28] J. Sese and S. Morishita, *Rank aggregation method for biological databases*, Genome Informatics, 12 (2001), pp. 506–507.

[29] R. R. Yager and V. Kreinovich, *On how to merge sorted lists coming from different web search tools*, Soft Computing Research Journal, 3 (1999), pp. 83–88.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 649–655

A COMBINATORIAL INTERPRETATION OF THE CHEBYSHEV POLYNOMIALS*

EMANUELE MUNARINI†

Abstract. We give a combinatorial interpretation of the Chebyshev polynomials in terms of the number of ideals of generalized fences and crowns.

Key words. Chebyshev polynomials, principle of inclusion-exclusion, generalized fences, generalized crowns, order ideals, multisets

AMS subject classifications. 05A15, 05A19, 06A07

DOI. 10.1137/S0895480103432283

1. Introduction. Chebyshev polynomials appear in several contexts, from analysis to combinatorics. Here we are interested in the combinatorial representations of these polynomials. For instance they are involved in the enumeration of certain permutations avoiding some patterns [10, 11, 6]. In [7] they arise by a commutative substitution into the *cd*-index of a special Eulerian partially ordered set. Similarly, in [8, 9] they are generalized by means of the *ce*-index of another special Eulerian partially ordered set.

In this paper we give a combinatorial interpretation of the Chebyshev polynomials which turns out to be much simpler than the ones recalled above. Such an interpretation is based on the enumeration of the ideals of certain posets which generalize fences (of even size) and crowns [12, 15]. The enumeration is obtained in an elementary way using the principle of inclusion-exclusion, generalizing the result obtained in [12] for ordinary fences and crowns.

Recall that an ideal of a poset P is a subset I such that for every $x, y \in P$, if $x \leq y$ and $y \in I$, then $x \in I$. Let $\mathcal{J}(P)$ be the set of all ideals of P and let $\mathcal{J}_k(P)$ be the set of all ideals of size k of P. The rank polynomial of $\mathcal{J}(P)$ is defined by

$$R(\mathcal{J}(P), x) = \sum_{k=0}^{|P|} |\mathcal{J}_k(P)|\, x^k.$$

Recall also [13, 14] that the Chebyshev polynomials of the first kind $T_n(x)$ are defined by the recurrence

$$T_{n+2}(x) = 2xT_{n+1}(x) - T_n(x) \tag{1}$$

with the initial conditions $T_0(x) = 1$, $T_1(x) = x$, while the Chebyshev polynomials of the second kind $U_n(x)$ are defined by the recurrence

$$U_{n+2}(x) = 2xU_{n+1}(x) - U_n(x) \tag{2}$$

with the initial conditions $U_0(x) = 1$, $U_1(x) = 2x$.

*Received by the editors July 28, 2003; accepted for publication (in revised form) October 20, 2004; published electronically September 5, 2006.
http://www.siam.org/journals/sidma/20-3/43228.html
†Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy (munarini@mate.polimi.it).

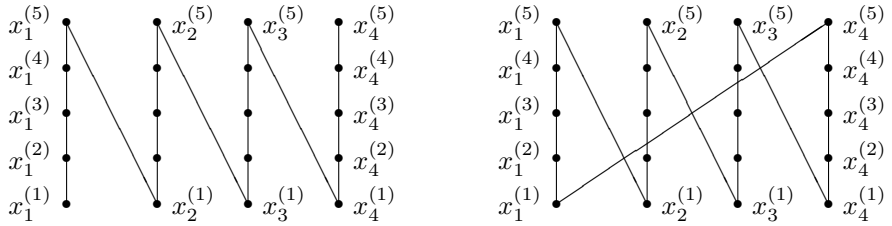


FIG. 1. (a) *The generalized fence* $\mathcal{F}_4^{(5)}$; (b) *the generalized crown* $\mathcal{C}_4^{(5)}$.

2. Ideals of generalized fences. Let $n,\ s \in \mathbb{N},\ s \neq 0$. The *generalized fence* $\mathcal{F}_n^{(s)}$ is the poset formed by $s\,n$ elements $x_1^{(1)}, \dots, x_1^{(s)}, \dots, x_n^{(1)}, \dots, x_n^{(s)}$ with cover relations $x_i^{(1)} < \cdots < x_i^{(s)}$ for $i = 1, \dots, n$, and $x_i^{(s)} > x_{i+1}^{(1)}$ for $i = 1, 2, \dots, n-1$ (see Figure 1(a) for an example).

To obtain an explicit form for the numbers $i_n^{(s)} = |\mathcal{J}(\mathcal{F}_n^{(s)})|$ and $i_{n,k}^{(s)} = |\mathcal{J}_k(\mathcal{F}_n^{(s)})|$ we first notice that the ideals of $\mathcal{F}_n^{(s)}$ are equivalent to particular multisets (a more elementary combinatorial structure) and then apply the principle of inclusion-exclusion.

Let $[n]$ be the set $\{1, 2, \dots, n\}$. Given an ideal I of $\mathcal{F}_n^{(s)}$ consider the multiset $\mu : [n] \to \mathbb{N}$ defined setting $\mu(i) = j$ when $x_i^{(j)} \in I$, $x_i^{(j+1)} \notin I$, and $\mu(i) = 0$ when $x_i^{(1)} \notin I$. This multiset is $(s+1)$-filtering, i.e., $\mu(i) \leq s$ for all $i \in [n]$. Since I is an ideal, μ satisfies the following condition:

$$\text{if} \quad \mu(i) = s \quad \text{then} \quad \mu(i+1) \neq 0, \quad \text{for all} \quad i = 1, 2, \dots, n-1\,. \tag{3}$$

Moreover, the order of μ, i.e., the sum $\mu(1) + \cdots + \mu(n)$, is equal to the size $|I|$ of the ideal I. For instance, the ideal $I = \{x_2^{(1)}, x_2^{(2)}, x_2^{(3)}, x_3^{(1)}\}$ of $\mathcal{F}_4^{(3)}$ corresponds to the multiset $\mu = 0310$. Let $M_n^{(s)}$ be the set of all $(s+1)$-filtering multisets on $[n]$ with property (3); similarly let $M_{n,k}^{(s)}$ be the set of all multisets in $M_n^{(s)}$ of order k.

It is easy to see that the correspondence $I \mapsto \mu$ just defined is a bijection between $\mathcal{J}(\mathcal{F}_n^{(s)})$ and $M_n^{(s)}$, and between $\mathcal{J}_k(\mathcal{F}_n^{(s)})$ and $M_{n,k}^{(s)}$. Hence $i_n^{(s)} = |M_n^{(s)}|$ and $i_{n,k}^{(s)} = |M_{n,k}^{(s)}|$.

Let A_i be the set of all $(s+1)$-filtering multisets μ on $[n]$ such that $\mu(i) = s$ and $\mu(i+1) = 0$. Then $M_n^{(s)} = A_1' \cap \cdots \cap A_{n-1}'$ (where the prime denotes complementation). Hence by the Sylvester formula [13, 15]

$$i_n^{(s)} = |A_1' \cap \cdots \cap A_{n-1}'| = \sum_{S \subseteq [n-1]} (-1)^{|S|} \left| \bigcap_{i \in S} A_i \right| .$$

Consider the set $A_S = \bigcap_{i \in S} A_i$, with $S \subseteq [n-1]$. If S contains two consecutive elements i and $i+1$, then $A_S = \emptyset$, since we have $0 = \mu(i+1) = s$ for every multiset μ in A_S and by hypothesis $s > 0$. On the contrary, if S is a sparse subset of $[n-1]$, i.e., it does not contain any two consecutive elements, then A_S is equivalent to the set of all $(s+1)$-filtering multisets on a set of $n-2|S|$ elements, since each $\mu \in A_S$ is already defined on i and $i+1$ for every $i \in S$. Hence we have the identity $|A_S| = (s+1)^{n-2|S|}$ which does not depend on the set S but only on its size. Since the number of all k-element sparse subsets of $[n-1]$ is given [13] by the binomial coefficient $\binom{n-k}{k}$, it

follows that

$$i_n^{(s)} = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} (-1)^k (s+1)^{n-2k} . \tag{4}$$

Next we want to determine a recurrence for the numbers $i_n^{(s)}$. Consider the poset $\mathcal{F}_{n+2}^{(s)}$ and the element $x_1^{(s)}$. An ideal I not containing $x_1^{(s)}$ is equivalent to a pair (I_1, I_2), where I_1 is an ideal of the chain $\{x_1^{(1)}, \ldots, x_1^{(s-1)}\}$ and I_2 is an ideal of the poset $G_{n+1}^{(s)}$ obtained by removing all the elements $x_1^{(1)}, \ldots, x_1^{(s)}$, which is isomorphic to $\mathcal{F}_{n+1}^{(s)}$. So there are $s \cdot i_{n+1}^{(s)}$ ideals of this kind. An ideal I containing $x_1^{(s)}$ contains also the elements $x_1^{(1)}, \ldots, x_1^{(s-1)}, x_2^{(1)}$. Hence I is equivalent to an ideal of $G_{n+1}^{(s)}$ containing $x_2^{(1)}$. The number of these ideals is given by the difference between the number of all ideals of $G_{n+1}^{(s)}$ and the number of all ideals of $G_{n+1}^{(s)}$ not containing $x_2^{(1)}$. But the ideals not containing $x_2^{(1)}$ do not contain any of the elements of the form $x_2^{(j)}$ and so are equivalent to the ideals of $\mathcal{F}_n^{(s)}$. Since $G_{n+1}^{(s)}$ and $\mathcal{F}_{n+1}^{(s)}$ are isomorphic, the number of all ideals in this second case is $i_{n+1}^{(s)} - i_n^{(s)}$. In conclusion we have the recurrence

$$i_{n+2}^{(s)} = (s+1)\, i_{n+1}^{(s)} - i_n^{(s)} . \tag{5}$$

Since the initial conditions are $i_0^{(s)} = 1$ and $i_1^{(s)} = s+1$, comparing (2) with (5), it follows that

$$i_n^{(s)} = U_n\left(\frac{s+1}{2}\right) . \tag{6}$$

In particular, setting $x = (s+1)/2$, from (6) and (4) we obtain the well-known [13] expansion of the Chebyshev polynomials of the second kind

$$U_n(x) = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} (-1)^k (2x)^{n-2k}.$$

A similar argument allows one to obtain $i_{n,k}^{(s)}$. It is sufficient to consider the sets A_i of all $(s+1)$-filtering multisets μ of order k on $[n]$ such that $\mu(i) = s$ and $\mu(i+1) = 0$. Again $A_S = \emptyset$ when S contains two consecutive elements and A_S is equivalent to the set of all $(s+1)$-filtering multisets of order $k - s|S|$ on a set of $n - 2|S|$ elements when S is sparse. Hence

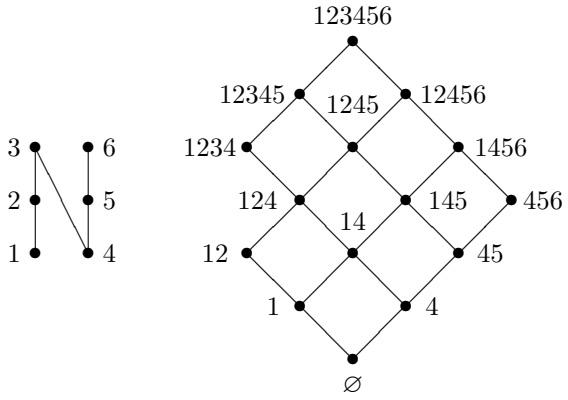
$$|A_S| = \binom{n-2|S|;\ s+1}{k-s|S|},$$

where $\binom{n;\, s+1}{k}$ is a polynomial coefficient and counts the $(s+1)$-filtering multisets of order k on a set of n elements [5]. In conclusion

$$i_{n,k}^{(s)} = \sum_{j \geq 0} \binom{n-j}{j} \binom{n-2j;\ s+1}{k-sj} (-1)^j . \tag{7}$$

Moreover, exactly as before, we can obtain the recurrence

$$i_{n+2,k+s}^{(s)} = i_{n+1,k+s}^{(s)} + \cdots + i_{n+1,k+1}^{(s)} + i_{n+1,k}^{(s)} - i_{n,k}^{(s)} . \tag{8}$$



$$F_2^{(3)}(x) = 1 + 2x + 3x^2 + 3x^3 + 3x^4 + 2x^5 + x^6\,, \qquad i_2^{(3)} = 15\,.$$

FIG. 2. *The generalized fence $\mathcal{F}_2^{(3)}$ and the associated lattice $\mathcal{J}(\mathcal{F}_2^{(3)})$.*

Using (7) and the identity [5]:

$$(1 + x + x^2 + \cdots + x^s)^n = \sum_{k=0}^{sn} \binom{n;\ s+1}{k} x^k\,, \tag{9}$$

it is straightforward to obtain

$$\sum_{k\geq 0} i_{n,k}^{(s)}\, x^k = (x^{s/2})^n \sum_{j\geq 0} \binom{n-j}{j} (-1)^j \frac{(1 + x + \cdots + x^s)^{n-2j}}{(x^{s/2})^{n-2j}}.$$

Then by (4) it follows that the rank polynomial $F_n^{(s)}(x) = \sum_{k\geq 0} i_{n,k}^{(s)}\, x^k$ of the lattice $J(\mathcal{F}_n^{(s)})$ is given by

$$F_n^{(s)}(x) = x^{sn/2}\, U_n\left(\frac{1 + x + \cdots + x^s}{2x^{s/2}}\right)\,. \tag{10}$$

It is also straightforward to obtain the recursion

$$F_{n+2}^{(s)}(x) = (1 + x + x^2 + \cdots + x^{s-1}) F_{n+1}^{(s)}(x) + x^s F_n^{(s)}(x)\,. \tag{11}$$

See Figure 2 for an example.

3. Ideals of generalized crowns. The *generalized crown* $\mathcal{C}_n^{(s)}$ is the poset obtained by $\mathcal{F}_n^{(s)}$ adding the cover relation $x_1^{(1)} < x_n^{(s)}$ (see Figure 1(b) for an example). To obtain $j_n^{(s)} = |\mathcal{J}(\mathcal{C}_n^{(s)})|$ and $j_{n,k}^{(s)} = |\mathcal{J}(\mathcal{C}_n^{(s)})|$ we proceed as in the previous case.

The ideals of $\mathcal{C}_n^{(s)}$ are equivalent to the $(s+1)$-filtering multisets on $[n]$ with the following property: if $\mu(i) = s$, then $\mu(i+1) \neq 0$ for all $i \in [n]$, where the indices are taken cyclically so that $\mu(n+1) = \mu(1)$. Let $N_n^{(s)}$ be the set of all these multisets and let $N_{n,k}^{(s)}$ be the set of all these multisets of order k. Hence $j_n^{(s)} = |N_n^{(s)}|$ and $j_{n,k}^{(s)} = |N_{n,k}^{(s)}|$.

Consider the set B_i of all the $(s+1)$-filtering multisets on $[n]$ such that $\mu(i) = s$ and $\mu(i+1) = 0$. Then $N_n^{(s)} = B_1' \cap \cdots \cap B_n'$ and

$$j_n^{(s)} = |B_1' \cap \cdots \cap B_n'| = \sum_{S\subseteq[n]} (-1)^{|S|} \left|\bigcap_{i\in S} B_i\right|\,.$$

Again the set $B_S = \bigcap_{i\in S} B_i$ is empty when S contains two consecutive elements (modulo n) and is isomorphic to the set of all the $(s+1)$-filtering multisets on a set of $n-2|S|$ elements otherwise. In the latter case $|B_S| = (s+1)^{n-2|S|}$. Since there are $\binom{n-k}{k}\frac{n}{n-k}$ k-element sparse subsets of $[n]$ (see [13]), it follows that

$$j_n^{(s)} = \sum_{k=0}^{\lfloor n/2\rfloor} \binom{n-k}{k} \frac{n}{n-k} (-1)^k (s+1)^{n-2k}. \tag{12}$$

To obtain a recurrence for these numbers, consider $\mathcal{C}_{n+2}^{(s)}$ and $x_1^{(s)}$. Given an ideal I of $\mathcal{C}_{n+2}^{(s)}$, then $x_1^{(s)} \notin I$ or $x_1^{(s)} \in I$. In the first case I is an ideal of the poset obtained by removing $x_1^{(s)}$. This poset is isomorphic to the generalized fence $\mathcal{F}_{n+2}^{(s)}$ with the element $x_{n+2}^{(s)}$ removed. Therefore I is an ideal of $\mathcal{F}_{n+2}^{(s)}$ not containing $x_{n+2}^{(s)}$ and so we have $i_{n+2}^{(s)} - i_{n+1}^{(s)}$ such ideals. In the second case, I contains $x_1^{(1)}, \ldots, x_1^{(s)}$ and $x_2^{(s)}$. By removing all the elements of the form $x_1^{(j)}$, we have that I is equivalent to an ideal of $\mathcal{F}_{n+1}^{(s)}$ containing $x_1^{(s)}$. It follows that there are $i_{n+1}^{(s)} - i_n^{(s)}$ such ideals. In conclusion, $j_{n+2}^{(s)} = (i_{n+2}^{(s)} - i_{n+1}^{(s)}) + (i_{n+1}^{(s)} - i_n^{(s)})$, that is

$$j_{n+2}^{(s)} = i_{n+2}^{(s)} + i_n^{(s)}.$$

Using recurrence (5) it is easy to see that the numbers $j_n^{(s)}$ satisfy the same recurrence, i.e.,

$$j_{n+2}^{(s)} = (s+1) j_{n+1}^{(s)} - j_n^{(s)}. \tag{13}$$

We do not define $\mathcal{C}_n^{(s)}$ for $n = 0$. However, we set $j_0^{(s)} = 2$ so that the sequence $\{j_n^{(s)}\}_n$ satisfies the recurrence (13) for each $n \geq 0$. Then, since the initial conditions are $j_0^{(s)} = 2$ and $j_1^{(s)} = s+1$, from (1) and (13) it follows that

$$j_n^{(s)} = 2\, T_n\left(\frac{s+1}{2}\right). \tag{14}$$

In particular, for $x = (s+1)/2$, by (14) and (12) we obtain the expansion [13] of the Chebyshev polynomials of the first kind

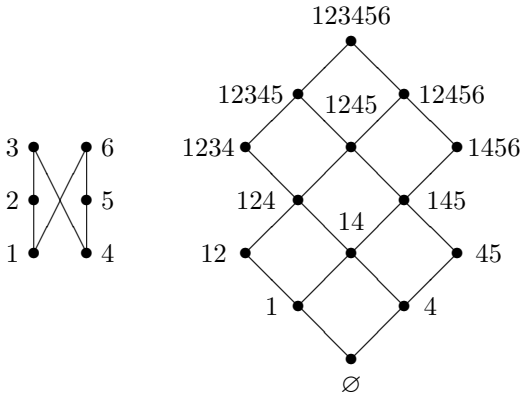
$$T_n(x) = \frac{1}{2} \sum_{k=0}^{\lfloor n/2\rfloor} \binom{n-k}{k} \frac{n}{n-k} (-1)^k\, (2x)^{n-2k}.$$

To obtain an explicit formula for the numbers $j_{n,k}^{(s)}$, consider the sets B_i of all $(s+1)$-filtering multisets μ of order k on $[n]$ such that $\mu(i) = s$ and $\mu(i+1) = 0$, for $i \in [n]$ (with $\mu(n+1) = \mu(1)$). Again, $B_S = \emptyset$ when S contains two consecutive elements (modulo n) and is equivalent to the set of all $(s+1)$-filtering multisets of order $k - s|S|$ on a set of $n - 2|S|$ elements when S is sparse. So

$$|B_S| = \binom{n - s|S|;\ s+1}{k - s|S|}$$

and

$$j_{n,k}^{(s)} = \sum_{j\geq 0} \binom{n-j}{j} \frac{n}{n-j} \binom{n - sj;\ s+1}{k - sj} (-1)^j. \tag{15}$$



$$C_2^{(3)}(x) = 1 + 2x + 3x^2 + 2x^3 + 3x^4 + 2x^5 + x^6\,, \qquad j_2^{(3)} = 14\,.$$

FIG. 3. *The generalized crown* $\mathcal{C}_2^{(3)}$ *and the associated lattice* $\mathcal{J}(\mathcal{C}_2^{(3)})$.

Finally, using identities (15), (9), and (12), the rank polynomial $\mathcal{C}_n^{(s)} = \sum_{j\geq 0} j_{n,k}^{(s)}\, x^k$ of the lattice $J(\mathcal{C}_n^{(s)})$ is given by

$$C_n^{(s)}(x) = 2\, x^{sn/2}\, T_n\left(\frac{1 + x + \cdots + x^s}{2x^{s/2}}\right)\,. \tag{16}$$

See Figure 3 for an example.

4. Concluding remarks. The poset $\mathcal{F}_n^{(s)}$ can be generalized as follows. Replace the kth column with a column of height s_k, for each k. In this way we obtain the generalized fence $\mathcal{F}_n^\sigma$, where $\sigma = (s_1, s_2, s_3, \dots)$. It is straightforward to obtain also in this case the recurrence

$$f_{n+2}^\sigma = (s_{n+2} + 1) f_{n+1}^\sigma - f_n^\sigma \tag{17}$$

with initial conditions $f_0^\sigma = 1$ and $f_1^\sigma = s_1 + 1$ for the numbers of all ideals of $\mathcal{F}_n^\sigma$. We still have a linear recurrence but with variable coefficients. Moreover, the principle of inclusion-exclusion is not useful since the cardinality of the sets A_S depends on S and not just on its size. However, recurrence (17) implies that the numbers f_n^σ can be expressed as continuants, i.e., as determinant of a tridiagonal matrix

$$\begin{vmatrix} s_1+1 & 1 & & & & \\ 1 & s_2+1 & 1 & & & \\ & 1 & s_3+1 & 1 & & \\ & & \cdots & \cdots & & \\ & & & s_{n-1}+1 & 1 \\ & & & 1 & s_n+1 \end{vmatrix} = f_n^\sigma.$$

These determinants appear in the context of Catalan-like numbers [1, 2]. Specifically, the Catalan-like numbers C_n^σ, where $\sigma = (s_0, s_1, \dots)$, are characterized as the unique sequence with Hankel determinants

$$\det[C_{i+j}^\sigma]_{i,j=0}^n = 1\,, \qquad \det[C_{i+j+1}^\sigma]_{i,j=0}^n = f_n^{\sigma-1},$$

where $\sigma - 1 = (s_0 - 1, s_1 - 1, \dots)$. In particular, when $\sigma = (s, s, s, \dots)$ we have that the Catalan-like numbers $C_n^{(s)}$ are characterized as the unique sequence with Hankel

determinants

$$\det[C^{\sigma}_{i+j}]^{n}_{i,j=0} = 1\,, \qquad \det[C^{\sigma}_{i+j+1}]^{n}_{i,j=0} = U_{n+1}(s/2).$$

In the enumeration of ideals of generalized fences and crowns sparse subsets of $\{1, 2, \ldots, n\}$ play a central role. As observed by the referee, these subsets play a crucial role also in the calculation of the *cd*-index [3, 4]. This connection seems to deserve to be explored more deeply.

REFERENCES

[1] M. Aigner, *Catalan-like numbers and determinants*, J. Combin. Theory Ser. A, 87 (1999), pp. 33–51.

[2] M. Aigner, *Catalan and other numbers: A recurrent theme*, in Algebraic Combinatorics and Computer Science, Springer Italia, Milan, 2001, pp. 347–390.

[3] L. J. Billera, R. Ehrenborg, and M. Readdy, *The cd-index of zonotopes and arrangements*, in Mathematical Essays in Honor of Gian-Carlo Rota, B. E. Sagan and R. P. Stanley, eds., Birkhäuser, Boston, 1998, pp. 23–40.

[4] L. J. Billera, S. K. Hsiao, and S. van Willigenburg, *Peak quasisymmetric functions and eulerian enumeration*, Adv. Math., 176 (2003), pp. 248–276.

[5] L. Comtet, *Advanced Combinatorics*, Reidel, Boston, 1974.

[6] E. S. Egge and T. Mansour, *Permutations which avoid* 1243 *and* 2143*, continued fractions, and Chebyshev polynomials*, Electron. J. Combin., 9 (2002/03), Research paper 7, 35 pp. (electronic).

[7] G. Hetyei, *Orthogonal polynomials represented by CW-spheres*, Electron. J. Combin., 11 (2004/06), no. 2, Research paper 4, 28 pp. (electronic).

[8] G. Hetyei, *Tchebyshev posets*, Discrete Comput. Geom., 32 (2004), pp. 493–520.

[9] G. Hetyei, *Matrices of formal power series associated to binomial posets*, J. Algebraic Combin., 22 (2005), pp. 65–104.

[10] T. Mansour, *Restricted* 132*-alternating permutations and Chebyshev polynomials*, Ann. Comb., 7 (2003), pp. 201–227.

[11] T. Mansour and Z. Stankova, 321*-polygon-avoiding permutations and Chebyshev polynomials*, Electron. J. Combin., 9 (2002/03), Research paper 5, 16 pp. (electronic).

[12] E. Munarini and N. Zagaglia Salvi, *On the rank polynomial of the lattice of order ideals of fences and crowns*, Discrete Math., 259 (2002), pp. 163–177.

[13] J. Riordan, *An Introduction to Combinatorial Analysis*, Princeton University Press, Princeton, NJ, 1978.

[14] T. J. Rivlin, *The Chebychev Polynomials*, John Wiley, New York, 1990.

[15] R. Stanley, *Enumerative Combinatorics*, Volume 1, Cambridge Stud. Adv. Math. 49, Cambridge University Press, Cambridge, UK, 1997.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 656–668

A NEW PERIODICITY LEMMA*

KANGMIN FAN†, SIMON J. PUGLISI‡, W. F. SMYTH†‡, AND ANDREW TURPIN§

Abstract. Given a string $\boldsymbol{x} = \boldsymbol{x}[1..n]$, a *repetition* of period p in $\boldsymbol{x}$ is a substring $\boldsymbol{u}^r = \boldsymbol{x}[i..i+rp-1]$, $p = |\boldsymbol{u}|$, $r \geq 2$, where neither $\boldsymbol{u} = \boldsymbol{x}[i..i+p-1]$ nor $\boldsymbol{x}[i..i+(r+1)p-1]$ is a repetition. The maximum number of repetitions in any string $\boldsymbol{x}$ is well known to be $\Theta(n \log n)$. A *run* or *maximal periodicity* of period p in $\boldsymbol{x}$ is a substring $\boldsymbol{u}^r\boldsymbol{t} = \boldsymbol{x}[i..i+rp+|\boldsymbol{t}|-1]$ of $\boldsymbol{x}$, where $\boldsymbol{u}^r$ is a repetition, $\boldsymbol{t}$ is a proper prefix of $\boldsymbol{u}$, and no repetition of period p begins at position $i-1$ of $\boldsymbol{x}$ or ends at position $i+rp+|\boldsymbol{t}|$. In 2000 Kolpakov and Kucherov [*J. Discrete Algorithms*, 1 (2000), pp. 159–186] showed that the maximum number $\rho(n)$ of runs in any string $\boldsymbol{x}$ is $O(n)$, but their proof was nonconstructive and provided no specific constant of proportionality. At the same time, they presented experimental data strongly suggesting that $\rho(n) < n$. Related work by Fraenkel and Simpson [*J. Combin. Theory Ser. A.*, 82 (1998), pp. 112–120] showed that the maximum number $\sigma(n)$ of *distinct* squares in any string $\boldsymbol{x}$ satisfies $\sigma(n) < 2n$, while experiment again encourages the belief that in fact $\sigma(n) < n$. In this paper, as a first step toward proving these conjectures, we present a periodicity lemma that establishes limitations on the number and range of periodicities that can occur over a specified range of positions in $\boldsymbol{x}$. We then apply this result to specify corresponding limitations on the occurrence of runs.

Key words. string, word, periodicity, square, repetition, run, maximal periodicity

AMS subject classification. 68R15

DOI. 10.1137/050630180

1. Introduction.

The study of strings began with an investigation of periodicity properties [23], and periodicity of various kinds still remains a central theme, important both in theory and practice—for example, in data compression, pattern matching, computational biology, and many other areas. In this paper we present results that specify restrictions on the nature and extent of periodic behavior in strings. Although these results are theoretical, their importance is very much a product of their practical application, as we explain below.

It will be convenient throughout to represent strings in boldface (for example, $\boldsymbol{x} = \boldsymbol{x}[1..n]$) and their lengths in italics (for example, $x = |\boldsymbol{x}|$).

If $\boldsymbol{w} = \boldsymbol{u}^r$ for some nonempty string $\boldsymbol{u}$ and some integer $r \geq 2$, then $\boldsymbol{w}$ is said to be a *repetition*. Further, a *repetition in* $\boldsymbol{x}$ is a substring $\boldsymbol{u}^r = \boldsymbol{x}[i..i+ru-1]$, $r \geq 2$, in $\boldsymbol{x}$, where $\boldsymbol{x}[i..i+u-1]$ is not a repetition and $\boldsymbol{x}[i..i+(r+1)u-1] \neq \boldsymbol{u}^{r+1}$. We call $\boldsymbol{u}$ the *generator* of the repetition, u its *period*, and r its *exponent*; and we represent it economically by an integer triple (i, u, r). In the early 1980s three quite different

*Received by the editors April 28, 2005; accepted for publication (in revised form) March 16, 2006; published electronically September 15, 2006. Preliminary versions of parts of this paper appeared in *Proceedings of the* 16*th Annual Symposium on Combinatorial Pattern Matching*, Lecture Notes in Comput. Sci. 3537, Springer-Verlag, Berlin, 2005, and in *Proceedings of the* 16*th Australasian Workshop on Combinatorial Algorithms*, University of Ballarat, Ballarat, Victoria, Australia, 2005.
http://www.siam.org/journals/sidma/20-3/63018.html

†Algorithms Research Group, Department of Computing and Software, McMaster University, Hamilton, ON L8S 4K1, Canada (fank@mcmaster.ca, smyth@mcmaster.ca, www.cas.mcmaster.ca/cas/research/algorithms.htm). The first and third authors were supported in part by grants from the Natural Sciences and Engineering Research Council of Canada.

‡Department of Computing, Curtin University, GPO Box U1987, Perth WA 6845, Australia (puglisi@computing.edu.au, smyth@computing.edu.au).

§Department of Computer Science and Information Technology, RMIT University, GPO Box 2476V, Melbourne V 3001, Australia (aht@cs.rmit.edu.au). The fourth author was supported by a grant from the Australian Research Council.

$O(x \log x)$ algorithms were published [2, 1, 17] for the computation of all the repetitions in a given string $\boldsymbol{x}$. In a sense these algorithms were all asymptotically optimal, since in [2] it was shown that in fact a Fibonacci string $\boldsymbol{f_n}$ contains $\Theta(f_n \log f_n)$ repetitions.

In [16] Main introduced a more compact encoding of repetitions: a *run* or *maximal periodicity* of period u in $\boldsymbol{x}$ was defined to be a substring $\boldsymbol{u}^r\boldsymbol{t} = \boldsymbol{x}[i..i+ru+t-1]$ of $\boldsymbol{x}$, where $\boldsymbol{u}^r$ is a repetition, $\boldsymbol{t}$ is a proper prefix of $\boldsymbol{u}$, and no repetition of period u begins at position $i-1$ of $\boldsymbol{x}$ or ends at position $i+ru+t$. $\boldsymbol{u}$ is called the *generator* of the run, $\boldsymbol{t}$ is called its *tail*, and a run is economically represented by a 4-tuple (i, u, r, t). Computing all the runs in $\boldsymbol{x}$ permits all the repetitions in $\boldsymbol{x}$ to be listed in an obvious way. Main [16] showed how to compute all the "leftmost" runs in $\boldsymbol{x}$ in time $\Theta(x)$, provided that the suffix tree [24, 18] and the Lempel–Ziv (LZ) factorization [14] of $\boldsymbol{x}$ were both available. In [4] it was shown that a suffix tree could be computed in linear time on an *indexed* (bounded integer) alphabet; since the LZ factorization is computable in linear time from the suffix tree, this meant that the overall worst-case time requirement of Main's algorithm was $\Theta(x)$ on an indexed alphabet. In [13] Kolpakov and Kucherov took matters a step further by extending Main's algorithm to also compute nonleftmost runs in $\boldsymbol{x}$ in time proportional to their number, and then by showing that the maximum number $\rho(x)$ of runs in any string $\boldsymbol{x}$ was at most

$$k_1 x - k_2 \log_2 x\sqrt{x}, \tag{1}$$

where k_1 and k_2 are positive constants. Thus, at least in principle, all the runs in $\boldsymbol{x}$ could be determined in linear time.

However, there is a problem with (1): The proof is nonconstructive and gives no information about the magnitude of the constants k_1 and k_2. Nevertheless Kolpakov and Kucherov provide convincing experimental evidence that

* $\rho(x) < x$;
* $\rho(x)$ is achieved by a cube-free string $\boldsymbol{x}$ on alphabet $\{a, b\}$;
* $\rho(x+1) \le \rho(x)+2$.

As far as we know, there are only two published works that address these fundamental questions of periodicity. In [7] an infinite family of strings $\boldsymbol{x}$ is constructed that is conjectured for sufficiently large x to achieve $\rho(x) < x$. This family thus provides a lower bound on $\rho(x)$. More recently, Rytter [21] has used interesting techniques to show that $\rho(n) \le 5n$, thus establishing an upper bound.

It was mentioned above that Main's algorithm computes all the leftmost runs in $\boldsymbol{x}$, that is, the leftmost occurrence of each distinct run, a collection that certainly includes the leftmost occurrence of each distinct square in $\boldsymbol{x}$. This suggests a connection with another well-known problem: the determination of $\sigma(x)$, the maximum number of distinct squares in any string $\boldsymbol{x}$, where again experiment strongly suggests that $\sigma(x) < x$. With this problem better progress has been made: Fraenkel and Simpson showed [6] that $\sigma(x) \le 2x-2$, a result recently proved somewhat more simply by Ilie [8], then later improved to $\sigma(x) \le 2x-\Theta(\log x)$ [9].

In order to show that in general $\rho(x) < x$ ($\sigma(x) < x$), it seems to be necessary to establish restrictions on the number of runs (squares) that can occur near a position in $\boldsymbol{x}$ at which one or two runs (squares) are already known to occur. Perhaps the most famous theoretical result available for such a purpose is the following "periodicity lemma."

LEMMA 1 (see [5]). *Let p and q be two periods of $\boldsymbol{x}$, and let $d = \gcd(p, q)$. If $p+q \le x+d$, then d is also a period of $\boldsymbol{x}$.*

Unfortunately this lemma provides no special information about runs or the squares with which runs must begin, and it places no restrictions on the positions at which periodic substrings may occur. To our knowledge the only result that provides such information is the following "three squares lemma."

LEMMA 2 (see [3, 15]). *Suppose* $\boldsymbol{u}$ *is not a repetition, and suppose* $\boldsymbol{w} \neq \boldsymbol{u}^j$ *for any* $j \geq 1$. *If* $\boldsymbol{u}^2$ *is a prefix of* $\boldsymbol{w}^2$, *in turn a proper prefix of* $\boldsymbol{v}^2$, *then* $w \leq v-u$.

Our main result in this paper is essentially a generalization of this result, which we call a "new periodicity lemma": We allow $\boldsymbol{w}$ to be offset by k positions from the start of $\boldsymbol{v}^2$, and we do not always require complete squares $\boldsymbol{v}^2$ and $\boldsymbol{w}^2$, only sufficiently long substrings of periods v and w. Moreover, as a corollary of our main result, we are able to specify exactly the periodic behavior in the string.

2. New periodicity lemma. In this section we prove results that establish restrictions on the squares that can occur in the neighborhood of positions in a string at which one or two squares already appear. We begin with three simple definitions.

DEFINITION 3. *A square* $\boldsymbol{u}^2$ *is said to be* irreducible *if* $\boldsymbol{u}$ *is not a repetition.*

DEFINITION 4. *A square* $\boldsymbol{u}^2$ *is said to be* regular *if no prefix of* $\boldsymbol{u}$ *is a square.*

DEFINITION 5. *A square* $\boldsymbol{u}^2$ *is said to be* minimal *if no proper prefix of* $\boldsymbol{u}^2$ *is a square.*

LEMMA 6. *If* $\boldsymbol{u}^2$ *is minimal, then* $\boldsymbol{u}^2$ *is regular; if* $\boldsymbol{u}^2$ *is regular, then* $\boldsymbol{u}^2$ *is irreducible.*

Proof. The proof of the first statement is immediate. To prove the second, observe that by Definition 4, no prefix of $\boldsymbol{u}$ is a square. Therefore $\boldsymbol{u}$ cannot be a repetition, and so by Definition 3 $\boldsymbol{u}^2$ is irreducible. □

The existence of a minimal square already imposes significant limitations on the nature of other squares that can exist, as the following result shows.

LEMMA 7. *If* $\boldsymbol{x} = \boldsymbol{u}^2$ *is minimal, then for all integers* $k \geq 0$ *and* $w \in u/2..u-1$,

(a) *if*

$$k + w \leq u, \quad k+3w \geq 2u, \tag{2}$$

$\boldsymbol{x}[k+1..k+2w]$ *is not a square;*

(b) *if*

$$k + w > u, \quad k+2w \leq 2u, \tag{3}$$

either $\boldsymbol{x}[k+1..k+2w]$ *is not a square or* $\boldsymbol{x}[w'+1..w'+u]$ *has period* $u-w$, *where*

$$w' = (k+w)-u.$$

Proof. Suppose that for some pair of integers k and w satisfying either (2) or (3), $\boldsymbol{x}[k+1..k+2w] = \boldsymbol{w}^2$.

First assume that $k = 0$. Then if (2) holds, either $w = u$, a contradiction, or else $w < u$, contradicting the minimality of $\boldsymbol{u}^2$. On the other hand, if (3) holds, then both $w > u$ and $w \leq u$ must hold, again a contradiction. Thus we can assume that $k \geq 1$.

(a) Suppose that (2) holds, let $w' = u-(k+w)$, and consider

$$\widehat{\boldsymbol{w}} = \boldsymbol{x}[1..w-w'] = \boldsymbol{x}[k+w'+1..k+w].$$

Since by (2)

$$(w-w') - (k+w') = k+3w-2u \geq 0,$$

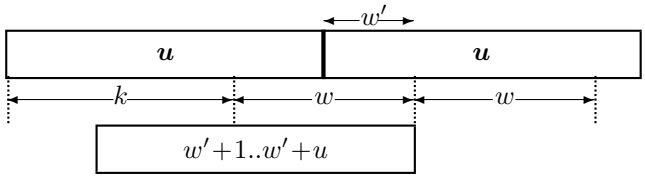


FIG. 1. *Lemma* 7(b).

the substring $\boldsymbol{x}[1..k+w]$ has period $k+w'$. Again by (2),

$$(k+w) - 2(k+w') = (k+w) - 2(u-w) \geq 0,$$

so that $\boldsymbol{x}[1..k+w]$ has prefix $\big(\boldsymbol{x}[1..k+w']\big)^2$, contradicting the minimality of $\boldsymbol{u}^2$. Thus in case (a) no such k and w can exist.

(b) Next we suppose that (3) holds so that $w < u$ and hence that $k–w' = u–w > 0$ (see Figure 1).
Consider

$$\boldsymbol{w} = \boldsymbol{x}[w'+1..w'+w] = \boldsymbol{x}[k+1..u+w'].$$

Since by (3) $w'+w = k+(2w-u) \geq k$, the substring $\boldsymbol{x}[w'+1..w'+u]$ of length u has period $k-w' = u-w$, as required. □

To show that in case (a) of Lemma 7 the assumption that $k+3w \geq 2u$ (as well as the weaker condition $w \geq u/2$) is necessary, consider the example $u = 14, k = 6, w = 5$:

$$\boldsymbol{x} = \boldsymbol{u}^2 = abbaba(babab)(bab\|ab)(babab)ababbab.$$

Here $\boldsymbol{w} = babab$, and $\boldsymbol{w}^3$ is a substring of $\boldsymbol{x}$.

To show that in case (b) of Lemma 7 the substring $\boldsymbol{w}^2$ can in fact exist, consider the example $u = 11, k = 4, w = 8$ with $w' = 1$:

$$\boldsymbol{x} = \boldsymbol{u}^2 = babc(abcabca\|b)(abcabcab)ca.$$

The substring $\boldsymbol{x}[2..12] = (abc)^3ab$ has period $u-w = 3$.

We turn now to the situation in which a regular square and an irreducible square occur at the same position. We first prove two basic lemmas that describe the relationship between regularity and irreducibility, and then go on to prove our main result.

LEMMA 8. *If $\boldsymbol{v}^2$ is irreducible with regular proper prefix $\boldsymbol{u}^2$, then*

$$v > \max\{u+1, 3u/2\}.$$

Proof. Observe that $1 \leq u < v$, and observe further that $u+1 \geq 3u/2$ if and only if $u \leq 2$.

For $u = 1$, $\boldsymbol{u}^2 = \lambda^2$ for some letter λ and the shortest irreducible square $\boldsymbol{v}^2 = (\lambda^2\mu)^2$ for some letter $\mu \neq \lambda$. Thus for $u = 1$, $v \geq 3 > u+1$, as required.

For $u = 2$, since $\boldsymbol{u}^2$ is regular, $\boldsymbol{u}^2 = (\lambda\mu)^2$ and the shortest irreducible square $\boldsymbol{v}^2 = (\lambda\mu\lambda\mu\nu)^2$ for some letter ν. Thus for $u = 2$, $v \geq 5 > u+1$, as required.

Suppose therefore that $u \geq 3$, and suppose further, without loss of generality, that $v < 2u$. Then

$$\boldsymbol{v} = \boldsymbol{u}\boldsymbol{u}[1..v-u] = \boldsymbol{u}[v-u+1..u]\boldsymbol{v}[2u-v+1..v],$$

where $\boldsymbol{y} = \boldsymbol{u}[1..v{-}u]$ of length $v{-}u$ is a prefix of $\boldsymbol{u}$, and hence of $\boldsymbol{v}$, and $\boldsymbol{z} = \boldsymbol{u}[v{-}u{+}1..u]$ of length $2u{-}v$ is a prefix of $\boldsymbol{v}$, and hence of $\boldsymbol{u}$. If we now assume $2v \leq 3u$, it follows that $v{-}u \leq 2u{-}v$, so that $\boldsymbol{y}$ is also a prefix of $\boldsymbol{z}$. Thus $\boldsymbol{u}$ has prefix $\boldsymbol{y}^2$ and so $\boldsymbol{u}^2$ cannot be regular, a contradiction. We conclude that $2v > 3u$, as required. ∎

Observe that if $\boldsymbol{u}^2$ is not regular, Lemma 8 may not hold: $\boldsymbol{u} = ababa$ allows $\boldsymbol{v} = ababaab$ with $v < 3u/2$.

LEMMA 9. *If $\boldsymbol{x} = \boldsymbol{v}^2$ is irreducible with regular proper prefix $\boldsymbol{u}^2$, $v < 2u$, then*

$$\boldsymbol{x} = \boldsymbol{u_1 u_2 u_1 u_1 u_2 u_1 u_2 u_1 u_1 u_2},$$

where $u_1 = 2u{-}v$, $u_2 = 2v{-}3u$.

Proof. Since $v < 2u$, $u \geq 3$ by Lemma 8. Let $\boldsymbol{u_1}$ be the suffix of $\boldsymbol{u}$ of length $u_1 = 2u{-}v$ that is a prefix of $\boldsymbol{v}$, and hence also a prefix of $\boldsymbol{u}$. By the regularity of $\boldsymbol{u}$ and Lemma 8, $u_1 < u/2$ and so $\boldsymbol{u} = \boldsymbol{u_1 u_2 u_1}$ for some nonempty $\boldsymbol{u_2}$. Then $\boldsymbol{v} = \boldsymbol{u_1 u_2 u_1 u_1 u_2}$, so that $u_2 = 2v{-}3u$, as required. ∎

For the proof of our main result, the following definitions will be helpful. If $\boldsymbol{x} = \boldsymbol{uv}$, $\boldsymbol{v}$ nonempty, then $\boldsymbol{vu} = R_u(\boldsymbol{x})$ is said to be the uth *rotation* of $\boldsymbol{x}$; also, if $\boldsymbol{u}$ is both a proper prefix and a suffix of $\boldsymbol{x}$, then it is said to be a *border* of $\boldsymbol{x}$.

We frequently make use of the following two well-known results.

LEMMA 10 (see [22, p. 76]). *Let $\boldsymbol{x}$ be a string of length n and minimum period p, and let $j \in 1..n{-}1$ be an integer. Then $R_j(\boldsymbol{x}) = \boldsymbol{x}$ if and only if $\boldsymbol{x}$ is a repetition and p divides j.*

LEMMA 11 (see [22, p. 76]). *If a string $\boldsymbol{x}$ is a repetition, then so is every rotation of $\boldsymbol{x}$.*

We first state the new periodicity lemma (NPL) in a rather general and easily understood form: Having gone through the proof, we will then be able to reexpress it to yield stronger conclusions based on weaker premises. A total of 14 cases arise in the proof (see Table 1). For each of these cases, we are able to identify a specific square prefix of $\boldsymbol{u}$ that is forced by the presence of $\boldsymbol{w}^2$ in the string $\boldsymbol{x}$, thus contradicting the assumption that $\boldsymbol{u}^2$ is regular; therefore, if $\boldsymbol{u}^2$ is *not* regular, the square prefix must exist.

For each of the main cases, we specify the range of values of k (either $k \in 0..u_1$ or $k \in u_1{+}1..u_1{+}u_2{-}1$) and the end position of $\boldsymbol{w}^{(1)}$ (first occurrence of $\boldsymbol{w}$) in $\boldsymbol{x}$. To facilitate this latter task, we introduce the notation $\boldsymbol{u_1}^{(j)}, \boldsymbol{u_2}^{(j)}$ to denote the jth occurrence of $\boldsymbol{u_1}, \boldsymbol{u_2}$, respectively, in $\boldsymbol{x}$. Thus "$\boldsymbol{w}^{(1)}$ ends in $\boldsymbol{u_2}^{(2)}$" means that the first occurrence of $\boldsymbol{w}$ in $\boldsymbol{x}$ ends in the second occurrence of $\boldsymbol{u_2}$ in $\boldsymbol{x}$. In most of the cases, it is useful to introduce a substring $\boldsymbol{s}$ that is both a prefix of $\boldsymbol{w}$ and a suffix of one of the substrings $\boldsymbol{u_1}^{(j)}$ or $\boldsymbol{u_2}^{(j)}$ in which $\boldsymbol{w}^{(1)}$ ends.

LEMMA 12 (NPL). *If $\boldsymbol{x}$ has regular prefix $\boldsymbol{u}^2$ and irreducible prefix $\boldsymbol{v}^2$, $u < v < 2u$, then for every $k \in 0..v{-}u{-}1$ and every $w \in v{-}u{+}1..v{-}1$, $w \neq u$, $\boldsymbol{x}[k{+}1..k{+}2w]$ is not a square.*

Proof. Suppose instead that for some k and w, $\boldsymbol{w}^2 = \boldsymbol{x}[k{+}1..k{+}2w]$. Recall the definitions of $\boldsymbol{u_1}$ and $\boldsymbol{u_2}$ given in Lemma 9, with $u_1{+}u_2 = v{-}u$.

A. $k \leq u_1$.

I. $\boldsymbol{w}^{(1)}$ ends in $\boldsymbol{u_1}^{(2)}$ ($k{+}w \leq u$, $s = u{-}(k{+}w)$).

(a) $\boldsymbol{w}^{(2)}$ ends in $\boldsymbol{u_1}^{(3)}$ ($k{+}2w \leq u{+}u_1$) (see Figure 2).

Define $\boldsymbol{q} = \boldsymbol{u_1}[1..q]$ and $\boldsymbol{z} = \boldsymbol{u_1}[1..z]$, which are both prefixes of $\boldsymbol{u_1}$ and suffixes of $\boldsymbol{w}$:

$$q = u_1{-}s = k{+}w{-}(u_1{+}u_2), \quad z = k{+}2w{-}u.$$

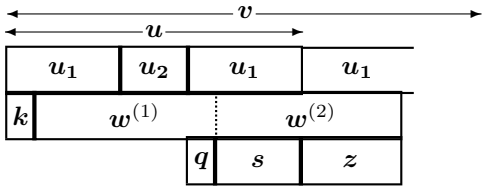


FIG. 2. *Case* A.I.(a).

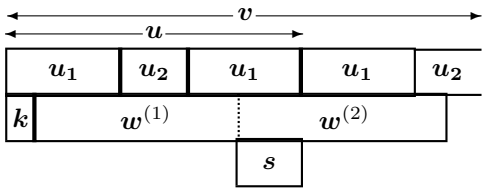


FIG. 3. *Case* A.I.(b).

Observe that

$$q-k = w-(u_1+u_2) > 0, \quad z-q = w-u_1 > 0.$$

Since $q < z$, $\boldsymbol{q}$ is a border of $\boldsymbol{z}$, and thus $\boldsymbol{z}$ has period $z-q$.

(i) $q \geq z/2$ ($k \geq u_2$). Here $\boldsymbol{z}$, and hence $\boldsymbol{u_1}$, has prefix

$$\boldsymbol{z}[1..z-q]^2 = \boldsymbol{z}[1..w-u_1]^2,$$

contradicting the regularity of $\boldsymbol{u}^2$.

(ii) $q < z/2$ ($k < u_2$). Here we can set $\boldsymbol{z} = \boldsymbol{qpq}$, where $p > 0$. Since $q > k$, we can also set $\boldsymbol{q} = \boldsymbol{kt}$, where, as noted above, $t = w-(u_1+u_2) > 0$. Hence $\boldsymbol{z} = \boldsymbol{ktpkt} = \boldsymbol{ktr}$ for $\boldsymbol{r} = \boldsymbol{pkt}$. Observe now that $\boldsymbol{tpkt}$ is a prefix of $\boldsymbol{w}^{(1)}$, while $\boldsymbol{r}$ is a prefix of $\boldsymbol{w}^{(2)}$. Thus $\boldsymbol{r} = R_t(\boldsymbol{r})$, so that by Lemmas 10 and 11, $\boldsymbol{r}$ and all of its rotations are repetitions of period t. It follows that $\boldsymbol{z}$, a prefix of $\boldsymbol{u_1}$, is a repetition of period $t = w-(u_1+u_2)$ and exponent at least 3, contradicting the regularity of $\boldsymbol{u}^2$.

(b) $\boldsymbol{w}^{(2)}$ ends in $\boldsymbol{u_2}^{(2)}$ ($k+2w > u+u_1$) (see Figure 3). Since $w > u_1+u_2$, $k+s < u_1$; since $w < u$, $k+s > 0$. Therefore $\boldsymbol{ks}$ is a prefix of $\boldsymbol{u_1}$, and since $\boldsymbol{su_1}$ is a prefix of $\boldsymbol{w}$, it follows that $\boldsymbol{u}$ has prefix $(\boldsymbol{ks})^2$, $k+s = u-w$, contradicting the regularity of $\boldsymbol{u}^2$.

II. $\boldsymbol{w}^{(1)}$ ends in $\boldsymbol{u_1}^{(3)}$ ($k+w \leq u+u_1,\ s = u+u_1-(k+w)$) (see Figure 4). Since $w \neq u$, $k+s \neq u_1$. Observe that $\boldsymbol{w}^{(1)}$ has prefix $R_k(\boldsymbol{u_1u_2})$, while $\boldsymbol{w}^{(2)}$ has prefix $R_{u_1-s}(\boldsymbol{u_1u_2})$. Since $u_1-s \neq k$ (otherwise $w = u$), it follows from Lemma 10 that $\boldsymbol{u_1u_2}$ is a repetition of period $|k-(u_1-s)| = |u-w|$, contradicting the regularity of $\boldsymbol{u}^2$. Note that if $\boldsymbol{u}^2$ is not regular, then $\boldsymbol{u}$ must also have period $|u-w|$.

III. $\boldsymbol{w}^{(1)}$ ends in $\boldsymbol{u_2}^{(2)}$ ($k+w \leq v,\ s = v-(k+w),\ k+s > 0$) (see Figure 5). $\boldsymbol{w}^{(1)}$ has prefix $R_k(\boldsymbol{u_1u_2})$, while $\boldsymbol{w}^{(2)}$ has prefix $R_t(\boldsymbol{u_1u_2})$, where $t = u_1+u_2-s$. Since $t = k+w-u > k$, it follows from Lemma 10 that $\boldsymbol{u_1u_2}$ is a repetition of period $t-k = w-u$, contradicting the regularity of $\boldsymbol{u}^2$. Note that if $\boldsymbol{u}^2$ is not regular, then $\boldsymbol{u}$ must also have period $w-u$.

IV. $\boldsymbol{w}^{(1)}$ ends in $\boldsymbol{u_1}^{(4)}$ ($k+w \leq 2u,\ s = 2u-(k+w)$) (see Figure 6).

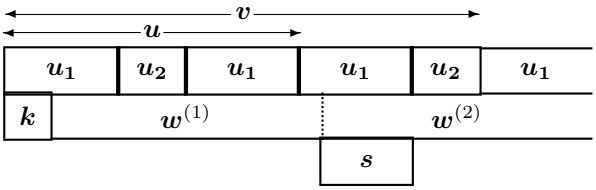


FIG. 4. *Case* A.II.

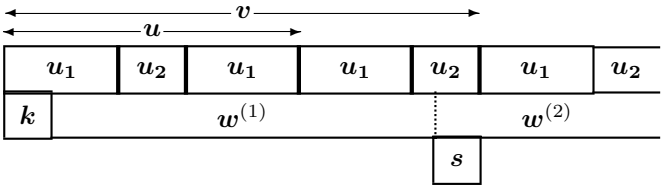


FIG. 5. *Case* A.III.

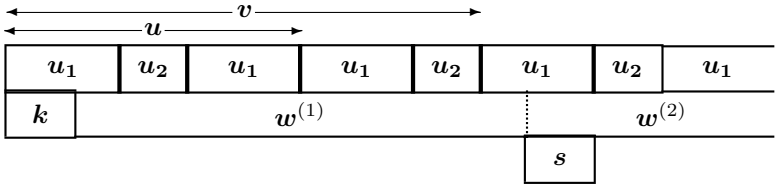


FIG. 6. *Case* A.IV.

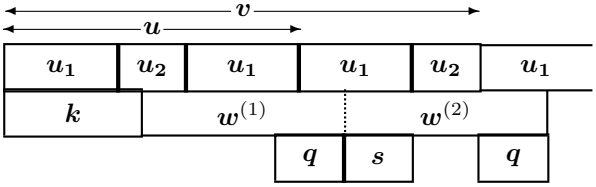


FIG. 7. *Case* B.I.(a).

As in case A.III., $\boldsymbol{w}^{(1)}$ has prefix $R_k(\boldsymbol{u_1u_2})$, while $\boldsymbol{w}^{(2)}$ has prefix $R_{u_1-s}(\boldsymbol{u_1u_2})$. It follows from Lemma 10 that $\boldsymbol{u_1u_2}$ is a repetition of period $k+s-u_1 = v-w$, contradicting the regularity of $\boldsymbol{u}^2$. Note that if $\boldsymbol{u}^2$ is not regular, then $\boldsymbol{u}$ must also have period $v-w$.

B. $k > u_1$.

I. $\boldsymbol{w}^{(1)}$ ends in $\boldsymbol{u_1}^{(3)}$ $(k+w \le u+u_1,\ s = u+u_1-(k+w),\ k+s < 2u_1)$.

(a) $\boldsymbol{w}^{(2)}$ ends in $\boldsymbol{u_1}^{(4)}$ $(k+2w \le 2u)$ (see Figure 7).

Let $\boldsymbol{q}$ be the prefix of $\boldsymbol{u_1}$ and suffix of $\boldsymbol{w}^{(2)}$ defined by

$$q = w-u_2-s = k+2w-v;$$

then, because it is a prefix of $\boldsymbol{u_1}$, $\boldsymbol{q}$ occurs at position $u+1$ of $\boldsymbol{x}$ and, because it is a suffix of $\boldsymbol{w}^{(1)}$, also at position $k+w-q+1$. These two copies of $\boldsymbol{q}$ are offset by period

$$t = u+q-(k+w) = w+u-v.$$

Since

$$\begin{aligned} q-2t &= k+2w-v-2w-2u+2v \\ &= k+v-2u > 0, \end{aligned}$$

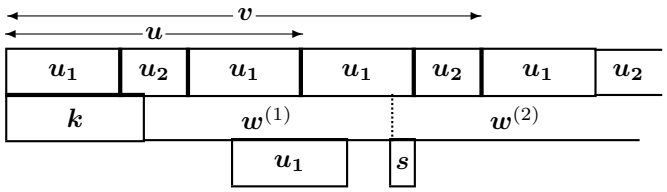


FIG. 8. *Case* B.I.(b).

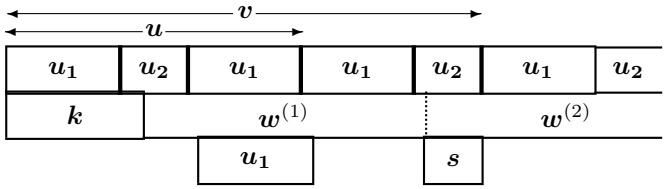


FIG. 9. *Case* B.II.(a).

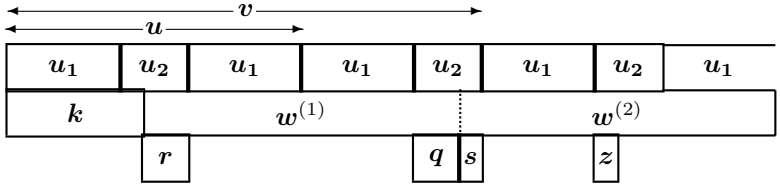


FIG. 10. *Case* B.II.(b)(i).

therefore $\boldsymbol{q}$, and hence $\boldsymbol{u_1}$, has a square prefix of period $w+u-v$, contradicting the regularity of $\boldsymbol{u}^2$.

(b) $\boldsymbol{w}^{(2)}$ ends in ${\boldsymbol{u_2}}^{(3)}$ $(k+2w > 2u)$ (see Figure 8).
Observe that since $\boldsymbol{u_1}$ occurs at position $s+u_2+1$ in $\boldsymbol{w}^{(2)}$, and since ${\boldsymbol{u_1}}^2$ begins at position u_1+u_2-k+1 in $\boldsymbol{w}^{(1)}$, therefore $\boldsymbol{u_1} = R_t(\boldsymbol{u_1})$ for $t = k+s-u_1 = u-w$. Hence by Lemma 10, $\boldsymbol{u_1}$ is a repetition of period $u-w$, contradicting the regularity of $\boldsymbol{u}^2$.

II. $\boldsymbol{w}^{(1)}$ ends in ${\boldsymbol{u_2}}^{(2)}$ $(k+w \le v,\ s = v-(k+w),\ k+s \ne u_1+u_2)$.

(a) $w < u$ $(k+s > u_1+u_2)$ (see Figure 9).
Observe that ${\boldsymbol{u_1}}^2$ occurs at position u_1+u_2-k+1 in $\boldsymbol{w}^{(1)}$, while $\boldsymbol{u_1}$ occurs at position $s+1$ in $\boldsymbol{w}^{(2)}$. Since $s > u_1+u_2-k$, this means that $\boldsymbol{u_1} = R_t(\boldsymbol{u_1})$ for $t = k+s-(u_1+u_2) = u-w > 0$. Hence $\boldsymbol{u_1}$ is a repetition of period $u-w$, contradicting the regularity of $\boldsymbol{u}^2$.

(b) $w > u$ $(k+s < u_1+u_2)$.

(i) $\boldsymbol{w}^{(2)}$ ends in ${\boldsymbol{u_1}}^{(5)}$ $(k+2w \le v+u,\ w-s \le u)$ (see Figure 10).
Let $\boldsymbol{r} = \boldsymbol{u_2}[k-u_1+1..u_2]$, where $r = u_1+u_2-k$ and $r-s = w-u > 0$. Observe that $\boldsymbol{w}^{(1)} = (\boldsymbol{r u_1})(\boldsymbol{u_1 q})$, where $\boldsymbol{q} = \boldsymbol{u_2}[1..u_2-s]$. Also $\boldsymbol{w}^{(2)}$ has prefix $\boldsymbol{s u_1 z}$, where $\boldsymbol{z} = \boldsymbol{u_2}[1..r-s]$, of length $r+u_1$. Since $w-s \le u$, the copy of $\boldsymbol{u}$ that begins at position $v+1$ of $\boldsymbol{x}$ has prefix $(\boldsymbol{u_1 z})(\boldsymbol{u_1 q})$, where $q-z = u_2-r > 0$. Thus $\boldsymbol{u}$ has prefix $(\boldsymbol{u_1 z})^2$ of period $u_1+r-s = w-(u_1+u_2)$, contradicting the regularity of $\boldsymbol{u}^2$.

(ii) $\boldsymbol{w}^{(2)}$ ends in ${\boldsymbol{u_1}}^{(6)}$ $(k+2w \le v+u+u_1,\ u < w-s \le u+u_1)$ (see Figure 11).
Observe that $\boldsymbol{w}^{(1)}$ has suffix $R_{u_2-s}(\boldsymbol{u_2}{\boldsymbol{u_1}}^2)$, while $\boldsymbol{w}^{(2)}$ has suffix

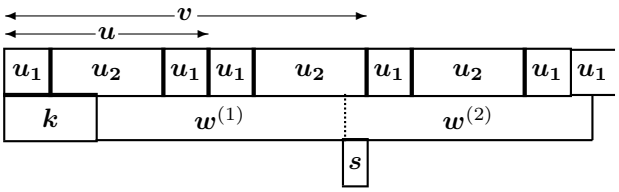


FIG. 11. *Case* B.II.(b)(ii).

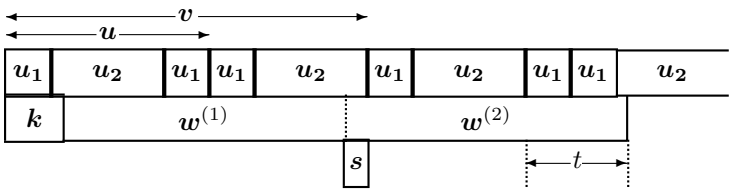


FIG. 12. *Case* B.II.(b)(iii).

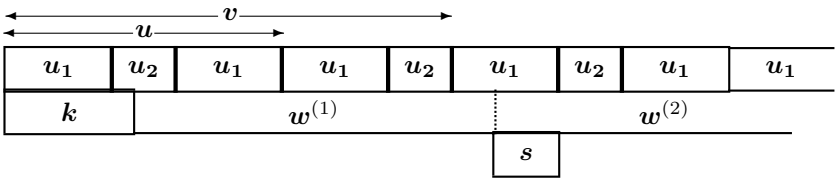


FIG. 13. *Case* B.III.

$R_t(\boldsymbol{u_2u_1}^2)$, where $t = u_1+u_2+w-s-u$. Since $t-(u_2-s) = w-(u_1+u_2) > 0$, it follows from Lemmas 10 and 11 that $\boldsymbol{u_2u_1}^2$, and hence $\boldsymbol{u}$, is a repetition of period $w-(u_1+u_2)$, contradicting the regularity of $\boldsymbol{u}^2$.

(iii) $\boldsymbol{w}^{(2)}$ ends in $\boldsymbol{u_2}^{(4)}$ $(k+2w < 2v,\ u+u_1 < w-s < v)$ (see Figure 12).
As in case B.II.(b)(ii), $\boldsymbol{w}^{(1)}$ has suffix $R_{u_2-s}(\boldsymbol{u_2u_1}^2)$, while now $\boldsymbol{w}^{(2)}$ has suffix $R_t(\boldsymbol{u_2u_1}^2)$, where $t = w-s-(u+u_1) > 0$. Since $u_2-s-t = v-w > 0$, it follows from Lemmas 10 and 11 that $\boldsymbol{u_2u_1}^2$, and hence $\boldsymbol{u}$, is a repetition of period $v-w$, contradicting the regularity of $\boldsymbol{u}^2$.

III. $\boldsymbol{w}^{(1)}$ ends in $\boldsymbol{u_1}^{(4)}$ $(k+w \le 2u,\ s = 2u-(k+w),\ k+s < u)$ (see Figure 13).
Observe that $\boldsymbol{w}^{(1)}$ has prefix $R_k(\boldsymbol{u})$, while $\boldsymbol{w}^{(2)}$ has prefix $R_{u_1-s}(\boldsymbol{u})$. Since $u_1 < k+s$, it follows by Lemma 10 that $\boldsymbol{u}$ is a repetition of period $k+s-u_1 = v-w$, contradicting the regularity of $\boldsymbol{u}^2$.
If $\boldsymbol{w}^{(2)}$ extends only to the end of $\boldsymbol{u_1}^{(5)}$, the argument of case B.II.(a) can instead be used to show that $\boldsymbol{u_1}$ is a repetition of period $v-w$, again contradicting the regularity of $\boldsymbol{u}^2$.

IV. $\boldsymbol{w}^{(1)}$ ends in $\boldsymbol{u_2}^{(3)}$ $(k+w \le 2u+u_2,\ s = 2u+u_2-(k+w))$ (see Figure 14).
The arguments of case B.III. apply: $\boldsymbol{u}$ (or $\boldsymbol{u_1}$) is a repetition of period $v-w$, contradicting the regularity of $\boldsymbol{u}^2$.

This completes the proof. □

In view of this result, and especially its proof, we realize that if $\boldsymbol{u}^2$ is not constrained to be regular, the existence of the three squares imposes severe conditions on

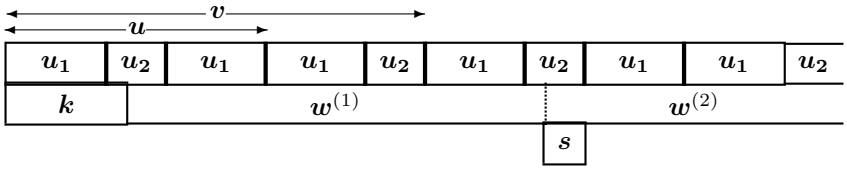


FIG. 14. *Case* B.IV.

the periodicity of $\boldsymbol{u}$, as shown in Table 1. In this table we specify, for each of the 14 cases identified in the proof, the prefix of $\boldsymbol{u}$ ($\boldsymbol{u_1}$, $\boldsymbol{u_1u_2}$, or $\boldsymbol{u}$ itself) that begins with a square, as well as the period of the square. We also indicate cases in which the entire prefix is in fact a repetition. Of course all copies of the prefix in $\boldsymbol{x}$ will have the same periodicity properties. Furthermore, in all cases (3–6 and 10–14) in which a period of $\boldsymbol{u}$ is identified, the periodicity lemma applies, since $\boldsymbol{u}$ also has period $u_1+u_2 = v-u$. For example, in cases 6 and 12–14 $\boldsymbol{u}$, hence $\boldsymbol{u}^2$, hence all of $\boldsymbol{v}^2$, will have period $\gcd(v-w, v-u)$; a similar result holds for cases 4–5 and 11. An alternative form of Lemma 12 may then be given as follows.

LEMMA 13. *Let* $\boldsymbol{u} = \boldsymbol{u_1u_2u_1}$, $\boldsymbol{v} = \boldsymbol{uu_1u_2}$, $\boldsymbol{u_1}$ *and* $\boldsymbol{u_2}$ *nonempty. If* $\boldsymbol{x} = \boldsymbol{v}^2 = \boldsymbol{ky}$, *where* $k \in 0..v-u-1$ *and* $\boldsymbol{y}$ *has period* $w \in v-u+1..v-1$, $w \neq u$, *then every occurrence of* $\boldsymbol{u}$ *in* $\boldsymbol{x}$ *is determined by cases* 1–6 *of Table* 1.

Observe that this result holds for *every* nonempty border $\boldsymbol{u_1}$ of $\boldsymbol{u}$ such that $u_1 < u/2$.

The rightmost column of Table 1 specifies the length of $\boldsymbol{x}$ that may be required in order to establish the periodicity of the prefix of $\boldsymbol{u}$. For example, in cases 1–3 not even all of $\boldsymbol{u}^2$ is required, and even in case 12 not all of $\boldsymbol{v}^2$ is required. This observation leads to the following weaker, but perhaps still interesting, corollary of the NPL that relates only to $\boldsymbol{u}^2$.

LEMMA 14. *Let* $\boldsymbol{u} = \boldsymbol{u_1u_2u_1}$, $\boldsymbol{u_1}$ *and* $\boldsymbol{u_2}$ *nonempty. If* $\boldsymbol{x} = \boldsymbol{u}^2\boldsymbol{u_1u_2} = \boldsymbol{ky}$, *where* $k \in 0..u_1$ *and* $\boldsymbol{y}$ *has period* $w \in u_1+u_2+1..u_1+u_2+u$, $w \neq u$, *then every occurrence of* $\boldsymbol{u}$ *in* $\boldsymbol{x}$ *is determined by Table* 1.

Again this result holds for every nonempty border $\boldsymbol{u_1}$ of $\boldsymbol{x}$.

We can state an equivalent of Lemma 12 for runs. Observe first that by definition every run is irreducible. Observe also that if a run of period u and tail $\boldsymbol{t}$ occurs at position i in $\boldsymbol{x}$, no run of the same period can occur at any position $j \in i..i+u+t$. Thus, if we define a *regular run* to be a run of generator $\boldsymbol{u}$ where $\boldsymbol{u}^2$ is a regular square, we can state the following lemma.

LEMMA 15. *Suppose* $\boldsymbol{x}$ *has a regular run of period* u *as prefix and another run of period* $v < 2u$ *as prefix. Then for every integer* $k \in 0..v-u-1$ *and for every* $w \in u..v$, *no run of period* w *(other than, for* $k = 0$, *the two given runs) occurs at position* $k+1$ *of* $\boldsymbol{x}$.

Finally, we remark that Lemmas 12 and 15 apply only trivially to the cases $u = 1$ and $u = 2$. As noted earlier for $u = 1$, $v \geq 3 > 2u$, while for $u = 2$, $v \geq 5 > 2u$, contrary to the requirement of the lemmas that $v < 2u$. However, for all $u \geq 3$, the hypothesis of the lemmas can be satisfied—for example, if $\boldsymbol{u} = aba$ of length 3, $\boldsymbol{v}$ may be *abaab* of length $5 < 2 \times 3$. More generally, we may think of such squares $\boldsymbol{v}^2$ as being "small," in contrast to those of period greater than $2u$ that are "large"; thus Lemmas 12 and 15 restrict the occurrences of squares/runs when the second square at some position is small. Note also that if $\boldsymbol{u}^2$ is in fact minimal (hence by Lemma 6 regular), then the irreducible square $\boldsymbol{v}^2$ must be regular.

TABLE 1
Periodicity table for k, $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{w}$ (unconstrained $\boldsymbol{u}$).

Case	$k \leq$	$k+w \leq$	Subcase	Sub subcase	Square prefix	Repetition?	Period	Required x
1	u_1	u	$k+2w \leq u+u_1$	$k \geq u_2$	$\boldsymbol{u_1}$		$w-u_1$	$u+u_1$
2				$k < u_2$	$\boldsymbol{u_1}$		$w-v+u$	$u+u_1$
3			$k+2w > u+u_1$		$\boldsymbol{u}$		$u-w$	v
4		$u+u_1$			$\boldsymbol{u_1 u_2}$ & $\boldsymbol{u}$	yes	$\lvert u-w \rvert$	$2u$
5		v			$\boldsymbol{u_1 u_2}$ & $\boldsymbol{u}$	yes	$w-u$	$2v-u$
6		$2u$			$\boldsymbol{u_1 u_2}$ & $\boldsymbol{u}$	yes	$v-w$	$v+u$
7	u_1+u_2-1	$u+u_1$	$k+2w \leq 2u$		$\boldsymbol{u_1}$		$w-v+u$	$2u$
8			$k+2w > 2u$		$\boldsymbol{u_1}$	yes	$u-w$	$2v-u$
9		v	$w < u$		$\boldsymbol{u_1}$	yes	$u-w$	$2v-u$
10			$w > u$	$k+2w \leq v+u$	$\boldsymbol{u}$		$w-v+u$	$v+u$
11				$k+2w \leq 2v-u_2$	$\boldsymbol{u}$	yes	$w-v+u$	$2v-u_2$
12				$k+2w < 2v$	$\boldsymbol{u}$	yes	$v-w$	$2v-1$
13		$2u$			$\boldsymbol{u_1}$ & $\boldsymbol{u}$	yes	$v-w$	$v+u$
14		$2u+u_2$			$\boldsymbol{u_1}$ & $\boldsymbol{u}$	yes	$v-w$	$v+u$

3. Discussion. We have proved two main lemmas (Lemmas 7 and 12) that restrict the periods w of squares that can occur at positions $i+k$ in $\boldsymbol{x}$ when at position i either one (Lemma 7) or two (Lemma 12) squares are known to occur. It seems that, with the exception of [15, Lemma 8.1.14], such properties have not been studied previously. In particular, we hope that with the help of Lemma 12, it will be possible to establish, or at least make progress with, the three conjectures arising out of [13].

The Main/Kolpakov–Kucherov algorithm [16, 13] is the only known linear-time algorithm for computing all the runs in a given string $\boldsymbol{x}$. It is complex and, until recently, depended for its worst-case linear behavior on the use of Farach's algorithm [4], also complex and not space-efficient, for linear-time computation of suffix trees. Since 2003 three worst-case linear-time suffix array construction algorithms [10, 11, 12] have been available for use in the computation of the LZ factorization, but even after the substitution of suffix arrays for suffix trees in the all-runs algorithm, significant complications remain. For instance, it seems clear [19, 20] that due to their recursive nature the linear-time algorithms are not in practice the fastest suffix array construction algorithms available. We hope that, with a more precise understanding of the periodicity of runs, it will become possible to design simpler algorithms that will compute all the runs in a string in a more direct and efficient manner.

Acknowledgment. The authors thank a referee for suggestions that have materially improved the presentation.

REFERENCES

[1] A. Apostolico and F. P. Preparata, *Optimal off-line detection of repetitions in a string*, Theoret. Comput. Sci., 22 (1983), pp. 297–315.

[2] M. Crochemore, *An optimal algorithm for computing the repetitions in a word*, Inform. Process. Lett., 12 (1981), pp. 244–250.

[3] M. Crochemore and W. Rytter, *Squares, cubes, and time-space efficient strings searching*, Algorithmica, 13 (1995), pp. 405–425.

[4] M. Farach, *Optimal suffix tree construction with large alphabets*, in Proceedings of the 38th Annual IEEE Symposium on Foundation of Computer Science, 1997, pp. 137–143.

[5] N. J. Fine and H. S. Wilf, *Uniqueness theorems for periodic functions*, Proc. Amer. Math. Soc., 16 (1965), pp. 109–114.

[6] A. S. Fraenkel and R. J. Simpson, *How many squares can a string contain?*, J. Combin. Theory Ser. A, 82 (1998), pp. 112–120.

[7] F. Franek, R. J. Simpson, and W. F. Smyth, *The maximum number of runs in a string*, M. Miller and K. Park, eds., in Proceedings of the 14th Annual Australasian Workshop on Combinatorial Algorithms, 2003, pp. 26–35.

[8] L. Ilie, *A simple proof that a word of length n has at most $2n$ distinct squares*, J. Combin. Theory Ser. A, 112 (2005), pp. 163–164.

[9] L. Ilie, *A note on the number of distinct squares in a word*, in Proceedings of the 5th Annual International Conference on Combinatorics on Words, S. Brlek and C. Reutenauer, eds., 2005, pp. 289–294.

[10] J. Kärkkäinen and P. Sanders, *Simple linear work suffix array construction*, in Proceedings of the 30th Annual International Conference on Automata, Languages, and Programming, 2003, pp. 943–955.

[11] D. K. Kim, J. S. Sim, H. Park, and K. Park, *Linear-time construction of suffix arrays*, in Proceedings of the 14th Annual Symposium on Combinatorial Pattern Matching, Lecture Notes in Comput. Sci. 2676, R. Baeza-Yates, E. Chávez, and M. Crochemore, eds., Springer-Verlag, Berlin, 2003, pp. 186–199.

[12] P. Ko and S. Aluru, *Space efficient linear time construction of suffix arrays*, in Proceedings of the 14th Annual Symposium on Combinatorial Pattern Matching, Lecture Notes in Comput. Sci. 2676, R. Baeza-Yates, E. Chávez, and M. Crochemore, eds., Springer-Verlag, Berlin, 2003, pp. 200–210.

[13] R. Kolpakov and G. Kucherov, *On maximal repetitions in words*, J. Discrete Algorithms, 1 (2000), pp. 159–186.
[14] A. Lempel and J. Ziv, *On the complexity of finite sequences*, IEEE Trans. Inform. Theory, 22 (1976), pp. 75–81.
[15] M. Lothaire, *Algebraic Combinatorics on Words*, Cambridge University Press, Cambridge, UK, 2002.
[16] M. G. Main, *Detecting leftmost maximal periodicities*, Discrete Appl. Math., 25 (1989), pp. 145–153.
[17] M. G. Main and R. J. Lorentz, *An $O(n \log n)$ algorithm for finding all repetitions in a string*, J. Algorithms, 5 (1984), pp. 422–432.
[18] E. M. McCreight, *A space-economical suffix tree construction algorithm*, J. ACM, 23 (1976), pp. 262–272.
[19] S. J. Puglisi, W. F. Smyth, and A. Turpin, *The performance of linear time suffix sorting algorithms*, in Proceedings of the IEEE Data Compression Conference, J. Storer and M. Cohn, eds., 2005, pp. 358–367.
[20] S. J. Puglisi, W. F. Smyth, and A. Turpin, *A taxonomy of suffix array construction algorithms*, ACM Comput. Surv., to appear.
[21] W. Rytter, *The number of runs in a string: Improved analysis of the linear upper bound*, in Proceedings of the 23rd Symposium on Theoretical Aspects of Computer Science, B. Durand and W. Thomas, eds., Lecture Notes in Comput. Sci. 2884, Springer-Verlag, Berlin, 2006, pp. 184–195.
[22] B. Smyth, *Computing Patterns in Strings*, Addison-Wesley, Reading, MA, 2003.
[23] A. Thue, *Über unendliche zeichenreihen*, Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiana, 7 (1906), pp. 1–22.
[24] P. Weiner, *Linear pattern matching algorithms*, in Proceedings of the 14th Annual IEEE Symposium on Switching and Automata Theory, 1973, pp. 1–11.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 669–681

CYCLE EXTENDABILITY AND HAMILTONIAN CYCLES IN CHORDAL GRAPH CLASSES*

ATIF ABUEIDA† AND R. SRITHARAN‡

Abstract. A cycle C in a graph is *extendable* if there exists a cycle C' such that $V(C) \subseteq V(C')$ and $| V(C') | = | V(C) | + 1$. A graph is *cycle extendable* if every non-Hamiltonian cycle in the graph is extendable. An unresolved question is whether or not every Hamiltonian chordal graph is cycle extendable. We show that Hamiltonian graphs in classes such as interval, split, and in some subclasses of strongly chordal graphs, are cycle extendable. We also address efficiently finding a Hamilton cycle in some cases. A unifying theme to our approach is the use of appropriate vertex elimination orders.

Key words. cycle, Hamiltonian, extendability, chordal graph

AMS subject classification. 05C38

DOI. 10.1137/S0895480104441267

1. Introduction.

The graphs that we consider are undirected and simple. A cycle C in a graph is *extendable* if there exists a cycle C' such that $V(C) \subseteq V(C')$ and $| V(C') | = | V(C) | + 1$. In this situation, we say C *extends to* C'. We also refer to C' as *an extension of* C. A graph is *cycle extendable* [6, 7] if every non-Hamiltonian cycle in the graph is extendable. Hendry [6, 7] raised the question of whether every Hamiltonian *chordal* graph is cycle extendable: A graph is chordal if every cycle with at least four vertices in the graph has a chord. The class of chordal graphs forms a well-studied [1] subclass of the class of perfect graphs.

Jiang [9] proved that every Hamiltonian chordal graph that is also planar is cycle extendable. The question of whether every Hamiltonian chordal graph is cycle extendable remains open. We show that Hamiltonian graphs in classes such as interval, split, and in some subclasses of strongly chordal graphs, are cycle extendable. We also address the problem of efficiently finding a Hamilton cycle in some subclasses of chordal graphs. It is known [1] that deciding whether a given graph is Hamiltonian is NP-complete even when the input belongs to severely restricted classes of chordal graphs.

A proof due to Chen et al. [2] that a Hamiltonian interval graph is cycle extendable appears in a companion paper in this issue of the journal. The technique employed in [2] is to derive a contradiction to the fact [2, 11] that an interval graph is Hamiltonian if and only if it is 1-tough. In contrast, our techniques rely on the theory of classes of chordal graphs. In particular, we make use of presence of vertices with special properties and elimination order of vertices with special properties. We use such properties to rearrange vertices on cycles so that they become conducive to our inductive arguments. Our inductive proofs employ deletion of vertices and deletion of edges, as well as addition of edges in some cases. In general, deletion and addition

*Received by the editors February 24, 2004; accepted for publication (in revised form) December 28, 2005; published electronically September 15, 2006. This research was supported by the Research Council, University of Dayton.
http://www.siam.org/journals/sidma/20-3/44126.html
†Department of Mathematics, The University of Dayton, Dayton, OH 45469 (atif.abueida@notes.udayton.edu).
‡Department of Computer Science, The University of Dayton, Dayton, OH 45469 (srithara@notes.udayton.edu).

of edges need not preserve membership in classes of chordal graphs. Hence, we first develop tools needed for these.

We next present some of the notation used in the paper. We then present relevant background, and the tools needed for our proofs, for the classes of graphs we study. The results on extending cycles follow. We conclude by presenting consequences of some of our proofs in regard to efficiently deciding whether a given graph is Hamiltonian and finding a Hamilton cycle.

2. Notation. Let $G = (V, E)$ be a graph. For a subset S of vertices of G, we use $G[S]$ to denote the subgraph of G induced by S. For vertices u, v, we use *u sees v* for *u is adjacent to v*, and *u misses v* for *u is not adjacent to v*. For a vertex v, $N(v)$ denotes the set of vertices adjacent to v. The *closed neighborhood of v*, $N[v]$, is $\{v\} \cup N(v)$. For vertex v, $d(v)$ denotes the degree of v, the number of vertices adjacent to v. We use n to refer to the number of vertices in G, m to refer to the number of edges in G, and $||G||$ to refer to $m + n$. $G - uv$ denotes the graph obtained from G by deleting the edge uv. Similarly, for nonadjacent vertices u, v, $G + uv$ is the graph obtained from G by adding the edge uv. For graphs G and F, we use *G contains F* or *G has F* to mean G has an induced subgraph isomorphic to F.

We introduce other definitions and notation as we need them. We present some relevant graphs in Figure 1.

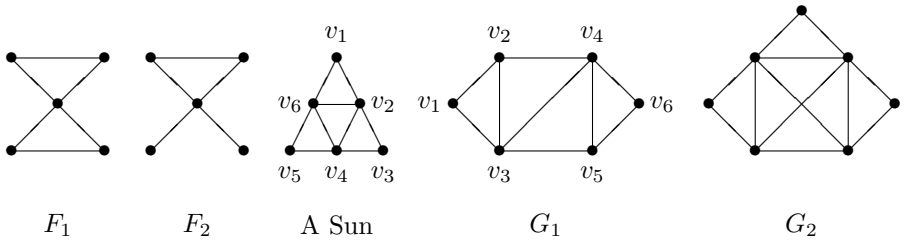


FIG. 1. *Some specific graphs.*

3. Background and tools. The classes of graphs that we deal with are *hereditary;* i.e., if a graph is in the class, then every induced subgraph of the graph is also in the class. However, some of our proofs, which are based on induction on $||G||$ (sum of the number of vertices and the number of edges in G), involve deletion of vertices, deletion of edges, and addition as well as deletion of edges. As deletion or addition of an arbitrary edge may not preserve membership in the class, special tools are needed for addition and deletion of edges. In this regard, we also provide the relevant tools for each class of graphs.

For vertices $v_1, v_2, \ldots, v_k$, by *$v_1 v_2 \ldots v_k$ is a segment of cycle C (path P)*, we mean that v_1 through v_k occur consecutively along C (P). For a cycle C in a graph such that $v \in V(C)$ and $|V(C)| \geq 4$, let uvw be a segment of C. In the case that u sees w, by C_{-v} we mean the cycle obtained from C by replacing the segment uvw with the segment uw.

Vertex v in a graph is *simplicial* if $N(v)$ induces a clique. It is well known [5] that graph G is chordal if and only if for every induced subgraph H of G, either H is a clique or there exist two nonadjacent simplicial vertices of H. A vertex is simplicial if and only if it belongs to exactly one maximal clique. Also, deletion of an edge incident on a simplicial vertex of a chordal graph results in a chordal graph.

3.1. Strongly chordal graphs. A graph is *strongly chordal* if it is chordal and every cycle in the graph on $2k$ vertices, $k \geq 3$, has a chord uv such that each segment of the cycle from u to v has an odd number of edges. Farber introduced [3] the class of strongly chordal graphs and gave several characterizations for the class; we need some definitions before we can present his theorem.

A vertex x in a graph is *simple* if it is simplicial, and also the vertices in $N(x)$ can be ordered as $x_1x_2 \ldots x_k$ such that $N[x_1] \subseteq N[x_2] \subseteq \cdots \subseteq N[x_k]$. For a simple vertex x, given such an ordering of its neighbors, we say $x_i < x_j$ *in* $N(x)$ whenever $i < j$. We also refer to x_1 and x_2 as the *smallest and second smallest vertices in* $N(x)$, respectively.

A *Sun* is a graph formed on the vertex-set $\{v_1, v_2, \ldots, v_{2k}\}$, $k \geq 3$, by starting with the cycle $v_1v_2 \ldots v_{2k}v_1$, and then making the vertex-subset $\{v_2, v_4, \ldots, v_{2k}\}$ induce a clique (see Figure 1). For a graph G, let $\mathcal{R} = v_1v_2 \ldots v_n$ be an ordering of vertices of G. Let $G_i = G[\{v_i, v_{i+1}, \ldots, v_n\}]$. $\mathcal{R}$ is a *simple elimination order* for G if v_i is simple in G_i, $1 \leq i \leq n$. $\mathcal{R}$ is a *strong elimination order* for G if $\mathcal{R}$ is a simple elimination order and, for every $i < j < k$ such that v_i sees v_j and v_k, $N[v_j] \subseteq N[v_k]$ in G_i. The following theorem is due to Farber [3].

THEOREM 3.1 (see [3]). *The following are equivalent for any graph G:*

- *G is strongly chordal.*
- *G is chordal and does not contain a Sun.*
- *Vertices of G admit a simple elimination order.*
- *Vertices of G admit a strong elimination order.*

The following lemma will be repeatedly used to rearrange neighbors of a simple vertex on a given cycle.

LEMMA 3.2. *Let G be a strongly chordal graph and x be a simple vertex of G. Suppose C is a cycle such that uxv is a segment of C, and xw is a chord of C such that $w < v$ in $N(x)$. Then there exists a cycle C' of G such that $V(C) = V(C')$ and uxw is a segment of C'.*

Proof of Lemma 3.2. Assume that in the clockwise direction along C, u, x, and v appear consecutively. Let z be the vertex next to w on C in the clockwise direction. Such a z exists as the segment $w \ldots x$ of C in the clockwise direction contains u. As $N[w] \subseteq N[v]$ in G, v sees z. Let $A = w \ldots v$ be the segment of C in the counterclockwise direction. Let $B = z \ldots x$ be the segment of C in the clockwise direction; observe that ux is a segment of B. Then, the required cycle C' is the one in which the segments xw, A, vz, and B occur consecutively in that order. As ux and xw are segments of C', uxw is a segment of C'. □

COROLLARY 3.3. *Let G be a strongly chordal graph and x be a simple vertex of G. Let p be the smallest vertex in $N(x)$ and q be the second smallest vertex in $N(x)$. Suppose C is a cycle such that $\{x, p, q\} \subseteq V(C)$. Then there exists a cycle C' of G such that $V(C) = V(C')$ and pxq is a segment of C'.*

Proof of Corollary 3.3. Suppose C has the segment axb where neither a nor b is p. Apply Lemma 3.2 to C with b and p in place of v and w, respectively. The resulting cycle has axp, or equivalently pxa, as a segment. If $a \neq q$, then apply Lemma 3.2 to this cycle with a and q in place of v and w, respectively, to get C'. □

The following is a direct corollary.

COROLLARY 3.4. *Suppose G is a strongly chordal graph, x is a simple vertex of G, and p, q are the two smallest vertices in $N(x)$. Then, G is Hamiltonian if and only if G has a Hamilton cycle in which pxq is a segment.*

LEMMA 3.5. *Let G be a strongly chordal graph, x be a simple vertex of G, and xu be any edge incident on x. Then $G - xu$ is strongly chordal.*

Proof of Lemma 3.5. We will show that $G - xu$ is chordal and does not contain a Sun. As x is simple (and hence simplicial) in G, $G - xu$ is chordal. Suppose deletion of the edge xu from G created a Sun on vertex-set $\{v_1, v_2, \ldots, v_{2k}\}$, $k \geq 3$. As x and u must be part of the Sun, let $x = v_i$ and $u = v_j$. Observe that no v_{2p} is simplicial in $G - xu$. In any Sun, between any two vertices whose subscripts are odd, there is a chordless path on 4 vertices (see Figure 1). Therefore, at least one of i, j is even, or else G would contain a chordless cycle on 4 vertices. As $\{ v_l \mid l \text{ is even } \}$ induces a clique in $G-xu$, and v_i misses v_j in $G-xu$, at least one of i, j must be odd. Therefore, one of i, j is even and the other is odd, and $\{ v_l \mid l \text{ is odd } \}$ induces an independent set in G also. Therefore, no v_{2p} is simplicial in G either. As v_i is simplicial in G, without loss of generality, we can assume $x = v_1$ and j is even. Observe that, in G and in $G - xu$, $\{v_2, v_{2k}\} \subseteq N(v_1)$. Therefore, $j \neq 2$ and $j \neq 2k$. Now, in G, v_3 sees v_2 but misses v_{2k}, and v_{2k-1} sees v_{2k} but misses v_2. Hence, in G, $N[v_{2k}] \not\subseteq N[v_2]$ and $N[v_2] \not\subseteq N[v_{2k}]$, contradicting x being a simple vertex. □

LEMMA 3.6. *Let G be a strongly chordal graph, x be a simple vertex of G, y be a simple vertex of $G\backslash x$, and xy be an edge. Then, for any edge yu incident on y such that $u \notin N(x)$, $G - yu$ is strongly chordal.*

Proof of Lemma 3.6. Suppose $G - yu$ is not chordal. Then, yu was the only chord of the cycle $yauby$ in G [14]. As y is simple (and hence simplicial) in $G\backslash x$, and a misses b in G, either $x = a$ or $x = b$; assume $x = a$. Then, $u \in N(x)$ in G, a contradiction.

Now, suppose deletion of the edge yu from G created a Sun on vertex-set $VS = \{v_1, v_2, \ldots, v_{2k}\}$, $k \geq 3$. As y and u must be part of the Sun, let $y = v_i$ and $u = v_j$. Observe that no v_{2p} is simplicial in $G-yu$. In any Sun, between any two vertices whose subscripts are odd, there is a chordless path on 4 vertices (see Figure 1). Therefore, at least one of i, j is even, or else G would contain a chordless cycle on 4 vertices. As $\{ v_l \mid l \text{ is even } \}$ induces a clique $G - yu$, and v_i misses v_j in $G - yu$, at least one of i, j must be odd. Therefore, one of i, j is even and the other is odd, and $\{ v_l \mid l \text{ is odd } \}$ induces an independent set in G also. Therefore, no v_{2p} is simplicial in G either. We next want to argue that i must be odd; suppose not and without loss of generality, let $y = v_2$. Then, in G, y sees v_3, y sees v_1, and v_1 misses v_3; therefore, y is not simplicial in G. However, as y is simplicial in $G\backslash x$, either $v_1 = x$ or $v_3 = x$; assume $v_1 = x$. Now, in $G\backslash x$, v_2 sees v_3, v_2 sees v_{2k}, but v_3 misses v_{2k}. Therefore, y is not simplicial in $G\backslash x$, a contradiction. Hence, $y = v_i$ where i is odd, and assume $y = v_1$ and j is even, where $j \neq 2$ and $j \neq 2k$. Let $G_S = G[VS]$. In G_S, v_3 sees v_2 but misses v_{2k}, and v_{2k-1} sees v_{2k} but misses v_2; therefore v_1 is not simple in G_S. However, $v_1 = y$ must be simple in $G\backslash x$ and hence in $G_S\backslash x$. Therefore, $x \in VS$, and assume $x = v_r$ for some r. As x is simplicial in G, it is also simplicial in G_S. However, as x sees y, r must be even. But, in G_S, $\{ v_l \mid l \text{ is odd } \}$ induces an independent set. Therefore, no v_{2p} is simplicial in G_S, and we have a contradiction. □

3.2. Interval graphs. A graph is an *interval graph* if each vertex of the graph can be mapped to an interval on the real line such that two vertices in the graph are adjacent if and only if their corresponding intervals have a nonempty intersection. Every interval graph is strongly chordal [1]. For an interval graph G, we use $\mathcal{I}(G)$ to refer to its *interval model*, the collection of intervals that represent G. For a vertex x, we use I_x to refer to the interval x is mapped to, $L(x)$ and $R(x)$ to refer to the left and right endpoints, respectively, of I_x, and $\mathcal{K}(G) = K_1 K_2 \ldots K_r$ to refer to a linear ordering of the maximal cliques of G such that for any vertex v of G, the maximal cliques that contain v occur consecutively in $\mathcal{K}(G)$ [4]. Every interval graph admits

an interval model in which all endpoints of intervals are distinct [5] and we always assume this. For an interval graph G, we use $\mathcal{I}(G)$ in which all endpoints are distinct, and $\mathcal{K}(G)$ that corresponds to $\mathcal{I}(G)$ interchangeably using the notation $(G, \mathcal{I}(G))$ or $(G, \mathcal{K}(G))$. Also, when the context is clear we simply use $\mathcal{I}$ and $\mathcal{K}$, omitting reference to G. Next we present some tools that are needed.

The following lemma is implicit in section 3.2 of [8].

LEMMA 3.7 (see [8]). *In an interval graph $(G, \mathcal{K}(G))$, let u and v be adjacent vertices such that some K_c in $\mathcal{K}(G)$ is the rightmost maximal clique containing u as well as the leftmost maximal clique containing v. Then, the graph $H = G - uv$ is an interval graph.*

The next lemma summarizes a few lemmata from [11].

LEMMA 3.8 (see [11]). *Suppose $(G, \mathcal{K} = K_1K_2 \dots K_r)$ is an interval graph and $P : x \dots z$ is a path in G such that $x \in K_1$ and $z \in K_r$ are simplicial vertices of G. Then, there is a path $P' : (x = w_1)w_2 \dots (w_k = z)$ in G such that the following hold:*

- *$V(P) = V(P')$.*
- *Every w_i can be mapped to a maximal clique $K_{f(w_i)}$ of $\mathcal{K}(G)$ that contains w_i in such a way that, for $1 \leq i \leq (k-1)$,*
  (1) *$K_{f(w_i)}$ also contains w_{i+1}, and*
  (2) *$f(w_i) \leq f(w_{i+1})$.*

In essence, the lemma says that the path P', which contains the same set of vertices as path P and also starts and ends at the same vertices as P, can be traversed along $\mathcal{K}(G)$, through the maximal cliques that w_i are mapped to, in such way that the traversal never moves to the left.

4. Cycle extendability. In this section, we present theorems pertaining to cycle extendability of Hamiltonian graphs in subclasses of chordal graphs. While a different proof is possible for the following lemma, we have chosen to prove it in the same spirit as our other inductive proofs.

LEMMA 4.1. *Suppose G is a Hamiltonian chordal graph and C is a non-Hamiltonian cycle in G such that $| V(C) | = 3$ or $| V(C) | = 4$. Then, C extends in G.*

Proof of Lemma 4.1. We consider the $| V(C) | = 3$ and $| V(C) | = 4$ cases separately. In both proofs, we assume that the given graph is not a clique, or else the lemma is trivially proved.

We first prove the lemma by induction on the number of vertices in the graph, when $| V(C) | = 3$. The lemma is easily verified for all Hamiltonian chordal graphs with at most 4 vertices. Suppose the lemma is true for all Hamiltonian chordal graphs with at most $(n-1)$ vertices, and let G be a Hamiltonian chordal graph with n vertices and C be a non-Hamiltonian cycle in G such that $| V(C) | = 3$. Let M be a Hamilton cycle of G. If $n = 4$, then C extends to M; therefore, we can assume $n \geq 5$. As G is not a clique, there exist nonadjacent simplicial vertices u and v of G. As $V(C)$ induces a clique in G, either $u \notin V(C)$ or $v \notin V(C)$; assume $v \notin V(C)$. As $G \backslash v$ has at least 4 vertices, C is a non-Hamiltonian cycle in $G \backslash v$. Further, $G \backslash v$ is chordal, the number of vertices in $G \backslash v$ is $(n-1)$, and M_{-v} is a Hamilton cycle of $G \backslash v$. Therefore, applying the inductive hypothesis to $G \backslash v$ and C, we get the required extension of C in G.

We next prove the lemma by induction on the number of vertices in the graph, when $| V(C) | = 4$. The lemma is easily verified for all Hamiltonian chordal graphs with at most 5 vertices. Suppose the lemma is true for all Hamiltonian chordal graphs with at most $(n-1)$ vertices, and let G be a Hamiltonian chordal graph with n vertices and C be a non-Hamiltonian cycle in G such that $| V(C) | = 4$. Let M be a Hamilton cycle of G. If $n = 5$, then C extends to M; therefore, we can assume $n \geq 6$. As G

is not a clique, there exist nonadjacent simplicial vertices u and v of G. If either $u \notin V(C)$ or $v \notin V(C)$, then, assuming $v \notin V(C)$, we can induct (as in the case of $| V(C) | = 3$) on $G\backslash v$ and C to get the required extension. Therefore, $C = uxvyu$. Now, at least one of u, v must have a degree of at least 3 in G; or else, each of u, v has degree 2 in G, implying that C is also a Hamilton cycle of G, contradicting G having at least 6 vertices. Without loss of generality, let $d(v) \geq 3$. Then, there exists vertex w such that w sees v, and $w \notin V(C)$. As v is simplicial in G, w must see both x and y. Therefore, the cycle $uxwvyu$ is a required extension of C. □

Remark 4.1. Suppose the Hamiltonian graph G with n vertices belongs to a hereditary class of chordal graphs. Let C be the non-Hamiltonian cycle of G that we wish to extend. Given Lemma 4.1 and Hamiltonicity of G, we can assume $5 \leq |V(C)| \leq (n-2)$. Further, for every simplicial vertex v of G, we can assume $v \in V(C)$ (else, induct on $G\backslash v$ and C to get an extension), and also every $w \in N(v)$ is on C (else, w can be inserted between v and a neighbor next to it on C to get an extension).

Recall that every interval graph is strongly chordal. We will repeatedly use Lemma 3.2 and Corollary 3.3 in proving the following theorem. Also, recall that $||G||$ refers to the sum of the number of vertices and the number of edges in G.

THEOREM 4.2. *A Hamiltonian interval graph is cycle extendable.*

Proof of Theorem 4.2. The proof is by induction on the sum of the number of vertices and the number of edges in the graph. The theorem is easily verified for every Hamiltonian interval graph G where $||G|| < 11$. Assume that the theorem holds for every Hamiltonian interval graph G where $||G|| \leq t-1$. Let $(G, \mathcal{K}(G) = K_1K_2 \dots K_r)$ be a Hamiltonian interval graph with $||G|| = t$, and let C be a given non-Hamiltonian cycle in G. Let M be a Hamilton cycle of G and n be the number of vertices in G. We assume as per Remark 4.1.

Suppose there exist vertices v and w of G such that each of v, w is contained only in the maximal clique K_1 in $\mathcal{K}(G)$. Then, each of v, w is simplicial in G. Also, it is the case that $N[v] = N[w]$. Now, M_{-v} is a Hamilton cycle of $G\backslash v$ and, as $5 \leq | V(C) | \leq (n-2)$, C_{-v} is a cycle in $G\backslash v$ such that $4 \leq | V(C_{-v}) | \leq (n-3)$. As $G\backslash v$ has $(n-1)$ vertices, C_{-v} is a non-Hamiltonian cycle in $G\backslash v$. Since $G\backslash v$ is a Hamiltonian interval graph with $||G\backslash v|| < ||G||$, we can apply the inductive hypothesis to $G\backslash v$ and C_{-v} to get an extension C' of C_{-v} in $G\backslash v$. Now, for any segment wz of C', v sees both w and z. Hence, by replacing the segment wz of C' with wvz, we can get the desired extension for C in G. Therefore, we can now assume that there is exactly one simplicial vertex of G in K_1.

Let $\mathcal{I}(G)$ be the interval model for G that corresponds to $\mathcal{K}(G) = K_1K_2 \dots K_r$. Recall that we can assume that endpoints of intervals in $\mathcal{I}(G)$ are distinct. Let $\mathcal{R} = (x = v_1)(y = v_2)v_3 \dots v_n$ be a strong elimination order of G that corresponds [10] to the increasing order of right endpoints of intervals in $\mathcal{I}(G)$. As x is the first vertex in $\mathcal{R}$, $R(x)$ is the smallest right endpoint in $\mathcal{I}(G)$; hence, $x \in K_1$. As x is simple (and hence simplicial) in G and $x \in K_1$, $x \notin K_i$, $i \geq 2$.

For the simple vertex x, let x_1 and x_2 be the smallest and second smallest vertices in $N(x)$. Recall that this means that when vertices in $N(x)$ are linearly ordered by containment of their closed neighborhoods, x_1 and x_2 will be the first and second vertices, respectively, in the order. Observe that as $\mathcal{R}$ is a strong elimination order for G, x_1 and x_2 naturally correspond to the first two neighbors of x with respect to $\mathcal{R}$ also. By Corollaries 3.3 and 3.4, we can assume that x_2xx_1 is a segment of M as well as C.

Suppose $d(x) \geq 3$ and let $\{x_1, x_2, z\} \subseteq N(x)$. Then, edge xz is such that $xz \notin E(C)$ and $xz \notin E(M)$. Clearly, the graph $G - xz$ is Hamiltonian.

As K_1 is the only maximal clique containing x, and as xz is an edge, K_1 is the rightmost maximal clique of $\mathcal{K}(G)$ containing x as well as the leftmost maximal clique of $\mathcal{K}(G)$ containing z. Therefore, by Lemma 3.7 the graph $G-xz$ is an interval graph. Also, as $|V(G-xz)| = |V(G)|$, but $G-xz$ has one less edge than G, $||G-xz|| = (||G||-1) < ||G||$. Therefore, by applying the inductive hypothesis to $G-xz$ and C, we can get the desired extension for C.

As G is Hamiltonian, $d(x) \geq 2$, and therefore, we can now assume that $d(x) = 2$ and $N(x) = \{x_1, x_2\}$ and $K_1 = \{x, x_1, x_2\}$. We consider two cases depending on the adjacency between x and y in G.

Case A. $(x = v_1)$ sees $(y = v_2)$ in G.

Clearly, $y = x_1$. Recall that $x_2x(x_1 = y)$ is a segment of both M and C, and $K_1 = \{x, x_1 = y, x_2\}$. Note that both x and x_2 see y. As C is a non-Hamiltonian cycle in G with $5 \leq |V(C)| \leq (n-2)$, G has at least 6 vertices. Therefore, $d(y) \geq 3$.

Let x_2xyp be a segment of M and x_2xyq be a segment of C where p and q may be different. If $d(y) \geq 5$, then there exist vertex z and edge yz such that $z \notin \{x, x_2, p, q\}$, $yz \notin E(M)$, and $yz \notin E(C)$. Clearly, $G - yz$ is Hamiltonian, and C is a non-Hamiltonian cycle in $G - yz$. We next want to show that we can then induct on $G - yz$ and C to get an extension for C. As x is the only vertex in K_1 that is not in any K_i, $i \geq 2$, $\{y, x_2\} \subseteq K_2$. As K_2 contains a vertex that is not in K_1, $K_1 - \{x\}$ is not a maximal clique in $G\backslash x$. For $2 \leq i \leq r$, as K_i is a maximal clique of G such that $x \notin K_i$, K_i is a maximal clique of $G\backslash x$ also. Therefore, $\mathcal{K}(G\backslash x) = K_2K_3 \ldots K_r$. As y is simplicial in $G\backslash x$, $y \notin K_i$, $i \geq 3$. Also, as y sees z, $z \notin K_1$, $z \in K_2$. Therefore, K_2 is the rightmost maximal clique in $\mathcal{K}(G)$ containing y as well as the leftmost maximal clique containing z. Then, by Lemma 3.7, $G - yz$ is an interval graph. As $||G-yz|| < ||G||$, we can indeed induct on $G-yz$ and C to get the required extension. We can now assume that $d(y) \leq 4$.

Suppose $d(y) = 3$ and for some vertex w, $N(y) = \{x, x_2, w\}$. As x is simplicial in G, $5 \leq |V(C)| \leq (n-2)$, and $G\backslash x$ has $(n-1)$ vertices, M_{-x} is a Hamilton cycle of the interval graph $G\backslash x$, and C_{-x} is a non-Hamiltonian cycle of $G\backslash x$. As $||G\backslash x|| < ||G||$, applying the inductive hypothesis to $G\backslash x$ and C_{-x} produces extension C' of C_{-x} in $G\backslash x$. As yx_2 is a segment of C', by replacing the segment yx_2 of C' with yxx_2, we can get an extension for C.

Hence, we can assume $d(y) = 4$, $N(y) = \{x, x_2, v, z\}$, the Hamilton cycle M contains the segment x_2xyv, and the edge yz is a chord of M.

If $yz \notin E(C)$ also, then an argument identical to the case of "$d(y) \geq 5$" can be employed to get an extension of C. Therefore, we assume $yz \in E(C)$ and hence x_2xyz is a segment of the cycle C.

Since y is simplicial in $G\backslash x$, vertex v sees y and z in G. If $v \notin V(C)$, then by replacing the segment x_2xyz in C with the segment x_2xyvz we can get an extension for C. Hence, we assume $v \in V(C)$.

We now consider three cases based on the relative order of the vertices x_2, v, and z in $\mathcal{R}$. Note that for the simple vertex y of $G\backslash x$, any such order will also correspond to an ordering of vertices of $N(y)$ in $G\backslash x$ according to containment of their closed neighborhoods. For vertices a and b, we use $a < b$ to mean a comes before b in $\mathcal{R}$.

Case A1. $x_2 < v$ and $x_2 < z$.

As x is simplicial in G, $G\backslash x$ is a Hamiltonian interval graph with $(n-1)$ vertices, and $||G\backslash x|| < ||G||$. As $5 \leq |V(C)| \leq (n-2)$, C_{-x} is a non-Hamiltonian cycle with the segment x_2y in $G\backslash x$. Apply the inductive hypothesis to $G\backslash x$ and C_{-x} to get an extension C' of C_{-x} in $G\backslash x$. As y is simple in $G\backslash x$, x_2 is

the smallest vertex in $N(y)$ in $G\backslash x$, and every neighbor of y in $G\backslash x$ is on C', by Corollary 3.3, we can assume x_2y is a segment of C'. Then, replacing the segment x_2y of C' with the segment x_2xy, we can get an extension for C.

Case A2. $v < x_2$ and $v < z$.

Recall that x_2xyv is a segment of M. We next want to show that vertices of cycle C can be rearranged so that x_2xyv is a segment of C also. Suppose x_2xyv is not a segment of C. Observe again that $5 \leq |\ V(C)\ | \leq (n-2)$.

Now, consider the cycle C_{-x} in $G\backslash x$. As y is simple in $G\backslash x$, every neighbor of y in $G\backslash x$ is on the cycle C_{-x}, x_2yz is a segment of C_{-x}, and v is the smallest vertex in $N(y)$ in $G\backslash x$, applying Lemma 3.2 with v and z in place of w and v, respectively, we can ensure that x_2yv is a segment of C_{-x}. Now, replacing the segment x_2y of C_{-x} with the segment x_2xy, we can get a cycle on the same set of vertices as $V(C)$ such that x_2xyv is a segment of that cycle. Therefore, we can assume C itself is a cycle in which x_2xyv is a segment.

Now, as the edge yz is such that $yz \notin E(C)$ and $yz \notin E(M)$, an argument identical to the case of "$d(y) \geq 5$" can be used to get an extension of C.

Case A3. $z < v$ and $z < x_2$.

Similar to the way vertices on C were rearranged in Case A2, rearrange vertices on M so that x_2xyz is a segment of M also. Now, as the edge yv is such that $yv \notin E(C)$ and $yv \notin E(M)$, an argument identical to the case of "$d(y) \geq 5$" can be used to get an extension of C.

Case B. $(x = v_1)$ misses $(y = v_2)$ in G.

As x misses y in G and y is simplicial in $G\backslash x$, y is simplicial in G also. Therefore, $y \in V(C)$. As $R(x)$ is the smallest right endpoint in $\mathcal{I}(G)$ and x misses y, $R(x) < L(y)$. As x sees x_1 and x_2, and $R(y)$ is the second smallest right endpoint in $\mathcal{I}(G)$, $L(x_i) < R(x) < L(y) < R(y) < R(x_i)$, $i = 1, 2$. Therefore, y sees x_1 and x_2. Note that $K_1 = \{x, x_1, x_2\}$ and $N(x) = \{x_1, x_2\}$.

The rightmost maximal clique of $\mathcal{I}(G)$, namely K_r, contains a vertex z such that $z \notin K_{r-1}$. As K_r is the only maximal clique containing z, z is simplicial in G and hence $z \in V(C)$. We make the following claims.

Claim 1. *Let T be a cycle in G such that $\{x, x_1, x_2, y, z\} \subseteq V(T)$. Then, there exists cycle T' in G such that $V(T) = V(T')$ and x_2xx_1y is a segment of T'.*

Claim 2. *Let $H = (G\backslash x_1) + xy$, i.e., the graph obtained from G by first deleting the vertex x_1 and then adding the edge xy. Then H is an interval graph.*

Before we prove the claims, we show how the claims can be used to complete the proof of the theorem.

As C and M are cycles of G that satisfy conditions of Claim 1, we can assume that x_2xx_1y is a segment of C as well as M. Let $H = (G\backslash x_1) + xy$. By Claim 2, H is an interval graph. Let C' be the cycle in H obtained from C by replacing the segment x_2xx_1y by x_2xy. Similarly, let M' be the cycle in H obtained from M by replacing the segment x_2xx_1y by x_2xy. Clearly, M' is a Hamilton cycle of H. As H has $(n-1)$ vertices and $4 \leq |\ V(C')\ | \leq (n-3)$, C' is a non-Hamiltonian cycle in H. As a vertex and at least three edges (namely, x_1x, x_1x_2, x_1y) were deleted, and the edge xy was added to derive H from G, $||H|| < ||G||$. By applying the inductive hypothesis to H and C', we get an extension C'' of C' in H. As x_2 and y are the only neighbors of x in H, x_2xy must be a segment of C''. As x_1 sees x and y in G, by replacing the segment xy of C'' with the segment xx_1y, we get the desired extension for C in G.

We now complete the proof of the theorem by presenting the proofs of Claims 1 and 2.

Proof of Claim 1. Recall that T is a cycle in the interval graph $(G, \mathcal{K}(G) = K_1K_2 \dots K_r)$, where $x \in K_1$ and $z \in K_r$. Let P_1 be the segment of T in the clockwise direction from x to z. Let P_2 be the segment of T in the counter-clockwise direction from x to z. As $N(x) = \{x_1, x_2\}$, without loss of generality, we can assume that $x_1 \in V(P_1)$ and $x_2 \in V(P_2)$.

Note that if G had exactly two maximal cliques, then as every simplicial vertex of G and every neighbor of every simplicial vertex of G is on C, $C = M$, a contradiction. Therefore, we can assume that $r \geq 3$.

Let P_s, $s = 1$ or $s = 2$, be the path such that $y \in V(P_s)$. We next show that the vertices on P_s can be rearranged so that xx_jy is a segment of P_s for $j = 1$ or $j = 2$.

Refer to Lemma 3.8. As vertices x, z and path P_s satisfy the hypotheses of Lemma 3.8, P_s satisfies the conclusions of the lemma; let f be as defined in the lemma. If xx_jy is a segment of P_s, then we are done; so, assume otherwise, and $P_s = xx_jp \dots y \dots z$, where $p \neq y$.

We next show that $y \in K_2$. Suppose not, and let i be the smallest integer such that $y \in K_i$; clearly, $i \geq 3$. Now, as there exists vertex $q \neq x$ such that $q \in K_2$ but $q \notin K_3$, it follows that $R(q) < R(y)$, a contradiction to $R(y)$ being the second smallest right endpoint in the model. Recall that as x misses y and y is simplicial in $G\backslash x$, y is simplicial in G also. We can conclude that K_2 is the only maximal clique of G containing y, and therefore $f(y) = 2$.

As $r \geq 3$ and K_r is the only maximal clique containing z, $y \neq z$.

Let w be the vertex that immediately follows y in P_s; such a w exists as $y \neq z$. Referring back to P_s, as $p \notin K_1$, $f(y) = 2$, and P_s is a path satisfying Lemma 3.8, it follows that for every vertex u in the segment $p \dots y$ of P_s, $f(u) = 2$. Further, as every neighbor of y must belong to K_2, $w \in K_2$. Therefore, all the vertices in the segment $p \dots yw$ of P_s belong to K_2. We can now simply delete y from its current position in P_s and insert it immediately before p to obtain a desired path in which xx_jy is a segment.

Finally, if $j = 1$, then as xx_1y is a segment of P_1 and xx_2 is a segment of P_2, x_2xx_1y is a segment of T, and $T' = T$ is the desired cycle. Otherwise, P_2 contains the segment xx_2y. Let xx_1q be a segment of P_1. As x is simple in G, and x_1 is the smallest vertex in $N(x)$, $N[x_1] \subseteq N[x_2]$ in G. Therefore, x_2 sees q. Let P' be the path in G in which the following are consecutive segments: segment xx_1 of P_1, edge x_1y, and segment $y \dots z$ of P_2. Let P'' be the path in G in which the following are consecutive segments: segment xx_2 of P_2, edge x_2q, and segment $q \dots z$ of P_1. Then, the union of P' and P'' is the desired cycle T' containing the segment x_2xx_1y. □

Proof of Claim 2. Recall that $(G, \mathcal{K}(G) = K_1K_2 \dots K_r)$ is the given interval graph. Let $\mathcal{I}(G)$ be the interval model that corresponds to $\mathcal{K}(G)$ in which endpoints are distinct. As x misses y in G and $R(x)$ is the smallest right endpoint, we have $R(x) < L(y)$. Derive interval model $\mathcal{I}'$ for H from $\mathcal{I}(G)$ as follows: First delete the interval I_{x_1} corresponding to vertex x_1. Then, place $L(y)$ so that $L(y) < R(x)$, leaving all other endpoints unchanged.

Since $R(y)$ is the second smallest right endpoint in $\mathcal{I}(G)$, there is no interval I_w in $\mathcal{I}(G)$ such that $R(x) < R(w) < L(y)$. Therefore, the only changes effected by $\mathcal{I}'$ are deletion of x_1, addition of y to the neighborhood of x, and addition of x to the neighborhood of y, as desired. □

The proof of Theorem 4.2 is now complete. □

Refer to Figure 1 for the graphs F_1 and F_2.

THEOREM 4.3. *A Hamiltonian strongly chordal graph with no F_1 or F_2 is cycle extendable.*

Proof of Theorem 4.3. The proof is by induction on the number of vertices in the graph. It is easily verified that the theorem is true for all relevant graphs with at most 5 vertices. Assume that the theorem is true for all Hamiltonian strongly chordal graphs with no F_1 or F_2 that have at most $(n-1)$ vertices. Let G be a Hamiltonian strongly chordal graph on n vertices that contains no F_1 or F_2, and C be a given non-Hamiltonian cycle in G. We assume as per Remark 4.1. Let x be a simple vertex of G, u the smallest vertex in $N(x)$, and v the second smallest vertex in $N(x)$. Let M be a Hamilton cycle of G.

By Corollary 3.3, we can now assume that vxu is a segment of the Hamilton cycle M as well as the cycle C. Let $vxup$ be a segment of M and $vxuq$ be a segment of C where p may be different from q. As $N[u] \subseteq N[v]$ in G, v sees both p and q in G.

Let $H = G\backslash\{x, u\}$, M_1 be the cycle obtained from M by replacing the segment $vxup$ with vp, and C_1 be the cycle obtained from C by replacing the segment $vxuq$ with vq. It follows that H is a strongly chordal graph with no F_1 or F_2, H has $(n-2)$ vertices, and M_1 is a Hamilton cycle of H. As $3 \leq \mid C_1 \mid \leq (n-4)$, C_1 is a non-Hamiltonian cycle in H. Therefore, by inductive hypothesis, C_1 extends to cycle C_2 in H. As $v \in V(C_2)$, there exists a segment wvz of C_2. We claim that u must see one of w, z in G. Suppose not, and u misses both w and z in G. As x is simple (and hence simplicial) in G, x must also miss both w and z. Then, depending on whether or not w sees z in G, the set $\{u, v, w, x, z\}$ of vertices induces either F_1 or F_2 in G, a contradiction.

Now, without loss of generality, let u see z in G. Starting from C_2, replacing the segment wvz with the segment $wvuz$, and then replacing the segment $wvuz$ in the resulting cycle with the segment $wvxuz$, we can obtain an extension for C. □

The graphs F_1 and F_2 are interval graphs. The graph G_2 in Figure 1 is not an interval graph. However, G_2 is a Hamiltonian strongly chordal graph that does not contain F_1 or F_2.

The following theorem essentially shows that the ideas employed in Case A of the proof of Theorem 4.2 hold for the larger class of Hamiltonian strongly chordal graphs also. Graph G_1 in Figure 1 is an example of a graph satisfying the conditions of Theorem 4.4.

THEOREM 4.4. *Suppose G is a Hamiltonian strongly chordal graph with the strong elimination order $v_1v_2 \ldots v_n$ of vertices such that v_iv_{i+1} is an edge, $1 \leq i \leq (n-1)$. Then, G is cycle extendable.*

Proof of Theorem 4.4. The proof is by induction on the sum of the number of vertices and the number of edges in the graph. We refer to a strong elimination order as in the statement of the theorem as a *special order*. The theorem is easily verified for every Hamiltonian strongly chordal graph G with a special order where $||G|| < 11$. Assume that the theorem holds for every Hamiltonian strongly chordal graph G with a special order where $||G|| \leq t-1$. Let G be a Hamiltonian strongly chordal graph on n vertices, with a special order $\mathcal{R} = v_1v_2 \ldots v_n$, with $||G|| = t$, and let C be a given non-Hamiltonian cycle in G. Let M be a Hamilton cycle of G. Observe that $\mathcal{R}\backslash v_1 = v_2v_3 \ldots v_n$ is a special order for $G\backslash v_1$. We assume as per Remark 4.1 that v_1 and every neighbor of v_1 are on C.

Clearly, v_2 is the smallest vertex in $N(v_1)$. Let v_k be the second smallest vertex in $N(v_1)$. By Corollaries 3.3 and 3.4, we can assume that $v_kv_1v_2$ is a segment of M as well as C.

Suppose $d(v_1) \geq 3$ and let $\{v_2, v_k, z\} \subseteq N(v_1)$. Then, edge v_1z is such that $v_1z \notin E(C)$ and $v_1z \notin E(M)$. Clearly, the graph $G - v_1z$ is Hamiltonian. By Lemma 3.5, $G - v_1z$ is strongly chordal. Vertex v_1 is simplicial in $G - v_1z$ also.

Suppose v_1 had neighbors v_i, v_j in $G - v_1z$ such that $i < j$, but $N[v_i] \not\subseteq N[v_j]$ in $G - v_1z$. Then, in $G - v_1z$, v_i sees v_p that v_j misses. Then, it must be that $\{v_p, v_j\} = \{v_1, z\}$. This is impossible as $v_1 \neq v_j$, and as v_1 sees v_j and v_p misses v_j in $G - v_1z$, $v_1 \neq v_p$ either. Finally, as $z \neq v_2$, $\mathcal{R}$ is a special order for $G - v_1z$ also. As $||G - v_1z|| < ||G||$, by applying the inductive hypothesis to $G - v_1z$ and C we can get the desired extension for C.

As G is Hamiltonian, $d(v_1) \geq 2$, and therefore, we can now assume that $d(v_1) = 2$ and $N(v_1) = \{v_2, v_k\}$. Therefore, $v_kv_1v_2$ is a segment of both M and C. Note that both v_1 and v_k see v_2. As C is a non-Hamiltonian cycle in G with $5 \leq |\, V(C)\, | \leq (n-2)$, G has at least 6 vertices. Therefore, $d(v_2) \geq 3$.

Suppose $d(v_2) = 3$ and for some vertex w, $N(v_2) = \{v_1, v_3, w\}$. Then, M_{-v_1} is a Hamilton cycle of $G\backslash v_1$, C_{-v_1} is a non-Hamiltonian cycle in $G\backslash v_1$, $\mathcal{R}\backslash v_1$ is a special order of $G\backslash v_1$, and $||G\backslash v_1|| < ||G||$. Applying the inductive hypothesis to $G\backslash v_1$ and C_{-v_1} produces extension C' of C_{-v_1} in $G\backslash v_1$. Then, replacing the segment v_kv_2 of C' with $v_kv_1v_2$, we can get an extension for C. Now, we can assume that $d(v_2) \geq 4$.

We now consider two cases based on the adjacency between v_1 and v_3.

Case 1. v_1 sees v_3 (and $v_k = v_3$).

As $N(v_1) = \{v_2, v_3\}$, $v_3v_1v_2$ is a segment of M as well as C. Then, M_{-v_1} is a Hamilton cycle of $G\backslash v_1$, C_{-v_1} is a non-Hamiltonian cycle in $G\backslash v_1$, $\mathcal{R}\backslash v_1$ is a special order of $G\backslash v_1$, and $||G\backslash v_1|| < ||G||$. Apply the inductive hypothesis to $G\backslash v_1$ and C_{-v_1} to get an extension C' of C_{-v_1} in $G\backslash v_1$. As v_2 is simple in $G\backslash v_1$, and v_3 is the smallest vertex in $N(v_2)$ in $G\backslash v_1$, by Lemma 3.2, we can assume v_3v_2 is a segment of C'. Then, replacing the segment v_3v_2 of C' with the segment $v_3v_1v_2$, we can get an extension for C.

Case 2. v_1 misses v_3 (and $v_k \neq v_3$).

Now, $v_k \neq v_3$ is the second smallest vertex in $N(v_1)$. Let $v_kv_1v_2p$ be a segment of M. We next show that if $p \neq v_3$, then vertices of M can be rearranged so that $v_kv_1v_2v_3$ is a segment of M. First apply Lemma 3.2 to $G\backslash v_1$, M_{-v_1}, and the simple vertex v_2 of $G\backslash v_1$, with v_3 and p in place of w and v, respectively, so that $v_kv_2v_3$ is a segment of C_{-v_1}. Now, replacing the segment v_kv_2 with $v_kv_1v_2$, we can make $v_kv_1v_2v_3$ a segment of M. Let $v_kv_1v_2q$ be a segment of C. If $v_3 \notin V(C)$, then as v_2 is simplicial in $G\backslash v_1$, v_3 sees consecutive vertices v_2 and q on C. We can then extend C by replacing the segment v_2q with v_2v_3q. Therefore, assume $v_3 \in V(C)$. Then an argument similar to the one used for M can be employed to show that $v_kv_1v_2v_3$ is a segment of C also.

As $d(v_2) \geq 4$, there now exist vertex $z \notin \{v_1, v_k, v_3\}$ and edge v_2z such that $v_2z \notin E(M)$ and $v_2z \notin E(C)$. Clearly, $G - v_2z$ is Hamiltonian and C is a non-Hamiltonian cycle in $G - v_2z$. By Lemma 3.6, $G - v_2z$ is strongly chordal. By a previous argument, $\mathcal{R}\backslash v_1$ is a special order of $(G - v_2z)\backslash v_1$. As $v_k \notin \{v_2, z\}$, no edge incident on v_k was deleted; therefore $N[v_2] \subseteq N[v_k]$ in $G - v_2z$ also, and $\mathcal{R}$ is a special order for $G - v_2z$. We can now apply the inductive hypothesis to $G - v_2z$ and C to get an extension for C.

The proof of Theorem 4.4 is now complete. □

THEOREM 4.5. *A Hamiltonian split graph is cycle extendable.*

Proof of Theorem 4.5. The proof is by induction on the number of vertices in the graph. The theorem is easily verified for Hamiltonian split graphs with at most 5 vertices and assume that the theorem is true for all Hamiltonian split graphs with at most $(n - 1)$ vertices. Let $G = (K, I, E)$ be a Hamiltonian split graph on n vertices and C be a given non-Hamiltonian cycle of G. We assume as per Remark 4.1.

Observe that for any $v \in I$, v is simplicial in G. We claim that for any non-Hamiltonian cycle C in G, of all vertices in $(K \cup I) - V(C)$, either there exists a simplicial vertex v of G or there exists a vertex u such that u sees two consecutive vertices w, z of C. In order to verify the claim, as every $v \in I$ is simplicial in G, if $I \not\subseteq V(C)$, then we are done; hence assume that $I \subseteq V(C)$. Now, consider $u \in (K - V(C))$. If u were simplicial in G, then we are done proving the claim; hence assume otherwise. As K induces a clique in G, it then follows that $(N(u) \cap I) \neq \emptyset$. Let $w \in I$ be a vertex that u sees and let wz be a segment of C. As w is simplicial in G, u sees z, and hence u sees two consecutive vertices on C.

To complete the proof of the theorem, if there exists $v \in ((K \cup I) - V(C))$ such that v is simplicial in G, we apply the inductive hypothesis to $G\backslash v$ and C to get the required extension of C; as $G\backslash v$ has $(n-1)$ vertices, C is a non-Hamiltonian cycle in $G\backslash v$. Also, as v is simplicial in G, $G\backslash v$ is Hamiltonian. Otherwise, we have $u \in ((K \cup I) - V(C))$ such that u sees consecutive vertices w, z of C. An extension of C can then be obtained by replacing the segment wz of C with the segment wuz. □

We conclude with some remarks on finding a Hamilton cycle in a strongly chordal graph that does not contain F_1 or F_2. Müller showed [12] that deciding whether a strongly chordal graph that is also a split graph is Hamiltonian is NP-complete. As a split graph cannot contain F_1, it follows that deciding whether a strongly chordal graph that does not contain F_1 is Hamiltonian is NP-complete. In contrast, Corollary 3.4 and the proof of Theorem 4.3 can be used to show that for a strongly chordal graph G on at least 5 vertices that does not contain F_1 or F_2, G is Hamiltonian if and only if $G\backslash\{x, u\}$ is Hamiltonian, where x is a simple vertex and u is the smallest vertex in $N(x)$. As a strong elimination order of a strongly chordal graph can be found in $O(n^2)$ time [13, 16, 15], it follows that whether a strongly chordal graph that does not contain F_1 or F_2 is Hamiltonian can be decided in $O(n^2)$ time.

Acknowledgments. We thank Carlin Sappenfield for timely help with the literature and the editors for their comments.

REFERENCES

[1] A. Brandstädt, V. B. Le, and J. P. Spinrad, *Graph classes: A survey*, SIAM Monogr. Discrete Math. Appl. 3, SIAM, Philadelphia, 1999.

[2] G. Chen, R. J. Faudree, R. J. Gould, and M. S. Jacobson, *Cycle extendability of Hamiltonian interval graphs*, SIAM J. Discrete Math., 20 (2006), pp. 682–689.

[3] M. Farber, *Characterizations of strongly chordal graphs*, Discrete Math., 43 (1983), pp. 173–189.

[4] P. C. Gilmore and A. J. Hoffman, *A characterization of comparability graphs and of interval graphs*, Canad. J. Math., 16 (1964), pp. 539–548.

[5] M. C. Golumbic, *Algorithmic graph theory and perfect graphs*, Academic Press, New York, 1980.

[6] G. R. T. Hendry, *Extending cycles in digraphs*, J. Combin. Theory Ser. B, 46 (1989), pp. 162–172.

[7] G. R. T. Hendry, *Extending cycles in graphs*, Discrete Math., 85 (1990), pp. 59–72.

[8] L. Ibarra, *Fully dynamic algorithms for chordal graphs and split graphs*, J. Algorithms, to appear.

[9] T. Jiang, *Planar Hamiltonian chordal graphs are cycle extendable*, Discrete Math., 257 (2002), pp. 441–444.

[10] H. Kaplan, R. Shamir, and R. E. Tarjan, *Tractability of parameterized completion problems on chordal, strongly chordal, and proper interval graphs*, SIAM J. Comput., 28 (1999), pp. 1906–1922.

[11] M. Keil, *Finding Hamilton circuits in interval graphs*, Inform. Process. Lett., 20 (1985), pp. 201–206.
[12] H. Müller, *Hamiltonian circuits in chordal bipartite graphs*, Discrete Math., 156 (1996), pp. 291–298.
[13] R. Paige and R. E. Tarjan, *Three partition refinement algorithms*, SIAM J. Comput., 16 (1987), pp. 973–989.
[14] 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.
[15] J. Sawada and J. P. Spinrad, *From a simple elimination ordering to a strong elimination ordering in linear time*, Inform. Process. Lett., 86 (2003), pp. 299–302.
[16] J. P. Spinrad, *Doubly lexical ordering of dense* 0-1 *matrices*, Inform. Process. Lett., 45 (1993), pp. 229–235.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 682–689

CYCLE EXTENDABILITY OF HAMILTONIAN INTERVAL GRAPHS*

GUANTAO CHEN†, RALPH J. FAUDREE‡, RONALD J. GOULD§, AND MICHAEL S. JACOBSON¶

Abstract. A graph G of order n is pancyclic if it contains cycles of all lengths from 3 to n. A graph is called cycle extendable if for every cycle C of less than n vertices there is another cycle C^* containing all vertices of C plus a single new vertex. Clearly, every cycle extendable graph is pancyclic if it contains a triangle. Cycle extendability has been intensively studied for dense graphs while little is known for sparse graphs, even very special graphs. We show that all Hamiltonian interval graphs are cycle extendable. This supports a conjecture of Hendry that all Hamiltonian chordal graphs are cycle extendable.

Key words. interval graph, Hamiltonian, cycle extendable

AMS subject classification. 05C38

DOI. 10.1137/S0895480104441450

1. Introduction.

All graphs considered in this paper are finite and simple. We will generally follow the notation and definitions of West [14]. Let G be a graph. We use $V(G)$ and $E(G)$ to denote its vertex set and edge set, respectively. For any vertex v of G, $N(v)$ (or $N_G(v)$) denotes the neighborhood of v (neighborhood of v in G) and $d(v)$ (or $d_G(v)$) denotes the degree of v (degree of v in G). For any $X \subseteq V(G)$, let $G[X]$ denote the subgraph induced by X. If H is a subgraph of G, we define $G[H] := G[V(H)]$.

A graph is chordal if every cycle of length at least 4 contains a chord. An *interval graph* is a graph whose vertices correspond to a family of intervals so that vertices are adjacent if and only if the corresponding intervals intersect. It is well known that all interval graphs are chordal graphs.

In a graph G, a *Hamiltonian cycle* is a cycle containing all vertices of G. A graph is *Hamiltonian* if it has a Hamiltonian cycle. Determining when graphs are Hamiltonian is one of the fundamental problems in graph theory. Although it is NP-hard to decide whether a graph is Hamiltonian, finding conditions sufficient to imply a graph is Hamiltonian has been intensively studied in the last thirty years. While studying Hamiltonicity, many related properties have also been heavily explored. For example, a graph G of order n is *pancyclic* if it contains cycles of all lengths from 3 to n. Clearly, every pancyclic graph is Hamiltonian, but the converse is not true. Being pancyclic provides a lot more cycle structure to graphs. Although there are many

*Received by the editors February 21, 2004; accepted for publication (in revised form) October 17, 2005; published electronically September 15, 2006.
http://www.siam.org/journals/sidma/20-3/44145.html

†Department of Computer Science and Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, and Faculty of Mathematics and Statistics, Huazhong Normal University, Wuhan, China (gchen@gsu.edu). The research of this author was partially supported by NSA grant H98230-04-1-0300 and NSF grant DMS-0500951.

‡University of Memphis, Memphis, TN 38152 (rfaudree@memphis.edu).

§Department of Math and Computer Science, Emory University, Atlanta, GA 30322 (rg@mathcs.emory.edu).

¶Department of Mathematics, University of Colorado at Denver, Denver, CO 80217-3364 (msj@math.cudenver.edu).

Hamiltonian graphs which are not pancyclic, the known sufficient degree conditions implying each of the properties are often similar. For example, the classic result of Ore [10] says that a graph G of order $n \geq 3$ is Hamiltonian if $d(u) + d(v) \geq n$ for every nonadjacent pair $u, v \in V(G)$. Bondy [2] showed the same condition implies that G is either pancyclic or a complete bipartite graph $K_{n/2,n/2}$. A common method of showing that a graph G is pancyclic is described below:

- Show that G has a triangle.
- Suppose that G has a cycle of length $k < n$, and find a special cycle of length k $(< n)$ and a special vertex $v \notin V(C)$ such that $G[V(C) \cup \{v\}]$ is Hamiltonian.

Motivated by the above observations, Hendry [7] gave the following definitions. In a graph G, a non-Hamiltonian cycle C is *extendable* if there exists a vertex $v \notin V(C)$ such that $G[V(C) \cup \{v\}]$ is Hamiltonian. A graph G is *cycle extendable* if all non-Hamiltonian cycles are extendable. In the same paper, Hendry showed that a graph G of order $n \geq 3$ is cycle extendable if $d(u) + d(v) \geq n + 1$ for every pair of nonadjacent vertices u and v. Graphs satisfying the above degree conditions must be very dense (in edges). To study the cycle structure of graphs less dense, usually some other structural properties are imposed, for example, planarity.

In 1931, Whitney [15] proved that every 4-connected plane triangulation contains a Hamiltonian cycle. In 1956, Tutte [13] extended that result to 4-connected planar graphs. Malkevitch [9] conjectured that every 4-connected graph containing a C_4 is pancyclic. Combining results from [12, 11, 3], we know that every 4-connected planar graph of order $n \geq 9$ contains cycles of length $n - i$ for $i = 1, \ldots, 6$. These results use the approach of finding shorter cycles from long cycles. However, this approach cannot demonstrate why C_4s should play an important role in 4-connected planar graphs being pancyclic. Thus, constructing larger cycles from smaller cycles might be a better approach. Hence, cycle extendable graphs take on added importance.

For any graph H, let $c(H)$ denote the number of connected components of H. Let $t > 0$ be a positive number. We say a graph is *t-tough* if $|A| \geq t \cdot c(G - A)$ for all cuts $A \subseteq V(G)$. Clearly, every Hamiltonian graph is 1-tough. On the other hand, a longstanding conjecture of Chvátal [5] states that there exists a constant t such that every t-tough graph is Hamiltonian. Although this conjecture remains open, Chen et al. [4] showed that all 18-tough chordal graphs are Hamiltonian. Note that a chordal graph containing a cycle C_k also contains a cycle C_{k-1} if $k \geq 4$. Repeating this argument, we see that all chordal Hamiltonian graphs are pancyclic. Hendry [7] gave the following conjecture.

CONJECTURE 1.1. *All Hamiltonian chordal graphs are cycle extendable.*

The purpose of this paper is to prove that Conjecture 1.1 is true for a special class of chordal graphs, namely interval graphs.

THEOREM 1.2. *All Hamiltonian interval graphs are cycle extendable.*

The proof of Theorem 1.2 will be given in section 3. In section 2 we will develop necessary properties of interval graphs.

Keil [8] designed a linear algorithm to find a Hamiltonian cycle in an interval graph. One consequence of his algorithm is that an interval graph is Hamiltonian if and only if it is 1-tough. We will heavily use this fact in our proof. For 1-tough Hamiltonian graphs, a cut A of G is called *critical* if $c(G - A) = |A|$. Let C be a Hamiltonian cycle of G and A be a critical cut of G; then the vertex sets of the components of $G - A$ are exactly those of the components of $C - A$. The following lemma regarding critical cuts on Hamiltonian graphs will be needed in our proof, and its proof is straightforward.

LEMMA 1.3. *Let G be a Hamiltonian graph with a Hamiltonian cycle C. If A is a cut of G such that all segments of $C - A$ induce components of $G - A$ and A does not contain two consecutive vertices of C, then A is a critical cut of G.*

For any two disjoint intervals A and B on the real number line, we let $d(A, B)$ denote the distance between A and B. Let G be an interval graph. For each vertex $v \in V(G)$, let $I(v)$ denote the corresponding interval called the *representation* of v. For each $W \subseteq V(G)$, let $I(W) = \bigcup_{v \in W} I(v)$. For each subgraph H of G, we define $I(H) = I(V(H))$. Clearly, $I(H)$ is also an interval of the real line if H is connected. Since only finite simple graphs will be considered in this paper, we assume that $I(v)$ is a closed interval for each $v \in V(G)$. For each interval $I = [a, b]$, we call a the left-end of I and b the right-end of I. We say a vertex v is on the *left side* of w (or equivalently w is on the *right side* of v) if $a \leq b$ for all $a \in I(v)$ and $b \in I(w)$. For any two vertex subsets U and W, we say that U is on the *left side* of W if u is on the left side of w for any $u \in U$ and $w \in W$.

2. Paths and cycles in interval graphs. In this section we will review some properties of interval graphs. Most of these properties are given in [8]. A clique D is a subgraph of G such that all vertices in D are mutually adjacent. This is equivalent to the property that the intersection of the corresponding intervals is not empty. Thus, a clique D can be represented by a point p which is contained in each of the intervals corresponding to the vertices of D. Note, however, that different cliques may have the same representative. A clique is *maximal* if there is no other clique containing this clique as a proper subgraph. It is not difficult to see that different maximal cliques must have different representatives. By selecting a representative p for each maximal clique D and ordering all maximal cliques from left to right on the real number line by their representative points, Gilmore and Hoffman [6] obtained the following property.

LEMMA 2.1. *The maximal cliques of an interval graph G can be linearly ordered, such that, for every vertex x of G, the maximal cliques containing x occur consecutively.*

We name such an ordering $D_1, D_2, \ldots, D_m$ the *linear order* of cliques, where a maximal clique is named D_i if its representative point p_i is the ith smallest representative of the maximal cliques of G.

A vertex v that appears in a maximal clique D_i is called a *conductor* for D_i if v also appears in the maximal clique D_{i+1}. Clearly, the interval corresponding to v contains the interval $[p_i, p_{i+1}]$. Let

$$L(D_i) := \{D_1, D_2, \ldots, D_i\} \quad \text{and } \tilde{L}(D_i) := \{D_{i+1}, \ldots, D_m\}.$$

A path P in G is *spanning* for $L(D_i)$ if P contains all vertices of G not appearing in $\tilde{L}(D_i)$ and P has two conductors of D_i as endvertices. Let R_i be the set of representatives of the maximal cliques containing vertex v_i. A point *embedding* Q of a path $P : v_1 v_2 \ldots v_n$ is an assignment of a real number $q(v_i) \in R_i$ to v_i such that $q(v_i) \in R_{i+1}$ for $1 \leq i \leq n-1$. A path is *straight* if it has a point embedding Q with the property that $q(v_r) \leq q(v_{r+1})$ for $1 \leq r \leq n-1$. The following lemma is due to Keil [8].

LEMMA 2.2. *Given a path P with point embedding Q, in an interval graph G, with an endpoint v_1 that appears only in D_1, there exists a straight path P', with v_1 as an endpoint, that has the same vertex set as P and has a point embedding Q' that has the same point set as Q.*

A path P, with endvertices u and v, that spans $L(D_i)$ is said to be *U-shaped* if there exists a vertex x in P that appears only in D_1 such that the two subpaths of

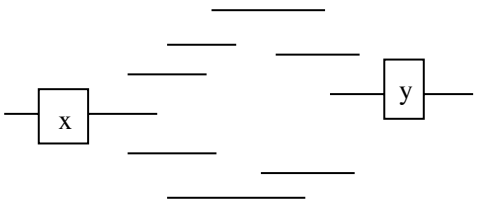


FIG. 1. *A standard cycle.*

P from x to u and from x to v are straight. Such a vertex x is called the *base* of the U-shaped path P. The point embedding of w in the U-shaped path P is the point embedding of v in the path from x to u if w lies on this path; otherwise it is the point embedding of v in the path from x to v. We denote the embedding by q_P. The following result is also due to Keil [8].

LEMMA 2.3. *If G is an interval graph with m maximal cliques, then G has a Hamiltonian cycle if and only if there exists a U-shaped spanning path for $L(D_i)$, $1 \leq i \leq m-1$.*

Based on Lemma 2.3, for every Hamiltonian interval graph there is a Hamiltonian cycle C and two vertices $x \in D_1$ and $y \in D_m$ such that both x-y paths induced by C are straight, x appears only in D_1, and y appears only in D_m. We name such a Hamiltonian cycle a *standard Hamiltonian cycle* (see Figure 1) and denote it by $(C : x, y)$ with distinguished vertices x and y. We also denote the embedding by q_C. Keil [8] also showed the following lemma.

LEMMA 2.4. *An interval graph with at least* 3 *vertices is Hamiltonian if and only if it is* 1*-tough.*

LEMMA 2.5. *Let G be a* 2*-connected chordal graph and e an edge of G. Then e is on a triangle of G.*

Proof. Let T be a smallest cycle containing e. Since every cycle of length at least 4 must contain a chord, T is a triangle. □

3. Proof of Theorem 1.2. Suppose, to the contrary, there is a Hamiltonian interval graph G and a non-Hamiltonian cycle C of G such that C is not extendable. Furthermore, we assume that $|G|$, the order of G, is minimum with respect to this assumption.

The strategy of the proof is to find a critical cut A of $H = G[V(C) \cup \{v\}]$ such that $H - A$ has $|A|$ components, there is a component of $G - V(H)$ adjacent only to vertices in A, and every other component of $G - V(H)$ is adjacent only to A and vertices in at most one component of $H - A$. Thus, $G - A$ has more components than $|A|$, a contradiction to the fact that G is 1-tough (violating Lemma 2.4).

Since C is a Hamiltonian cycle in $G[V(C)]$, we can assume that there exist two vertices x and y such that $(C : x, y)$ is a standard Hamiltonian cycle of $G[V(C)]$. Further, x appears only in D_1 and y appears only in D_m, where the ordering of $D_1, D_2, \ldots, D_m$ is the linear ordering of maximal cliques of $G[V(C)]$. Let P_1 and P_2 be the two x-y paths induced by C. Let q_i be an embedding of P_i for each $i = 1$, 2, respectively. Since x appears only in D_1, all neighbors of x are adjacent. So, without loss of generality, we assume that $q_1(x) = q_2(x)$. Similarly, we assume that $q_1(y) = q_2(y)$. For convenience, we define $q_C(v) = q_i(v)$ if $v \in P_i$.

Let B be a Hamiltonian cycle of G and assume that B has a given orientation. Since B is a cycle, $B - V(C)$ is a union of disjoint segments. Let $B(a_i, b_i), i = 1, 2, \ldots,$ denote those nonempty segments, where a_i and b_i are in $V(C)$. A segment $B(a_i, b_i)$ is

a *type*-1 segment if a_i and b_i are adjacent in G. Otherwise, we call $B(a_i, b_i)$ a *type*-2 *segment*.

CLAIM 3.1. *If $B(a_i, b_i)$ is a type-1 segment, then there is a vertex $c_i \in B(a_i, b_i)$ such that c_i is adjacent to both a_i and b_i.*

Proof. Since $a_i B(a_i, b_i] a_i$ is a cycle and G is a chordal graph, by Lemma 2.5, $a_i b_i$ is on a triangle in the subgraph induced by this cycle. Let c_i be the other vertex of this triangle. Clearly, c_i is adjacent to both a_i and b_i. □

CLAIM 3.2. *All $B(a_i, b_i)$ are type-2 segments.*

Proof. Suppose, to the contrary, that $B(a_1, b_1)$ is a type-1 segment. Let $G^* = G - V(B(a_1, b_1))$ and $B^* = B \cup \{a_1 b_1\} - V(B(a_1, b_1))$. Clearly, B^* is a Hamiltonian cycle of G^* and $V(G^*) \supset V(C)$. If $|G^*| > |C|$, G^* is cycle extendable by the induction hypothesis. Thus, C is extendable in G^*, so it is extendable in G, a contradiction. If $|G^*| = |C|$, then by Claim 3.1, there exists $c_1 \in B(a_1, b_1)$ such that $a_1 c_1$, $b_1 c_1 \in E$. Then, $C^* = B[b_1, a_1] c_1 b_1$ is an extension of C, a contradiction. □

Let $H := G[V(C)]$ and for any $v \in V(G) - V(C)$ let $H_v := G[V(C) \cup \{v\}]$. The following claim is a direct consequence of the fact that cycles are 1-tough.

CLAIM 3.3. *If A is a critical cut of H, then A does not contain two consecutive vertices of C and all segments of $C - A$ induce components of $H - A$. Thus, all segments of $C - A$ induce disjoint intervals on the real line.*

CLAIM 3.4. *If $v \notin V(H)$ has at least two neighbors in H, there exists a nontrivial critical cut A of H such that $N(v) \subseteq A$.*

Proof. Since H_v is a non-Hamiltonian interval graph, it is not 1-tough. Hence, there is a cut A of H_v such that $c(H_v - A) \geq |A| + 1$. Since H is a Hamiltonian interval graph, it is 1-tough. Thus, v itself is a component of $H_v - A$, A is a critical cut of H, and $N(v) \subseteq A$. □

CLAIM 3.5. *For each segment $B(a_i, b_i)$ there exists $c_i \in B(a_i, b_i)$ such that c_i has two neighbors on C. Thus, H has a nontrivial critical cut.*

Proof. Since $I(a_i) \cap I(b_i) = \emptyset$, let I denote the interval between $I(a_i)$ and $I(b_i)$. Then, $I \subseteq I(B(a_i, b_i))$. Let $c_i \in B(a_i, b_i)$ such that $I(c_i) \cap I(a_i) \neq \emptyset$ and $I(c_i) \cap I \neq \emptyset$. Since C is connected, $I \subseteq I(C)$, so that there exists $d_i \in V(C) - \{a_i, b_i\}$ such that $c_i d_i \in E$. Thus, $|N_C(c_i)| \geq 2$, and we are done by Claim 3.4. □

CLAIM 3.6. *Let A be a nontrivial critical cut of H. Then, $I(S)$ are disjoint intervals for all components $S \subseteq H - A$. If there exists a path P in $G - V(C)$ connecting two components S and T of $H - A$, then $I(S)$ and $I(T)$ must be two consecutive intervals in $I(H - A)$.*

Proof. The first part of Claim 3.6 is trivial. To prove the second part of the claim, suppose, to the contrary, there is a component R of $H - A$ such that $I(R)$ is between $I(S)$ and $I(T)$. So $I(R) \subset I(P)$. Let $r \in R$. Then, $q_C(r) \in I(r) \subset I(P)$, so that there is a vertex $w \in P$ such that $q_C(r) \in I(w)$. Since $q_C(r)$ is contained in two consecutive vertices of C, w can be inserted into cycle C to make a larger cycle, which is a contradiction. □

Recall that $(C : x, y)$ is a standard Hamiltonian cycle in H. If A is a critical cut of H, A does not contain two consecutive vertices of C and each component of $C - A$ induces a component of $H - A$.

CLAIM 3.7. *For any nontrivial critical cut A of H, $x \notin A$ and $y \notin A$.*

Proof. Since $x \in D_1$, all neighbors of x in H are adjacent. Thus, $x \notin A$. Similarly, $y \notin A$. □

CLAIM 3.8. *For every component D of $G - V(C)$, there exists a nontrivial critical cut A of H such that $N(D) \subseteq A$; i.e., all neighbors of D are in A.*

Proof. Let D be a component of $G - V(C)$ and $v \in D$. We assume, without loss of generality, $v \in B(a_1, b_1)$. By Claim 3.4, let $A := A_v$ be a critical cut of H such that $N_C(v) \subseteq A$.

Note that $I(H - A)$ is a union of disjoint intervals and each such interval corresponds to a component of $H - A$. Let L be the component of $H - A$ such that $I(L)$ is the closest interval of $I(H - A)$ on the left side of v and let R be the component of $H - A$ such that $I(R)$ is the closest interval of $I(H - A)$ on the right side of v. Since A is critical and $N_C(v) \subseteq A$, such components L and R exist. We assume that $|V(L)| + |V(R)|$ is at its minimum over all nontrivial critical cuts $A := A_v$.

We claim that $N_H(D) \subseteq A$. Suppose, to the contrary, that $N_H(D) \not\subseteq A$; then we have $N_H(D) \cap V(L \cup R) \neq \emptyset$. Assume, without loss of generality, that for $w \in D$, we have $N_C(w) \cap V(R) \neq \emptyset$ and $\text{dist}_D(v, w)$ is minimum with this property. Let $P[v, w]$ be a shortest path in D connecting v and w. Then, $N(P[v, w)) \cap R = \emptyset$.

Since $N(w) \cap R \neq \emptyset$ and $N(P[v, w)) \cap R = \emptyset$, $I(w)$ must contain the left-end of $I(R)$. Since there are two paths from x to R along C, then $|N_C(w)| \geq 2$. By Claim 3.4, let $A^* := A_w$, be a nontrivial critical cut of H such that $N_C(w) \subseteq A^*$. Let

$$A_L = \{a \in A \ : \ a \text{ is on the left side of } w\},$$

$$A^*_R = \{a^* \in A^* : \ a^* \text{ is not on the left side of } a\},$$

$$X = A_L \cup A^*_R.$$

We will show that X is a critical cut of H. Note that

- each component S of $H - X$ such that $I(S)$ is on the left side of $I(w)$ is a component of $H - A$,
- each component S of $H - X$ such that $I(S)$ is on the right side of $I(w)$ is a component of $C - A^*$, and
- there is no component S of $H - X$ such that $I(S)$ is between $I(v)$ and $I(w)$.

Thus, X is a cut of H, and, by Claim 3.3, in order to show that X is a critical cut, we need only show that X does not contain two consecutive vertices of C. Suppose, to the contrary, there are two consecutive vertices a and b on C and $a, b \in X$. Without loss of generality, we assume that $a \in A \setminus A^*$ and $b \in A^* \setminus A$. By the definition of X, a is on the left side of w and b is not on the left side of w. Thus, $b \in R$. Since $q_C(a)$ is on the left side of w and $q_C(a) \in I(b)$ (because a and b are consecutive on C), $I(P[v, w)] \cap I(b) \neq \emptyset$, which contradicts the minimality of $P[v, w]$.

Let R^* be the component of $H - X$ such that $I(R^*)$ is the closest interval of $I(H - X)$ on the right side of w and let $I := I(P[v, w])$. Note that if $x \in V(C)$ such that $I(x) \cap I \neq \emptyset$, then either $x \in A^*$ or $x \in A$. In any case, we have that $x \in X$. Note that R is induced by a segment of C. Let y_0 be the first vertex along the segment of R from left to right such that $y_0 w \in E(G)$. Without loss of generality, we assume that $y_0 \in P_1$. Let x_0 be the predecessor of the segment R along P_1 from x to y and let x_0^- be the predecessor of x_0. Since X does not contain two consecutive vertices of C and $x_0 \in X$, $q_C(x_0^-)$ must lie on the left side of $I(R)$. Since C is a standard cycle of H, $q_C(x_0^-) \notin I$. Thus, $q_C(x_0^-)$ is on the left side of the interval of c_1. Thus, $x_0 \in A \cap A^*$. Let S be the segment of R from the first vertex of R to the predecessor of y_0. We first note that $S \neq \emptyset$ (since X does not contain two consecutive vertices). Thus, S is a component of $H - X$.

We claim that $|V(R^*)| < |V(R)|$, which leads to a contradiction of the minimality of $|V(L)| + |V(R)|$. This is certainly true if $R^* = S \subset R$. Suppose $R^* \neq S$. Then,

$I(R^*)$ is between $I(w)$ and $I(S)$. From the definition of R and S, we have $R^* \subseteq A$. Since A does not contain two consecutive vertices of C, $|V(R^*)| = 1$. Since $|V(R)| \geq |V(S)| + 1 \geq 2$, we have $|V(R^*)| < |V(R)|$, as desired. □

Let $D_1, D_2, \ldots, D_m$ be the components of $G - V(C)$. Assume, without loss of generality, that $I(D_i)$ is on the left side of $I(D_j)$ whenever $i < j$. By Claim 3.8, for each D_i, $I(D_i) \subseteq I(C)$ and there exists a nontrivial critical cut A_i of H such that $N(D_i) \subseteq A_i$. Let $LA_i = \{a \in A_i \mid a \text{ is on the left side of } A_i\}$ and $RA_i = A - LA_i$. We now inductively define B_i for each $i = 1, 2, \ldots, m$ as follows: $B_1 = A_1$ and, for each $i > 1$, if D_i is adjacent to at most one component of $H - B_{i-1}$, let $B_i = B_{i-1}$. Otherwise, let

$$B_i = \{b \in B_{i-1} \mid b \text{ is on the left side of } D_i\} \cup RA_i.$$

CLAIM 3.9. *B_i is a nontrivial critical cut for each $i = 1, 2, \ldots, m$.*

Proof. Claim 3.9 is true for $i = 1$. Suppose it is true for $i - 1 \geq 1$. If $B_i = B_{i-1}$, then it is also true for i. So, we assume that $B_i \neq B_{i-1}$. In this case, let L and R be two components of $H - B_{i-1}$ such that $N(D_i) \cap L \neq \emptyset$ and $N(D_i) \cap R \neq \emptyset$. By Claim 3.6, $I(L)$ and $I(R)$ are two consecutive intervals of $I(H - B_{i-1})$. Furthermore, $I(D_i)$ contains the interval between L and R as a subinterval. Note that components of $H - B_i$ on the left side of D_i are those of $H - B_{i-1}$ and components of $H - B_i$ on the right side of D_i are those of $H - A_i$. In order to show that B_i is a critical cut, we only need show that B_i does not contain two consecutive vertices of C. Suppose, to the contrary, a and b are two consecutive vertices on C such that $a \in B_{i-1} \setminus A_i$ and $b \in A_i \setminus B_{i-1}$. Since $b \notin B_{i-1}$, $q_C(a) \in I(a) \cap I(b)$ must be on the right side of L. Similarly, $q_C(a)$ must be on the left side of R. Thus, $q_C(a) \in I(D_i)$, so that there exists $w \in I(D_i)$ adjacent to both a and b. Then, C is extendable, which is a contradiction. □

By the definition, we have $N(D_1) \subseteq B_m$ and, for each $i > 1$, either $N(D_i) \subseteq B_m$ or D_i is adjacent to at most one component of $H - B_m$. Since $H - B_m$ has exactly $|B_m|$ components, $G - B_m$ has at least $|B_m| + 1$ components, which contradicts the fact that G is 1-tough. This contradiction completes the proof. □

Note: Just at the time of submission we were informed of another proof of this result in [1].

Acknowledgment. We thank the referees for their useful suggestions which led to a simpler proof.

REFERENCES

[1] A. ABUEIDA AND R. SRITHARAN, *Cycle extendable and Hamilton cycles in chordal graph classes*, SIAM J. Discrete Math., 20 (2006), pp. 669–681.
[2] J. A. BONDY, *Pancyclic graphs*, J. Combin. Theory Ser. B, 11 (1977), pp. 80–84.
[3] G. CHEN, G. FAN, AND X. YU, *Cycles in 4-connected planar graphs*, European J. Combin., 25 (2004), pp. 763–780.
[4] G. CHEN, M. S. JACOBSON, A. KÉZDY, AND J. LEHEL, *Tough enough chordal graphs are Hamiltonian*, Networks, 31 (1998), pp. 29–38.
[5] V. CHVÁTAL, *Tough graphs and Hamiltonian circuits*, Discrete Math., 5 (1973), pp. 215–228.
[6] P. C. GILMORE AND A. J. HOFFMAN, *A characterization of cocomparability graphs and of interval graphs*, Canad. J. Math., 16 (1964), pp. 539–548.
[7] G. R. T. HENDRY, *Extending cycles in graphs*, Discrete Math., 85 (1990), pp. 59–72.
[8] J. M. KEIL, *Finding Hamiltonian circuits in interval graphs*, Inform. Process. Lett., 20 (1985), pp. 201–206.
[9] J. MALKEVITCH, *Polytopal graphs*, in Selected Topics in Graph Theory 3, Beneike and R. Wilson, eds., Academic Press, New York, 1988, pp. 169–188.

[10] O. Ore, *Note on Hamilton circuits*, Amer. Math Monthly, 67 (1960), p. 55.
[11] D. P. Sanders, *On paths in planar graphs*, J. Graph Theory, 24 (1997), pp. 341–345.
[12] R. Thomas and X. Yu, 4-*connected projective-planar graphs are Hamiltonian*, J. Combin. Theory Ser. B, 62 (1994), pp. 114–132.
[13] T. Tutte, *A theorem on planar graphs*, Trans, Amer. Math. Soc., 82 (1956), pp. 99–116.
[14] D. West, *Introduction to Graph Theory*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 2001.
[15] H. Whitney, *A theorem on graphs*, Ann. of Math. (2), 32 (1931), pp. 378–390.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 690–704

THE CHANNEL ASSIGNMENT PROBLEM WITH VARIABLE WEIGHTS*

DANIEL KRÁL'†

Abstract. A λ-graph G is a (finite or infinite) graph with k types of edges, x_1-edges, ..., x_k-edges. A labeling c of the vertices of G by nonnegative reals is proper with respect to reals $x_1, \ldots, x_k$ if the labels of the end-vertices of an x_i-edge differ by at least x_i. The span of the labeling c is the supremum of the labels used by c. The λ-function $\lambda_G(x_1, \ldots, x_k)$ is the infimum of the spans of all the proper labelings with respect to $x_1, \ldots, x_k$. We show that the λ-function of any graph G is piecewise linear in $x_1, \ldots, x_k$ with finitely many linear parts (unless the λ-function is infinite). Moreover, we show that for all integers k and χ, there exist constants $C_{k,\chi}$ and $D_{k,\chi}$ such that the λ-function of every λ-graph G with k types of edges and chromatic number at most χ is comprised of at most $C_{k,\chi}$ linear parts, and that the coefficients of $x_1, \ldots, x_k$ of the linear functions comprising $\lambda_G(x_1, \ldots, x_k)$ are integers between 0 and $D_{k,\chi}$. Among others, our results yield proofs of the piecewise linearity conjecture, coefficient bound conjecture, and delta bound conjecture of Griggs and Jin [*SIAM J. Discrete Math.*, 20 (2006), pp. 302–327].

Key words. channel assignment problem, graph labeling with distance conditions

AMS subject classification. 05C15

DOI. 10.1137/040619636

1. Introduction.

Radio frequency problems can be expressed as various graph labeling problems [14, 20]. A prominent role among such problems is played by the notion of $L(p_1, \ldots, p_k)$-labelings, also referred to as graph labelings with distance constraints. So far, the study of the dependence of optimal $L(p_1, \ldots, p_k)$-labelings on the parameters $p_1, \ldots, p_k$ has not been studied too intensively; however, several new approaches to studying this dependence have recently been proposed: An approach based on real-value relaxation of $L(p_1, \ldots, p_k)$-labelings can be found in the work of Griggs and Jin [9, 10, 11], and, more recently, another approach based on the notion of λ-graphs was developed in [2]. In the present paper, we generalize the notion of λ-graphs introduced in [2] from $k = 2$ to arbitrary k and provide structural results for the general model. The results we obtain yield proofs of the piecewise linearity conjecture, coefficient bound conjecture, and delta bound conjecture of Griggs and Jin [9].

A labeling c of the vertices of a (finite or infinite) graph G by nonnegative integers is an $L(p_1, \ldots, p_k)$-*labeling* for positive integers $p_1, \ldots, p_k$ if the labels of any two vertices u and v at distance (exactly) i differ by at least p_i. Let us remark here that all graphs as well as λ-graphs considered in this paper can be finite or infinite unless stated otherwise. The supremum of the labels used by c is said to be the *span* of c, and the least span of an $L(p_1, \ldots, p_k)$-labeling of a graph G is denoted by $\lambda_G(p_1, \ldots, p_k)$

*Received by the editors November 25, 2004; accepted for publication (in revised form) April 14, 2006; published electronically September 15, 2006.
http://www.siam.org/journals/sidma/20-3/61963.html

†Institute for Mathematics, Technical University Berlin, Strasse des 17. Juni 136, D-10623 Berlin, Germany. The author was a postdoctoral fellow at TU Berlin within the framework of the European training network COMBSTRU from October 2004 to July 2005. Department of Applied Mathematics and Institute for Theoretical Computer Science (ITI), Faculty of Mathematics and Physics, Charles University, Malostranské nám. 25, 118 00 Prague, Czech Republic (kral@kam.mff.cuni.cz). Institute for Theoretical Computer Science is supported by Ministry of Education of Czech Republic as project 1M0545. The author is now a Fulbright scholar at School of Mathematics, Georgia Institute of Technology, 686 Cherry St., Atlanta, GA 30332-0160 (kral@math.gatech.edu).

(we deviate here from the standard notation in order to emphasize the dependence on the parameters $p_1, \ldots, p_k$). In the extensive literature on $L(p_1, \ldots, p_k)$-labelings, one can find many papers on algorithms for $L(p_1, \ldots, p_k)$-labelings of (finite) graphs [1, 4, 7, 8, 17, 21]. From the structural point of view, the attention of researchers focused mainly on the case of $L(2,1)$-labelings, partly because of the following conjecture of Griggs and Yeh [12].

CONJECTURE 1.1 (Δ^2 conjecture). *If G is a finite graph of maximum degree $\Delta \geq 2$, then $\lambda_G(2,1) \leq \Delta^2$.*

Conjecture 1.1 was verified for several special classes of graphs, including graphs of maximum degree two, chordal graphs [22] (see also [5, 18]), hamiltonian cubic graphs [15, 16], and planar graphs with maximum degree $\Delta \neq 3$ [3]. In the general case, the original bound $\lambda_G(2,1) \leq \Delta^2 + 2\Delta$ from [12] has been improved to $\lambda_G(2,1) \leq \Delta^2 + \Delta$ in [6]. A recent more general result of the author and Škrekovski [19] yields $\lambda_G(2,1) \leq \Delta^2 + \Delta - 1$. The present record $\lambda_G(2,1) \leq \Delta^2 + \Delta - 2$ has been recently established by Gonçalves [13].

In order to capture the dependence of the optimum spans on the parameters, Griggs and Jin [9] allow the parameters $p_1, \ldots, p_k$ and the labels used by a labeling c to be any nonnegative reals. Similarly to the original notion, they define the *span* of a labeling c as the supremum of the labels used by c, and $\lambda_G(p_1, \ldots, p_k)$ denotes the span of an optimum labeling of a graph G, i.e., the minimum (that is always attained if $\lambda_G(p_1, \ldots, p_k)$ is finite) of the spans of all $L(p_1, \ldots, p_k)$-labelings of G. The function λ_G is a function from $\mathbb{R}_+^k$ to $\mathbb{R}_+$, where $\mathbb{R}_+$ is the set of nonnegative reals. Note that $\lambda_G(p_1, \ldots, p_k)$ is finite if the maximum degree of G is bounded. In this setting, Griggs and Jin [9] prove that for any reals $p_1, \ldots, p_k$, the value of $\lambda_G(p_1, \ldots, p_k)$ of a graph G with bounded maximum degree is equal to $\sum_{i=1}^k \alpha_i p_i$ for some nonnegative integers α_i, and if all $p_1, \ldots, p_k$ are integers, the values $\lambda_G(p_1, \ldots, p_k)$ in the original and the relaxed settings coincide. Among others, Griggs and Jin [9] also show that the function $\lambda_G(p_1, \ldots, p_k)$ is a continuous function piecewise linear in the parameters $p_1, \ldots, p_k$; i.e., the set $\mathbb{R}_+^k$ can be partitioned into parts (of positive measure) such that $\lambda_G(p_1, \ldots, p_k)$ is linear on each of the parts. They do not prove that the number of such parts is finite for $k > 2$ and conjecture that the following more general statements actually hold [9].

CONJECTURE 1.2 (piecewise linearity conjecture). *For every graph G with bounded maximum degree, the function $\lambda_G(p_1, \ldots, p_k)$ is comprised of finitely many linear parts; i.e., $\mathbb{R}_+^k$ can be partitioned into finitely many parts such that $\lambda_G(p_1, \ldots, p_k)$ is linear on each of the parts.*

CONJECTURE 1.3 (coefficient bound conjecture). *For every graph G with bounded maximum degree and every integer k, there exists a constant $D_{k,G}$ such that the following holds for all reals $p_1, \ldots, p_k$: The value of the function $\lambda_G(p_1, \ldots, p_k)$ is equal to $\sum_{i=1}^k \alpha_i p_i$ for some integer coefficients $\alpha_1, \ldots, \alpha_k$ between 0 and $D_{k,G}$ (the integers $\alpha_1, \ldots, \alpha_k$ depend on $p_1, \ldots, p_k$). Moreover, there is a labeling c with span $\lambda_G(p_1, \ldots, p_k)$ such that $c(v) = \sum_{i=1}^k \alpha_i(v) p_i$, where $\alpha_1(v), \ldots, \alpha_k(v)$ are between 0 and $D_{k,G}$.*

CONJECTURE 1.4 (delta bound conjecture). *For all integers Δ and k, there exists a constant $D_{k,\Delta}$ such that for every graph G with maximum degree at most Δ and all reals $p_1, \ldots, p_k$ the following holds: The value of the function $\lambda_G(p_1, \ldots, p_k)$ is equal to $\sum_{i=1}^k \alpha_i p_i$ for some integer coefficients $\alpha_1, \ldots, \alpha_k$ between 0 and $D_{k,\Delta}$ (the integers $\alpha_1, \ldots, \alpha_k$ depend on $p_1, \ldots, p_k$). Moreover, there is a labeling c with span $\lambda_G(p_1, \ldots, p_k)$ such that $c(v) = \sum_{i=1}^k \alpha_i(v) p_i$, where $\alpha_1(v), \ldots, \alpha_k(v)$ are integers*

between 0 *and* $D_{k,\Delta}$.

Note that the delta bound conjecture implies the coefficient bound conjecture. Griggs and Jin [9] proved the three Conjectures 1.2, 1.3, and 1.4 for $k = 2$ (for $k = 1$, the conjectures are trivial) as well as Conjecture 1.2 for finite graphs G (for all values of k).

In the present paper, we consider the problems posed in [9] in the more general setting of λ-graphs introduced for $k = 2$ in [2]. A *λ-graph G with k types of edges* is a graph G whose edges are labeled by variables $x_1, \ldots, x_k$. An edge labeled by a variable x_i is called an *x_i-edge*. Two vertices of G may be joined by edges of several types. A *proper labeling* c of G with respect to real numbers $x_1, \ldots, x_k$ is a labeling of the vertices of G by nonnegative reals such that the labels of the end-vertices of an x_i-edge uv differ by at least x_i; i.e., $|c(u) - c(v)| \geq x_i$. The *span* of the labeling c is the supremum of the labels used by c, and $\lambda_G(x_1, \ldots, x_k)$ is defined to be the infimum of the spans of all proper labelings with respect to $x_1, \ldots, x_k$. Using the technique developed in [2], we show in section 2 that for all reals $x_1, \ldots, x_k$, if $\lambda_G(x_1, \ldots, x_k)$ is finite, then there exists a proper labeling c with span $\lambda_G(x_1, \ldots, x_k)$ and the span of c is equal to the maximum label used by c; i.e., both the infimum and the supremum in the definitions are attained. The *λ-function* of a λ-graph G is $\lambda_G(x_1, \ldots, x_k)$ viewed as a function of variables $x_1, \ldots, x_k$; i.e., λ_G is a function from $\mathbb{R}_+^k$ to $\mathbb{R}_+$. The *chromatic number* $\chi(G)$ of a λ-graph G is the chromatic number of the underlying graph; i.e., $\chi(G) = \lambda_G(1, \ldots, 1) + 1$.

$L(p_1, \ldots, p_k)$-labelings of graphs can be modeled as λ-graphs as follows: If G is a graph, form a λ-graph $G^{(k)}$ with the vertex set $V(G)$ such that two vertices u and v are joined by an x_i-edge in $G^{(k)}$, $i = 1, \ldots, k$, if their distance in G is exactly i. Clearly, the optimum span $\lambda_G(p_1, \ldots, p_k)$ is equal to the value $\lambda_{G^{(k)}}(x_1, \ldots, x_k)$ of the λ-function of $G^{(k)}$ for $x_i = p_i$, $i = 1, \ldots, k$. Because of this close relation, we decided to use the notation $\lambda_G(\ldots)$ for both the spans of optimum $L(p_1, \ldots, p_k)$-labelings and the λ-functions of λ-graphs. Since it is always clear throughout the paper whether G is a graph (in which case $\lambda_G(p_1, \ldots, p_k)$ stands for the span of an optimum $L(p_1, \ldots, p_k)$-labeling) or a λ-graph (in which case λ_G stands for the λ-function of G), the confusion of the notations is avoided.

As in the case of $L(p_1, \ldots, p_k)$-labeling [9], λ-functions of λ-graphs have the *scaling property*; i.e., for all nonnegative reals $x_1, \ldots, x_k$ and β, the following holds: $\lambda_G(\beta x_1, \ldots, \beta x_k) = \beta \lambda_G(x_1, \ldots, x_k)$. Therefore, the λ-function of any λ-graph is linear on every ray through the origin in $\mathbb{R}_+^k$. Let us remark that we implicitly assume that the λ-function is finite on its domain (see the remark after Proposition 2.3). In section 4, we show that the λ-function λ_G of any λ-graph G is comprised of finitely many linear parts; i.e., $\mathbb{R}_+^k$ can be partitioned into finitely many parts such that λ_G is linear on each of the parts. Because of the scaling property, it is easy to see that there is such a partition of $\mathbb{R}_+^k$ where all the parts are infinite polyhedral cones with the tips at the origin.

Our main result is Theorem 4.1, which asserts the existence of the constants $C_{k,\chi}$ and $D_{k,\chi}$ such that the λ-function of any λ-graph G with k types of edges and chromatic number at most χ is comprised of at most $C_{k,\chi}$ linear parts, and the coefficients of $x_1, \ldots, x_k$ of the linear functions comprising the λ-function are integers between 0 and $D_{k,\chi}$. In fact, there exists a partition of $\mathbb{R}_+^k$ into $C_{k,\chi}$ infinite polyhedral cones such that the λ-function of each λ-graph G with k types of edges and chromatic number at most χ is linear on each of the cones; i.e., there is such a partition that does not depend on G. In this paper, we solely focus on proving the existence of the constants $C_{k,\chi}$ and $D_{k,\chi}$ without attempting to optimize their

growth. Let us remark that the existence of the constants $C_{k,\chi}$ and $D_{k,\chi}$ for $k = 2$ follows from the results of [2]. However, the technique used in [2] does not seem to generalize to $k > 2$. As demonstrated in section 5, our main result yields the proofs of the piecewise linearity conjecture, coefficient bound conjecture, and delta bound conjecture for $L(p_1, \ldots, p_k)$-labelings (Conjectures 1.2, 1.3, and 1.4).

2. Preliminaries. The very first natural questions on λ-graphs are whether the infimum in the definition of the function $\lambda(x_1, \ldots, x_k)$ is always attained and whether the span of every optimal labeling c is equal to the maximum label (the supremum in the definition of the span is attained). The positive answers to these two questions are provided by an analogue of the Gallai–Roy theorem for infinite graphs with edges of finitely many different weights proved in [2]. We will not state this theorem in its full generality but just in the form restricted to λ-graphs. An orientation of an infinite graph G is said to be *finitary* if the maximum length of its directed walks is bounded. In particular, a finitary orientation of G is acyclic. A *weight* of a finite directed path P in an orientation of a λ-graph G with respect to $x_1, \ldots, x_k$ is the sum of the variables assigned to its edges, i.e., $\sum_{i=1}^{k} \alpha_i x_i$ if P contains α_i x_i-edges. The *weight of a finitary orientation* $\vec{G}$ of a λ-graph G is the maximum weight of a directed path in $\vec{G}$ (note that the maximum is always attained since the lengths of directed paths in $\vec{G}$ are bounded in a finitary orientation and there are only finitely many different types of edges in G). We can now state the version of the Gallai–Roy theorem for λ-graphs.

THEOREM 2.1. *Let G be a λ-graph with k types of edges. For any real numbers $x_1, \ldots, x_k$, if the value of $\lambda_G(x_1, \ldots, x_k)$ is finite, then it is equal to the minimum weight of a finitary orientation $\vec{G}$ of G (in particular, there exists a finitary orientation of weight $\lambda_G(x_1, \ldots, x_k)$).*

Let us remark that the proof of Theorem 2.1 involves the axiom of choice.

If $\vec{G}$ is a finitary orientation of a λ-graph G, then the labeling c, where $c(v)$ is the maximum weight of a directed path ending at a vertex v, is a proper labeling of G with respect to $x_1, \ldots, x_k$ (note that there is always a path with maximum weight). We say that the labeling c, defined in this way, *corresponds* to the orientation $\vec{G}$. Clearly, the span of the labeling corresponding to $\vec{G}$ is the weight of $\vec{G}$. On the other hand, for a proper labeling c of G for positive reals $x_1, \ldots, x_k$ that has a finite span, one may define a (finitary) orientation $\vec{G}$ of G such that an edge uv is directed from u to v if $c(u) < c(v)$. We say that this orientation $\vec{G}$ *corresponds* to the labeling c. Observe that the weight of the orientation corresponding to a proper labeling c is at most its span and, in general, the weight can be strictly smaller.

We now answer the two questions posed at the beginning of our discussion.

PROPOSITION 2.2. *Let G be a λ-graph with k types of edges and $x_1, \ldots, x_k$ parameters such that $\lambda_G(x_1, \ldots, x_k)$ is finite. There exists a proper labeling c of G with respect to $x_1, \ldots, x_k$ with span $\lambda_G(x_1, \ldots, x_k)$. Moreover, every labeling c with span $\lambda_G(x_1, \ldots, x_k)$ that is proper with respect to $x_1, \ldots, x_k$ contains a vertex v with $c(v) = \lambda_G(x_1, \ldots, x_k)$.*

Proof. We can assume without loss of generality that all the parameters $x_1, \ldots, x_k$ are positive: If this is not true, then we consider the λ-graph formed by x_i-edges for $x_i > 0$. By Theorem 2.1, there exists a finitary orientation $\vec{G}$ of G with weight equal to $\lambda_G(x_1, \ldots, x_k)$. Let c be the labeling corresponding to the orientation $\vec{G}$. Since the span of c is equal to $\lambda_G(x_1, \ldots, x_k)$, the first part of the statement of the proposition follows.

Let c be a labeling with span $\lambda_G(x_1,\ldots,x_k)$. Assume to the contrary that $c(v) < \lambda_G(x_1,\ldots,x_k)$ for all vertices v. Let $\vec{G}$ be a finitary orientation corresponding to c and let c' be the labeling corresponding to $\vec{G}$ with respect to $x_1,\ldots,x_k$. It is easy to observe that $c'(v) \leq c(v)$ for each vertex v. Since the weight of $\vec{G}$ is equal to $\max c'(v)$ and $c'(v) \leq c(v) < \lambda_G(x_1,\ldots,x_k)$, we infer that the span of c' is smaller than $\lambda_G(x_1,\ldots,x_k)$. However, this is impossible by the definition of $\lambda_G(x_1,\ldots,x_k)$. □

The definition of λ-functions does not guarantee that the function is finite for all values of $x_1,\ldots,x_k$. But at least the following proposition holds.

PROPOSITION 2.3. *The λ-function of a λ-graph G with k types of edges is finite for all $x_1,\ldots,x_k \in \mathbb{R}^k_+$ if and only if it is finite for some positive reals $x_1,\ldots,x_k$.*

Proof. Clearly, it is enough to prove that if $\lambda(x_1,\ldots,x_k)$ is finite for some positive reals $x_1,\ldots,x_k$, then $\lambda(y_1,\ldots,y_k)$ is finite for all nonnegative reals $y_1,\ldots,y_k$. Let c be a labeling with span at most $\lambda(x_1,\ldots,x_k)$ (it exists by Proposition 2.2). Let us define a new labeling c' as follows:

$$c'(v) = \frac{\max\{y_1,\ldots,y_k\}}{\min\{x_1,\ldots,x_k\}} c(v) \ .$$

It is straightforward to verify that c' is a proper labeling of G with respect to $y_1,\ldots,y_k$ and its span is equal to $\frac{\max\{y_1,\ldots,y_k\}}{\min\{x_1,\ldots,x_k\}}\lambda(x_1,\ldots,x_k)$. Hence, $\lambda(y_1,\ldots,y_k)$ is finite. □

In the rest of this paper, we always implicitly assume that the λ-function of a considered λ-graph is finite for all nonnegative reals.

If G is a λ-graph, we say that an edge uv is an $x_{\leq\ell}$-edge if uv is an x_i-edge where $i \leq \ell$. Similarly, we use the terms $x_{<\ell}$-edges, $x_{\geq\ell}$-edges, etc. We demonstrate the notation introduced in the next auxiliary lemma that will be used later. Though the lemma is quite easy to prove, we decided to include its proof to demonstrate our notation.

LEMMA 2.4. *Let G be a λ-graph with k types of edges and with chromatic number at most χ, and let $0 \leq \ell < k$. If there exist an integer D and a finitary orientation $\vec{G}$ of G such that every directed path in $\vec{G}$ contains at most D $x_{\leq\ell}$-edges, then*

$$\lambda_G(x_1,\ldots,x_k) \leq d_{\max} + (\ell+1)^D \cdot \chi \cdot \max\{x_{\ell+1},\ldots,x_k\},$$

where $d_{\max}$ is the maximum sum of weights of $x_{\leq\ell}$-edges on a directed path in $\vec{G}$; i.e., $d_{\max}$ would be the weight of $\vec{G}$ if the parameters $x_{\ell+1},\ldots,x_k$ were equal to zero.

In particular, it holds that $\lambda_G(x_1,\ldots,x_k) \leq \chi \cdot \max\{x_1,\ldots,x_k\}$.

Proof. Fix a finitary orientation $\vec{G}$ that has the properties described in the statement of the lemma. If $\ell = 0$, fix any finitary orientation $\vec{G}$ of G (note that G has a finitary orientation because its chromatic number is finite). Let $d(v)$ be the maximum sum of the weights of $x_{\leq\ell}$-edges on a directed path in $\vec{G}$ ending at a vertex v; i.e., $d(v)$ would be the label of v corresponding to the orientation $\vec{G}$ if the parameters $x_{\ell+1},\ldots,x_k$ were equal to zero. Clearly, $d_{\max} = \max_{v\in V(G)} d(v)$. Let $\mathcal{D}$ be the set of all different values of $d(v)$ and let $\delta(v)$ be the number of the elements of $\mathcal{D}$ smaller than $d(v)$. Since every directed path in $\vec{G}$ contains at most D $x_{\leq\ell}$-edges, it holds that $|\mathcal{D}| \leq (\ell+1)^D$. Hence, $0 \leq \delta(v) < |\mathcal{D}| \leq (\ell+1)^D$ for every vertex v of G. Finally, let μ be a coloring of the vertices of G with colors $1,\ldots,\chi$.

Let us define a labeling c' of the vertices of G as follows:

$$c'(v) = d(v) + (\delta(v)\chi + \mu(v)) \cdot \max\{x_{\ell+1},\ldots,x_k\} \ .$$

Since $\delta(v) < |\mathcal{D}| \leq (\ell+1)^D$ for every vertex v of G, the span of c' does not exceed

$$d_{\max} + |\mathcal{D}|\chi \cdot \max\{x_{\ell+1}, \ldots, x_k\} \leq d_{\max} + (\ell+1)^D \chi \cdot \max\{x_{\ell+1}, \ldots, x_k\}.$$

In the rest, we show that c' is a proper labeling with respect to $x_1, \ldots, x_k$.

Consider an x_i-edge uv of G. By symmetry, we may assume that the edge uv is directed from u to v in $\vec{G}$. In particular, it holds that $d(u) \leq d(v)$. Hence, $\delta(u) \leq \delta(v)$. We distinguish two major cases: The first one is $i \leq \ell$. In this case, $d(u) + x_i \leq d(v)$ and thus $\delta(u) < \delta(v)$. We can immediately conclude that

$$\begin{aligned} c'(v) - c'(u) &= d(v) - d(u) + ((\delta(v) - \delta(u))\chi + \mu(v) - \mu(u)) \cdot \max\{x_{\ell+1}, \ldots, x_k\} \\ &\geq d(v) - d(u) + (\chi + \mu(v) - \mu(u)) \cdot \max\{x_{\ell+1}, \ldots, x_k\} \\ &\geq d(v) - d(u) \geq x_i. \end{aligned}$$

In particular, the edge uv is properly colored in this case.

The other case is that $i > \ell$. If $d(u) = d(v)$, then $\delta(u) = \delta(v)$ and the following holds (similarly to the first case):

$$|c'(u) - c'(v)| = |\mu(u) - \mu(v)| \cdot \max\{x_{\ell+1}, \ldots, x_k\} \geq x_i.$$

If $d(u) < d(v)$, then $\delta(u) < \delta(v)$ and we infer the following:

$$\begin{aligned} c'(v) - c'(u) &= d(v) - d(u) + ((\delta(v) - \delta(u))\chi + \mu(v) - \mu(u)) \cdot \max\{x_{\ell+1}, \ldots, x_k\} \\ &\geq (\chi + \mu(v) - \mu(u)) \cdot \max\{x_{\ell+1}, \ldots, x_k\} \geq x_i. \end{aligned}$$

Hence, the labels of u and v always differ by at least x_i. □

We remark that χ can be replaced by $\chi - 1$ in the estimate on $\lambda_G(x_1, \ldots, x_k)$ of Lemma 2.4—we decided to state the lemma with the slightly worse bound in order to try to keep the formulas simple.

3. Orientations with minimum weight. In this section, we construct orientations of λ-graphs with minimum weight such that the maximum length of a directed path in the constructed orientation is bounded. First, let us define numbers $D_{i,\chi}$ and $K_{i,\chi}$ for integer χ and i recursively as follows:

$$\begin{aligned} D_{1,\chi} &= \chi, \\ K_{i,\chi} &= (i+1)^{D_{i,\chi}}, \text{ and} \\ D_{i+1,\chi} &= (2K_{i,\chi})^{K_{i,\chi}^2+3} \cdot \chi. \end{aligned}$$

Next, we state several propositions that can be verified directly from the definitions of $D_{i,\chi}$ and $K_{i,\chi}$. Their proofs are left to the reader.

PROPOSITION 3.1. *For integers $\chi \geq 2$ and $i \geq 1$, the number of multisets that consist of at most $D_{i,\chi}$ numbers $1, \ldots, i$ does not exceed $K_{i,\chi} - 1$.*

PROPOSITION 3.2. *The following holds for all integers $\chi \geq 2$ and $i \geq 1$:*

$$D_{i+1,\chi} \geq (2K_{i,\chi})^{K_{i,\chi}^2+2} \cdot \chi + K_{i,\chi} \cdot \chi\,.$$

Before proceeding with introducing further notation, let us provide some insights into the statement and the proof of the main lemma of this section (Lemma 3.5). The lemma asserts that for every k and χ, every λ-graph G with k types of edges and chromatic number at most χ has an optimal finitary orientation $\vec{G}$ such that every

directed path of $\vec{G}$ contains at most $D_{\ell,\chi}$ $x_{\leq\ell}$-edges for $\ell = 1, \ldots, k$. In the proof of our main result, we apply the lemma with $\ell = k$ to show that the length of the longest directed walk in an optimal finitary orientation can be bounded by constant that does not depend on $x_1, \ldots, x_k$.

The lemma is proved by induction on ℓ. In the ℓth step, we have already found an optimal finitary orientation $\vec{G}$ and the corresponding labeling c such that every directed path of $\vec{G}$ contains at most $D_{\ell-1,\chi}$ $x_{\leq\ell-1}$-edges. We let $d(v)$ be the labeling of G corresponding to $\vec{G}$ if the parameters $x_\ell, \ldots, x_k$ were equal to zero. As the next step of the proof, we define a new labeling c' as follows: If the labels $c(v)$ and $d(v)$ are close, we define $c'(v) = c(v)$, and we define $c'(v)$ to be approximately $d(v) + \mu(v)x_\ell$, otherwise, where μ is a fixed coloring of G with colors $1, \ldots, \chi$. In the latter case, the value of $c'(v)$ is not exactly $d(v) + \mu(v)x_\ell$, but it is shifted by a value that depends on $d(v)$. We then establish that c' is a proper labeling and the corresponding orientation $\vec{G}'$ does not have directed paths with more than $D_{\ell,\chi}$ $x_{\leq\ell}$-edges.

The notion of "being close" is defined in terms of possible differences between the labels $d(v)$; the set of such differences is the set $|\Gamma'_{D_{\ell-1,\chi}}|$ that is defined later. The size of this set is $K^2_{\ell-1,\chi}$ (see Proposition 3.3). Two labels are considered to be close if their difference is at most $C \cdot t$, where C is a constant (that depends on $d(v)$) and the value of t is chosen to have Property $(\star)$ introduced in the proof of Lemma 3.5. Property $(\star)$ can be rephrased vaguely as follows: "If the difference of labels $d(v)$ and $d(v')$ is larger than t, then it is much larger than t." The value of t can be chosen not to exceed $(2K_{\ell-1,\chi})^{K^2_{\ell-1,\chi}}\chi x_\ell \approx D_{\ell,\chi}x_\ell$. Since the ratio $t/x_\ell \approx D_{\ell,\chi}$ essentially determines the maximum number of x_ℓ-edges contained in a directed path of $\vec{G}'$, the lemma will be established.

We now introduce additional notation used in the proof of the main lemma of this section. For an integer M and positive reals $x_1, \ldots, x_k$, $\Gamma_M(x_1, \ldots, x_k)$ denotes the set of all combinations of $x_1, \ldots, x_k$ with nonnegative integer coefficients whose sum does not exceed M; i.e.,

$$\Gamma_M(x_1, \ldots, x_k) = \left\{ \sum_{j=1}^{k} \alpha_j x_j \text{ for } 0 \leq \alpha_1, \ldots, \alpha_k \text{ and } \sum_{j=1}^{k} \alpha_j \leq M \right\}.$$

Note that $\Gamma_{K_{\ell-1,\chi}}(x_1, \ldots, x_{\ell-1})$ is the set of all possible labels that can be assigned to the vertices of G by the labeling d corresponding to $\vec{G}$. The set $\Gamma'_M(x_1, \ldots, x_k)$ is then defined to be the set of all nonnegative reals that can be expressed as a difference of two numbers from $\Gamma_M(x_1, \ldots, x_k)$; i.e.,

$$\Gamma'_M(x_1, \ldots, x_k) = \{\alpha - \beta | \alpha, \beta \in \Gamma_M(x_1, \ldots, x_k) \text{ and } \alpha - \beta \geq 0\}.$$

Since $0 \in \Gamma_M(x_1, \ldots, x_k)$, the set $\Gamma_M(x_1, \ldots, x_k)$ is a subset of $\Gamma'_M(x_1, \ldots, x_k)$. The following estimates on the sizes of $\Gamma_M(x_1, \ldots, x_k)$ and $\Gamma'_M(x_1, \ldots, x_k)$ directly follow from Proposition 3.1.

PROPOSITION 3.3. *Let $x_1, \ldots, x_k$ be any positive real numbers and let $\chi \geq 2$ be a positive integer. The following two estimates hold:*

$$|\Gamma_{D_{k,\chi}}(x_1, \ldots, x_k)| < K_{k,\chi} \text{ and}$$
$$|\Gamma'_{D_{k,\chi}}(x_1, \ldots, x_k)| < K^2_{k,\chi}.$$

We now establish an auxiliary lemma that is needed in the proof of Lemma 3.5 to define the notion of "being close." As already explained, a label $c(v)$ is close from the label $d(v) < c(v)$ if the difference $c(v) - d(v)$ is at most $C \cdot t$. Lemma 3.5 guarantees the existence of a suitable number t that is bounded from below and above by a constant multiple of a parameter y (chosen later to be x_ℓ). The value of t has to satisfy an additional property (later referred to as Property $(\star)$) that is crucial in establishing that the modified labeling c' is proper.

LEMMA 3.4. *Let $k \geq 1$ and $\chi \geq 2$ be positive integers, S a set of at most $K_{k,\chi}^2 - 1$ positive real numbers, and y another positive real number. There exists a real number t,*

$$K_{k,\chi}\chi y \leq t \leq (2K_{k,\chi})^{K_{k,\chi}^2}\chi y,$$

such that the set S contains no element strictly between t and $K_{k,\chi}(t + \chi y)$. In particular, the real t has the following property:

$$\text{If } \gamma \in S \text{ and } \gamma > t, \text{ then } \gamma \geq K_{k,\chi}(t + \chi y).$$

Proof. Let us define reals t_j, $j = 0, \ldots, K_{k,\chi}^2$, as follows:

$$t_j = (2K_{k,\chi})^j \chi y.$$

Since $t_j < K_{k,\chi}(t_j + \chi y) = K_{k,\chi}(t_j + t_0) \leq 2K_{k,\chi}t_j = t_{j+1}$ for all $j = 1, \ldots, K_{k,\chi}^2 - 1$, all the open intervals I_j,

$$I_j = (t_j, K_{k,\chi}(t_j + \chi y)) \text{ with } j = 1, \ldots, K_{k,\chi}^2,$$

are disjoint. Since all the $K_{k,\chi}^2$ intervals I_j are disjoint and $|S| < K_{k,\chi}^2$, there exists $j_0 \geq 1$ such that no element of S is contained in I_{j_0}. The number t_{j_0} is the desired number t. □

We are now ready to state and prove the key lemma of this section.

LEMMA 3.5. *Let G be a (finite or infinite) λ-graph G with k types of edges and with chromatic number at most χ. Fix real numbers $x_1 \geq \cdots \geq x_k > 0$. For each $\ell = 1, \ldots, k$, there exists a finitary orientation $\vec{G}$ of G of weight $\lambda_G(x_1, \ldots, x_k)$ such that every directed path in $\vec{G}$ contains at most $D_{\ell,\chi}$ $x_{\leq \ell}$-edges.*

Proof. If $\chi = 1$, there is nothing to prove since G contains no edges and the statement of the lemma holds vacuously. Therefore, we assume $\chi \geq 2$ in the remaining. For the rest of the proof, let us fix a proper coloring μ (in the usual sense) of the vertices of G with colors $1, \ldots, \chi$.

The proof of the lemma proceeds by induction on the number ℓ. First, we have to deal with the case $\ell = 1$. Consider any finitary orientation $\vec{G}$ of G of weight $\lambda_G(x_1, \ldots, x_k)$. Such an orientation exists by Theorem 2.1. By Lemma 2.4, we get that $\lambda_G(x_1, \ldots, x_k) \leq \chi x_1$. Hence, every directed path in $\vec{G}$ contains at most $D_{1,\chi} = \chi$ x_1-edges.

We now deal with the case $\ell > 1$. By the induction, there exists a finitary orientation $\vec{G}$ of G of weight $\lambda_G(x_1, \ldots, x_k)$ such that any directed path contains at most $D_{\ell-1,\chi}$ $x_{\leq \ell-1}$-edges. Let $c(v)$ be the labeling corresponding to $\vec{G}$, and let $d(v)$ be the maximum sum of weights of $x_{\leq \ell-1}$-edges on a directed path in $\vec{G}$ ending at a vertex v. Clearly, $d(v) \leq c(v)$ for every vertex v of G. Finally, let $\delta(v)$ be the number of the elements of $\Gamma_{D_{\ell-1,\chi}}$ smaller than or equal to $d(v)$. Since $|\Gamma_{D_{\ell-1,\chi}}| < K_{\ell-1,\chi}$ by Proposition 3.3, $1 \leq \delta(v) \leq K_{\ell-1,\chi} - 1$ for every vertex v of G.

Since $\Gamma'_{D_{\ell-1,\chi}}(x_1,\ldots,x_{\ell-1})$ contains at most $K^2_{\ell-1,\chi}-1$ real numbers by Proposition 3.3, we infer from Lemma 3.4 (applied for $k=\ell-1$, $S=\Gamma'_{D_{\ell-1,\chi}}(x_1,\ldots,x_{\ell-1})$, and $y=x_\ell$) that there exists a real number t,

$$K_{\ell-1,\chi}\chi x_\ell \leq t \leq (2K_{\ell-1,\chi})^{K^2_{\ell-1,\chi}}\chi x_\ell\ , \tag{3.1}$$

such that the set $\Gamma'_{D_{\ell-1,\chi}}(x_1,\ldots,x_{\ell-1})$ contains no element strictly between t and $K_{\ell-1,\chi}(t+\chi x_\ell)$; i.e., t has the following property:

If $\gamma\in\Gamma'_{D_{\ell-1,\chi}}(x_1,\ldots,x_{\ell-1})$ and $\gamma>t$, then $\gamma\geq K_{\ell-1,\chi}(t+\chi x_\ell)$.

We refer to this property throughout the proof as Property $(\star)$.

As the next step of the proof, we define a new labeling c' and show that it is a proper labeling with respect to $x_1,\ldots,x_k$:

1. If $c(v)-d(v)\leq(K_{\ell-1,\chi}-\delta(v))t$, then $c'(v)=c(v)$.
2. Otherwise, $c'(v)=d(v)+(K_{\ell-1,\chi}-1)t+\delta(v)\chi x_\ell+\mu(v)x_\ell$.

Note that the definition of being "close" depends on the value of $d(v)$. For values of $d(v)$ near $\lambda_G(x_1,\ldots,x_k)$, the difference is required to be a small multiple of t (it can be as small as t or $2t$ for maximal possible values of $d(v)$). The threshold difference is increased by t by each step "down" in the order of the numbers contained in $\Gamma_{D_{\ell-1,\chi}}(x_1,\ldots,x_{\ell-1})$ (note that each $d(v)$ is contained in this set). The labeling c' is a combination of the original labeling c for vertices, where $d(v)$ and $c(v)$ are "close," and a completely different labeling for the remaining vertices—the additional space among labels considered to be "close" for small values of $d(v)$ is needed so that the combination of the two labelings is proper.

We now prove that the labeling c' is proper with respect to $x_1,\ldots,x_k$. Let us consider an x_i-edge uv of G. In order to verify that c' is a proper labeling on the edge uv, we distinguish five major cases:

- *Both the labels $c'(u)$ and $c'(v)$ are defined by the first rule.*
  Since $c'(u)=c(u)$ and $c'(v)=c(v)$, we have $|c'(u)-c'(v)|=|c(u)-c(v)|\geq x_i$.
- *The label $c'(u)$ is defined by the first rule, the label $c'(v)$ is defined by the second rule, and $i<\ell$.*
  We distinguish two subcases according to the orientation of the edge uv in $\vec{G}$. If the edge is directed from u to v, we have $d(u)+x_i\leq d(v)$. Because the label of u is defined by the first rule, the label $c'(u)=c(u)$ is at most $d(u)+(K_{\ell-1,\chi}-1)t$. On the other hand, the label $c'(v)$ is larger than $d(v)+(K_{\ell-1,\chi}-1)t$. We infer that $c'(v)-c'(u)\geq d(v)-d(u)\geq x_i$.
  The other subcase is that the edge uv is directed from v to u. In particular, $d(v)+x_i\leq d(u)$, $\delta(v)<\delta(u)$ and $c(v)\leq c(u)$. First, we show that $d(u)-d(v)-x_i>t$. Assume to the contrary that

  $$d(u)-d(v)-x_i\leq t. \tag{3.2}$$

  Since c is a proper labeling of G, it holds that $c(v)\leq c(u)-x_i$. And since the label of u was defined by the first rule, we have

  $$c(u)\leq d(u)+(K_{\ell-1,\chi}-\delta(u))t. \tag{3.3}$$

  Therefore, the following holds:

$$\begin{aligned} c(v) &\leq c(u)-x_i \\ &\leq d(u)+(K_{\ell-1,\chi}-\delta(u))t-x_i \quad \text{(by (3.3))} \\ &\leq d(v)+(K_{\ell-1,\chi}-\delta(u))t+t \quad \text{(by (3.2))} \\ &\leq d(v)+(K_{\ell-1,\chi}-\delta(v))t. \end{aligned}$$

However, this yields that the label of v should have been defined by the first rule. We conclude that $d(u) - d(v) - x_i > t$.
Since $d(u) - d(v) - x_i \in \Gamma'_{D_{\ell-1,\chi}}(x_1, \ldots, x_{\ell-1})$, it holds that $d(u) - d(v) - x_i \geq K_{\ell-1,\chi}(t + \chi x_\ell)$ by Property $(\star)$.
We now bound the label $c'(v)$ assigned to the vertex v from above (recall that $\mu(v)x_\ell \leq \chi x_\ell \leq t$ and $\delta(v) \leq K_{\ell-1,\chi} - 1$):

$$\begin{aligned} c'(v) &= d(v) + (K_{\ell-1,\chi} - 1)t + \delta(v)\chi x_\ell + \mu(v)x_\ell \\ &\leq d(v) + K_{\ell-1,\chi}t + K_{\ell-1,\chi}\chi x_\ell \\ &\leq d(u) - x_i \;\leq\; c(u) - x_i \;=\; c'(u) - x_i. \end{aligned}$$

Hence, the labels of the vertices u and v differ by at least x_i as required.

- *The label $c'(u)$ is defined by the first rule, the label $c'(v)$ is defined by the second rule, and $i \geq \ell$.*
  If $d(u) \leq d(v)$, then $c'(u) \leq d(u) + (K_{\ell-1,\chi} - 1)t$ and $c'(v) \geq d(v) + (K_{\ell-1,\chi} - 1)t + \mu(v)x_\ell \geq d(u) + (K_{\ell-1,\chi} - 1)t + x_\ell$. Therefore, $c'(v) - c'(u) \geq x_i$ as desired.
  In the rest, we focus on the case $d(u) > d(v)$. In particular, the edge uv is directed from v to u. By the definition of $\delta(u)$ and $\delta(v)$, we have $\delta(u) > \delta(v)$. Since the edge uv is directed from v to u, it also holds that $c(u) > c(v)$.
  First, we exclude the case $d(u) - d(v) \leq t$. Since the label of the vertex u was defined by the first rule, we have $c(u) \leq d(u) + (K_{\ell-1,\chi} - \delta(u))t$. We infer the following upper bound on $c(v)$:

$$\begin{aligned} c(v) \leq c(u) &\leq d(u) + (K_{\ell-1,\chi} - \delta(u))t \\ &\leq d(v) + t + (K_{\ell-1,\chi} - \delta(u))t \\ &\leq d(v) + (K_{\ell-1,\chi} - \delta(v))t. \end{aligned}$$

  However, the label to v should then have been defined by the first rule, not by the second one. We conclude that $d(u) - d(v) > t$.
  Since the difference $d(u) - d(v)$ is contained in $\Gamma'_{D_{\ell-1,\chi}}(x_1, \ldots, x_{\ell-1})$, it holds that $d(u) - d(v) \geq K_{\ell-1,\chi}(t + \chi x_\ell)$ by Property $(\star)$. The following upper bound on $c'(v)$ readily follows (recall that $\delta(v) \leq K_{\ell-1,\chi} - 1$ and $\mu(v) \leq \chi$):

$$\begin{aligned} c'(v) &= d(v) + (K_{\ell-1,\chi} - 1)t + \delta(v)\chi x_\ell + \mu(v)x_\ell \\ &\leq d(v) + K_{\ell-1,\chi}t - t + K_{\ell-1,\chi}\chi x_\ell \\ &\leq d(u) - t \;\leq\; c(u) - x_\ell \;=\; c'(u) - x_\ell \;\leq\; c'(u) - x_i. \end{aligned}$$

  Hence, the labels of the end-vertices of the x_i-edge uv differ by at least x_i as required.
- *Both the labels $c'(u)$ and $c'(v)$ are defined by the second rule and $i < \ell$.*
  By symmetry, we may assume that the edge uv is directed from u to v in $\vec{G}$. In this case, $d(u) + x_i \leq d(v)$ and $\delta(u) < \delta(v)$. The difference of the labels $c'(u)$ and $c'(v)$ can be easily estimated:

$$\begin{aligned} c'(v) - c'(u) &= d(v) - d(u) + (\delta(v) - \delta(u))\chi x_\ell + (\mu(v) - \mu(u))x_\ell \\ &\geq x_i + \chi x_\ell - |\mu(v) - \mu(u)|x_\ell \;\geq\; x_i. \end{aligned}$$

  We conclude that the edge uv is properly labeled.

- *Both the labels $c'(u)$ and $c'(v)$ are defined by the second rule, and $i \geq \ell$.*
  By symmetry, it can be assumed that the edge uv is directed from u to v in $\vec{G}$. In this case, $d(u) \leq d(v)$. If $d(u) = d(v)$, then

  $$|c'(v) - c'(u)| = |\mu(v) - \mu(u)|x_\ell \geq x_\ell \geq x_i.$$

  In the rest, we deal with the case $d(u) < d(v)$. In particular, $\delta(u) < \delta(v)$. The difference between $c'(u)$ and $c'(v)$ as follows (recall that $1 \leq \mu(u), \mu(v) \leq \chi$) can then be bounded as follows:

  $$\begin{aligned} c'(v) - c'(u) &= d(v) - d(u) + (\delta(v) - \delta(u))\chi x_\ell + (\mu(v) - \mu(u))x_\ell \\ &\geq \chi x_\ell - |\mu(v) - \mu(u)|x_\ell \ \geq \ x_\ell \ \geq \ x_i. \end{aligned}$$

  Hence, the difference of the labels $c'(u)$ and $c'(v)$ is at least x_i as desired.

As the next step, we show that the span of c' is equal to $\lambda_G(x_1, \ldots, x_k)$. In order to do so, it is enough to show that $c'(v) \leq \lambda_G(x_1, \ldots, x_k)$ for every vertex v of G. Let $c_{\max}$ and $d_{\max}$ be the maximums of the values $c(v)$ and $d(v)$ taken over all the vertices v of G. Clearly, $c_{\max} = \lambda_G(x_1, \ldots, x_k)$. By Lemma 2.4 and (3.1), the following holds:

$$c_{\max} \leq d_{\max} + \ell^{D_{\ell-1,\chi}}\chi \cdot x_\ell = d_{\max} + K_{\ell-1,\chi}\chi \cdot x_\ell \leq d_{\max} + t.$$

Fix a vertex v of G. In order to show that $c'(v) \leq c_{\max}$, we distinguish three cases according to the difference between $d(v)$ and $d_{\max}$:

- $d(v) = d_{\max}$
  Since $c(v) \leq c_{\max} \leq d_{\max} + K_{\ell-1,\chi}\chi x_\ell \leq d_{\max} + t = d(v) + t \leq d(v) + (K_{\ell-1,\chi} - \delta(v))t$, the label of the vertex v was defined by the first rule. Consequently, $c'(v) = c(v) \leq c_{\max}$.
- $0 < d_{\max} - d(v) \leq t$
  First, observe that $\delta(v) \leq K_{\ell-1,\chi} - 2$. Again, we bound the original label $c(v)$ from above:

  $$c(v) \leq c_{\max} \leq d_{\max} + t \leq d(v) + 2t \leq d(v) + (K_{\ell-1,\chi} - \delta(v))t.$$

  Therefore, the label of the vertex v was defined by the first rule, and $c'(v) = c(v) \leq c_{\max}$.
- $d_{\max} - d(v) > t$
  Since $d_{\max} - d(v) \in \Gamma'_{D_{\ell-1,\chi}}(x_1, \ldots, x_{\ell-1})$, we infer from Property $(\star)$ that $d_{\max} - d(v) \geq K_{\ell-1,\chi}(t + \chi x_\ell)$. If the first rule applies to the vertex v, then $c'(v) = c(v) \leq c_{\max}$. If the second rule applies, then the following estimate on $c'(v)$ holds (recall that $\delta(v) \leq |\Gamma_{D_{\ell-1,\chi}}| \leq K_{\ell-1,\chi} - 1$):

  $$\begin{aligned} c'(v) &= d(v) + (K_{\ell-1,\chi} - 1)t + \delta(v)\chi x_\ell + \mu(v)x_\ell \\ &\leq d(v) + K_{\ell-1,\chi}t + K_{\ell-1,\chi}\chi x_\ell \\ &\leq d_{\max} \ \leq \ c_{\max}. \end{aligned}$$

  Hence, the label $c'(v)$ does not exceed $c_{\max}$.

Let $\vec{G}'$ be the orientation of G corresponding to the labeling c'. Since all $x_1, \ldots, x_k$ are positive, the orientation of G is finitary and its weight is at most the span of c'. Since the span of c' is $\lambda_G(x_1, \ldots, x_k)$, the weight of $\vec{G}'$ is exactly $\lambda_G(x_1, \ldots, x_k)$. In order to finish the proof of the lemma, we establish that each directed path in $\vec{G}'$ contains at most $D_{\ell,\chi}$ $x_{\leq \ell}$-edges.

All the labels $c'(v)$ defined by the first rule are contained in the following union of intervals by (3.1):

$$\bigcup_{\gamma\in\Gamma_{D_{\ell-1,\chi}}(x_1,\dots,x_{\ell-1})} [\gamma,\gamma+K_{\ell-1,\chi}t\rangle$$

$$\subseteq \bigcup_{\gamma\in\Gamma_{D_{\ell-1,\chi}}(x_1,\dots,x_{\ell-1})} \left[\gamma,\gamma+(2K_{\ell-1,\chi})^{K^2_{\ell-1,\chi}+1}\chi x_\ell\right).$$

The labels $c'(v)$ assigned by the second rule are from the following set:

$$\bigcup_{\gamma_i\in\Gamma_{D_{\ell-1,\chi}}(x_1,\dots,x_{\ell-1})} \{\gamma_i+(K_{\ell-1,\chi}-1)t+i\chi x_\ell+jx_\ell, j=1,\dots,\chi\},$$

where $\gamma_1,\gamma_2,\dots$ are all the elements of $\Gamma_{D_{\ell-1,\chi}}$ listed in increasing order. Consider a directed path P in $\vec{G}'$ and let C be the set of labels $c'(v)$ of the end-vertices of $x_{\le\ell}$-edges on P. Since any two labels in C differ by at least x_ℓ and $|\Gamma_{D_{\ell-1,\chi}}(x_1,\dots,x_{\ell-1})| < K_{\ell-1,\chi}$ by Proposition 3.3, we have the following upper bound on the number of labels $c'(v)\in C$, $v\in P$ that were defined by the first rule:

$$|\Gamma_{D_{\ell-1,\chi}}(x_1,\dots,x_{\ell-1})|\frac{(2K_{\ell-1,\chi})^{K^2_{\ell-1,\chi}+1}\chi x_\ell}{x_\ell} \le (2K_{\ell-1,\chi})^{K^2_{\ell-1,\chi}+2}\chi\,.$$

Similarly, the number of such labels defined by the second rule does not exceed

$$|\Gamma_{D_{\ell-1,\chi}}(x_1,\dots,x_{\ell-1})|\chi \le K_{\ell-1,\chi}\chi\,.$$

Combining both bounds, we infer from Proposition 3.2 that the size of C does not exceed $D_{\ell,\chi}=(2K_{\ell-1,\chi})^{K^2_{\ell-1,\chi}+3}\chi$. Therefore, every directed path in $\vec{G}'$ contains at most $D_{\ell,\chi}$ $x_{\le\ell}$-edges as desired. □

We modify Lemma 3.5 to a version used in section 4.

LEMMA 3.6. *Let G be a (finite or infinite) λ-graph with k types of edges and chromatic number at most χ. For any k-tuple of nonnegative reals $x_1,\dots,x_k$, there exists a finitary orientation of G with weight $\lambda_G(x_1,\dots,x_k)$ for which the maximum length of a directed path is at most $D_{k,\chi}$.*

Proof. By symmetry, we can assume that $x_1\ge\cdots\ge x_k$ (otherwise, permute the types of the edges of G). If $x_k>0$, the statement of the lemma follows directly from Lemma 3.5. In the rest, we deal with the case when $x_{k'}>0$ and $x_{k'+1}=\cdots=x_k=0$.

Fix a coloring μ of G with χ colors $1,\dots,\chi$. Let G' be the subgraph of G formed by $x_{\le k'}$-edges. By Lemma 3.5, there exists a finitary orientation $\vec{G}'$ of G' with weight $\lambda_{G'}(x_1,\dots,x_{k'})=\lambda_G(x_1,\dots,x_k)$ and with maximum path length at most $D_{k',\chi}$. Let $c(v)$ be the labeling of G' corresponding to $\vec{G}'$. Observe that $c(v)\in\Gamma_{D_{k',\chi}}(x_1,\dots,x_{k'})$ for every vertex v of G.

We extend $\vec{G}'$ to a finitary orientation $\vec{G}$ of G. An $x_{>k'}$-edge uv is directed from u to v if $c(u)<c(v)$, and from v to u if $c(u)>c(v)$. If $c(u)=c(v)$, then the edge uv is directed from u to v if $\mu(u)<\mu(v)$, and from v to u otherwise. Clearly, the weight of $\vec{G}$ is the same as the weight of $\vec{G}'$.

Let P be a directed path in $\vec{G}$. The labels $c(v)$ of vertices v do not decrease along the path P. Moreover, each subpath of P formed by vertices v with the same label

assigned by c has length at most $\chi-1$ as the colors $\mu(v)$ of the vertices comprising the subpath strictly increase. Since all the labels $c(v)$ are from the set $\Gamma_{D_{k',\chi}}(x_1,\ldots,x_{k'})$, P contains at most $K_{k',\chi}$ such monochromatic subpaths. Hence, the length of a directed P in $\vec{G}$ does not exceed $K_{k',\chi}\cdot\chi$, and the maximum length of a directed path in $\vec{G}$ is at most $K_{k',\chi}\cdot\chi\le D_{k,\chi}$. □

4. Main result. In this section, we prove our main result on the structure of the λ-functions of λ-graphs. Before doing so, we introduce several definitions. As we see later, the (finite) set $\mathcal{F}^{\rm minmax}_{k,\chi}$ of piecewise-linear functions, which is defined in what follows, is a superset of all λ-functions of λ-graphs G with k types of edges and chromatic number at most χ.

Let $\mathcal{F}_{k,\chi}$ be the set of all linear functions of k variables with integer coefficients between 0 and $D_{k,\chi}$; i.e.,

$$\mathcal{F}_{k,\chi}=\left\{\sum_{i=1}^{k}\alpha_i x_i, 0\le\alpha_i\le D_{k,\chi}\right\}.$$

Next, $\mathcal{F}^{\rm max}_{k,\chi}$ is the set of all functions φ that are equal to the maximum of some of the functions from $\mathcal{F}_{k,\chi}$; i.e.,

$$\mathcal{F}^{\rm max}_{k,\chi}=\{\varphi(x_1,\ldots,x_p)=\max_{f\in F}f(x_1,\ldots,x_p)\text{ for }F\subseteq\mathcal{F}_{k,\chi},F\neq\emptyset\}\,.$$

Finally, $\mathcal{F}^{\rm minmax}_{k,\chi}$ is the set of all functions that are equal to the minimum of some of the functions from $\mathcal{F}^{\rm max}_{k,\chi}$; i.e.,

$$\mathcal{F}^{\rm minmax}_{k,\chi}=\{\varphi(x_1,\ldots,x_p)=\min_{f\in F}f(x_1,\ldots,x_p)\text{ for }F\subseteq\mathcal{F}^{\rm max}_{k,\chi},F\neq\emptyset\}\,.$$

Observe $\mathcal{F}_{k,\chi}\subseteq\mathcal{F}^{\rm max}_{k,\chi}\subseteq\mathcal{F}^{\rm minmax}_{k,\chi}$. Clearly, all three sets $\mathcal{F}_{k,\chi}$, $\mathcal{F}^{\rm max}_{k,\chi}$, and $\mathcal{F}^{\rm minmax}_{k,\chi}$ are finite. Therefore, $\mathbb{R}^k_+$ can be partitioned into finitely many polyhedral cones (with the tips at the origin) such that every function contained in $\mathcal{F}^{\rm minmax}_{k,\chi}$ is linear on each of the cones. Let $C_{k,\chi}$ be the number of such cones.

We now state and prove the main result of the paper (recall that both the numbers $C_{k,\chi}$ and $D_{k,\chi}$ just depend on k and χ).

THEOREM 4.1. *For every λ-graph G with k types of edges and chromatic number at most χ, $\lambda_G(x_1,\ldots,x_k)$ is a piecewise linear function of $x_1,\ldots,x_k$ with at most $C_{k,\chi}$ linear parts formed by linear functions with integer coefficients between 0 and $D_{k,\chi}$. Moreover, $\mathbb{R}^k_+$ can be partitioned into at most $C_{k,\chi}$ polyhedral cones such that for each of the cones the following holds: There exist integers $\alpha_i(v)$ between 0 and $D_{k,\chi}$ such that the labeling c, $c(v)=\sum_{i=1}^{k}\alpha_i(v)x_i$, is a proper labeling of G with respect to $x_1,\ldots,x_k$ and the span of c is $\lambda_G(x_1,\ldots,x_k)$. For fixed k and χ, this partition of $\mathbb{R}^k_+$ is independent of a λ-graph G.*

Proof. Let $\mathcal{D}$ be the set of all finitary orientations of G with maximum length of a directed path at most $D_{k,\chi}$. For an orientation $\vec{G}\in\mathcal{D}$, let $F(\vec{G})$ be the set of all the functions $\sum_{i=1}^{k}\alpha_i x_i$ such that $\vec{G}$ contains a directed path with precisely α_i x_i-edges (for all i). Since the maximum length of a directed path in $\vec{G}$ does not exceed $D_{k,\chi}$, the set $F(\vec{G})$ is a subset of $\mathcal{F}_{k,\chi}$, i.e., $F(\vec{G})\subseteq\mathcal{F}_{k,\chi}$. By the definition, the weight of the orientation $\vec{G}$ with respect to $x_1,\ldots,x_k$ is the following:

$$w_{\vec{G}}(x_1,\ldots,x_k)=\max_{f\in F(\vec{G})}f(x_1,\ldots,x_k).$$

Let W be the set of all the functions $w_{\vec{G}}(x_1,\dots,x_k)$, where $\vec{G}$ ranges through all the orientations contained in $\mathcal{D}$. Clearly, $W \subseteq \mathcal{F}^{\max}_{k,\chi}$. For $w \in W$, let $\vec{G}_w$ be one of the orientations in $\mathcal{D}$ with $w_{\vec{G}} = w$. By Theorem 2.1 and Lemma 3.6, the following equality holds:

$$\lambda_G(x_1,\dots,x_k) = \min_{\vec{G}\in\mathcal{D}} w_{\vec{G}}(x_1,\dots,x_k) = \min_{w\in W} w(x_1,\dots,x_k).$$

Similarly as before, we have $\lambda_G(x_1,\dots,x_k) \in \mathcal{F}^{\text{minmax}}_{k,\chi}$.

Consider the partition of $\mathbb{R}^k_+$ into $C_{k,\chi}$ polyhedral cones such that every function of $\mathcal{F}^{\text{minmax}}_{k,\chi}$ is linear on each of the cones. In particular, $\lambda_G(x_1,\dots,x_k) \in \mathcal{F}^{\text{minmax}}_{k,\chi}$ is linear on each of the cones.

Fix one such cone and let $w \in W$ be a function such that $\lambda_G(x_1,\dots,x_k) = w(x_1,\dots,x_k)$ on the fixed cone. Let c be the labeling corresponding to the orientation $\vec{G}_w$. Since no directed path of $\vec{G}_w \in \mathcal{D}$ has length more than $D_{k,\chi}$, the corresponding label $c(v)$, when viewed as a function of $x_1,\dots,x_k$, belongs to the set $\mathcal{F}^{\max}_{k,\chi}$. In particular, each $c(v)$ is a linear function on the fixed cone. Therefore, $c(v)$ can be expressed as a combination of $x_1,\dots,x_k$ with integer coefficients between 0 and $D_{k,\chi}$. The statement of the theorem now follows. □

Since only the functions from $\mathcal{F}^{\text{minmax}}_{k,\chi}$ could be the λ-function of a λ-graph with k types of edges and with chromatic number at most χ, we have the following corollary.

COROLLARY 4.2. *There exist only finitely many piecewise linear functions that could be the λ-function of a λ-graph with k types of edges and with chromatic number at most χ.*

Another immediate corollary is the following somewhat surprising statement.

COROLLARY 4.3. *Let $x_1,\dots,x_k$ be a fixed k-tuple of positive reals and let γ be a nonnegative real. There exist only finitely many different k-parameter λ-functions λ_G such that $\lambda_G(x_1,\dots,x_k) \le \gamma$.*

Proof. If G is a λ-graph with k types of edges such that $\lambda_G(x_1,\dots,x_k) \le \gamma$, then the chromatic number of G does not exceed $\frac{\lambda_G(x_1,\dots,x_k)}{\min\{x_1,\dots,x_k\}} + 1 \le \frac{\gamma}{\min\{x_1,\dots,x_k\}} + 1$ by the scaling property. By Corollary 4.2, λ-graphs with k types of edges with bounded chromatic number have only finitely many different λ-functions. □

Note that Corollary 4.3 includes the result of [2] that the number of λ-functions with prescribed boundary values is finite.

5. Labelings with distance conditions. In this section, we infer from Theorem 4.1 the piecewise linearity conjecture, coefficient bound conjecture, and delta bound conjecture stated in [9]. First, let us state the following simple proposition that can be found in [9] (note that its proof employs the axiom of choice).

PROPOSITION 5.1. *If G is a graph of maximum degree Δ and k is a positive integer, then the chromatic number of the kth power of G does not exceed $\Delta^k + 1$.*

We can now state the theorem from which the three conjectures mentioned above readily follow.

THEOREM 5.2. *For all integers $\Delta \ge 1$ and $k \ge 2$, there exist constants $C'_{k,\Delta}$ and $D'_{k,\Delta}$ with the following property: For any graph G with maximum degree Δ, there exists a partition of $\mathbb{R}^k_+$ into at most $C'_{k,\Delta}$ polyhedral cones with the tips at the origin such that the function $\lambda_G(p_1,\dots,p_k)$ is linear in each of the cones. Moreover, for each of the cones the following is true: There exist integers $\alpha_i(v)$ between 0 and $D_{k,\chi}$ such that the labeling c, $c(v) = \sum_{i=1}^k \alpha_i(v)p_i$, is a proper $L(p_1,\dots,p_k)$-labeling of G with respect to $x_1,\dots,x_k$ and the span of c is $\lambda_G(p_1,\dots,p_k)$.*

Proof. Set $C'_{k,\Delta} = C_{k,\Delta^k+1}$ and $D'_{k,\Delta} = D_{k,\Delta^k+1}$. Let G be a graph with maximum degree Δ, and form the λ-graph $G^{(k)}$ as described in section 1. By Proposition 5.1, the chromatic number of $G^{(k)}$ does not exceed $\Delta^k + 1$. Theorem 5.2 now follows from Theorem 4.1. □

An immediate corollary of Theorem 5.2 (alternatively, of Corollary 4.2) is the following corollary.

COROLLARY 5.3. *Let $\Lambda_{k,\Delta}$ for $k \geq 2$ be the set that consists of all the functions $\lambda_G(p_1, \ldots, p_k)$ of graphs G with maximum degree at most Δ. The set $\Lambda_{k,\Delta}$ is finite.*

Acknowledgments. The author would like to thank Jerrold R. Griggs for interesting and helpful discussions on real number graph labelings with distance conditions. The comments of the three anonymous referees that helped to improve the presentation of the results contained in this paper are greatly appreciated.

REFERENCES

[1] G. AGNARSSON, R. GREENLAW, AND M. M. HALLDÓRSSON, *Powers of chordal graphs and their coloring*, Congr. Numer., to appear.
[2] R. BABILON, V. JELÍNEK, D. KRÁL', AND P. VALTR, *Labelings of graphs with fixed and variable edge-weights*, submitted.
[3] P. BELLA, D. KRÁL', B. MOHAR, AND K. QUITTNEROVÁ, *Labeling planar graphs with a condition at distance two*, European J. Combin., to appear.
[4] H. L. BODLAENDER, T. KLOKS, R. B. TAN, AND J. VAN LEEUWEN, *λ-coloring of graphs*, in STACS 2000 (Lille), Lecture Notes in Comput. Sci. 1770, G. Goos, J. Hartmanis, and J. van Leeuwen, eds., Springer, Berlin, 2000, pp. 395–406.
[5] G. J. CHANG, W.-T. KE, D. D.-F. LIU, AND R. K. YEH, *On $L(d,1)$-labellings of graphs*, Discrete Math., 3 (2000), pp. 57–66.
[6] G. J. CHANG AND D. KUO, *The $L(2,1)$-labeling problem on graphs*, SIAM J. Discrete Math., 9 (1996), pp. 309–316.
[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, Berlin, 2000, pp. 363–372.
[9] J. R. GRIGGS AND X. T. JIN, *Real number graph labellings with distance conditions*, SIAM J. Discrete Math., 20 (2006), pp. 302–327.
[10] J. R. GRIGGS AND X. T. JIN, *Real number graph labellings of paths and cycles*, submitted.
[11] J. R. GRIGGS AND X. T. JIN, *Real number graph labellings of infinite graphs*, submitted.
[12] J. R. GRIGGS AND R. K. YEH, *Labelling graphs with a condition at distance* 2, SIAM J. Discrete Math., 5 (1992), pp. 586–595.
[13] D. GONÇALVES, *On the $L(p,1)$-labelling of graphs*, Discrete Math. Theor. Comput. Sci., AE (2005), pp. 81–86.
[14] W. K. HALE, *Frequency assignment: Theory and applications*, Proc. IEEE, 68 (1980), pp. 1497–1514.
[15] J.-H. KANG, *$L(2,1)$-labeling of* 3*-regular Hamiltonian graphs*, submitted.
[16] J.-H. KANG, *$L(2,1)$-Labelling of* 3*-Regular Hamiltonian Graphs*, Ph.D. thesis, University of Illinois, Urbana-Champaign, IL, 2004.
[17] D. KRÁL', *An exact algorithm for the channel assignment problem*, Discrete Appl. Math. 145 (2005), pp. 326–331.
[18] D. KRÁL', *Coloring powers of chordal graphs*, SIAM J. Discrete Math., 18 (2004), pp. 451–461.
[19] D. KRÁL' AND R. ŠKREKOVSKI, *A theorem about the channel assignment problem*, SIAM J. Discrete Math., 16 (2003), pp. 426–437.
[20] C. MCDIARMID, *Discrete mathematics and radio channel assignment*, in Recent Advances in Algorithms and Combinatorics, C. Linhares-Sales and B. Reed, eds., Springer, New York, 2003, pp. 27–63.
[21] C. MCDIARMID, *On the span in channel assignment problems: Bounds, computing and counting*, Discrete Math., 266 (2003), pp. 387–397.
[22] D. SAKAI, *Labeling chordal graphs: Distance two condition*, SIAM J. Discrete Math., 7 (1994), pp. 133–140.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 705–726

COMPACT ROUTING WITH NAME INDEPENDENCE*

MARTA ARIAS†, LENORE J. COWEN†, KOFI A. LAING†, RAJMOHAN RAJARAMAN‡, AND ORJETA TAKA†

Abstract. This paper is concerned with compact routing schemes for arbitrary undirected networks in the name-independent model first introduced by Awerbuch, Bar-Noy, Linial, and Peleg. A compact routing scheme that uses local routing tables of size $\tilde{O}(n^{1/2})$, $O(\log^2 n)$-sized packet headers, and stretch bounded by 5 is obtained, where n is the number of nodes in the network. (We use the notation $\tilde{O}(f(n))$ to represent $O(f(n)\log^c n)$, where c is an arbitrary nonnegative real number, independent of n.) Alternative schemes reduce the packet header size to $O(\log n)$ at the cost of either increasing the stretch to 7 or increasing the table size to $\tilde{O}(n^{2/3})$. For smaller table-size requirements, the ideas in these schemes are generalized to a scheme that uses $O(\log^2 n)$-sized headers and $\tilde{O}(k^2 n^{2/k})$-sized tables, and achieves a stretch of $\min\{1+(k-1)(2^{k/2}-2), 16k^2-8k\}$, improving the best previously known name-independent scheme due to Awerbuch and Peleg.

Key words. compact routing, name independence, stretch, routing table, networks

AMS subject classifications. 05C12, 68W15, 68M10

DOI. 10.1137/04062053

1. Introduction.

Consider an undirected (weighted) n-node network in which nodes are labeled with an arbitrary permutation P of the labels $\{0,\ldots,n-1\}$. A packet labeled i can arrive at any node in the network and must then be delivered to the node that P assigned label i. This is called *name-independent* routing, since the labels are unrelated to network topology. Consider the scheme in which each node stores an entry for each destination i in its local routing table, containing the name of the outgoing link for the first edge along the shortest path from itself to i. This uses $O(n\log n)$ space at every node and routes along shortest paths.

In this paper, we study the design of name-independent compact routing schemes. More formally, let us define the *stretch* of a path $p(u,v)$ from node u to node v as $\frac{|p(u,v)|}{d(u,v)}$, where $d(u,v)$ is the length of the shortest u-v path and $|p(u,v)|$ is the length of $p(u,v)$. We consider the following question: Can a routing scheme be designed that uses *sublinear* space per node for the routing tables, and routes packets along paths with bounded stretch? All results in this paper, except one in section 2, which considers the special case of trees, are *universal* in the sense that they apply to any undirected n-node network with positive edge weights.

The study of compact routing schemes is the study of routing-time and space tradeoffs for approximate shortest paths—one of the most fundamental problems in distributed algorithms. The design of compact routing algorithms was originally motivated by the need for scalable routing in communication networks and has recently

*Received by the editors December 9, 2004; accepted for publication (in revised form) January 30, 2006; published electronically September 29, 2006. A preliminary version of this paper appeared in Proceedings of the 15th Annual ACM Symposium on Parallelism in Algorithms and Architectures [2].
http://www.siam.org/journals/sidma/20-3/62053.html

†Department of Computer Science, Tufts University, Medford, MA 02155 (marias@cs.tufts.edu, cowen@cs.tufts.edu, laing@cs.tufts.edu, otaka@cs.tufts.edu). The work of the first author was supported in part by NSF grant IIS-0099446. The work of the second and fifth authors was supported in part by NSF grant CCR-0208629. The work of the third author was supported in part by NSF grant EHR-0227879.

‡College of Computer & Information Science, Northeastern University, Boston, MA 02115 (rraj@ccs.neu.edu). The work of this author was supported in part by NSF CAREER award CCR-9983901.

been evaluated for routing in Internet-like graphs [15]. Compact routing has also recently gathered interest in the contexts of efficient searching of distributed hash tables, distributed dictionaries, and peer-to-peer systems [1].

Though the name-independent version of the compact routing problem was first introduced in 1989 by Awerbuch et al. [5], progress has been slow. Much recent work [4, 11, 9, 21] has occurred on the easier but related compact routing problem, where the compact routing scheme designer may assign his/her own polylogarithmic-sized node labels (generally $O(\log n)$- or $O(\log^2 n)$-bit), dependent on network topology. That is, when a packet destined for i arrives, "i" has been renamed, not by some arbitrary permutation P but by the routing scheme designer, in order to give maximum information about the underlying topology of the network. (An alternative but equivalent formulation is that a packet destined for i arrives also with a short (up to) $O(\log^2 n)$-bit address chosen by the compact routing scheme designer, dependent on network topology.) For example, if the underlying network were a planar grid in the topology-dependent (also called the *name-dependent*) model, then the algorithm designer could require that a packet destined for a node comes addressed with its (x, y) coordinates, whereas in the name-independent model under consideration here, the packet would come with a destination name, independent of its (x, y) coordinates, and would have to learn information about its (x, y) coordinates from its name as it wandered the network.

In [5], Awerbuch et al. argued that even though topology-dependent node labels might be fine for static networks, they make less sense in a dynamic network, where the network topology changes over time. There are serious consistency and continuity issues if the identifying label of a node changes as network connectivity evolves. In such a model, a node's identifying label needs to be decoupled from network topology. In fact, network nodes should be allowed to choose arbitrary names (subject to the condition that node names are unique), and packets destined for a particular node name enter the network with this name only, with no additional topological address information.[1] Routing information relating this name to the location of the destination node is distributed in the routing tables of the network, which can be updated if network topology changes.

The scheme of Awerbuch et al. in [5, 6] showed, perhaps surprisingly, that the problem of compact routing with name-independent node names was not impossible. They presented the first universal compact routing scheme to achieve all of the following four properties: (1) sublinear-space routing tables at every node; (2) constant size stretch; (3) polylogarithmic-sized routing headers; and (4) topology-independent node names. We note that [5, 6] also studied routing schemes for minimizing *total* space, over all nodes, as opposed to the *maximum* space at a node, which is our measure of interest with respect to space. Bounds for routing schemes in terms of the degree of the network were also derived in [6].

While the Awerbuch et al. scheme achieved constant stretch with sublinear space, it was of theoretical interest only, because the stretch they achieved was far too large. Exploring the different stretch-space tradeoffs of [5], we obtain that the minimum stretch any of their schemes use when achieving sublinear space is 486 (calculated from Corollary 6.5 in their paper, setting $k = 3$, and noting that the constant in the big-O notation is in fact 2). That is, their schemes produce paths that are at

[1] Notice that this is a slightly stronger condition than having the nodes labeled with an arbitrary permutation P, since that assumes that the labels are precisely the integers $\{0, \ldots, n-1\}$. We talk about how to get around this in section 6.

TABLE 1
A comparison of our results (shown in boldface) to prior results on name-independent compact routing.

	Table size	Header size	Stretch
[5]	$\tilde{O}\left(n^{1/2}\right)$	$O(\log n)$	2592
[5]	$\tilde{O}\left(n^{2/3}\right)$	$O(\log n)$	486
[3]	$\tilde{O}\left(n^{1/2}\right)$	$O(\log n)$	1088
[3]	$\tilde{O}\left(n^{2/3}\right)$	$O(\log n)$	634
This paper	$\tilde{O}\left(n^{1/2}\right)$	$O(\log^2 n)$	**5**
This paper	$\tilde{O}\left(n^{1/2}\right)$	$O(\log n)$	**7**
This paper	$\tilde{O}\left(n^{2/3}\right)$	$O(\log n)$	**5**
Lower bound [13]	$o(n)$	$\log_2 n$	3

most 486 times the optimal length of the shortest path. A paper by Awerbuch and Peleg [3] that appeared a year later presented an alternate scheme with a polynomial space/stretch tradeoff that achieves superior stretch to the [5] construction when space is $\leq \tilde{O}(n^{1/2})$ (achieving a stretch of 1088 by substituting $k = 2$ into Lemma 3.2, whereas meeting this space bound in [5] requires setting $k = 4$ in Corollary 6.5 with a resulting stretch bound of 2592).

Gavoille and Gengler proved a lower bound of 3 for the stretch of any compact routing scheme that uses sublinear space at every node. Their result applies when there are up to $\log_2 n$ bits of topology-dependent routing information, and therefore applies also to the name-independent model [13].

Since the conference version of this paper [2], the gap between the Gavoille and Gengler lower bound of 3 [13] and this paper's upper bound of 5 has been closed by Abraham et al. [7] for polylogarithmic-sized headers. They obtained a stretch of 3 with $\tilde{O}\left(n^{1/2}\right)$ space and $O(\log^2 n/\log\log n)$-sized headers. Also, our upper bound with polynomial stretch tradeoff has been improved to a linear tradeoff by Abraham et al. [1], who obtained stretch $O(k)$ for space $\tilde{O}(n^{1/k}\log D)$, where D is the normalized diameter of the network.

1.1. Our results. This paper presents the first practical universal compact routing algorithms that achieve constant stretch with sublinear-sized routing tables, polylogarithmic packet headers, and name independence. Our first results substantially improve the best known stretch achievable with the sublinear-space constraint, as listed in Table 1. We then present tradeoff schemes that obtain increased but still bounded stretch, while decreasing the space to $\tilde{O}(n^{1/k})$ for each integral $k > 1$. The principal ingredients of our schemes include the following: the $O(\log n)$ greedy approximation to dominating set, used in the same fashion as in [6, 11, 9, 22] for most of the constructions; the sparse neighborhood covers of [3] for the construction in section 5; a distributed dictionary, as first defined by Peleg [19]; the schemes of [9] and [21, 12] for compact routing on trees; and a new randomized block assignment of ranges of addresses.

We note that our algorithms can be easily modified to determine either the name-dependent name of the destination or the results of a "handshaking scheme" in the sense of [21]. Therefore, if there is a whole stream of packets from a single origin headed for the same destination, once routing information is learned and the first packet is sent, an acknowledgment packet can be sent back with topology-dependent

address information so that subsequent packets can be sent to the destination using name-dependent routing—that is, without the overhead in stretch incurred due to the name-independent model, which arises partly from the need to perform lookups.

Stretch 5 and 7 schemes with different resource requirements are presented in section 3. In sections 4 and 5, we generalize the ideas in our stretch 5 and 7 constructions to two separate schemes that produce different stretch/space tradeoffs parameterized by an integer k. The scheme in section 4 uses space $\tilde{O}(kn^{1/k})$ and achieves stretch bounded by $1 + (2k-1)(2^k - 2)$. It achieves our best stretch/space tradeoff for $3 \leq k \leq 8$ (for $k = 2$, use the stretch 5 scheme of section 2; for $k \geq 9$, use the scheme in section 5). The scheme in section 5 uses space $\tilde{O}(k^2 n^{2/k})$ for graphs in which the edge weights are polynomial in n, and has a stretch bound of $16k^2 - 8k$. Combining the two bounds together yields the result given in the abstract, which improves on the best previously known stretch bounds for all integers $k > 1$ in the name-independent model. (The previous Awerbuch–Peleg scheme [3] uses space $\tilde{O}(k^2 n^{2/k})$ and achieves stretch bounded by $64k^2 + 16k$ for graphs whose edge weights are polynomial in n.)

1.2. Remarks on the model. Before we go into the details of the constructions, we make a few remarks about the model. We assume the nodes are labeled precisely with a permutation of the integers $\{0, \ldots, n-1\}$, but we refer the reader to section 6 for how to extend this to a more arbitrary set of distributively self-chosen node names. Each node v is also assumed to have a unique name from the set $\{1, \ldots, deg(v)\}$ assigned to each outgoing edge, but these names are assumed to be assigned locally with no global consistency. The model in which the names of the port numbers are chosen by the routing algorithm (based on network topology) was called the designer-port model by [12]. When the names of the port numbers are arbitrarily assigned by the network, the model is called the fixed-port model [12]. All of the results in this paper assume the more difficult fixed-port model.

Second, we point out that all our schemes in the name-independent model use writable packet headers; packets that are told only a topology-independent name may, in the course of their route, discover and then store topology-dependent routing information (of length at most $O(\log n)$ or $O(\log^2 n)$) to route to their destination. This is in contrast to the topology-dependent routing schemes, where in some of those schemes the fixed-topology information is "hardwired in" as the address of the packet and need never be changed.

2. Preliminaries. In this section we review two previously known name-dependent results on compact routing in tree networks, which we will use as subroutines, analyze the time taken to precompute the routing tables in these schemes, and also present a new name-independent compact routing scheme for trees and single-source routing in general graphs. We will use the following two compact routing results in the name-dependent model. We note that Lemma 2.2 is not the only tree-routing scheme in [21, 12], and that each of these papers presents different general schemes in the stronger designer-port model.

LEMMA 2.1 (see Cowen [9]). *There is a name-dependent routing scheme for any tree T with root l such that given any node v in T, the scheme routes along the optimal path of length $d(l, v)$ in the fixed-port model. The space per node is $O(\sqrt{n} \log n)$, and the address size is $O(\log n)$.*

LEMMA 2.2 (see Thorup and Zwick [21], Fraigniaud and Gavoille [12]). *There is a name-dependent routing scheme for any tree T such that given any pair of nodes u and v in T, the scheme routes along the optimal path of length $d(u, v)$ in the fixed-port model. The space per node is $O(\log n)$, and the address size is $O(\log^2 n)$.*

2.1. Precomputation running time. The following result is not proven explicitly in the original paper, but its running time is necessary for our running time analysis.

LEMMA 2.3. *The task of precomputing the routing tables and node labels of the name-dependent scheme of Cowen* [9] *in Lemma* 2.1 *runs in linear time.*

Proof. The identification of the big nodes $BN(T)$ in the tree T can be trivially accomplished in an $O(n)$ pass by reading the adjacency list of the tree and counting the degree of each node.

Having identified $BN(T)$, one $O(n)$ depth-first traversal through the tree is sufficient to determine the depth-first labels of the nodes, as well as the routing tables. During this pass we maintain a stack of pairs (u, p), where each u is the name of a big node on the path from the currently visited node to the root l, and p is the local port number at u for descending to the current node being visited by the depth-first traversal. The pairs (u, p) are ordered on the stack by distance from the root; a new pair $(u, \perp)$ is pushed each time we descend into a new big node u (but after we have chosen a label for the node u), and it is popped when we finish traversing the subtree rooted at u. The right element in the pair is initially blank (indicated $\perp$) but it is set to a different value for every child of u that we visit.

The routing labels $R(v)$ assigned to the currently traversed node v includes the depth-first number of v and the item (u, p) currently on top of the stack.

For the nodes that are not big nodes, at most one routing table entry (corresponding to an interval) is created in constant time for each end of an undirected edge in the tree, so creating the set of routing tables $Tab(u)$ for the non-big nodes u takes $O(n)$ time and can be accomplished without increasing the running time of the depth-first traversal.

We create the routing tables for the big nodes in $BN(T)$ as follows. We use the stack of big nodes containing pairs of the form (u, p) described previously. Each time a new pair $(u', \perp)$ is pushed onto the stack, we traverse the stack nondestructively from top to bottom (that is, towards the root l) and set the routing table $Tab(u)$ of u in each pair (u, p) to contain a pointer to the newly discovered big node u'. That is, we indicate that at node u, in order to reach node u', we use port p. Note that even though the big nodes may be scattered within the tree, having the stack enables us to create each table entry (in an ancestor of u') in $O(1)$ time. Since there are at most $O(\sqrt{n})$ big nodes each with at most $\sqrt{n}$ big descendants, the total running time for this phase is $O(n)$. Note that for clarity we have presented these two $Tab(u)$ computations for big and non-big nodes separately, but in practice they can be interleaved in one pass with the same total asymptotic running time. □

We now turn to the schemes of Thorup and Zwick [21] and Fraigniaud and Gavoille [12]. In [12], it is shown that the tables for their name-dependent scheme can be precomputed in $O(n \log n)$ time. It can be easily shown that the scheme of Thorup and Zwick [21] can also be computed in $O(n \log n)$ time.

2.2. Single-source name-independent compact routing. In this section we present a new name-independent compact routing scheme for single-source routing in arbitrary networks. The intuition for name-independent compact routing can be explained by a simple analogy. In the real world a directory maps names to contact information. A compact routing table is modeled in this analogy by being unable to store the entire directory at any single location. We therefore split up the directory into equal-sized consecutive blocks and distribute these close to the node that needs to refer to them. We must be able to predict which nearby node will contain the

entry we are interested in. Then, given a name, we will locate the nearby node with the relevant portion of the directory, read the directory entry, then use the contact information to locate the person. The main problem is showing how to do this in such a way that the lookup process is not excessive compared with the cost of finally using the contact information to locate the person.

Let T be a (weighted) rooted n-node tree with root r, whose nodes are labeled $\{0, \ldots, n-1\}$ according to some arbitrary permutation P (T could be a shortest path tree in a general graph, for single-source routing). For simplicity, we assume that $\sqrt{n}$ is an integer.[2] We first prove the following.

LEMMA 2.4. *Given a tree T with a weight function w defined on its edges, there exists a name-independent routing scheme that*

1. *requires $O(\sqrt{n} \log n)$ space per node;*
2. *remembers at most $O(\log n)$ bits in the packet header;*
3. *routes a packet from the root r to the node with label j (for any $j \in \{0, \ldots, n-1\}$) along a path of length at most $3d(r, j)$.*

Proof. Let r denote the root of T. For each i and j, let e_{ij} denote the port name of the first edge along the shortest path from i to j. Denote by $N(i)$ the set of the $\sqrt{n}$ closest nodes to i in T, including i, and breaking ties lexicographically by node name. Furthermore, divide the space of node labels $\{0, \ldots, n-1\}$ into blocks of size $\sqrt{n}$, so that block B_0 consists of the addresses from $0 \ldots \sqrt{n}-1$ and block B_i consists of the node labels $i\sqrt{n}$ to $(i+1)\sqrt{n}-1$ (recall that $\sqrt{n}$ is assumed to be an integer). Let $CR(x)$ denote the address label that a node x would be assigned under the tree-routing scheme of Lemma 2.1, and let $CTab(x)$ denote the corresponding routing table stored by node x.

Let $v_{\phi(0)}, v_{\phi(1)}, \ldots, v_{\phi(\sqrt{n}-1)}$ be the names assigned to the nodes in $N(r)$, ordered by distance from the root r, with ties broken lexicographically. The following are stored at each node i in T:

- (r, e_{ir}) for the root node r.
- If $i \in N(r)$, then $i = v_{\phi(t)}$ for some unique index t. For each $j \in B_t$, $(j, CR(j))$ is stored. Call this the block table.
- $CTab(i)$.

In addition, the following extra information is stored at the root node r:

- For each node x in $N(r)$, $(x, CR(x))$ is stored. Call this the root table.
- For $0 \le k < \sqrt{n}$, the pair $(k, v_{\phi(k)})$ is stored. Call this the dictionary table.

Now suppose a packet destined for j arrives at r. If $(j, CR(j))$ is in the root table, the packet writes $CR(j)$ into its header and routes optimally to j with stretch 1 using the $CTab(x)$ tables. Otherwise, let t be the index such that j is in B_t, and look up $(t, v_{\phi(t)})$ in the dictionary table, followed by $(v_{\phi(t)}, CR(v_{\phi(t)})$ in the root table, and write $CR(v_{\phi(t)})$ into the packet header (where we note that there is guaranteed to be an entry for $v_{\phi(t)}$ in the root table because $v_{\phi(t)} \in N(r)$). As illustrated in Figure 1, we route optimally to $v_{\phi(t)}$, look up $(j, CR(j))$ in its block table, write $CR(j)$ into the packet header, and route optimally back to the root using the (r, e_{xr}) entries found at intermediate nodes x. Then we route optimally from the root to j using $CR(j)$ and the $CTab(x)$ tables. Since $v_{\phi(t)}$ is among r's closest $\sqrt{n}$ nodes and j is not, we have $d(r, v_{\phi(t)}) \le d(r, j)$ and thus the total route length is $\le 3d(r, j)$.

$CTab(x)$ is of size $O(\sqrt{n} \log n)$ by Lemma 2.1. Since there are exactly $\sqrt{n}$ nodes in $N(i)$ for every n, every block table has $\sqrt{n}$ entries, each of size $O(\log n)$ bits. The

[2] When $\sqrt{n}$ is not an integer we round n up to the smallest larger perfect square, at a cost of (less than) doubling the length of a $\log n$ bit node identifier.

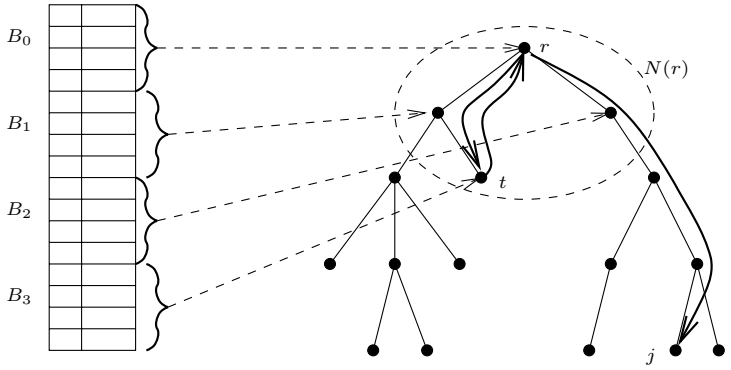


FIG. 1. *An illustration of the derivation of the stretch bound for single-source routing in a tree T. A table of the name-dependent labels keyed by the topology-independent node names (shown on the left) is broken into $\sqrt{n}$ equal-sized blocks and distributed in the $\sqrt{n}$-sized neighborhood $N(r)$ of the root r. To route to a destination node j outside $N(r)$, we look up the name-dependent label for j at t, where $t \in N(r)$ satisfies $j \in B_t$ (routing to t and back is done optimally). Finally we route optimally to j, obtaining a stretch of* 3.

additional information stored at the root consists of two $\sqrt{n}$-entry tables, each with $O(\log n)$-bit entries. The maximum space requirement is therefore $O(\sqrt{n}\log n) = \tilde{O}(\sqrt{n})$ at every node. □

Note that if we substitute a name-dependent tree-routing scheme that satisfies Lemma 2.2 for the one in Lemma 2.1 in the construction above, we get the same stretch bounds, but the packet header size increases to $O(\log^2 n)$.

2.3. Hitting set algorithm. Given an undirected (weighted) network G with n nodes and m edges, we determine for each node u a *neighborhood ball* $N(u)$ of the $n^{1/2}$ nodes closest to u including u and breaking ties lexicographically by node name. Next we define a hitting set L of *landmarks*, such that for every node v, $N(v)$ contains a node in L. The following well-known result appears in [18]; it follows from the general $(1 + \ln n)$-approximation algorithm for set-cover.

LEMMA 2.5 (see Lovász [18]). *Let $G = (V, E)$ be an undirected graph of n nodes and m edges. Let $N(v)$ denote the set of v's $n^{1/2}$ closest neighbors (with ties broken lexicographically by node name). There exists a set $L \subset V$ such that $|L| = O(n^{1/2}\log n)$ and $\forall v \in V, L \bigcap N(v) \neq \emptyset$. A greedy algorithm exists that computes L in $\tilde{O}(m + n^{3/2})$ time.* □

3. Name-independent routing with stretch 5 and 7 in general networks. Let V be labeled with unique addresses $\{0, \ldots, n-1\}$. We divide the address space into blocks B_i for $i = 0, \ldots, \sqrt{n}-1$, so that block B_i consists of the node labels $i\sqrt{n}$ to $(i+1)\sqrt{n}-1$. A polylogarithmic number of blocks will be assigned to each node such that each neighborhood contains an instance of every block (see Lemma 3.1)—let S_i be the set of blocks assigned to node i.

Let T_l denote a single-source shortest path tree rooted at l that spans all the nodes of the network. Also, partition the nodes of G into sets H_l according to their closest landmarks, so that $H_l = \{v | v\text{'s closest landmark is } l\}$. Let $T_l[H_l]$ be a single-source shortest path tree rooted at l spanning just the nodes of H_l. Let l_u denote u's closest landmark in L.

In what follows, we present three compact routing schemes A, B, and C in the name-independent model. Scheme A uses $\tilde{O}(n^{1/2})$-sized routing tables and $O(\log^2 n)$-

sized routing headers, while achieving a stretch bound of 5. Scheme B improves on header size at the expense of stretch—it uses $\tilde{O}(n^{1/2})$-sized routing tables and $O(\log n)$-sized routing headers, while achieving a stretch bound of 7. Finally, compared to Scheme A, Scheme C trades table size for header size, and uses $\tilde{O}(n^{2/3})$-sized routing tables and $O(\log n)$-sized routing headers, while achieving a stretch bound of 5.

3.1. Common data structures. In this subsection we present some data structures common to all three routing schemes and analyze their precomputation time. All three schemes utilize the sets of blocks S_v, whose properties are described by the following lemma, which is a special case of Lemma 4.1. To avoid repetition, the latter is proved in section 4, after building up more general definitions which are not necessary for the following case (substituting $k = 2$ in Lemma 4.1 immediately yields the following).

LEMMA 3.1. *Let G be a graph on n nodes, and let $N(v)$ denote the set of v's closest $\sqrt{n}$ neighbors (including v itself) with ties broken lexicographically by node name. Let $\{B_i | 0 \leq i < \sqrt{n}\}$ denote a set of blocks. There exists an assignment of sets S_v of blocks to nodes v, such that*

- *$\forall v \in G$, $\forall B_i$ $(0 \leq i < \sqrt{n})$, there exists a node $j \in N(v)$ with $B_i \in S_j$;*
- *$\forall v \in G$, $|S_v| = O(\log n)$.*

This assignment can be computed in $\tilde{O}(n^2)$ expected time or $\tilde{O}(n^3)$ deterministic time.

Given such an assignment of blocks, the routing tables of all three schemes contain the following for every node u:

1. For every node v in $N(u)$, (v, e_{uv}).
2. For every i, $0 \leq i < \sqrt{n}$, (i, t), where $t \in N(u)$ satisfies $B_i \in S_t$ (such a node t exists by our construction of S_u in Lemma 3.1).

Clearly (1) takes $O(\sqrt{n} \log n)$ space. Note that since $N(u)$ is of size $\sqrt{n}$, and since (2) takes $O(\log n)$ space for each of $\sqrt{n}$ values i, these common data structures require a total of $O(\sqrt{n} \log n)$ space.

To compute the neighborhood $N(u)$ of each node u, we run a truncated Dijkstra algorithm. This takes $\tilde{O}(n)$ time per node [10], for a total of $\tilde{O}(n^2)$ time. During the computation of the truncated Dijkstra algorithm, we also obtain the table of entries (v, e_{uv}) in item (1) by overloading the relaxation operation that sets a new parent for each node, so that it also derives a new tentative port number from its new parent, and this is stored in the root of that particular truncated Dijkstra run. Since the extra work done by each relaxation takes $O(1)$ time, the asymptotic running time remains the same. Finally, on identifying all the neighborhoods we create a sorted list of the neighbors of each node, in a total of $O(n^{3/2} \log n)$ time.

The assignment of sets satisfying Lemma 3.1 is computed in expected $\tilde{O}\left(n^2\right)$ or deterministic $\tilde{O}\left(n^3\right)$ time. After determining the assignment, we compute the pairs (i, t) in item (2). For each node u and each neighbor $t \in N(u)$, for every block B_i assigned to node t, we store the pair (i, t) in an array of length $\sqrt{n}$ associated with node u. The total time for this procedure is $\tilde{O}(n^{3/2})$. Thus we have shown the following.

LEMMA 3.2. *The common data structures described above are of size $O(\sqrt{n} \log n)$ and can be computed in $\tilde{O}(n^2)$ expected or $\tilde{O}(n^3)$ deterministic time.*

3.2. Scheme A.

3.2.1. Data structures. Let L be any set of landmarks that satisfies Lemma 2.5, and let $Tab(x)$ and $R(x)$ refer to the routing table and address, respectively, of node

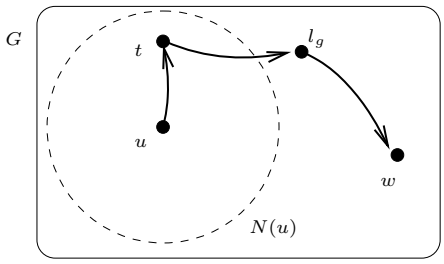


FIG. 2. *An illustration of the worst-case route taken by a packet under Scheme* A. *Only relevant nodes and paths are depicted. This route is shown to satisfy a stretch bound of* 5 *in Theorem* 3.3.

x, under a tree-routing scheme that satisfies the requirements of Lemma 2.2. Recall that e_{uv} denotes the first edge along a shortest path from u to v. Each node u stores the following, in addition to the common data structures described in section 3.1:

1. For every node $l \in L$, (l, e_{ul}).
2. For every block B_k in S_u, and for each node j in B_k, the triple $(j, l_g, R(j))$, where l_g is a landmark that minimizes, over all landmarks in L, the quantity $d(u, l_g) + d(l_g, j)$, and $R(j)$ is the tree-routing address j in the tree T_{l_g}.
3. For every landmark $l \in L$, u stores the routing table $Tab(u)$ for the tree T_l.

3.2.2. Routing algorithm. Consider two cases for the location of the destination node w relative to the source node u.

(i) $w \in N(u) \bigcup L$: Then the entry (w, e_{vw}) is stored at every node v on the shortest path from u to w and we route directly to w with a stretch of 1.

(ii) $w \notin N(u) \bigcup L$: (The route for this case is illustrated in Figure 2.) On failing to find (w, e_{uw}) stored at u, it must be the case that $w \notin N(u) \bigcup L$. Compute the index i for which $w \in B_i$, and look up the node $t \in N(u)$ that stores entries for all nodes in B_i. Next, route optimally to the node t using (t, e_{xt}) information at intermediate nodes x. At node t, we look up l_g, route optimally to l_g, following the (l_g, e_{vl_g}) entries in the routing tables in nodes v on the shortest path from t to l_g, and then optimally from l_g to w, using the address $R(w)$ and the tree-routing tables $Tab(x)$ stored at all nodes for the tree rooted at l_g.

THEOREM 3.3. *Given a graph G with n nodes and m positive-weighted edges, there exists a stretch-*5 *compact routing scheme that uses $O(\sqrt{n}\log^3 n)$-sized local routing tables and $O(\log^2 n)$ headers which can be precomputed in $\tilde{O}(n^2 + m\sqrt{n})$ expected or $\tilde{O}(n^3)$ deterministic time.*

Proof. First we show that the stretch of Scheme A is bounded by 5. If $w \in N(u) \bigcup L$, we route optimally with stretch 1. Otherwise, the route taken is of length $d(u,t) + d(t, l_g) + d(l_g, w)$. We have $d(u,t) + d(t, l_g) + d(l_g, w) \leq d(u,t) + d(t, l_u) + d(l_u, w)$, because l_g was chosen to minimize precisely the quantity $d(t, l) + d(l, w)$, for all $l \in L$. Now $d(t, l_u) \leq d(t, u) + d(u, l_u)$ by the triangle inequality, and similarly $d(l_u, w) \leq d(l_u, u) + d(u, w)$. Since $t \in N(u)$ by construction, $w \notin N(u)$ implies $d(u,t) \leq d(u,w)$. Similarly, L being a hitting set for $N(u)$ implies $l_u \in N(u)$, thus $d(u, l_u) \leq d(u, w)$. Thus the route taken is of length $\leq 2d(u,t) + 2d(u, l_u) + d(u,w) \leq 5d(u,w)$.

Next we show that the data structures of section 3.2.1 require $O(n^{1/2}\log^3 n)$ space. The space of (1) is $O(\log n)$ bits for each landmark in the set L which is of size $O(\sqrt{n}\log n)$. For (2) we need $O(\log^2 n)$ space for each of the $\sqrt{n}$ nodes in each block

(note that (2) includes an $O(\log^2 n)$-sized tree-routing address for each node in each block), times the number of blocks in S_u, which is $O(\log n)$ by Lemma 3.1, for a total of $O(n^{1/2}\log^3 n)$ space. (3) takes $O(\sqrt{n}\log^2 n)$ space because the number of trees is equal to the number of landmarks, which is $O(\sqrt{n}\log n)$, and u stores $O(\log n)$ bits for each tree.

Finally we analyze the running time required for computing the routing tables. Recall from Lemma 2.5 that computing the set of landmarks L takes $\tilde{O}(m+n^{3/2})$ time [18]. To obtain the pointers (l, e_{ul}) at each node (item (1)), we run the full Dijkstra algorithm for single-source shortest path trees from each landmark $l \in L$ in a total of $\tilde{O}(n^{3/2}+m\sqrt{n})$ time (using a Fibonacci heap implementation of Dijkstra's algorithm). We can overload the relaxation operation so that when the parent of b is set to a, b is added to the adjacency list of a (and deleted from that of its previous parent if there was one). This gives us an adjacency-list representation of the single-source shortest path tree. We could also sort the list of landmarks to facilitate the use of a binary search for quickly obtaining (l, e_{ul}), but this can be done in $\tilde{O}(\sqrt{n})$ time.

For each tree subgraph T_l with n nodes (a single-source shortest path tree rooted at $l \in L$), we can compute the routing labels $R(u)$ and name-dependent routing tables $Tab(u)$ (of item (3)) according to the scheme of [12] or [21] in $\tilde{O}(n)$ time (by [12], for example). Since there are $\tilde{O}(\sqrt{n})$ landmarks in all, the total running time is $\tilde{O}(n^{3/2})$.

That done, we fill in the contents of the blocks (item (2)) in $O(n^2\log^2 n)$ time (n nodes times $O(\log n)$ blocks per node, times $\sqrt{n}$ entries per block, times $O(\sqrt{n}\log n)$ landmark candidates, times $O(1)$ for looking up $d(u, l_i)$ and $d(l_i, j)$ from the precomputed tables).

The total expected running time for Scheme A is therefore $\tilde{O}(n^2 + m\sqrt{n})$. It is interesting to note that this is less than the best known running time for all-pairs shortest path algorithms. □

3.3. Scheme B.

3.3.1. Data structures. Again we define the set L as any set of landmarks that satisfies Lemma 2.5. Recall that $CTab(x)$ and $CR(x)$ refer to the routing table and address, respectively, of node x in a scheme that satisfies the requirements of Lemma 2.1. In addition to the common data structures in section 3.1, each node u stores the following:

1. For every node $l \in L$, (l, e_{ul}).
2. For every node j in B_k, where B_k is a block in S_u, the name of the closest landmark l_j to j, and the tree-routing address $CR(j)$ for j in the tree $T_{l_j}[H_{l_j}]$.
3. If l_u is u's closest landmark, then u stores its routing table $CTab(u)$ for the tree T_{l_u}.

3.3.2. Routing algorithm. Again, consider two possible cases on the location of the destination node w relative to the source node u.

(i) $w \in N(u) \bigcup L$: Then the entry (w, e_{vw}) is stored at every node v on the shortest path from u to w and we route directly to w with a stretch of 1.

(ii) $w \notin N(u) \bigcup L$: (This case is illustrated in Figure 3.) On failing to find (w, e_{uw}) stored at u, it must be that $w \notin N(u) \bigcup L$. Compute the index i for which $w \in B_i$, and let $t \in N(u)$ be the node that stores entries for all nodes in B_i. Use the entries (t, e_{xt}) at intermediate nodes x to route optimally to the node t. At node t, we look up l_w, route optimally to l_w, following the (l_w, e_{vl_w}) entries in the routing tables in nodes v on the shortest path from t to l_w, and then optimally from l_w to w, using the address $CR(w)$ in the tree

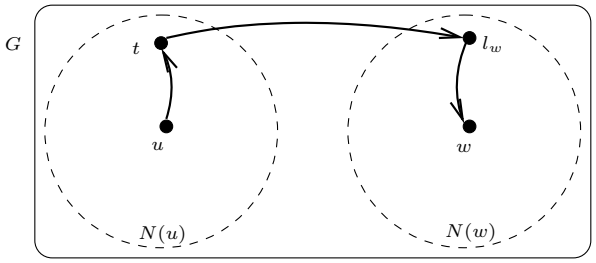


FIG. 3. *An illustration of the worst-case route taken by a packet under Schemes* B *and* C. *This route is shown to satisfy a stretch bound of* 7 *in Theorem* 3.4. *For Scheme* C, *we use the name-dependent scheme of Cowen* [9], *so we have the extra condition that* $d(l_w, w) = d(w, l_w) < d(u, w)$ *and this reduces the stretch bound to* 5, *as shown in Theorem* 3.6.

T_{l_w}, coupled with the tree-routing tables $CTab(x)$ stored for all nodes x that chose l_w as their closest landmark.

THEOREM 3.4. *Given a graph* G *with* n *nodes and* m *positive-weighted edges, there exists a stretch-*7 *compact routing scheme that uses* $O(\sqrt{n}\log^2 n)$*-sized local routing tables and* $O(\log n)$ *headers which can be precomputed in* $\tilde{O}(n^2+m\sqrt{n})$ *expected or* $\tilde{O}(n^3)$ *deterministic time.*

Proof. First we show that the stretch of Scheme B is bounded by 7. If $w \in N(u) \bigcup L$, we route optimally with a stretch of 1. Otherwise, the route taken by the algorithm is of length $d(u,t) + d(t,l_w) + d(l_w,w)$. Now $d(t,l_w) \le d(t,w) + d(w,l_w) \le d(t,u) + d(u,w) + d(w,l_w)$, by repeated applications of the triangle inequality, so the route taken by the algorithm is of length $\le 2d(u,t) + d(u,w) + 2d(l_w,w)$. But $d(u,t) \le d(u,w)$ because $t \in N(u)$ and w is not. Also, $d(l_w,w) \le d(l_u,w)$, since l_w is w's closest landmark. So $d(l_w,w) \le d(w,l_u) \le d(w,u)+d(u,l_u) \le 2d(u,w)$, where the second inequality follows from the triangle inequality and the third from the fact that $l_u \in N(u)$ (since L is a hitting set), while w is not. So $2d(u,t)+d(u,w)+2d(l_w,w) \le 7d(u,w)$, proving the result.

Next we show that the items in section 3.3.1 require $O(n^{1/2}\log^2 n)$ space. The space for (1) is exactly the same as for Scheme A, which we have already shown is $O(\sqrt{n}\log^2 n)$. (2) takes $O(\sqrt{n}\log n)$ space per block times the number of blocks that are stored at a node. This is because we are storing $O(\log n)$ bits for the tree-routing address of each of the $\sqrt{n}$ nodes in every block. So (2) takes $\sqrt{n}$ space times the number of blocks in S_u, which is $O(\log n)$ by Lemma 3.1. (3) takes $O(\sqrt{n}\log n)$ space in Scheme B because the trees T_l partition the nodes and each node participates in only one tree, thus requiring space $O(\sqrt{n}\log n)$.

Finally we consider the running time. The computation of item (1) for Scheme B is identical to that in Scheme A, and runs in $\tilde{O}(n^{3/2} + m\sqrt{n})$ time. By Lemma 2.3, given a tree subgraph T_l with n nodes (a single-source shortest path tree rooted at $l \in L$), we can compute the routing labels $R(u)$ and name-dependent routing tables $Tab(u)$ (of item (3)) according to the scheme of Cowen [9] in $O(n)$ time per landmark. Therefore the total running time for computing item (3) and the routing labels used in item (4) is $\tilde{O}(n^{3/2})$, since there are $\tilde{O}(\sqrt{n})$ landmarks and the running time is $O(n)$ per landmark.

We compute the closest landmark to each node using the tables of shortest distances from each landmark to every node in the graph, in $\tilde{O}(n^{3/2})$ time, by considering each node and, for each node, every possible landmark.

We can now fill in the contents of the blocks (item (2)): as in Scheme A, there are $O(n^{3/2} \log n)$ entries in all the blocks in all the nodes, times $O(1)$ to look up the closest landmark l_j to node j, and the routing label $CR(j)$ for j in the tree T_{l_j} spanning the set of nodes whose closest landmark is l_j. This comes to a total of $\tilde{O}(n^{3/2})$. The total expected running time is therefore $\tilde{O}(n^2 + m\sqrt{n})$. Again, this is less than the best known running time for all-pairs shortest path algorithms. □

3.4. Scheme C.

3.4.1. Data structures. Let L be the set of landmarks constructed in the following topology-dependent compact routing scheme.

LEMMA 3.5 (see Cowen [9]). *There is a name-dependent compact routing algorithm with $O(n^{2/3} \log^{4/3} n)$-sized tables and $O(\log n)$-bit headers at each node which achieves stretch* 3.

Also let $LTab(x)$ and $LR(x)$ denote the corresponding routing table and address for node x that the scheme of [9] constructs. Recall that $CTab(x)$ and $CR(x)$ refer to the same parameters in a scheme that satisfies the requirements of Lemma 2.1. Each node u stores the following:

1. For every node j in B_k, where B_k is a block in S_u, the name of the closest landmark l_j to j, and the tree-routing address $CR(j)$ for j in the tree $T_{l_j}[H_{l_j}]$.
2. The routing table $LTab(u)$ and for every node $v \in N(u)$, $LR(v)$.

3.4.2. Routing algorithm. If u has stored an entry for w that gives w's address $LR(w)$, we use Cowen's compact routing scheme of [9] to route to w, with stretch bounded by 3. So suppose u has no address $LR(w)$ stored for w in its local table. It must be that $w \notin N(u) \bigcup L$. Compute the index i for which $w \in B_i$.

(i) If $u \in L$, look up the node $t \in N(u)$ that stores entries for all nodes in B_i, and use (t, e_{xt}) to route optimally to t. At t, write $LR(w)$ into the packet header, and then use the landmark pointers in the routing tables to route optimally back from t to u. Then, use $LR(w)$ and Cowen's compact routing scheme (see [9]) to route to w with stretch bounded by 3. The cost of the round trip to t and back is less than $2d(u, w)$, because $t \in N(u)$ and $w \notin N(u)$ implies $d(u, t) < d(u, w)$, so the total stretch is bounded by 5.

(ii) If $u \notin L$, by Cowen's construction, since u has no address $LR(w)$ stored for w in its local table, it must be that $d(l_w, w) < d(u, w)$. In this case (illustrated in Figure 3), we look up (t, e_{xt}) to route optimally to the node $t \in N(u)$ that stores entries for all nodes in B_i. We determine the identity of l_w, and the address of w in the tree routed at l_w from t's entry for w in its local table. Then we route optimally from t to l_w, and then from l_w to w.

THEOREM 3.6. *Given a graph G with n nodes and m positive-weighted edges, there exists a stretch-5 compact routing scheme that uses $O(n^{2/3} \log^{4/3} n)$-sized local routing tables and $O(\log n)$ headers which can be precomputed in $\tilde{O}\left(n^2 + mn^{2/3}\right)$ expected or $\tilde{O}(n^3)$ deterministic time.*

Proof. First we show the stretch bound of 5. In this regard, it remains to analyze the case when $w \notin N(u) \bigcup L$ and $u \notin L$. Then, as remarked above, the absence of an entry for w in Cowen's scheme implies $d(l_w, w) \leq d(u, w)$, and the route taken is of length $d(u, t) + d(t, l_w) + d(l_w, w)$. Now $d(t, l_w) \leq d(t, u) + d(u, l_w)$, and $d(u, l_w) \leq d(u, w) + d(w, l_w)$. So the route is of length $\leq 2d(u, t) + d(u, w) + 2d(w, l_w) \leq 5d(u, w)$, since $w \notin N(u)$ and $t \in N(u)$ implies $d(u, t) \leq d(u, w)$.

The space requirements are as follows. As in Schemes A and B, item (1) takes $O(\sqrt{n} \log^2 n)$ space. The items in (2) are the tables of Cowen's scheme, which are

proved to be of size $O(n^{2/3}\log^{4/3} n)$ in [9]; this clearly dominates the space requirements of (1).

Finally the precomputation time is obtained as follows. In Scheme C, we use the name-dependent routing scheme for general graphs of [9]. Therefore we require $O(n^{5/3} + n^{2/3}m)$ time for precomputing the name-dependent routing labels $LR(u)$ and tables $LTab(u)$ [9].

Item (1) is computed similarly to item (2) of Scheme B except that it refers to $LR(u)$ and $LTab(u)$, and again the running time is $\tilde{O}(n^{3/2})$.

Finally item (2) stores, for each node u and each node $v \in N(u)$, the value $LR(v)$. This takes a total of $\tilde{O}(n^{3/2})$. Clearly the running time for Scheme C is dominated by the precomputations for the name-dependent scheme and the $\tilde{O}(n^2)$ time required for computing the common data structures, and altogether this takes $O(n^2 + n^{2/3}m)$ expected running time. □

4. A generalized routing scheme for $\tilde{O}(n^{1/k})$ space. In this section we present compact routing schemes that provide tradeoffs between the amount of space available at a node and the stretch obtained. In the process, we prove Lemma 4.1, of which Lemma 3.1 is a special case (obtained by setting $k = 2$).

4.1. Preliminaries. Given a graph G with $V = \{0, \ldots, n-1\}$, we assume for simplicity that $n^{1/k}$ is an integer, and define the alphabet $\Sigma = \{0, \ldots, n^{1/k} - 1\}$. For each $0 \le i \le k$, Σ^i is the set of words over Σ of length i. Let $\langle u \rangle \in \Sigma^k$ be the base $n^{1/k}$ representation of u, padded with leading zeros so that it is of length exactly k. For each $0 \le i \le k$, we also define functions $\sigma^i : \Sigma^k \longrightarrow \Sigma^i$, such that $\sigma^i((a_0, \ldots, a_{k-1})) = (a_0, \ldots, a_{i-1})$. That is, σ^i extracts the prefix of length i from a string $\alpha \in \Sigma^k$.

For each $\alpha \in \Sigma^{k-1}$, define a set $B_\alpha = \{u \in V | \sigma^{k-1}(\langle u \rangle) = \alpha\}$. We will call these sets *blocks*. Clearly $\forall \alpha \in \Sigma^{k-1}, |B_\alpha| = n^{1/k}$. We abuse notation slightly by defining $\sigma^i(B_\alpha) = \sigma^i(\alpha 0)$, where $\alpha 0$ is the word in Σ^k obtained by appending a zero to α. Note that by this definition, $\sigma^{k-1}(B_\alpha) = \sigma^{k-1}(\langle u \rangle)$ whenever $u \in B_\alpha$.

For every node u, we define the neighborhoods $N^i(u)$ as the set of $n^{i/k}$ nodes closest to u including u itself, breaking ties lexicographically by node name. We first prove the following.

LEMMA 4.1. *Given a graph G, there exists an assignment of sets of blocks S_v to nodes v, so that*

- *$\forall v \in G$, $\forall 0 \le i < k$, $\forall \tau \in \Sigma^i$, there exists a node $w \in N^i(v)$ with $B_\alpha \in S_w$ such that $\sigma^i(B_\alpha) = \tau$;*
- *$\forall v \in G$, $|S_v| = O(\log n)$.*

Such an assignment can be computed in expected $\tilde{O}\left(n^{3-2/k}\right)$ time or in deterministic $\tilde{O}\left(n^{4-2/k}\right)$ time.

Proof. Our existence proof is by the probabilistic method. Let n be the number of nodes in the graph G. Consider a random assignment of $f(n)$ blocks to each node, each block chosen independently and uniformly at random from $\mathcal{B}$, the set of all blocks. Here $f(n)$ will be defined later to ensure the result.

For $u \in G$ and $\tau \in \Sigma^i$ for some i, $0 \le i < k$, we say that (u, τ) is *covered* if there exists a node w in $N^i(u)$ such that w is assigned a block B_α for which $\sigma^i(B_\alpha) = \tau$. Since for every i, $|\Sigma^i| = |N^i(u)| = n^{i/k}$, the number of times a node in $N^{|\tau|}(u)$ is assigned a block is $n^{|\tau|/k} \cdot f(n)$; in each instance, the probability that (u, τ) is covered is $1/n^{|\tau|/k}$. Thus, the probability that (u, τ) is uncovered at the end of the assignment

is

$$\left(1 - \frac{1}{n^{|\tau|/k}}\right)^{n^{|\tau|/k} f(n)} \leq e^{-f(n)}.$$

The total number of different pairs (u, τ) is

$$n \cdot \sum_{0 \leq i < k} n^{i/k} = \frac{n(n-1)}{n^{1/k} - 1} < n^2,$$

since $n^{1/k} \geq 2$. The expected number of pairs that remain uncovered at the end of the assignment is less than $n^2 e^{-f(n)}$. If we choose $f(n) = \lceil 2 \ln n \rceil$, then the expected number of uncovered pairs is strictly less than 1, thus guaranteeing the existence of an assignment that covers all pairs. If we choose $f(n) = \lceil 2 \ln n + \ln 2 \rceil$, then the failure probability is at most $1/2$. Thus repeating this procedure an expected $O(1)$ times would yield the desired assignment.

We now calculate the expected running time of the above randomized algorithm. We can calculate $N^i(u)$ for $0 \leq i < k$ and u using a truncated Dijkstra algorithm [10] in $\tilde{O}\left(n^{(2k-2)/k}\right)$ time per node. During this calculation, we can also compute for each w and $0 \leq i < k$ the list of u such that $w \in N^i(u)$. Given a complete assignment, we go over each node w and each block B_α assigned to w and mark all pairs (u, τ) that are covered due to this block. The total time taken for this procedure is

$$\sum_{w \in V} O(\log n) \cdot \sum_{0 \leq i < k} |\{u : w \in N^i(u)\}| = O(\log n) \cdot \sum_{u \in V} \sum_{0 \leq i < k} |N^i(u)| = O(n^{2-1/k} \log n).$$

Thus the expected total running time is $\tilde{O}\left(n^{3-2/k}\right)$.

We now derandomize the above probabilistic assignment by deterministically assigning blocks to the nodes one at a time, subject to the constraint that the number of blocks assigned to each node at the end of the procedure is $f(n) = \lceil 2 \ln n \rceil$. The procedure consists of $nf(n)$ steps, numbered from 1. In each step, we arbitrarily select a node u which can be assigned at least one more block. We assign to u a block that minimizes the expected number of uncovered pairs, conditioned on the partial assignment chosen thus far, assuming that the blocks assigned in subsequent steps are chosen independently and uniformly at random from $\mathcal{B}$. Let A_i represent the partial assignment at the end of step i and let U be the random variable representing the number of uncovered pairs at the end of the complete assignment. For convenience, let A_0 denote the empty assignment.

By our argument above, we know that $E[U \mid A_0] < 1$. We will now show that for $1 \leq j \leq nf(n)$, $E[U \mid A_j] \leq E[U \mid A_{j-1}]$. This would imply that $E[U \mid A_{nf(n)}] < 1$; since $A_{nf(n)}$ is the complete assignment, the random variable $U \mid A_{nf(n)}$ is, in fact, deterministic and, being an integer, has to equal 0.

Consider the jth step. Let u be the node to which a block is assigned in this step. We have

$$E[U \mid A_{j-1}] = \frac{1}{|\mathcal{B}|} \sum_{B \in \mathcal{B}} E[U \mid A_{j-1} \cup \{B \to u\}].$$

(Here, $B \to u$ denotes that block B is assigned to node u and we represent a partial assignment as a multiset of elements of the form block $\to$ node.) Clearly, there exists $B \in \mathcal{B}$ such that $E[U \mid A_j \cup \{B \to u\}] \leq E[U \mid A_j]$. By our choice of the block in

each step, it follows that $E[U \mid A_j] \leq E[U \mid A_{j-1}]$. This completes the proof that the final assignment has the property that all of the pairs are covered.

It remains to establish that our block assignment procedure is polynomial time. As for the randomized algorithm, we first compute the neighborhoods $N^i(\cdot)$ and their inverses for each node u in $\tilde{O}\left(n^{3-2/k}\right)$ time using a truncated Dijkstra algorithm [10]. The total number of block assignment steps is $n\lceil 2 \ln n \rceil$. In step j, we examine all blocks B and compute $E[U \mid A_{j-1} \cup \{B \to u\}]$. We maintain the set of uncovered pairs (u, τ) and, for each pair (u, i) where $0 \leq i < k$, a count $c(u, i)$ of the total number of blocks that remain to be assigned to nodes in $N^i(u)$. Maintaining the set of uncovered pairs and the count $c(\cdot, \cdot)$ for the uncovered pairs takes overall time $O(n^{(k-1)/k} \log n \cdot n^{2-1/k})$ since we consider $O(n^{(k-1)/k} \log n)$ blocks per node, and $n^{2-1/k}$ is the sum of the sizes of all of the neighborhoods $N^i(u)$, over all i and all u.

Given a partial assignment A, the conditional expectation can be calculated as follows:

$$E[U \mid A] = \sum_{\text{uncovered}(u,\tau)} \left(1 - \frac{1}{n^{|\tau|/k}}\right)^{c(u,|\tau|)}.$$

This takes time $O(n^{2-1/k})$, which is a bound on the number of total pairs (u, τ). Since the total number of blocks equals $|\Sigma^{k-1}| = n^{(k-1)/k}$, the total number of conditional expectations calculated overall is $n^{1-1/k} \times nf(n) = O(n^{2-1/k} \log n)$. Thus, the total time for calculating conditional expectations is $O(n^{4-2/k} \log n)$, and this is precisely the asymptotic bound on the total running time. □

4.2. Space. A component of the algorithm is the following name-dependent routing algorithm.

THEOREM 4.2 (see Thorup and Zwick [21]). *Given an integer $k \geq 2$, there exist name-dependent routing schemes which use $O(n^{1/k} \log^{1-1/k} n) \times o(\log^2 n) = \tilde{O}(n^{1/k})$ space per node, $o(\log^2 n)$-sized headers, and which deliver messages with stretch $2k - 1$.*

We note that this is the version of their algorithm which requires handshaking, but our scheme stores the precomputed handshaking information with the destination address. Let $TZR(u, v)$ denote the address required for routing from u to v (that is, the final $o(\log^2 n)$-bit header Thorup and Zwick determine from u and v after executing the handshaking protocol), and let $TZTab(u)$ denote the routing table their algorithm stores at node u.

Let $\{S_u | u \in V\}$ be a collection of sets of blocks that satisfies Lemma 4.1. For each node u, let $S'_u = S_u \bigcup \{B_\beta\}$, where $u \in B_\beta$ (that is, each node always stores the block to which its own address belongs). Each node u stores the following:

1. $TZTab(u)$.
2. For every $v \in N^1(u)$, the pair (v, e_{uv}), where e_{uv} is the first edge on a shortest path from u to v.
3. The set S'_u of $O(\log n)$ blocks B_α, and for each block $B_\alpha \in S'_u$, the following:
   (a) For every $0 \leq i < k - 1$, and for every $\tau \in \Sigma$, let v be the nearest node containing a block B_β such that $\sigma^i(B_\beta) = \sigma^i(B_\alpha)$ and the $(i+1)$st symbol of $\sigma^{k-1}(B_\beta)$ is τ. If $i = 0$, we store the node name v; otherwise we store the routing address $TZR(u, v)$.
   (b) Corresponding to $i = k-1$, for every $\tau \in \Sigma$, we store the routing address $TZR(u, v)$, where $\langle v \rangle = \alpha\tau$. Note that, consistently with 3(a), the node

v satisfies $\sigma^{k-1}(B_\alpha) = \alpha = \sigma^{k-1}(\langle v \rangle)$ and the kth symbol of $\sigma^k(\langle v \rangle)$ is τ.

LEMMA 4.3. *The space requirement of our algorithm is $o(kn^{1/k} \log^3 n)$ bits, which is simply $\tilde{O}(n^{1/k})$ bits for fixed constant k.*

Proof. By Theorem 4.2, we need $o(n^{1/k} \log^{3-1/k} n)$ space per node, for (1). Since $|N^1(u)| = n^{1/k}$ for all u, it is clear that (2) uses $O(n^{1/k} \log n)$ space. For (3) we note that $|S'_u| = O(\log n)$ blocks. For each block, we store $kn^{1/k}$ values $TZR(u, v)$, where the size of $TZR(u, v)$ in bits is $o(\log^2 n)$. Therefore the space requirement for (3) is $o(kn^{1/k} \log^3 n)$, and this dominates the other two terms. ∎

4.3. Routing algorithm. We denote by $Hop(u, v)$ the Thorup–Zwick route from a node u that stores the routing information $TZR(u, v)$ to the node v. For source node s and destination node t, our algorithm routes a packet through a sequence of nodes $s = v_0, v_1, \ldots, v_k = t$. For any two successive nodes v_i and v_{i+1} in this sequence that are distinct (except for v_0 and v_1), the transition between them is made through the path $Hop(v_i, v_{i+1})$. The sequence $s = v_0, v_1, \ldots, v_k = t$ has the property that each v_i (except v_k) contains a block B_{β_i} for which $\sigma^i(B_{\beta_i}) = \sigma^i(\langle t \rangle)$. The case when $v_i = v_{i+1}$ occurs when node v_i coincidentally contains a block that matches the destination in at least $i + 1$ digits.

Figure 4 diagrams an example sequence of nodes v_i. The following is the pseudocode for the algorithm.

ALGORITHM 4.4.

if *($t \in N^1(s)$):*
    route to t using shortest path pointers e_{ut}
else*:*
    $i \leftarrow 0$
    while *($i \neq k$):*
        $\tau \leftarrow \sigma^{i+1}(\langle t \rangle)$
        if *($i+1 < k$):*
            $v_{i+1} \leftarrow$ *closest $v \in N^{i+1}(v_i)$ such that $\exists B_\beta \in S_v, \sigma^{i+1}(B_\beta) = \tau$*
        else*:*
            $v_k \leftarrow t$
        if *($v_i \neq v_{i+1}$):*
            if *($i = 0$):*
                route to v_1 by shortest path pointers e_{uv_1}
            else: *($i \geq 1$)*
                route to v_{i+1} along $Hop(v_i, v_{i+1})$ using $TZR(v_i, v_{i+1})$
        $i \leftarrow i + 1$

LEMMA 4.5. *Algorithm 4.4 always delivers a given packet successfully from a source node s to a destination t.*

Proof. At each v_i we have sufficient routing information to route to node v_{i+1}, and delivery to node v_{i+1} is guaranteed by the Thorup–Zwick algorithm. The algorithm terminates on finding t, because in the worst case we have stored information for routing to a node v in $N^k(v_{k-1}) = V$ such that $\sigma^k(\langle v \rangle) = \sigma^k(\langle t \rangle)$, and the latter condition implies $v = t$. ∎

We note that the idea of matching increasing prefixes of node names appears in the parallel algorithms literature for multidimensional array routing (see [17]); it has also been cleverly used in recent schemes proposed in the context of locating replicated objects in peer-to-peer systems [23, 16, 20, 14].

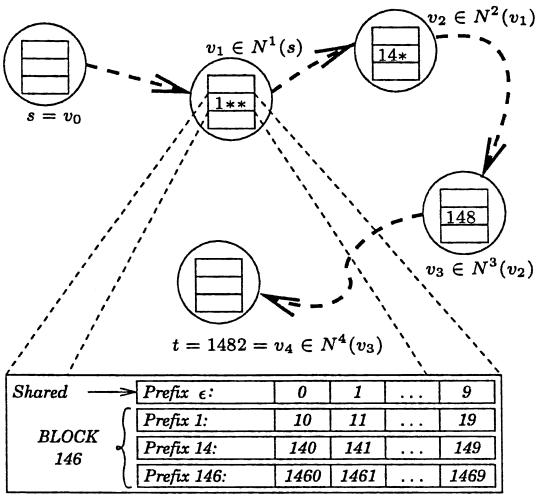


FIG. 4. *A schematic of how the prefix-matching algorithm of Theorem* 4.8 *works. The figure includes only the sequence of nodes where the distributed dictionary is read; the other nodes in the path are not shown. For illustration purposes each node contains only* 3 *blocks, and the contents of each block are illustrated in the magnified table. Asterisks stand for arbitrary digits in block labels. Notice that the blocks that are actually consulted (shown labeled) have prefixes that increasingly match the destination* 1482.

4.4. Stretch analysis. In this section we complete the analysis of Algorithm 4.4 by analyzing the stretch.

LEMMA 4.6. *For* $0 \le i \le k-1$, $d(v_i, v_{i+1}) \le 2^i d(s,t)$.

Proof. Recall that v_i is the first node that is found to match the ith prefix of the destination t by the routing algorithm, as defined above. For each $0 \le i \le k$, let v_i^* be the closest node to node s such that $\sigma^i(\langle v_i^* \rangle) = \sigma^i(\langle t \rangle)$. The proof is by induction.

For the basis case, we note that based on the algorithm $d(s,v_1) = d(v_0,v_1) \le 2^0 d(s,t)$, since t itself is a candidate to be v_1. If $d(s,t) < d(s,v_1)$, then t would have been chosen to be node v_1, because t contains a block B_β such that $\sigma^1(B_\beta) = \sigma^1(\langle t \rangle)$.

The inductive hypothesis is that for all i such that $0 \le i \le r-1 < k-1$, we have $d(v_i, v_{i+1}) \le 2^i d(s,t)$. We bound $d(v_r, v_{r+1})$ as follows:

$$\begin{aligned}
d(v_r, v_{r+1}) &\le d(v_r, v_{r+1}^*) && (1)\\
&\le d(v_r, s) + d(s, v_{r+1}^*) && (2)\\
&\le d(s,t) + d(v_r, s) && (3)\\
&\le d(s,t) + d(s, v_r) && (4)\\
&\le d(s,t) + \sum_{i=0}^{r-1} d(v_i, v_{i+1}) && (5)\\
&\le d(s,t)\left[1 + \sum_{i=0}^{r-1} 2^i\right] && (6)\\
&\le 2^r d(s,t).
\end{aligned}$$

(1) follows by definition of v_{r+1} and v_{r+1}^*, and (2) follows since $d(v_r, v_{r+1}^*)$ is a shortest distance. We obtain (3) by commutativity, and since t is a candidate to be the node v_{r+1}^*. By symmetry we get (4), and (5) follows since $d(s,v_r)$ is a shortest distance. Finally (6) is obtained by applying the inductive hypothesis, and the result follows. ∎

In this context let $p'(s,t)$ be the path obtained by routing from s to t, using a *shortest* path between each pair of distinct v_i and v_{i+1}.

COROLLARY 4.7. *For all s,t, $p'(s,t) \leq (2^k-1)d(s,t)$.*

Proof. $p'(s,t) = \sum_{i=0}^{k-1} d(v_i, v_{i+1}) \leq \sum_{i=0}^{k-1} 2^i d(s,t) \leq (2^k-1)d(s,t)$. ∎

THEOREM 4.8. *For fixed constant $k \geq 2$, Algorithm 4.4 uses space $\tilde{O}(n^{1/k})$, and delivers packets correctly with stretch $1+(2k-1)(2^k-2)$ using packet headers of size $o(\log^2 n)$. The routing tables can be computed in polynomial time.*

Proof. The space bound and termination are established in Lemmas 4.3 and 4.5, respectively.

While routing from $s = v_0$ to v_1, we do not use the name-dependent algorithm, since we have shortest path pointers within each ball of size $n^{1/k}$, so the stretch for that segment is 1. The stretch for the remaining segments, based on the previous corollary, is (2^k-2) times the stretch factor of $2k-1$ from the Thorup–Zwick name-dependent scheme. ∎

We note that for the special case when $k = 2$, our earlier specialized algorithm (Scheme A) with a stretch of 5 is better than the generalized algorithm of this section, which has stretch 7 when $k = 2$.

5. A generalized routing scheme with a polynomial tradeoff. In this section we present a universal name-independent compact routing scheme that, for every $k \geq 2$, uses space $\tilde{O}(k^2 n^{\frac{2}{k}} \log n)$ and achieves a stretch of $16k^2 - 8k$, with $O(\log^2 n)$-bit headers, on any undirected graph with edge weights whose size is polynomial in n. The scheme is very similar to Awerbuch and Peleg's scheme [3]. Like [3], we use an underlying topology-dependent routing scheme with low stretch and build on top of that a dictionary to retrieve topology-dependent information. Our dictionary is based on the prefix matching idea of section 4.

5.1. Preliminaries. Given an undirected network $G = (V, E)$ with n nodes and polynomial-sized edge weights, we define $\hat{N}^m(v)$ as the set of nodes in V that are within distance m from $v \in V$; $Diam(G)$ is the ratio of the maximum distance between any pair of nodes in G to the minimum distance in G; $Rad(v,G)$ is the ratio of the maximum distance between any node in G and v to the minimum distance in G; $Rad(G)$ is $\min\{Rad(v,G) | v \in V\}$; and $Center(G)$ is any vertex $v \in V$ such that $Rad(v,G) = Rad(G)$.

A *cluster* C is a subset of the nodes in the graph, and a *cover* is a collection of clusters $\mathcal{C} = \{C_i\}_i$ covering all the vertices of G, that is, such that $\bigcup_i C_i = V$. We extend our definition of $Diam()$, $Rad()$, and $Center()$ to clusters C by considering the subgraph induced by the vertices in C. Finally, these definitions are extended to covers $\mathcal{C}$ by taking the maximum over the values of every cluster in the cover, e.g., $Rad(\mathcal{C}) = \max\{Rad(C) | C \in \mathcal{C}\}$.

Let C be a connected set of vertices, and $v = Center(C)$ its center. We define $Tree(C)$ as the shortest path tree rooted at v that spans all the vertices in C. Define $Height(T)$ where T is a tree as the maximum distance from the root of T to any vertex in T. Notice that by construction, we have $Height(Tree(C)) = Rad(C)$. We use the following result.

THEOREM 5.1 (see Awerbuch and Peleg [3]). *Given an integer $k > 1$, a weighted graph $G = (V, E)$ with $|V| = n$, and a distance r such that $1 \leq r \leq Diam(G)$, it is possible to construct a tree cover $\mathcal{T}$ satisfying the following:*

1. *For every node $v \in V$, there is a tree $T \in \mathcal{T}$ spanning all the vertices in $\hat{N}^r(v)$.*

2. *For every tree* $T \in \mathcal{T}$, *Height*$(T) \leq (2k-1)r$.
3. *For any* $v \in V$, v *appears in at most* $2kn^{1/k}$ *trees.*

We use the same hierarchy of covers as in [3]. For every $i = 1, \ldots, \lceil \log(Diam(G)) \rceil$, we apply Theorem 5.1 with $r = 2^i$ and construct a tree cover $\mathcal{T}_i$ such that (1) for every $v \in V$, there exists a tree in the cover that includes $\hat{N}^{2^i}(v)$; (2) the height of such a tree is at most $(2k-1)2^i$; and (3) every vertex appears in no more than $2kn^{\frac{1}{k}}$ trees. For every cluster C_i we define a tree T_i on the nodes of C_i. Then at every level $i = 1, \ldots, \lceil \log(Diam(G)) \rceil$, every node v in the network chooses a tree T_i that contains $\hat{N}^{2^i}(v)$. Following the terminology of [3], we refer to that tree as v's home tree at level i. Notice that the existence of such a tree is guaranteed by the construction.

We use a name-dependent tree-routing scheme that satisfies Lemma 2.2 to route within trees in the covers. Let $Tab(T, x)$ denote the routing table for x in the shortest path tree T, and let $R(T, x)$ denote x's topology-dependent address for that tree.

5.2. Space. Let Σ and the set of functions σ be defined as in section 4. For every level $i = 1, \ldots, \lceil \log(Diam(G)) \rceil$, every vertex u stores the following:

1. An identifier for u's home tree at level i.
2. For every tree T_i in the ith level tree cover that vertex u is in, u stores the following:
   (a) $Tab(T_i, u)$.
   (b) For every $\tau \in \Sigma$ (notice there are $n^{\frac{1}{k}}$ choices) and for every $j \in \{0, \ldots, k-1\}$ (k choices), $R(T_i, v)$, where $v \in C_i$ is the nearest node such that $\sigma^j(\langle u \rangle) = \sigma^j(\langle v \rangle)$ and the $(j+1)$st symbol of $\langle v \rangle$ is τ if such a node v exists. It also stores the root of T_i.

LEMMA 5.2. *The space requirement of our scheme is* $O(k^2 n^{2/k} \log^2 n \log Diam(G))$.

Proof. Notice first that $O(\log n)$ bits are sufficient to identify a tree in a given level since there are at most $2kn^{1+\frac{1}{k}}$ such trees. Fix some level i of the hierarchy. The space that level i imposes on any node u in the graph is

$$\underbrace{O(\log n)}_{(1)} + \underbrace{2kn^{1/k}}_{(2)}\Big(\underbrace{O(\log n)}_{(3)} + \underbrace{n^{1/k} k O(\log^2 n)}_{(4)}\Big) = O(k^2 n^{2/k} \log^2 n),$$

where (1) is the length of u's home tree identifier, (2) accounts for the number of trees u appears in, (3) is the space needed for the routing table of u within a tree, and (4) is the space required to specify the prefix table. The total space requirement for any node in the graph is therefore $O(k^2 n^{\frac{2}{k}} \log^2 n \log(Diam(G)))$. ∎

5.3. Routing algorithm. To route from u to v we do the following. For increasing values of $i = 1$ up to $\lceil \log(Diam(G)) \rceil$, u attempts to route to v in its home tree T_i at level i, until the destination is reached. Notice that success is guaranteed because in level $i = \lceil \log(Diam(G)) \rceil$ trees span the entire graph.

To route a message from u to v within cluster T_i we go through a series of nodes in T_i (this is illustrated in Figure 5). The message always carries the tree-routing label of the origin u and an identifier of the current tree T_i. From any intermediate node, say w, in this series (u is the first such node), it is routed to a node in T_i that matches the largest prefix of the name of the destination v. If no such node exists in T_i, then the message is returned to u by using the tree-routing label of u (this is when failure to deliver is detected). Otherwise, the message reaches the destination after at most k such trips. Notice that while node w might appear in different clusters,

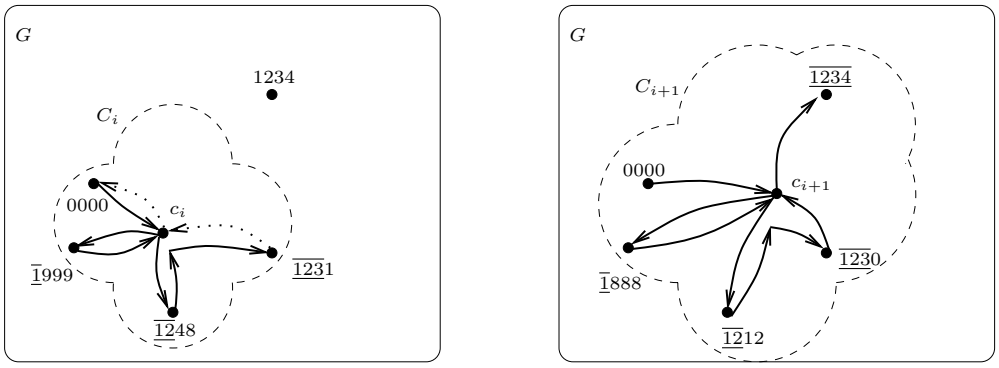


FIG. 5. *An illustration of the routing scheme with polynomial tradeoff. The left figure illustrates how a routing is attempted from node* 0000 *to node* 1234 *at level* i. *On each successive step we route to another node in the current home cluster* C_i, *which matches the destination in one more digit—note that the prefix which matches the destination is marked. Each of these steps may or may not go through the center of tree* c_i. *In this case, since* 1234 *is not in the home tree* C_i *of* 0000, *this process eventually fails, and the packet is returned to the source (shown as the dotted line). At the next level, a larger home tree* C_{i+1} *is used, and this time delivery is successful.*

we retrieve the information corresponding to the appropriate tree T_i; we can do this because an identifier for the current cluster T_i is included in the message.

5.4. Stretch analysis. Let the distance between u and v be d. There exists a level $i \leq \log(2d)$ such that u's home tree T_i contains v. When routing within tree T_i, there are at most k nodes visited, and the distance between nodes is no more than $2Height(T_i)$. The total distance traveled within T_i is at most

$$\begin{aligned} k2Height(T_i) &\leq k2(2k-i)2^i \quad \text{(by Theorem 5.1)} \\ &\leq k2(2k-1)2d \quad \text{(since } i \leq \log(2d)\text{)} \\ &= (8k^2-4k)d. \end{aligned}$$

The total distance traveled in the whole process is at most twice the distance in the last level visited, i.e., $(16k^2-8k)d$. The stretch is therefore $16k^2-8k$. Thus we have the following.

THEOREM 5.3. *For every* $k \geq 2$, *there is a universal name-independent compact routing scheme that uses* $O(k^2 n^{\frac{2}{k}} \log^2 n \log(Diam(G)))$ *space,* $O(\log^2 n)$*-bit headers, and achieves stretch* $16k^2-8k$, *where* D *is the normalized diameter of the network.* □

6. A remark on node names. We have thus far assumed that the node names form an arbitrary permutation of $\{0,\ldots,n-1\}$. We argue here that this assumption can be made without loss of generality. Suppose we have a set of n nodes, each having a unique name from an arbitrary universe $\mathcal{U}$. We use a hash function h that maps $\mathcal{U}$ to the set $\{0,\ldots,p-1\}$, where $p \geq n$ is a prime. The hash function is chosen such that (1) it can be computed fast; (2) it requires small space; and (3) the probability of collision is small. A natural candidate for this hash function is from the class proposed by Carter and Wegman. We draw a polynomial H from a class of integer polynomials $\mathcal{H}$ of degree $O(\log n)$. For any node u, we rename u to $\text{name}(u) = H(\text{int}(u)) \bmod p$,

where $\text{int}(u)$ is the integer representation in $\mathbf{Z}_p$. The following lemma, which directly follows from [8], guarantees low collision probabilities.

LEMMA 6.1 (Carter and Wegman [8]). *Let $m \in [p]$ be an integer. For every collection of ℓ nodes u_1 through u_ℓ, we have*

$$\mathbf{Pr}\left[name(u_1) = name(u_2) = \cdots = name(u_\ell) = m\right] \leq \left(\frac{2}{p}\right)^{\ell}.$$

By setting $p = \Theta(n)$, we ensure that the number of bits in the new name is $\log n + O(1)$, and that the probability of $\Omega(\log n)$ nodes mapping to the same name is inverse-polynomial. Furthermore, the representation of the hash function H only requires storing $O(\log n)$ words of $O(\log n)$ bits each at a node, amounting to $O(\log^2 n)$ bits at each node.

We now describe how the routing schemes proposed in the paper can be modified to handle two consequences of the above hashing mechanism: (1) the node names are chosen from $[0, \Theta(n))$ rather than $[0, n)$; and (2) there may be $\Theta(\log n)$ collisions for a given name. We first note that all of our routing schemes easily adapt to the case where the unique names are drawn from the integers in the interval $[0, \Theta(n))$. In the new schemes, there will be no routing-table entries containing names from the interval $[0, \Theta(n))$ that do not exist. The adapted schemes yield the same respective stretch factors at the cost of a constant-factor increase in space.

In order to accommodate the event that the hashed names of two nodes are the same, a small modification to the routing tables suffices. Suppose, in our original scheme with unique names, the table at a node contains an entry for a node with name X, satisfying a property (e.g., lying within a certain neighborhood). In the new scheme, the table will contain an entry for *all* nodes with hashed name X that satisfy the property; the entries may be distinguished by also storing the original names of the nodes, or by comparing the result of applying another hash function (or a series of hash functions, for increasing confidence) to the node names. Specifically, for Schemes A, B, and C, the primary change will be that a block may have more than $\sqrt{n}$ nodes (but $O(\sqrt{n})$ with high probability), thus increasing the space at each node by at most a constant factor with high probability. For the schemes of sections 4 and 5, the primary change will be owing to the following: for a node u and a given k-bit sequence μ, there may be multiple nodes whose prefix matches that of u in all but the last k bits and has μ in the last k bits. Thus in step 3(b) of section 4.2 and in step 2(b) of section 5.2, the modified scheme will store the hashed names and the original name (for resolving conflicts) of all such nodes, rather than the unique node under the earlier uniqueness assumption. Note that the increase in size of the namespace and the collisions result in the increase in space of $O(\log n)$ per node with high probability, while maintaining the same stretch factor.

7. Concluding remarks. In this paper, we have developed low-stretch schemes that decouple node names from network topology. An important next step is to study this problem on fully dynamic networks, where routing tables must be updated online as nodes and edges arrive and depart from the network.

REFERENCES

[1] I. ABRAHAM, C. GAVOILLE, AND D. MALKHI, *Routing with improved communication-space trade-off*, in Proceedings of the 18th International Symposium on Distributed Computing

(DISC), R. Guerraoui, ed., Lecture Notes in Comput. Sci. 3274, Springer, Berlin, 2004, pp. 305–319.

[2] M. Arias, L. Cowen, K. Laing, R. Rajaraman, and O. Taka, *Compact routing with name independence*, in Proceedings of the 15th Annual ACM Symposium on Parallelism in Algorithms and Architectures, ACM, New York, 2003, pp. 184–192.

[3] B. Awerbuch and D. Peleg, *Sparse partitions*, in Proceedings of the 31st Annual IEEE Symposium on Foundations of Computer Science, 1990, pp. 503–513.

[4] B. Awerbuch and D. Peleg, *Routing with polynomial communication-space trade-off*, SIAM J. Discrete Math., 5 (1992), pp. 151–162.

[5] B. Awerbuch, A. Bar-Noy, N. Linial, and D. Peleg, *Compact distributed data structures for adaptive network routing*, in Proceedings of the 21st ACM Symposium on Theory of Computing, ACM, New York, 1989, pp. 479–489.

[6] B. Awerbuch, A. Bar-Noy, N. Linial, and D. Peleg, *Improved routing strategies with succinct tables*, J. Algorithms, 11 (1990), pp. 307–341.

[7] I. Abraham, C. Gavoille, D. Malkhi, N. Nisan, and M. Thorup, *Compact name-independent routing with minimum stretch*, in Proceedings of the 16th Annual ACM Symposium on Parallelism in Algorithms and Architectures (SPAA 2004), ACM, New York, 2004, pp. 20–24.

[8] J. L. Carter and M. N. Wegman, *Universal classes of hash functions*, J. Comput. System Sci., 18 (1979), pp. 143–154.

[9] L. Cowen, *Compact routing with minimum stretch*, J. Algorithms, 38 (2001), pp. 170–183.

[10] D. Dor, S. Halperin, and U. Zwick, *All pairs almost shortest paths*, in Proceedings of the 37th Annual IEEE Symposium on Foundations of Computer Science, 1996, pp. 452–461.

[11] T. Eilam, C. Gavoille, and D. Peleg, *Compact routing schemes with low stretch factor*, in Proceedings of the 17th Annual ACM Symposium on Principles of Distributed Computing (PODC), ACM, New York, 1998, pp. 11–20.

[12] P. Fraigniaud and C. Gavoille, *Routing in trees*, in 28th International Colloquium on Automata, Languages and Programming (ICALP), F. Orejas, P. G. Spirakis, and J. van Leeuwen, eds., Lecture Notes in Comput. Sci. 2076, Springer, Berlin, 2001, pp. 757–772.

[13] C. Gavoille and M. Gengler, *Space-efficiency of routing schemes of stretch factor three*, J. Parallel Distrib. Comput., 61 (2001), pp. 679–687.

[14] K. Hildrum, J. Kubiatowicz, S. Rao, and B. Zhao, *Distributed object location in a dynamic network*, in Proceedings of the 14th Annual ACM Symposium on Parallel Algorithms and Architectures, ACM, New York, 2002, pp. 41–52.

[15] D. Krioukov, K. Fall, and X. Yang, *Compact routing on Internet-like graphs*, in Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (Infocom 2004), 2004, pp. 209–219.

[16] J. Kubiatowicz, D. Bindel, Y. Chen, S. Czerwinski, P. Eaton, D. Geels, H. Weatherspoon, R. Gummadi, S. Rhea, W. Weimer, C. Wells, and B. Zhao, *Oceanstore: An architecture for global-scale persistent storage*, SIGPLAN Not., 35 (2000), pp. 190–201.

[17] T. Leighton, *Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes*, Morgan Kaufmann, San Mateo, CA, 1992.

[18] L. Lovász, *On the ratio of optimal integral and fractional covers*, Discrete Math., 13 (1975), pp. 383–390.

[19] D. Peleg, *Distance-dependent distributed directories*, Inform. Comput., 103 (1993), pp. 270–298.

[20] C. G. Plaxton, R. Rajaraman, and A. W. Richa, *Accessing nearby copies of replicated objects in a distributed environment*, Theory Comput. Syst., 32 (1999), pp. 241–180.

[21] M. Thorup and U. Zwick, *Compact routing schemes*, in Proceedings of the 13th Annual ACM Symposium on Parallel Algorithms and Architectures, ACM, New York, 2001, pp. 1–10.

[22] M. Thorup and U. Zwick, *Approximate distance oracles*, in Proceedings of the 33rd Annual ACM Symposium on Theory of Computing, ACM, New York, 2001, pp. 183–192.

[23] M. van Steen, F. J. Hauck, and A. S. Tanenbaum, *A model for worldwide tracking of distributed objects*, in Proceedings of the 1996 Conference on Telecommunications Information Networking Architecture (TINA 96), 1996, pp. 203–212.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 727–740

RECONSTRUCTING CHAIN FUNCTIONS IN GENETIC NETWORKS*

IRIT GAT-VIKS†, RICHARD M. KARP‡, RON SHAMIR†, AND RODED SHARAN†

Abstract. The following problems arise in the analysis of biological networks: We have a boolean function of n variables, each of which has some default value. An *experiment* fixes the values of any subset of the variables, the remaining variables assume their default values, and the function value is the result of the experiment. How many experiments are needed to determine (reconstruct) the function? How many experiments that involve fixing at most q values are needed? What are the answers to these questions when an unknown subset of the variables are actually involved in the function? In the biological context, the variables are genes and the values are gene expression intensities. An experiment measures the gene levels under conditions that perturb the values of a subset of the genes. The goal is to reconstruct the particular logic (regulation function) by which a subset of the genes together regulate one target gene, using few experiments that involve minor perturbations. We study these questions under the assumption that all functions belong to a biologically motivated set of so-called chain functions. We give optimal reconstruction schemes for several scenarios and show their application in reconstructing the regulation of galactose utilization in yeast.

Key words. network reconstruction, experimental design

AMS subject classifications. 90B10, 62K99, 06E30

DOI. 10.1137/S089548010444376X

1. Introduction.

In this paper we study the problem of function reconstruction. We have a set of N boolean variables. Each variable has a default value, and an *experiment* can change (fix to 0 or 1) its value. The *order* of an experiment is the number of variables fixed during the experiment. The value of one variable of interest (the output) is determined by a boolean function of n other variables. The *output* of an experiment is the value of the function, where all fixed variables attain their respective values and the rest attain their default values. The problem of *function reconstruction* is to determine this function using a minimum number of experiments of the smallest possible order.

The motivation to studying the problem arises in molecular biology: The regulation of biological entities is key to cellular function. The genes are expressed (transcribed) into mRNAs, which are translated into proteins. The regulatory factors which control (regulate) gene expression are themselves protein products of other genes. The result is a complex network of regulatory relations among genes. A *genetic network* consists of a set of variables that correspond to *genes*, attaining real values, called *states*. The state of a gene indicates the discretized expression level of the gene. A gene may be *regulated* by several other genes, implying that its state

*Received by the editors May 16, 2004; accepted for publication (in revised form) January 24, 2006; published electronically October 4, 2006. The work of the first author was supported by a Colton fellowship. The second and third authors were supported by a grant from the US-Israel Binational Science Foundation (BSF). The fourth author was supported by a Fulbright grant and by NSF ITR grant CCR-0121555. A preliminary version of this paper appeared in *Proceedings of the Ninth Pacific Symposium on Biocomputing* [10].

http://www.siam.org/journals/sidma/20-3/44376.html

†School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel (iritg@tau.ac.il, rshamir@tau.ac.il, roded@tau.ac.il).

‡International Computer Science Institute, 1947 Center St., Berkeley, CA 94704 (karp@icsi.berkeley.edu).

is a function of the states of its regulating genes, or its *regulators*. An experiment involves *perturbations* such as knocking out certain genes (fixing their states to some low value) or overexpressing them (fixing their states to some high value) and measuring the expression levels of all other genes. The measurement of gene expression levels is facilitated by high throughput technologies, such as DNA microarrays (e.g., [6]). The order of an experiment is the number of genes that are perturbed. In order to reconstruct the regulatory relations among genes, we need to infer the set of genes that cooperate in the regulation of a given gene and the particular logical function by which this regulation is determined. This paper studies the number and order of experiments that are needed in order to infer the regulatory function that governs a specific gene.

A key obstacle in the inference of regulation relations is the large number of possible solutions and, consequently, the unrealistically large amount of data needed to identify the right one. A common and simple model for genetic networks is the *boolean* model, in which the state of a gene is 0 (off) or 1 (on). The boolean assumption is a drastic simplification of real biology, yet it captures important features of biological systems and was frequently used in previous studies [16].

There is a large body of previous work on learning boolean functions from a random sample of their output values (see [3] for a review). Those studies focus on devising efficient probably approximately correct (PAC) learning algorithms for subclasses of boolean functions using a polynomial-size sample. Another body of work is devoted to exact learning of certain classes of boolean functions using a polynomial number of queries (see, e.g., [4] and references thereof). For the specific problem of exact boolean function reconstruction in a genetic network, Akutsu et al. [1] have shown that the number of experiments (or queries) that are needed for reconstructing a function of N genes is prohibitive: The lower and upper bounds on the number of experiments of order $N-1$ that are needed are $\Omega(2^{N-1})$ and $O(N \cdot 2^{N-1})$, respectively. When the function involves only d regulators, the number of required experiments of order d is still $\Omega(N^d)$ and $O(N^{2d})$, respectively [1].

The inherent complexity of this problem led researchers to seek ways around this problem. Ideker, Thorson, and Karp [16] studied how to dynamically design experiments so as to maximize the amount of information extracted. Friedman et al. [8] used Bayesian networks to reveal parts of the genetic network that are strongly supported by the data. Tanay and Shamir [24] suggested a method of expanding a known network core using expression data. Several studies used prior knowledge about the network structure, or restrictive models of the structure, in order to identify relevant processes in gene expression data [12, 15, 23, 22].

Recently, a biologically motivated, boolean model of regulation relations based on *chain functions* was suggested in order to cope with the problem of function reconstruction in biological context [9]. In a chain function, the state of the regulated gene depends on the influence of its direct regulator, whose activity may in turn depend on the influence of another regulator, and so on, in a chain of dependencies (we defer formal definitions to the next section). The class of chain functions has several important advantages [9]: These functions reflect common biological regulation behavior, so many real biological regulatory relations can be elucidated using them (examples include the SOS response mechanism in *E. coli* [21] and galactose utilization in yeast [18]). Moreover, by restricting consideration to chain functions, the number of candidate functions drops from double exponential to single exponential only.

In this paper we study several computational problems arising when wishing to

reconstruct chain functions using a minimum number of experiments of the smallest possible order. We address both the question of finding the set of regulators of a chain function, which is typically much smaller than the entire set of genes, and the question of reconstructing the function given its regulators. We give optimal reconstruction schemes for several scenarios and show their application on real data. Our analysis focuses on the theoretical complexity of reconstructing regulation relations (number and order of experiments), assuming that experiments provide accurate results and that the target function can be studied in isolation from the rest of the genetic network.

The paper is organized as follows: Section 2 contains basic definitions related to chain functions. In section 3 we give worst-case and average-case analyses of the number of experiments needed in order to reconstruct a chain function. Both low-order and high-order experimental settings are considered. In section 4 we study the reconstruction of composite regulation functions that combine several chains. Finally, in section 5 we describe a biological application of our analysis to reconstruct the regulation mechanism of galactose utilization in yeast.

2. Chain functions. Chain functions were introduced by Gat-Viks and Shamir [9]. In the following we define these functions and describe their main properties. Our presentation differs from the original one to allow succinct description of the reconstruction schemes in later sections.

Variables, regulators and states. Let U denote the set of all variables in a network, where $|U| = N + 1$. These variables correspond to genes, mRNAs, proteins, or metabolites. Each variable may attain one of two *states*: 1 or 0. The state of gene g, denoted by $state(g)$, indicates the discretized expression level of the gene. A variable normally attains its *wild-type* state, but perturbations such as gene knockouts may change its state. We say that a variable $g_0 \in U$ is *regulated by* a set $S = \{g_1, \ldots, g_n\} \subset U$ if $state(g_0) = f^{g_0}(state(g_n), \ldots, state(g_1))$ and S is a minimal set with that property. In that case we say that S is the *regulator set* of g_0, and g_0 is called the *regulatee*. Associated with each regulator g_i is a binary constant y_i which dictates the *control* property of g_i. If $y_i = 0$ then g_i is an *activator*; otherwise g_i is a *repressor*. This is an intrinsic property of the regulator and is not subject to change. The *control pattern* of f^{g_0} is the binary vector $(y_n, \ldots, y_1)$.

Given a certain order $g_n, \ldots, g_1$ of the regulators, we call g_i a *predecessor* of g_j for $i > j$ and a *successor* of g_k for $i < k$. We also say that g_i is to the *left* of g_j and to the *right* of g_k. Each regulator transmits a signal to its immediate successor, and this chain of events enables a signal to propagate from g_n to g_0 in a manner defined by a chain function (see Figure 1, top part).

Chain function definition. The chain function model assumes that the functional relations are deterministic. The chain function f^{g_0} on the regulators $g_n, \ldots, g_1$ determines the state of the regulatee g_0.

The function f^{g_0} can be defined using two n-long boolean vectors attributing *activity* and *influence* to each g_i. Let $a(g_i)$ denote the activity of g_i, and let $infl(g_i)$ denote the influence signal from g_i to g_{i-1}. The definitions of activity and influence on the other regulators are recursive: The influence on g_n is always 1. g_i is active ($a(g_i) = 1$) iff it exists ($state(g_i) = 1$) and it receives a positive influence from its predecessor ($infl(g_{i+1}) = 1$). The influence $infl(g_i)$ transmitted from g_i to g_{i-1} is a xor ($\oplus$) of $a(g_i)$ and y_i: $infl(g_i)$ is 1 if g_i is an activator and is itself activated or if g_i is a repressor and is not activated (so that it fails to repress g_{i-1}). Formally,

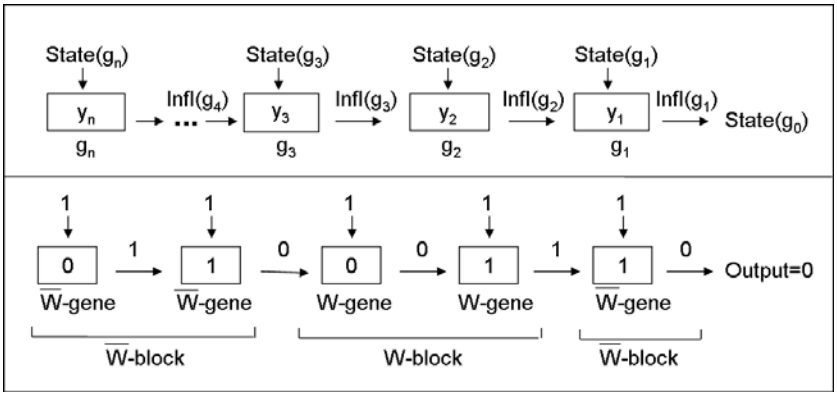


FIG. 1. *The chain function model. Top: A chain function model. Bottom: An illustration of a chain function with five regulators.* g_1, g_2, g_4 *are repressors, and* g_3, g_5 *are activators. The state of all regulators is* 1. *Influences are indicated on the horizontal arrows. Regulator types and blocks are indicated below.*

$$a(g_i) \;=\; infl(g_{i+1}) \;\wedge\; state(g_i), \tag{1}$$

$$infl(g_i) \;=\; y_i \oplus a(g_i). \tag{2}$$

Finally, the state of the regulatee g_0 is simply the influence of g_1. We define the *output* of f^{g_0} to be $state(g_0)$.

A chain function is uniquely determined by its set of regulators, their order, and the control pattern. For example, if g_0 is regulated by (g_3, g_2, g_1) via a chain function with control pattern 010, then $f(1,1,1) = 0$ and $f(0,1,1) = 1$.

3. Reconstruction of chain functions. In this section we study the question of uniquely determining the chain function which operates on a known regulatee, using a minimum number of experiments. We assume throughout that all variable states in wild type are known. We further assume that all regulator states in wild type are 1, except possibly g_n. The latter assumption is motivated by the observation that in many biological examples, all regulators are expressed in wild type, and the state of the regulatee is determined by the presence or absence of a metabolite g_n. (Examples include the Trp, lac, and araBAD operons in *E. coli* [21], the regulation of galactose utilization [18] in yeast, and human MAPK cascades [17]).

An *experiment* is defined by a set of variables that are externally perturbed (knocked-out or overexpressed). The states of the perturbed variables are thus fixed, and the states of all nonperturbed regulators are assumed to remain at the wild-type values. The state of the regulatee is determined by the chain function. The *order* of an experiment is the number of externally perturbed variables in it.

Our reconstruction algorithms are based on performing various experiments and observing their effect on the state of the regulatee. The algorithms implicitly assume that the regulation function is indeed a chain function and do not explicitly test this property.

We now devise a simple set of equations that characterize the output of a chain function as a function of the control pattern and the states of the regulators, both

in the wild-type state and in states produced by perturbing some regulators. These equations are the foundation of all the subsequent reconstruction schemes:

PROPOSITION 1. *Let f be a chain function on $g_n, \ldots, g_1$. If $state(g_i) = 1$ for $1 \le i < n$, then $state(g_0) = state(g_n) \oplus (\oplus_{i=1}^{n} y_i)$. For any other state vector, if the least index of a state-0 regulator is $j \le n$, then $f^{g_0}(g_n, \ldots, g_1) = \oplus_{i=1}^{j} y_i$.*

Proof. By definition, $a(g_n) = state(g_n)$. For $i < n$, $state(g_i) = 1$ implies that $a(g_i) = a(g_{i+1}) \oplus y_{i+1}$. It follows by induction that $state(g_0) = state(g_n) \oplus (\oplus_{i=1}^{n} y_i)$. Similarly, if $state(g_j) = 0$ and $state(g_i) = 1$ for all $i < j$, it follows by induction that $f^{g_0}(g_n, \ldots, g_1) = \oplus_{i=1}^{j} y_i$. ∎

Under the above assumptions on regulator states, a chain function can be viewed as a series of inversion and identity gates, whose input is the state of g_n. Each identity gate corresponds to an activator, whose output is equal to its input. Each inversion gate corresponds to a repressor, whose output is opposite to its input. The output of the last gate in the chain is the state of the regulatee.

3.1. Types and blocks. A *perturbation* is an experiment that changes the state of a variable to the opposite of its state in wild type. By our assumption on the regulator states in wild type (all regulator states in wild type are 1, except possibly g_n), the perturbation of a regulator in $\{g_{n-1}, \ldots, g_1\}$ is a knockout. For $S \subseteq U$, an *S-perturbation* is an experiment in which the states of all the variables in S are perturbed.

Let w be $state(g_0)$ in wild type. Let $\bar{w}$ be the opposite state. For the reconstruction, we first classify the variables in U into two *types*: W and $\bar{W}$ (see Figure 1, bottom part). A variable is in W ($\bar{W}$) if its perturbation produces output w ($\bar{w}$). Typically, the majority of the genes have type W, since in particular all the genes that are not part of the chain function are such. By Proposition 1 we have $g_n \in \bar{W}$, and $g_{n-1} \in W$ iff $state(g_n) \oplus y_n = 0$. We call a gene that belongs to W ($\bar{W}$) a *W-gene* (*$\bar{W}$-gene*). Similarly, we call a regulator of type W ($\bar{W}$) a *W-regulator* (*$\bar{W}$-regulator*). For a given gene, we call a successor of type W ($\bar{W}$) of that gene a *W-successor* (*$\bar{W}$-successor*).

The type of a gene can be determined by a single perturbation of the gene. Such an experiment will be referred to as a *typing experiment* throughout.

COROLLARY 2. *Given an ordered set of regulators $g_n, \ldots, g_1$, their control pattern can be reconstructed using $n - 1$ typing experiments.*

Proof. Perform typing experiments for $g_1, \ldots, g_{n-1}$ (by definition $g_n \in \bar{W}$). By Proposition 1, for every $1 < i < n$, $y_i = 1$ iff the types of g_i and g_{i-1} differ. Also, $y_n = 1$ iff either $state(g_n) = 0$ and the types of g_n, g_{n-1} are equal, or $state(g_n) = 1$ and the two types differ. Finally, we can use Proposition 1 to deduce y_1. ∎

Any control pattern $(y_n, \ldots, y_1)$ may be separated into *blocks* of consecutive regulators by truncating the control pattern after each 1. The first block (rightmost, ending at g_1) has two possible forms: $0 \ldots 0$ or $0 \ldots 01$. All other blocks are of the form $0 \ldots 01$, so the right boundary of a block corresponds to a regulator g_j with $y_j = 1$, and any other regulator g_i in the block has $y_i = 0$.

LEMMA 3. *Each block contains regulators of a single type, and two adjacent blocks contain regulators of opposite types.*

The proof follows from the fact that the type of g_i, $i < n$ differs from the type of g_{i-1} iff $y_i = 1$. Thus, we can refer to a block as either a *W-block* or a *$\bar{W}$-block*, and the two types of blocks alternate. For convenience, we shall refer to g_n as forming a $\bar{W}$-block of its own.

3.2. Reconstructing the regulator set and the function. Consider a chain function with control pattern $(y_n, \ldots, y_1)$ and let $g_j, \ldots, g_i$ be a block. Then $infl(g_i) = [infl(g_{j+1}) \wedge (\bigwedge_{h=i}^{j} state(g_h))] \oplus y_i$. Thus, the effect of the block on the function is determined by the boolean variable $infl(g_{j+1})$, by the control pattern, and by the conjunction of the states of its regulators. Since this conjunction is independent of the order of occurrence of these genes, no experiment based on perturbing the states of the genes can determine the order of the genes within the block. In view of this limitation, we shall aim to find the equivalence class of chain functions as detectable by perturbation experiments, i.e., our goal is to reconstruct the control pattern, the set of genes within each block (but not their order), and the ordering of the blocks. Correspondingly, in the following we will use the term *successor* of a gene to denote a regulator that succeeds that gene in the chain and is not a member of its block. For convenience, we shall refer to gene (in fact, W-genes) that are not regulators of g_0 as predecessors of g_n.

The above discussion implies that once we have typed each gene, it remains to determine, for each pair consisting of a W-gene and a $\bar{W}$-gene, which one precedes the other in the chain. Let k_W and $k_{\bar{W}}$ denote the number of regulators of type W and $\bar{W}$, respectively. Note that $k_W + k_{\bar{W}} = n \leq N$, and in fact, typically, $n \ll N$ as $k_W \ll |W|$.

Suppose we perform a $\{i, k\}$-perturbation with $g_i \in W$ and $g_k \in \bar{W}$. If the result is w, then g_k precedes g_i. Otherwise, g_i precedes g_k. A 2-order experiment for determining the relative order of a W-gene and a $\bar{W}$-gene will be called a *comparison* throughout.

PROPOSITION 4. *Given the set of regulators of a chain function and their types, $k_W k_{\bar{W}}$ comparisons are necessary and sufficient to reconstruct the function.*

Proof. The upper bound follows by comparing every W-regulator with every $\bar{W}$-regulator. The lower bound follows from the fact that, in the special case where every $\bar{W}$-regulator precedes every W-regulator, no set of comparisons can determine the relative order of a given pair consisting of a W-regulator and a $\bar{W}$-regulator, unless it includes a direct comparison between the pair. Therefore, all such comparisons must be performed. □

Note that the problem of reconstructing a chain function by comparisons, once the regulators have been typed, can be viewed as a sorting problem: The input is a list of n elements of two types, such that the set of elements of each type consists of several equivalence classes, and there is a linear order of all these classes. The objective is to find the equivalence classes and their order, using only queries that compare two elements of distinct types. In the special case that each equivalence class consists of one element, the problem is related[1] to the well-studied problem of matching nuts and bolts [2] and has an optimal $\Theta(n \log n)$ deterministic solution [19].

We now turn to the question of reconstructing a chain function without prior knowledge of the identity of its regulators. The discussion above suggests a way to solve the problem: First, we find the gene types using N typing experiments. Next, we reconstruct the block structure by performing all possible comparisons between a W-gene and a $\bar{W}$-gene.

A more efficient reconstruction is possible when g_n is known. This is often the case when the chain function models a signal transduction pathway, where g_n represents

[1] The difference between the problem of matching nuts and bolts and our problem is that in our case we have strict linear order among all the elements and there is no notion of matching between W-regulators and $\bar{W}$-regulators.

a known stimulator of the corresponding biological response. If g_n is known, then since $g_n \in \bar{W}$, all W-regulators can be identified by comparing every W-gene with g_n, using a total of $N - k_{\bar{W}}$ comparisons. Since every $\bar{W}$-gene is a regulator, these experiments are sufficient to identify all the regulators, and we can apply Proposition 4 to complete the reconstruction in $N - k_{\bar{W}} + k_W(k_{\bar{W}} - 1)$ comparisons. In summary, we have the following proposition.

PROPOSITION 5. *A chain function can be reconstructed using at most N typing experiments and $k_{\bar{W}}(N - k_{\bar{W}})$ comparisons. Given g_n, a chain function can be reconstructed using at most $N - 1$ typing experiments and $N - n + k_W k_{\bar{W}}$ comparisons.*

We can prove a matching lower bound by generalizing the argument in Proposition 4.

PROPOSITION 6. *At least $k_{\bar{W}}(N - k_{\bar{W}})$ comparisons are necessary to reconstruct a chain function.*

Proof. Consider the case where all $\bar{W}$-regulators precede the W-regulators. In this case, no set of comparisons can determine the relative order of a given pair consisting of a W-gene and a $\bar{W}$-gene unless it includes a direct comparison between the pair. Therefore, all such comparisons must be performed. □

Propositions 4 and 5 provide a worst-case analysis. Next, we describe another reconstruction algorithm, whose *expected* number of required experiments is lower. The analysis of the running time is similar to that of quick-sort (cf. [5]) and assumes that the chain to be reconstructed has $\bar{W}$-blocks of bounded size. Denote by D_g the set of W-successors of $g \in \bar{W}$ in f.

PROPOSITION 7. *A chain function with $\bar{W}$-blocks of size bounded by d can be reconstructed using N typing experiments and an expected number of $O(Nd \log k_{\bar{W}} + k_W k_{\bar{W}})$ comparisons.*

Proof.

Algorithm: First, we perform N typing experiments. Next, we apply a randomized scheme to reconstruct the chain: Each time we pick a gene $g \in \bar{W}$ at random, find its successors and their order, and remove g and all its successors from further consideration. We stop when no $\bar{W}$ genes are left. In order to find the successors of g, we first identify the members of D_g using at most $N - k_{\bar{W}}$ comparisons. Using D_g, we then reconstruct the part of the chain that spans g and its successors by at most $|D_g|(k_{\bar{W}} - 1)$ comparisons, as in Proposition 4.

Complexity: The set of comparisons can be divided into two parts: those that are required to identify the sets D_g and those required to reconstruct the chain parts induced by these sets. For the latter, at most $k_W k_{\bar{W}}$ comparisons are needed in total, since every pair consisting of a W-regulator and a $\bar{W}$-regulator is compared at most once. Thus, it suffices to compute the expectation of the first part. Let $T(x)$ be this expectation, given that the current $\bar{W}$ set (i.e., the set of $\bar{W}$-genes that were not removed in previous iterations) contains x elements, where $T(0) = 0$. For $x \geq 1$, with probability $\frac{1}{x}$ the qth rightmost element of $\bar{W}$ is chosen in the current iteration. Hence, $T(x) \leq \frac{1}{x} \sum_{q=1}^{x} (d(N - k_{\bar{W}}) + T(x - q))$. By induction, $T(x) \leq d(N - k_{\bar{W}})(\log x + 1)$. Substituting $x = k_{\bar{W}}$, we obtain the required bound. □

The expected number of experiments improves over the upper bound of Proposition 5 for $d < k_{\bar{W}}$, which is the case in many real biological regulations, e.g., the filamentous-invasion pathway ($n = 9$, $k_{\bar{W}} = 2$, and $d = 1$, illustrated in [11, Figure 3]), and the HOG signaling pathway ($n = 6$, $k_{\bar{W}} = 3$, and $d = 2$ [13]) in yeast.

3.3. Using high-order experiments. In this section we show how to improve the above results when using experiments of order $q > 2$. The results in this section

are mainly of theoretical interest, since high-order experiments may not be practical.

PROPOSITION 8. *Given the set of n regulators of a chain function, the function can be reconstructed using $O(\frac{n^2}{q}\log q)$ experiments of order at most $q \le n$. This is optimal up to constant factors for $q = \Theta(n)$.*

Proof. The number of possible chain functions with n regulators is $\Theta((\log_2 e)^{n+1}n!)$ [9]. Since each experiment provides one bit of information, the information lower bound is $\Omega(n \log n)$ experiments.

Suppose at first that $q = n$. Let n_i be the number of regulators in block i, where blocks are indexed in right-to-left order. Our reconstruction algorithm is as follows: First, we perform n typing experiments. Next, we identify the type of the first block using one experiment of order n, in which all regulators are perturbed (this way we perturb also the genes in the first block, and thus its type is identical to the output). We proceed to reconstruct the blocks one by one, according to their order along the chain. Note that the type of each block is now known, since the two types alternate. Suppose we have already reconstructed blocks $1, \ldots, i-1$. For reconstructing the ith block we only consider the set of regulators that do not belong to the first $i-1$ blocks. Out of this set, let A be the subset of regulators that have the same type as block i, and let B be the subset of regulators of the opposite type. In order to identify the members of the ith block we use a binary-search-like procedure: We divide A into two halves. For each half we perform a perturbation that includes that half and all regulators in B. If the result is the type of block i, we continue recursively with that half. Otherwise, we discard it. The search requires $O(n_i \log n)$ experiments. Thus, altogether we perform $O(n \log n)$ experiments.

When $q < n$, we use the above algorithm as a component in our reconstruction scheme, allowing us to reconstruct a subchain of size q within a chain of size n using $O(q \log q)$ experiments of order at most q. Our reconstruction scheme is based on Proposition 4, which shows that for reconstruction it suffices to compare every W-regulator with every $\bar{W}$-regulator. To this end we form $O(\frac{n^2}{q^2})$ regulator subsets, each of size at most q, such that every pair consisting of a W-regulator and a $\bar{W}$-regulator appears in one of the subsets. To compute these subsets we form a $k_W \times k_{\bar{W}}$ matrix, whose entries are in 1-1 correspondence with $(W, \bar{W})$-regulator pairs. We then cover this matrix using $O(\frac{k_W k_{\bar{W}}}{q^2})$ disjoint submatrices of dimension at most $\lfloor q/2 \rfloor \times \lceil q/2 \rceil$, each identifying a regulator subset of the required size.

Next, we reconstruct the subchain of size q associated with each subset using $O(q \log q)$ experiments of order at most q. After this process, each $(W, \bar{W})$-regulator pair appears in one of the subchains, and thus its relative order has been determined. This is sufficient in order to computationally reconstruct the chain (as in Proposition 4). Altogether we use $O(\frac{k_W k_{\bar{W}}}{q} \log q) = O(\frac{n^2}{q} \log q)$ experiments for reconstructing the chain from its regulators. □

We now provide a reconstruction scheme for the case that the set of regulators is not known. Let f be a chain function. For a gene $g \in \bar{W}$, denote as before by D_g its set of W-successors in f. A building block in our reconstruction scheme is a method to efficiently identify the members of D_g using $O(|D_g| \log q + N/q)$ experiments of order at most q. The process is as follows: We partition the W-genes into $\lceil \frac{N}{q-1} \rceil$ subsets of size at most $q-1$. For each subset R we test whether it contains some successor of g using an $(R \cup \{g\})$-perturbation, in which g and the subset members are perturbed. If as a result of the perturbation the output changes to w, then at least one of the members in R succeeds g. In this case we use standard binary search to identify all the m successors in R by performing additional $O(m \log q)$ experiments of order at

most $(\lfloor q/2 \rfloor + 1)$. Otherwise, all the subset members precede g and we discard R. Each of the successors of g is discovered exactly once, which gives the required bound.

PROPOSITION 9. *For $q \leq n$, a chain function can be reconstructed using $O(nN/q + n^2 \log q/q)$ experiments of order at most q. For $q > n$, $O(N + n \log q)$ experiments of order at most q are sufficient.*

Proof. The reconstruction is done in three stages. First, we perform N typing experiments. Second, we discover all W-regulators as follows: For each regulator $b \in \bar{W}$ we use the scheme described above to identify its successors in W, and remove them from further consideration. Each W-regulator is discovered exactly once and, thus, we need $O(k_{\bar{W}} N/q + k_W \log q)$ experiments of order at most q altogether. Last, we reconstruct the chain, given the regulators and their types, in $O(n^2 \log t/t)$ experiments, using the method given in Proposition 8, where $t = \min\{q, n\}$. In total $O(N + k_{\bar{W}} N/q + k_W \log q + n^2 \log t/t)$ experiments are used. □

A lower bound on the number of experiments that are required is given in the following proposition.

PROPOSITION 10. $\Omega(\max\{N/q, nN/q^2, n \log N\})$ *experiments of order at most q are necessary to reconstruct a chain function.*

Proof. We give three different lower bounds, whose union yields the required result. First, $\Omega(N/q)$ experiments are required to identify at least one $\bar{W}$-regulator. Second, $\Omega(nN/q^2)$ experiments are required to cover every pair of a W- and a $\bar{W}$-gene. Third, the number of possible chain functions is $\Theta(\binom{N}{n}(\log_2 e)^{n+1} n!)$ [9]. Hence, the information theoretic lower bound on the reconstruction is $\Omega(n \log N)$. □

Finally, we give an optimal reconstruction scheme when g_n is known and $q = \lceil N/2 \rceil + 1$.

PROPOSITION 11. *In case g_n is known, there is an optimal reconstruction scheme that uses $\Theta(n \log N)$ experiments of order at most $\lceil N/2 \rceil + 1$.*

Proof. We perform the reconstruction in two stages. In the first stage we discover the set of regulators and their types. In the second stage we apply Proposition 8 to reconstruct the chain function. To discover the set of regulators we perform a binary-search-like process as follows: We partition all variables excluding g_n and g_0 into two halves, H_1 and H_2. For $i = 1, 2$ we apply an $H_i \cup \{g_n\}$-perturbation. Since g_n is perturbed, all nonregulator effects are masked, and we get the result w iff H_i contains some W-regulators. Therefore, for each set that gives the results w, we continue recursively until we reach single genes. In this way we have identified a subset T of the W-regulators, including all those in the first (rightmost) block. We now repeat the recursive process on $U \setminus (T \cup g_n \cup g_0)$, but this time do not include g_n in the perturbations. This process identifies a subset T' of the $\bar{W}$ regulators, including the first $\bar{W}$-block. By repeating these two recursive processes (with and without including g_n in the perturbations) we eventually identify all regulators. The total effort is $O(n \log N)$ since each path that identifies one of the n regulators is a binary search in N variables and thus takes $O(\log N)$ experiments. □

4. Combining several chains. In this section we extend the notion of a chain function to cover common biological examples in which the regulatee state is a boolean function of several chains. Frequently, a combination of several signals influences the transcription of a single regulatee via several pathways that carry these signals to the nucleus, and a regulation function that combines them together. Here, we formalize this situation by modeling each signal transduction pathway by a chain function, and letting the outputs of these paths enter a boolean gate.

Define a *k-chain function* f as a boolean function which is composed of k chain

functions over disjoint sets of regulators that enter a boolean gate $G(f)$. Let f^i be the ith chain function and let g^i_j denote the jth regulator in f^i. The output of the function is $G(infl(g^1_1), \ldots, infl(g^k_1))$.

In the following we present several biological examples for k-chain functions that arise in transcriptional regulation in different organisms: The lac operon [21] codes for lactose utilization enzymes in *E. coli*. It is under both negative and positive transcriptional control. In the absence of lactose, lac-repressor protein binds to the promoter of the lac operon and inhibits transcription. In the absence of glucose, the level of cAMP in the cell rises, which leads to the activation of CAP, which in turn promotes transcription of the lac operon. In our formalism, the lac operon is controlled by a 2-chain function with an AND gate. The chains are $f^1(g^1_2, g^1_1) = f^1$(lactose, lac-repressor), with control pattern 11, and $f^2(g^2_3, g^2_2, g^2_1) = f^2$(glucose, cAMP, CAP), with control pattern 100. Other examples of 2-chains with AND gates are the regulation of arginine metabolism and galactose utilization in yeast [18]. A 2-chain with an OR gate regulates lysine biosynthesis pathway enzymes in yeast [18].

These examples motivate us to restrict attention to gates that are either OR or AND. We first show that we can distinguish between OR and AND gates. We then show how to reconstruct k-chain functions in the case of OR and later extend our method to handle AND gates.

Denote the output of f^i by O_i. If $O_i = 1$ in wild type, we call f^i a 1-*chain* and, otherwise, a 0-*chain*. A regulator g^i_j is called a 0-*regulator* (1-*regulator*) if its perturbation produces $O_i = 0$ ($O_i = 1$). Let k_0 (k_1) be the number of 0-regulators (1-regulators) in f. A block is called a 0-*block* (1-*block*), if it consists of 0-regulators (1-regulators).

LEMMA 12. *Given a k-chain function f with gate $G(f)$ which is either AND or OR, $k \geq 2$, we can determine, using $O(N^2)$ experiments of order at most 2, whether $G(f)$ is an AND gate or an OR gate.*

Proof. We perform N typing experiments. If $w = 0$ and $\bar{W} = \emptyset$, then $G(f)$ is an AND gate. If $w = 1$ and $\bar{W} = \emptyset$, then $G(f)$ is an OR gate. Otherwise, $\bar{W} \neq \emptyset$. In this situation the cases of $w = 0$ and $w = 1$ are similarly analyzed. We describe only the former.

If $w = 0$, we have to differentiate between the case of an OR gate, whose inputs are all 0-chains, and the case of an AND gate, whose inputs are one 0-chain and $(k-1)$ 1-chains. To this end we perform all comparisons of a W-gene and a $\bar{W}$-gene. Let T be the set of genes g such that the result of a $\{g, g'\}$-perturbation is w for every $g' \in \bar{W}$. Then $T \neq \emptyset$ iff $G(f)$ is an AND gate. □

We now study the reconstruction of an OR gate. Let S be the (possibly empty) set of regulators that reside in one of the first blocks (i.e., the blocks containing g^i_1), that are also 1-blocks. We observe that a perturbation of any regulator in S results in $state(g_0) = 1$ regardless of any other simultaneous perturbations we may perform. Hence, determining the specific chain to which an element from S belongs is not possible. Therefore, our reconstruction will be unique up to the ordering within blocks and up to the assignment of the regulators in S to their chains. The next lemma handles the case $w = 0$. The subsequent lemma treats the case $w = 1$.

LEMMA 13. *Given a k-chain function f with an OR gate and assuming that $w = 0$, we can reconstruct f using N typing experiments and $(N - k_1)k_1$ comparisons.*

Proof. We perform N typing experiments. Then, for each 1-regulator b, we perform all possible comparisons, thereby identifying all 0-regulators that succeed b in its chain. This completes the reconstruction. □

LEMMA 14. *Let f be a k-chain function with an OR gate. Assume that $w = 1$, and let r be the number of 1-chains entering the OR gate. Then f can be reconstructed using $O(N^r + Nn)$ experiments of order at most $r + 2$.*

Proof. First, we determine r, the minimum order of an experiment that will produce output 0 for f. For $i = 1, 2, \ldots$ we perform all possible i-order experiments; r is determined as the smallest i for which we obtain output 0. In total we perform $O(N^r)$ experiments. We call the set of perturbed genes in an r-order experiment which results in output 0, a *reset combination*.

Next, we reconstruct the 1-chains. Fix an arbitrary reset combination R. For every $a \in R$ we perform a set of experiments of order $r + 1$ as follows: For every reset combination $R' \supset R \setminus \{a\}$ with $a \notin R'$, we perturb R' and in addition each other gene, one at a time, recording those that produce output 1 as 1-regulators. For every a, the sets of 1-regulators discovered in these experiments form a linear order under set inclusion. The 1-regulators that are *not* common to all these sets are exactly the 1-regulators (that are not in S) of the chain that includes a. For each 0-regulator in $R' \setminus R$ our experiments determine the 1-regulators that succeed it in this chain. Thus, we can infer all the 1-chains. The total number of experiments performed is $O(Nk_0)$.

Finally, we reconstruct the 0-chains. To this end we perturb the 1-regulators in R, thereby deactivating the 1-chains and reducing the problem of reconstructing the 0-chains to that of reconstructing a $(k - r)$-chain function with an OR gate and $w = 0$ (removing the already discovered regulators of the 1-chains from consideration). This is done by applying the reconstruction method of Lemma 13 using $O(Nk_1)$ experiments of order at most $r + 2$. The assignment of 1-regulators in S will remain uncertain. ∎

Note that for $k = 1$ the above algorithms will reconstruct a single chain. Indeed, for $w = 0$ the algorithm of Lemma 13 coincides with that of section 3, and for $w = 1$, applying the algorithm of Lemma 14 we shall discover that $r = k = 1$. Further note that for every reconstructed chain we can identify whether its first block is a 1-block (i.e., contains genes in S). This is simply done by computing for that chain the value of $state(g_n) \oplus (\oplus_i y_i)$ on its known members and comparing it to the chain's output. Last, note that if k is known and $r = k$, then the order of the experiments that are required to reconstruct the k-chain is at most $r + 1$, since f contains no 0-chains.

The reconstruction method for the case of an OR gate can be used for the reconstruction of an AND gate as well, by exchanging the roles of 0 and 1 in the above description. This gives rise to the following result:

THEOREM 15. *A k-chain function with an OR or an AND gate can be reconstructed using $O(N^k)$ experiments of order at most $k+1$. The reconstruction requires $\Omega(\binom{N}{k}/k)$ experiments of this order.*

Proof. The upper bound follows from Lemmas 12, 13, and 14 and the duality of AND and OR gates. For the lower bound consider a k-chain function with an OR gate consisting of k 1-chains, each of which contains a single 0-regulator. Such a function has a single reset combination, which must be identified in the process of reconstructing the chain. Since each experiment of order $k + 1$ can test at most k combinations, $\Omega(\binom{N}{k}/k)$ experiments are required for the reconstruction. ∎

5. A biological application. The methods we presented above can be applied to reconstruct chain functions from biological data. We describe one such application to the reconstruction of the yeast galactose regulation function, for which some of the required perturbations have been performed. We show that one additional experiment suffices to fully reconstruct the regulation function.

The galactose utilization in the yeast *Saccharomyces cerevisiae* [18] occurs in a biochemical pathway that converts galactose into glucose-6-phosphate. The transporter gene gal2 encodes a protein that transports galactose into the cell. A group of enzymatic genes, gal1, gal7, gal10, gal5, and gal6, encode the proteins responsible for galactose conversion. The regulators gal4p, gal3p, and gal80p control the transporter, the enzymes, and to some extent each other (Xp denotes the protein product of gene X). In the following, we describe the regulatory mechanism. gal4p is a DNA binding factor that activates transcription. In the absence of galactose, gal80p binds gal4p and inhibits its activity. In the presence of galactose in the cell, gal80p binds gal3p. This association releases gal4p, promoting transcription. This mechanism can be viewed as a chain function, where $f^1(g_4^1, g_3^1, g_2^1, g_1^1) = f^1(galactose, gal3, gal80, gal4)$, and the corresponding control pattern is 0110 (see also [9]). The gal7, gal10, and gal1 regulatees are also negatively controlled by another chain $f^2(g_2^2, g_1^2) = f^2(glucose, mig1)$ with control pattern 01. The two chains are combined by an AND gate (see Figure 2(A)).

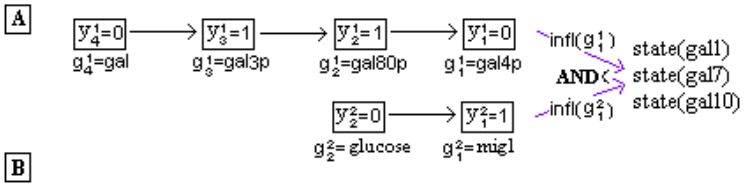
Ideker et al. [14] performed several experiments to interrogate the galactose utilization mechanism. In these experiments glucose was absent from the media. Consequently, the output of f^2 was always 1, and hence we shall focus on the reconstruction of f^1 using the experimental data of [14]. Using the discretization procedure employed by Ideker et al. [14], the measured wild-type levels of gal3, gal80, and gal4 were 1, in accordance with our model assumption. The wild-type level of galactose was also 1.

Assuming we know the group of four regulators, we need, according to Proposition 4, a total of 4 typing experiments and 3 comparisons (since only gal80 is of type W) to reconstruct the chain. Notably, all 4 typings and 2 of the 3 comparisons[2] were performed by Ideker et al. [14] (see Figure 2(B)). Using the same discretization procedure, the experiments yielded the correct results for all three regulatees. The results suggest two possible chain functions: $f^1(g_4^1, g_3^1, g_2^1, g_1^1) = f^1(galactose, gal3, gal80, gal4)$ or $f^1(g_4^1, g_3^1, g_2^1, g_1^1) = f^1(galactose, gal80, gal3, gal4)$, both with control pattern 0110. The missing experiment is a comparison of gal80 and gal3. A correct result of this experiment will lead to full and unique reconstruction of the chain function.

6. Concluding remarks. In this paper we studied the computational problems arising when wishing to reconstruct regulation relations using a minimum number of experiments, assuming that the experiments' results are noiseless. We restricted attention to common biological relations, called chain functions, and exploited their special structure in the reconstruction. We also suggested an extension of that model, which combines several chain functions, and studied some of the same reconstruction questions for the extended model. On the practical side, we have shown an application of our reconstruction scheme for inferring the regulation of galactose utilization in yeast.

The task of designing optimal experimental settings is fundamental in meeting the great challenge of regulatory network reconstruction. While this task entails coping with complex interacting regulation functions and noisy biological data, we chose here to focus on the reconstruction of a single regulation relation of a single regulatee and assume that the function can be studied in isolation. Hence, upon any perturbation, none of the other regulators change their states. Another major

[2]In fact, the gal80Δgal4Δ-gal experiment was of order 3 but allowed the comparison of gal80 and gal4.



B

Ideker et al. Experiment name	galactose	gal3	gal80	gal4	regulatee	
wt-gal	0	1	1	1	$\bar{w}$	Typing experiments
gal3Δ+gal	1	0	1	1	$\bar{w}$	
gal80Δ+gal	1	1	0	1	w	
gal4Δ+gal	1	1	1	0	$\bar{w}$	
gal80Δ-gal	0	1	0	1	w	Comparison experiments
gal80Δgal4Δ-gal	0	1	0	0	$\bar{w}$	
not perfomed	1	0	0	1	?	

FIG. 2. *Galactose pathway regulation.* (A) *The* 2*-chain function regulating gal*1*, gal*7*, and gal*10 *transcription.* (B) *Typing and comparison experiments performed by Ideker et al.* [14].

assumption is that the wild-type state of all regulators (except possibly g_n) is 1. This assumption, which is necessary for the analysis (e.g., Lemma 3) is commonly held in undelayed biological systems, where all the regulators exist in a certain basal level and the signal can propagate fast (e.g., MAPK systems in unicellular organisms such as yeast and multicellular organisms including humans, reviewed in [17]). Regulations that involve production of absent regulators are typically (slow) temporal processes. Our analysis should be extended in order to deal with such complex regulations and temporal processes.

This analysis focuses on theoretical complexity of regulation reconstruction, assuming perturbation experiments that measure (accurately) only gene states. It is clear, however, that other experimental techniques (e.g., interaction measurements [7, 20]) might help to constrain the reconstruction and reduce the solution space. In a practical approach, diverse data sources should be incorporated, and the experiments should be designed dynamically and take into consideration the experimental noise. The theoretical analysis here could hopefully serve as a component in such a practical experimental design.

REFERENCES

[1] T. AKUTSU, S. KUHARA, O. MARUYAMA, AND S. MIYANO, *Identification of genetic networks by strategic gene disruptions and gene overexpressions under a boolean model*, Theoret. Comput. Sci., 298 (2003), pp. 235–251.

[2] N. ALON, M. BLUM, A. FIAT, S. KANNAN, M. NAOR, AND R. OSTROVSKY, *Matching nuts and bolts*, in Proceedings of the 5th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 1994, pp. 690–696.

[3] M. ANTHONY, *The Sample Complexity and Computational Complexity of Boolean Function Learning*, Tech. Report LSE-CDAM-2002-13, London School of Economics and Political Science, London, UK, 2002.

[4] N. H. BSHOUTY, *Exact learning Boolean function via the monotone theory*, Inform. and Comput., 123 (1995), pp. 146–153.

[5] T. H. CORMEN, C. E. LEISERSON, AND R. L. RIVEST, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 1990.

[6] J. Derisi, V. Iyer, and P. Brown, *Exploring the metabolic and genetic control of gene expression on a genomic scale.*, Science, 282 (1997), pp. 699–705.
[7] A. H. Tong et al., *Global mapping of the yeast genetic interaction network*, Science, 303 (2004), pp. 808–13.
[8] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, *Using Bayesian networks to analyze expression data*, J. Comp. Biol., 7 (2000), pp. 601–620.
[9] I. Gat-Viks and R. Shamir, *Chain functions and scoring functions in genetic networks*, Bioinformatics, 19, Supplement 1 (2003), pp. 108–117.
[10] I. Gat-Viks, R. Shamir, R. M. Karp, and R. Sharan, *Reconstructing chain functions in genetic networks*, in Proceedings of the Ninth Pacific Symposium on Biocomputing (PSB'04), 2004.
[11] M. C. Gustin, J. Albertyn, M. Alexander, and K. Davenport, *Map kinase pathways in the yeast Saccharomyces cerevisiae*, Microbiol. Mol. Biol. Rev., 62 (1998), pp. 1264–1300.
[12] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer, *Co-clustering of biological networks and gene expression data*, Bioinformatics, 18, Supplement 1 (2002), pp. 145–154.
[13] S. Hohmann, *Osmotic stress signaling and osmoadaptation in yeasts.*, Microbiol. Mol. Biol. Rev., 66 (2002), pp. 300–372.
[14] T. Ideker et al., *Integrated genomic and proteomic analyses of systematically perturbed metabolic network*, Science, 292 (2001), pp. 929–933.
[15] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, *Discovering regulatory and signaling circuits in molecular interaction networks.*, Bioinformatics, 18, Supplement 1 (2002), pp. 233–240.
[16] T. Ideker, V. Thorsson, and R. M. Karp, *Discovery of regulatory interaction through perturbation: Inference and experimental design*, in Proceedings of Pacific Symposioum in Biocomputing, 2000, pp. 305–316.
[17] G. L. Johnson and R. Lapadat, *Motigen-activated protein kinase pathways mediated by ERK, JNK, and p*38 *protein kinases*, Science, 298 (2002), pp. 1911–12.
[18] E. W. Jones, J. R. Pringle, and J. R. Broach, eds., *The Molecular and Cellular Biology of the Yeast Saccharomyces: Gene Expression*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992.
[19] J. Komlós, Y. Ma, and E. Szemerédi, *Matching nuts and bolts in* $o(n \log n)$ *time*, SIAM J. Discrete Math., 11 (1998), pp. 347–372.
[20] T. I. Lee et al., *Transcriptional regulatory networks in Saccharomyces Cerevisiae*, Science, 298 (2002), pp. 799–804.
[21] F. C. Neidhardt, ed., *Escherichia coli and Salmonella: Cellular and Molecular Biology*, ASM Press, 1996.
[22] D. Pe'er, A. Regev, and A. Tanay, *Minreg: Inferring an active regulator set*, Bioinformatics, 18, Supplement 1 (2002), pp. 258–267.
[23] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller, *Rich probabilistic models for gene expression*, Bioinformatics, 17, Supplement 1 (2001), pp. 243–252.
[24] A. Tanay and R. Shamir, *Computational expansion of genetic networks*, Bioinformatics, 17, Supplement 1 (2001), pp. 270–278.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 741–747

ON THE STRONG CHROMATIC NUMBER OF GRAPHS*

MARIA AXENOVICH† AND RYAN MARTIN†

Abstract. The strong chromatic number, $\chi_S(G)$, of an n-vertex graph G is the smallest number k such that after adding $k\lceil n/k\rceil - n$ isolated vertices to G and considering *any* partition of the vertices of the resulting graph into disjoint subsets $V_1, \ldots, V_{\lceil n/k\rceil}$ of size k each, one can find a proper k-vertex-coloring of the graph such that each part V_i, $i = 1, \ldots, \lceil n/k\rceil$, contains exactly one vertex of each color. For any graph G with maximum degree Δ, it is easy to see that $\chi_S(G) \geq \Delta + 1$. Recently, Haxell proved that $\chi_S(G) \leq 3\Delta - 1$. In this paper, we improve this bound for graphs with large maximum degree. We show that $\chi_S(G) \leq 2\Delta$ if $\Delta \geq n/6$ and prove that this bound is sharp.

Key words. strong chromatic number, triangle factors, transversals

AMS subject classifications. 05C35, 05C15

DOI. 10.1137/050633056

1. Introduction.

An n-vertex graph G is *strongly r-colorable* if after adding $r\lceil n/r\rceil - n$ isolated vertices to G and considering *any* partition of the vertices of the resulting graph into disjoint subsets $V_1, \ldots, V_{\lceil n/r\rceil}$ of size r each, one can find a proper r-vertex-coloring of the graph such that each part V_i, $i = 1, \ldots, \lceil n/r\rceil$, contains exactly one vertex of each color. In [5], it was shown that if a graph G is strongly r-colorable, then it is strongly $(r+1)$-colorable.

The *strong chromatic number* of G, denoted $\chi_S(G)$, is the smallest positive integer k such that G is strongly k-colorable.

The famous "cycle plus triangles" problem of Erdős [4], asking whether $\chi_S(C_{3m}) = 3$, was answered affirmatively by Fleischner and Stiebitz [7], [8]; see also [15]. In general, Alon [1] proved that for any graph G with maximum degree Δ, $\chi_S(G) \leq c\Delta$, where c is a very large constant (as the author remarks, c could be reduced to 10^8). Recently, Haxell [12] improved the bound by Alon drastically, proving that $\chi_S(G) \leq 3\Delta - 1$ for any graph G with maximum degree Δ.

As far as the lower bound is concerned, it is easy to see that the strong chromatic number of a graph with maximum degree Δ is at least $\Delta + 1$ by taking one of the V_i's to be the neighborhood of a vertex of maximum degree.

Let

$$f(\Delta, n) = \max\{\chi_S(G) : G \text{ has maximum degree } \Delta \text{ and order } n\}.$$

Therefore, the best known general bounds are

$$\Delta + 1 \leq f(\Delta, n) \leq 3\Delta - 1$$

for any Δ and any $n \geq \Delta + 1$.

The following theorem is our main result which gives an exact value for $f(\Delta, n)$ when $\Delta \geq n/6$. It also provides a minimum degree condition for the existence of a K_3-factor in tripartite graphs.

*Received by the editors June 3, 2005; accepted for publication (in revised form) April 17, 2006; published electronically October 4, 2006.

http://www.siam.org/journals/sidma/20-3/63305.html

†Department of Mathematics, Iowa State University, Ames, IA 50011 (axenovic@math.iastate.edu, rymartin@iastate.edu). The research of the second author was partially supported by NSA grant H98230-05-1-0257.

THEOREM 1.1. *Let G be a graph on n vertices with maximum degree Δ, $\Delta \geq n/6$. Then $\chi_S(G) \leq 2\Delta$. Moreover, for any positive integers Δ and n, such that $\Delta \leq n/2$ there is a graph G_0 on n vertices, maximum degree Δ and $\chi_S(G_0) \geq 2\Delta$.*

COROLLARY 1.2. *For any positive integer Δ and any n such that $n/6 \leq \Delta \leq n/2$, $f(\Delta, n) = 2\Delta$. Moreover, $f(\Delta, n) \geq 2\Delta$ when $\Delta \leq n/2$.*

2. Proof of Theorem 1.1. In [7], [8], and other sources, it was noted that for specific values of n depending on Δ, there is a graph G such that $\chi_S(G) \geq 2\Delta$. We observe here that a similar but general construction gives the same bound for arbitrary n. Let $\Delta \leq n/2$, and let G_0 be a graph formed by a disjoint union of a complete bipartite graph $K_{\Delta,\Delta}$ and $n - 2\Delta$ isolated vertices. Assume that $\chi_S(G_0) \leq 2\Delta - 1$; that is, for $r = 2\Delta - 1$, any partition of $V(G_0)$ and $r\lceil n/r \rceil - n$ isolated vertices into $t = \lceil n/r \rceil$ sets of equal sizes, $V_1, \ldots, V_t$, allows a proper r-coloring of the resulting graph such that each V_i uses all the colors. Note that $t \geq \lceil 2\Delta/(2\Delta - 1) \rceil = 2$. Now, let A, B be the partite sets of a complete bipartite subgraph of G_0 with $|A| = |B| = \Delta$, and let $A \subseteq V_1$ and $B \subseteq V_2$. Then it is easy to see that it is impossible to find the desired r-coloring.

Together with the upper bound which we prove below, we shall have that $\chi_S(G_0) = 2\Delta$ when $n/6 \leq \Delta \leq n/2$.

Now we shall prove the main statement of Theorem 1.1 by providing an upper bound on the strong chromatic number. Let G be a graph on n vertices with maximum degree $\Delta \geq n/6$.

Let $\Delta \geq n/2$. Then $2\Delta \geq n$ and we trivially have that $\chi_S(G) \leq n \leq 2\Delta$.

Let $n/4 \leq \Delta < n/2$. Thus, $n/2 \leq 2\Delta < n$ and we have to partition $V(G)$ and needed isolated vertices arbitrarily into two sets V_1 and V_2, $|V_1| = |V_2|$. Each vertex in V_1 is nonadjacent to at least $|V_2|/2$ vertices in V_2 and vice versa. Consider the bipartite complement G' of this graph. That is, the edge set of G' consists of all pairs $\{v_1, v_2\}$, $v_1 \in V_1$ and $v_2 \in V_2$ such that $\{v_1, v_2\} \notin E(G)$. We claim that for each $S \subseteq V_1$, $|N(S)| \geq |S|$. Indeed, assume that there is a set $S' \subseteq V_1$ for which $|N(S')| < |S'|$. We have then that $|S'| > |V_1|/2$; thus, for any vertex $v \in V_2 \setminus N(S')$, v is adjacent to at most $|V_1| - |S'| < |V_1|/2$ vertices, a contradiction. Applying the König–Hall theorem to G' gives a perfect matching, which provides a proper coloring of the original graph, G, with 2Δ colors, each represented exactly once in V_1 and exactly once in V_2.

Let $n/6 \leq \Delta < n/4$. As before, in order to verify that $\chi_S(G) \leq r = 2\Delta$, we need to add $r\lceil n/r \rceil - n$ isolated vertices to G and partition the resulting vertex set arbitrarily into parts V_1, V_2, V_3 of equal sizes. We shall be treating this case by analyzing and extending partial colorings.

A *partial strong coloring* of G with respect to V_1, V_2, V_3 is a proper coloring of a subset of the vertices of G such that no two colored vertices in the same part V_i, $i = 1, 2, 3$, have the same color and each color class contains exactly 3 vertices. For a set S of vertices and a vertex coloring χ, we say that S is *partially multicolored* by χ if any two vertices in S, which are colored by χ, have distinct colors. Let χ be a maximal partial strong coloring of G with respect to V_1, V_2, V_3. We will show that we can always *enlarge* such partial strong coloring; i.e., create another partial strong coloring with more colors, until we color all the vertices. For a color c, we denote the vertices of this color $\{c_1, c_2, c_3\}$, where $c_i \in V_i$ for $i = 1, 2, 3$. We fix $v_1 \in V_1$, $v_2 \in V_2$, $v_3 \in V_3$ such that none of v_1, v_2, v_3 are colored by χ. For $i = 1, 2, 3$, define the following set:

$$X_i \stackrel{\text{def}}{=} \{u \in V_i : v_i \text{ is not adjacent to a vertex of color } \chi(u)\} \cup \{u \in V_i : u \text{ is not colored by } \chi\}.$$

Observe that any colored vertex in X_i can be replaced by v_i, $i = 1, 2, 3$, to create another strong partial coloring. Note also that

$$|X_i| \geq |V_i| - \deg(v_i) + t_i \geq \Delta, \quad i = 1, 2, 3,$$

where t_i is the number of neighbors of v_i in $\{v_1, v_2, v_3\}$.

To simplify the notation, we shall assume that no color of χ is labeled by x, v, or w, we reserve x_i or w_i to denote a vertex in X_i (it might be colored or not colored), and v_i are the vertices fixed above. We shall write $z \sim y$, $z \not\sim y$ if $zy \in E(G)$, $zy \notin E(G)$, respectively. For disjoint subsets S_1, S_2 of vertices of G and a vertex z, $z \notin S_1$, we write $S_1 \sim S_2$ if each vertex in S_1 is adjacent to all vertices in S_2, $S_1 \not\sim S_2$ if there are no edges between S_1 and S_2, and $z \sim S_1$, $z \not\sim S_1$ if $\{z\} \sim S_1$, $\{z\} \not\sim S_1$, respectively.

To start the proof, we give two lemmas which allow us either to enlarge χ or to replace χ with another partial strong coloring such that some three specific vertices become uncolored and the number of colors remains the same.

LEMMA 2.1. *Let $x_i \in X_i$, $i = 1, 2, 3$. If $\{x_1, x_2, x_3\}$ is partially multicolored, then there is a strong partial coloring with as many color classes as χ and with x_i's being uncolored.*

Proof. Suppose each x_i, $i = 1, 2, 3$, is colored; i.e., $x_1 = a_1$, $x_2 = b_2$, $x_3 = c_3$ with distinct colors a, b, c. Replace color classes a, b, and c with new color classes $\{v_1, a_2, a_3\}$, $\{b_1, v_2, b_3\}$, and $\{c_1, c_2, v_3\}$; see Figure 1.

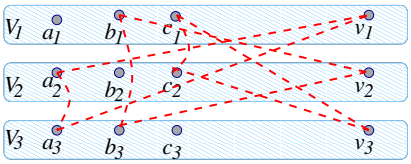


FIG. 1. *Color switches for Lemma* 2.1.

Now, suppose exactly one x_i is uncolored, without loss of generality, $x_1 = a_1$, $x_2 = b_2$, and x_3 is uncolored. Replace color classes a and b with new color classes $\{v_1, a_2, a_3\}$, $\{b_1, v_2, b_3\}$. Suppose exactly one x_i is colored, without loss of generality, $x_1 = a_1$, and x_2, x_3 are uncolored. Replace color class a with the new color class $\{v_1, a_2, a_3\}$. Each case makes $\{x_1, x_2, x_3\}$ uncolored. □

LEMMA 2.2.

(1) *If there is a set $\{x_1, x_2, x_3\}$, with $x_i \in X_i$, $i = 1, 2, 3$, which induces an independent set and is partially multicolored, then χ can be enlarged.*
(2) *If there is a set $\{x_i, x_i', x_j, x_k\}$, with $x_i, x_i' \in X_i$, $x_j \in X_j$, $x_k \in X_k$, $\{i, j, k\} = \{1, 2, 3\}$ such that $\{x_j, x_k\}$ is partially multicolored, and both $\{x_i, x_j, x_k\}$ and $\{x_i', x_j, x_k\}$ induce independent sets, then χ can be enlarged.*
(3) *Let a set $\{x_1, x_2, x_3\}$, with $x_i \in X_i$, $i = 1, 2, 3$, induce an independent set and the set $\{v_1, v_2, v_3\}$ induce neither an independent set nor a clique. Then either χ can be enlarged or one can find another partial strong coloring with as many color classes as in χ and with three uncolored vertices $x_i' \in X_i$, $i = 1, 2, 3$, that induce a K_3.*

Proof. (1) By Lemma 2.1 there is a partial strong coloring with as many color classes as in χ and such that x_1, x_2, x_3 are uncolored. We can give these vertices a new color, thus enlarging the coloring.

(2) If either $\{x_i, x_j, x_k\}$ or $\{x'_i, x_j, x_k\}$ is partially multicolored then we can use (1); otherwise assume, without loss of generality, that $i = 1$, $j = 2$, $k = 3$ and $x_1 = a_1$, $x'_1 = b_1$, $x_2 = b_2$, $x_3 = a_3$ for distinct colors a, b. Consider the following sets of vertices: $\{v_1, b_2, a_3\}$, $\{b_1, v_2, b_3\}$, and $\{a_1, a_2, v_3\}$. They are independent because of the definition of X_i's, $i = 1, 2, 3$. We can color vertices in each of these sets with the same new color, which replaces color classes a, b and saturates vertices $\{v_1, v_2, v_3\}$, thus enlarging χ.

(3) We can assume, without loss of generality, that $v_1 \sim v_2$ and $v_2 \not\sim v_3$.

Case 1. $\chi(x_1) = \chi(x_2) = \chi(x_3) = a$. Replace the color class a with two new color classes, $\{x_1, v_2, v_3\}$ and $\{v_1, x_2, x_3\}$, thus enlarging χ.

Case 2. $\chi(x_2) = \chi(x_3) = a$. If x_1 is not colored by χ, replace color class a with two new color classes: $\{a_1, v_2, v_3\}$ and $\{x_1, x_2, x_3\}$. If x_1 is colored b, replace color classes a and b with the following three new color classes: $\{a_1, v_2, v_3\}$, $\{v_1, b_2, b_3\}$, $\{x_1, x_2, x_3\}$. This enlarges χ.

Case 3. $\chi(x_1) = \chi(x_2) = a$. If $x_3 \sim \{v_1, v_2\}$, then replace x_3 with v_3 in its color class if x_3 is colored by χ. Then v_1, v_2, x_3 are three uncolored vertices inducing a clique. Let $\{x'_1, x'_2, x'_3\} = \{v_1, v_2, x_3\}$. If $x_3 \not\sim v_1$, then replace x_3 by v_3 in its color class (if x_3 is colored) and replace a color class a with two new color classes, $\{v_1, x_2, x_3\}$, $\{x_1, v_2, a_3\}$, thus enlarging χ. If $x_3 \not\sim v_2$, then replace x_1 by v_1 in its color class, replace x_3 by v_3 in its color class (if x_3 is colored), and give a new color to the independent set $\{x_1, v_2, x_3\}$, thus enlarging χ. Note that the case when $\chi(x_1) = \chi(x_3) = a$ is symmetric.

Case 4. $\{x_1, x_2, x_3\}$ is partially multicolored. This is part (1) of this lemma. ∎

Next, we consider three cases depending on how many edges the set $\{v_1, v_2, v_3\}$ induces in G. We shall greedily choose appropriate $x_i \in X_i$, $i = 1, 2, 3$, and enlarge the coloring. The proof begins with Case 1, where $\{v_1, v_2, v_3\}$ induces three edges. In this case, the coloring can be enlarged. In Case 2, $\{v_1, v_2, v_3\}$ induces two edges, without loss of generality $v_2 \not\sim v_3$, and either the coloring can be enlarged or another coloring with the same number of colors can be found so that there are three pairwise adjacent uncolored vertices reducing the analysis to Case 1. Finally, in Case 3 there is only one edge, without loss of generality v_1v_2, induced by $\{v_1, v_2, v_3\}$. In this case, either the coloring can be enlarged or we can find a coloring that puts us in Case 2 or Case 1. See Figure 2.

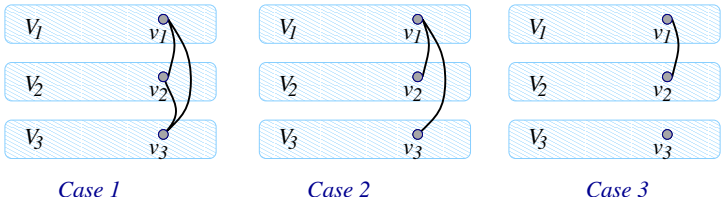


FIG. 2. *Cases* 1, 2, 3.

In Cases 1 and 2 we shall need the following parameter:

$$q \stackrel{\text{def}}{=} \max\{|N(x) \cap X_j| : \quad x \in X_i;\ i \neq j \text{ with } i, j \in \{1, 2, 3\}\}.$$

Case 1: $v_1 \sim v_2$, $v_1 \sim v_3$ *and* $v_2 \sim v_3$. We have $|X_i| \geq |V_i| - (\deg(v_i) - 2) \geq \Delta + 2$ for $i = 1, 2, 3$. Without loss of generality, assume that $q = |N(x_1) \cap X_2|$ for $x_1 \in X_1$. Let $x_2 \in X_2 \setminus N(x_1)$ be a vertex not of color $\chi(x_1)$. Consider $S = X_3 \setminus (N(x_1) \cup N(x_2))$. By the choice of x_1, $|S| \geq |X_3| - (\Delta - q) - q \geq (\Delta + 2) - \Delta = 2$; thus there are two vertices $x_3, x_3' \in X_3$ nonadjacent to both x_1 and x_2. Therefore, Lemma 2.2 (2) can be applied to the four vertices x_1, x_2, x_3, x_3' to enlarge the coloring.

Case 2: $v_1 \sim v_2$, $v_1 \sim v_3$ *and* $v_2 \not\sim v_3$. In this case, $|X_1| \geq \Delta + 2$ and $|X_2|, |X_3| \geq \Delta + 1$. Let $q = |N(x_i) \cap X_j|$, $x_i \in X_i$, and let $k \in \{1, 2, 3\} \setminus \{i, j\}$. Let $x_j \in X_j \setminus N(x_i)$, and let $x_k \in X_k \setminus (N(x_i) \cup N(x_j))$. Note that such x_j and x_k exist since $|X_j| \geq \Delta + 1$ and $|X_k \setminus (N(x_i) \cup N(x_j))| \geq \Delta + 1 - q - (\Delta - q) \geq 1$.

Therefore, we can apply Lemma 2.2 (3) to an independent set $\{x_i, x_j, x_k\}$. This either enlarges χ or reduces Case 2 to Case 1.

Case 3: $v_1 \sim v_2$, $v_1 \not\sim v_3$ *and* $v_2 \not\sim v_3$. We show that in each of the Cases 3.1–3.3 one can enlarge the coloring, either directly or by finding a coloring with the same number of colors that satisfies either the conditions of Case 2 or the conditions of Case 1. These subcases are arranged according to the presence of specific paths in $X_1 \cup X_2 \cup X_3$; see Figure 3.

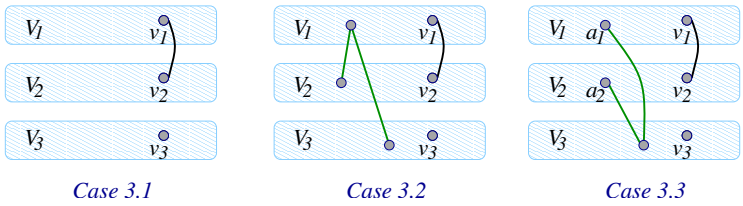


FIG. 3. *Subcases of Case* 3.

Case 3.1. There is no path with three vertices w_1, w_2, w_3; $w_i \in X_i$, $i = 1, 2, 3$.

We have that $|X_1|, |X_2| \geq \Delta + 1$ and $|X_3| \geq \Delta$. Let $G_{i,j}$ be the bipartite subgraph of G induced by the edges of G between X_i and X_j, with $i \neq j$ and $i, j \in \{1, 2, 3\}$. Note that $G_{i,j} = G_{j,i}$. Moreover, the distinct graphs $G_{i,j}$ are pairwise vertex-disjoint. If one of $G_{i,j}$ has a nonedge $x_i \not\sim x_j$, then for any $x_k \in X_k$, $k \in \{1, 2, 3\} \setminus \{x_i, x_j\}$, $\{x_1, x_2, x_3\}$ is an independent set. Thus, we can assume that each $G_{i,j}$ is a complete bipartite graph. It is easy to see that in this case there is also an independent set $\{x_1, x_2, x_3\}$, $x_i \in X_i$, $i = 1, 2, 3$. Now, we can apply Lemma 2.2 (3) to $\{x_1, x_2, x_3\}$ and either enlarge the coloring or reduce the analysis to Case 1.

Case 3.2. There is a path P with three vertices w_1, w_2, w_3, $w_i \in X_i$, $i = 1, 2, 3$, such that either the vertices of P are partially multicolored or the middle vertex of P is in $X_1 \cup X_2$.

If P is partially multicolored, we can apply Lemma 2.1 immediately to obtain a partial strong coloring with as many colors as χ and with vertices of P being uncolored. We can now choose $v_i = w_i$, $i = 1, 2, 3$, and use Case 2 or Case 1.

If P has repeated colors on its vertices, these can be only endvertices of P. Without loss of generality, let the midpoint of P be $w_1 \in X_1$, and let $a_2 = w_2$, $a_3 = w_3$ be the endpoints of P. If w_1 is not colored, replace color class a with an independent set $\{a_1, v_2, v_3\}$. If w_1 has color b, then, in addition, replace the color class b with an independent set $\{v_1, b_2, b_3\}$. This uncolors w_1, w_2, w_3 and brings us to Case 2.

Case 3.3. There is a path (w_1, w_3, w_2) with $w_i \in X_i$ for $i = 1, 2, 3$ and w_1, w_2 of the same color. Moreover, there are no paths satisfying the conditions of Case 3.2.

Note that there is no independent set $\{x_1, x_2, x_3\}$, $x_i \in X_i$; otherwise we can

either enlarge the coloring or reduce the analysis to Case 1 by Lemma 2.2 (3). Note also that if $x_1 \sim x_2$, $x_i \in X_i$, $i = 1, 2$, then $\{x_1, x_2\} \not\sim X_3$; otherwise it is Case 3.2. Therefore, we have that the bipartite subgraph of G with parts X_1, X_2 induces one nontrivial connected component F which must be a complete bipartite graph. Since $v_i \in X_i$, $i = 1, 2, 3$, and $v_1 \sim v_2$, $v_1, v_2 \in V(F)$. Let $B_1 \subseteq X_1$, $B_2 \subseteq X_2$ be the partite sets of F. Let $A_i = X_i \setminus B_i$, $i = 1, 2$. Then we have that $B_1 \sim B_2$, $A_1 \cup A_2 \sim X_3$, $A_1 \not\sim A_2$. Then, in particular, we have that $a_i \in A_i$, $i = 1, 2$, and $|A_1| = |A_2| = 1$; otherwise we shall find a path satisfying Case 3.2. Since $|X_1|, |X_2| \geq \Delta + 1$, we have that $|B_1| = |B_2| = \Delta$. Therefore, we can conclude that $|X_1| = |X_2| = \Delta + 1$ and $|X_3| = \Delta$.

Claim. The vertices v_1, v_2, v_3 are the only uncolored vertices and every color class other than a has exactly one member in $X_1 \cup X_2$.

Proof of claim. Let b be a color used by χ, $b \neq a$, not present on vertices of X_1. $N(v_2) = B_1$, so $v_2 \not\sim b_1$ and $v_2 \not\sim b_3$. This implies that $b_2 \in X_2$. Thus, any color b, $b \neq a$, is used on some vertex in $X_1 \cup X_2$.

Let t be the number of uncolored vertices in each V_i, $i = 1, 2, 3$, i.e., the number of color classes in χ is $2\Delta - t$. The fact that each color class other than a contains at least one member of $X_1 \cup X_2$ and a contains two such members gives that $|X_1| + |X_2| \geq (2\Delta - t + 1) + 2t$. Here, the expression in parenthesis gives the lower bound on number of colored vertices in X_1 and X_2 and $2t$ is the number of uncolored vertices in X_1 and X_2. Because $|X_1| + |X_2| = 2\Delta + 2$, we have that $t = 1$. As a result, every vertex other than v_1, v_2, v_3 is colored and every color class other than a contains exactly one vertex from $X_1 \cup X_2$.

By claim, there are $2\Delta - 2$ colors different from a in χ. Let ν be the number of neighbors of v_3 colored differently than a. Since $v_3 \sim \{a_1, a_2\}$, we have that $\deg(v_3) \geq \nu + 2$. For a color c, $c \neq a$, the conditions $v_3 \not\sim c_1$ and $v_3 \not\sim c_2$ imply that $c_3 \in X_3$. Using also the fact that $v_3 \in X_3$, we have that $|X_3| \geq (2\Delta - 2 - \nu) + 1$. Since $|X_3| = \Delta$, we have that $(2\Delta - 2 - \nu) + 1 \leq \Delta$, thus $\nu \geq \Delta - 1$. Therefore, $\deg(v_3) \geq \nu + 2 \geq \Delta + 1$, a contradiction.

This concludes Case 3.3, and the proof of Theorem 1.1. ∎

3. Concluding remarks. It should be noted that Theorem 1.1 is equivalent to the following.

COROLLARY 3.1. *Let G be a tripartite graph with parts of size n each. If the minimum degree of G is at least $3n/2$ then G has a K_3-factor.*

This result provides another sufficient condition for the existence of K_3-factors. For other results in this area, see, for example, [3, 11, 2, 6, 13, 14, 9]. It also came to author's attention after this paper was submitted that this problem has been considered independently in [10] by treating r-factors in multipartite graphs under maximum degree conditions.

Acknowledgments. The authors are indebted to anonymous referees for their careful work. They are especially thankful to a referee who provided Lemma 2.2 (3), which helped shorten the proofs.

REFERENCES

[1] N. ALON, *The strong chromatic number of a graph*, Random Structures Algorithms, 3 (1992), pp. 1–7.

[2] N. ALON AND R. YUSTER, *H-factors in dense graphs*, J. Combin. Theory Ser. B, 66 (1996), pp. 269–282.

[3] K. Corrádi and A. Hajnal, *On the maximal number of independent circuits in a graph*, Acta Math. Acad. Sci. Hungar., 14 (1963), pp. 423–439.
[4] P. Erdős, *On some of my favourite problems in graph theory and block designs*, Matematiche (Catania), 45 (1990), pp. 61–73.
[5] M. R. Fellows, *Transversals of vertex partitions in graphs*, SIAM J. Discrete Math., 3 (1990), pp. 206–215.
[6] E. Fischer, *Variants of the Hajnal-Szemerédi theorem*, J. Graph Theory, 31 (1999), pp. 275–282.
[7] H. Fleischner and M. Stiebitz, *A solution to a colouring problem of P. Erdős*, Discrete Math., 101 (1992), pp. 39–48.
[8] H. Fleischner and M. Stiebitz, *Some remarks on the cycle plus triangles problem*, in The Mathematics of Paul Erdős, II, Algorithms Combin. 14, Springer, Berlin, 1997, pp. 136–142.
[9] R. Johansson, *Triangle factors in a balanced blown-up triangle*, Discrete Math., 211 (2000), pp. 249–254.
[10] A. Johansson, R. Johansson, and K. Markström, *Factors of r-Partite Graphs*, personal communication.
[11] A. Hajnal and E. Szemerédi, *Proof of a conjecture of P. Erdős*, in Combinatorial Theory and Its Applications, II (Balatonfüred, 1969), North-Holland, Amsterdam, 1970, pp. 601–623.
[12] P. E. Haxell, *On the strong chromatic number*, Combin. Probab. Comput., 13 (2004), pp. 857–865.
[13] Cs. Magyar and R. Martin, *Tripartite version of the Corrádi-Hajnal theorem*, Discrete Math., 254 (2002), pp. 289–308.
[14] R. Martin and E. Szemerédi, *Quadripartite version of the Hajnal-Szemerédi theorem*, Discrete Math., to appear.
[15] H. Sachs, *Elementary proof of the cycle-plus-triangles theorem*, in Combinatorics, Paul Erdős Is Eighty, Vol. 1, Bolyai Soc. Math. Stud., János Bolyai Math. Soc., Budapest, 1993, pp. 347–359.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 748–768

APPROXIMATION ALGORITHMS FOR RECTANGLE STABBING AND INTERVAL STABBING PROBLEMS*

SOFIA KOVALEVA† AND FRITS C. R. SPIEKSMA‡

Abstract. In the weighted rectangle stabbing problem we are given a grid in $\mathbb{R}^2$ consisting of columns and rows each having a positive integral weight, and a set of closed axis-parallel rectangles each having a positive integral demand. The rectangles are placed arbitrarily in the grid with the only assumption being that each rectangle is intersected by at least one column or row. The objective is to find a minimum-weight (multi)set of columns and rows of the grid so that for each rectangle the total multiplicity of selected columns and rows stabbing it is at least its demand. A special case of this problem, called the interval stabbing problem, arises when each rectangle is intersected by exactly one row. We describe an algorithm called *STAB*, which is shown to be a constant-factor approximation algorithm for different variants of this stabbing problem.

Key words. rectangle stabbing, approximation algorithms, combinatorial optimization

AMS subject classifications. 68W25, 68R05, 90C27

DOI. 10.1137/S089548010444273X

1. Introduction.

The *weighted rectangle stabbing problem* (WRSP) can be described as follows: given are a grid in $\mathbb{R}^2$ consisting of columns and rows each having a positive integral weight, and a set of closed axis-parallel rectangles each having a positive integral demand. The rectangles are placed arbitrarily in the grid with the only assumption being that each rectangle is intersected by at least one column or row. The objective is to find a minimum-weight (multi)set of columns and rows of the grid so that for each rectangle the total multiplicity of selected columns and rows stabbing this rectangle equals at least its demand. (A column or row is said to stab a rectangle if it intersects it.)

A special case of the WRSP is the case where each rectangle is intersected by exactly one row; we will refer to the resulting problem as the *weighted interval stabbing problem* (WISP), or ISP in the case of unit weights (see Figure 1 for an example of an instance of the ISP).

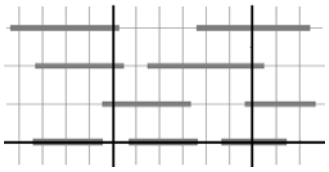


FIG. 1. *An instance of ISP with unit demands. The rectangles (or intervals in this case) are in grey; the columns and row in black constitute a feasible solution.*

*Received by the editors March 31, 2004; accepted for publication (in revised form) August 16, 2005; published electronically October 12, 2006. This work grew out of the Ph.D. thesis [7]; a preliminary version of this paper appeared in the *Proceedings of the 12th Annual European Symposium on Algorithms* [10]. This research was supported by EU-grant APPOL, IST 2001-30027.
http://www.siam.org/journals/sidma/20-3/44273.html
†Corresponding author. Department of Quantitative Economics, Maastricht University, P.O. Box 616, NL-6200 MD Maastricht, The Netherlands (sonja.kovaleva@mail.com).
‡Department of Applied Economics, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000, Leuven, Belgium (frits.spieksma@econ.kuleuven.be).

Motivation. Although at first sight the WRSP may seem rather specific, it is not difficult to see that the following two problems can be reduced to WRSP.

- *Solving special integer programming problems.* The following type of integer linear programming problem can be reformulated as instances of WRSP: $minimize\{wx|\ (B|C)x \geq b,\ x \in \mathbb{Z}^l\}$, where B and C are both 0,1-matrices with consecutive 1's in the rows (so-called interval matrices; see, e.g., Schrijver [11]), $b \in \mathbb{Z}_+^n$, $w \in \mathbb{Z}_+^l$. Indeed, construct a grid which has a column for each column in B and a row for each column in C. For each row i of matrix $B|C$, draw a rectangle i such that it intersects only the columns and rows of the grid corresponding to the positions of 1's in row i. Observe that this construction is possible since B and C have consecutive 1's in the rows. To complete the construction, assign demand b_i to each rectangle i and a corresponding weight w_j to each column and row of the grid. Let the decision variables x describe the multiplicities of the columns and rows of the grid. In this way we have obtained an instance of WRSP. In other words, integer programming problems where the columns of the constraint matrix A can be permuted such that $A = (B|C)$, with B and C each being an interval matrix, are special cases of WRSP.
- *Stabbing geometric figures in the plane.* Given a set of arbitrary connected closed geometric sets in the plane, use a minimum number of straight lines of two given directions to stab each of these sets at least once. Indeed, by introducing a new coordinate system specified by the two directions and by replacing each closed connected set by a closed rectangle defined by the projections of the set to the new coordinate axes, we obtain an instance of the problem of stabbing rectangles using a minimum number of axis-parallel lines. More specifically, we define a grid whose rows and columns are axis-parallel lines containing the rectangles' edges. We can restrict attention to those lines since any axis-parallel line stabbing some set of rectangles can be replaced by a line stabbing this set and containing a rectangle's edge. Therefore, the problem of stabbing the rectangles with axis-parallel lines reduces to the problem of stabbing them with the rows and columns of the grid.

Literature. The WRSP and its special case WISP have already received attention in the literature. Motivated by an application in parallel processing, Gaur, Ibaraki, and Krishnamurti [3] present a 2-approximation algorithm for the WRSP with unit weights and demands, which admits an easy generalization to arbitrary weights and demands. Furthermore, Hassin and Megiddo [4] (mentioning military and medical applications) study a number of special cases of the problem of stabbing geometric figures in $\mathbb{R}^2$ by a minimum number of straight lines. In particular, they present a 2-approximation algorithm for the task of stabbing connected figures of the same shape and size with horizontal and vertical lines. Moreover, they study the case of stabbing horizontal line segments of length K, whose endpoints have integral x-coordinates, with a minimum number of horizontal and vertical lines, and give a $2 - \frac{1}{K}$-approximation algorithm for this problem. In our setting this corresponds to the ISP with unit demands, where each rectangle in the input is intersected by exactly K columns. Finally, Călinescu et al. [2], mentioning applications in embedded sensor networks, show that the problem of separating n points in the plane with a minimum number of axis-parallel lines is a special case of the unweighted rectangle stabbing problem.

Concerning computational complexity, a special case of ISP where each rectangle is stabbed by at most two columns is shown to be APX-hard in [9].

Our results. We present here an approximation algorithm called *STAB* for different variants of WISP (see, e.g., Vazirani [12] for an overview on approximation algorithms). First, we show that *STAB* is a $\frac{1}{(1-(1-1/k)^k)}$-approximation algorithm for ISP_k, the variant of ISP where each row intersects at most k rectangles (e.g., the instance depicted in Figure 1 is an instance of ISP_3). Observe that *STAB* is a $\frac{4}{3}$-approximation algorithm for the case $k = 2$, and that *STAB* is an $\frac{e}{e-1}$-approximation algorithm for the case where the number of rectangles sharing a row is unlimited ($k = \infty$). Thus, *STAB* improves upon the results described in Hassin and Megiddo [4] (for $K \geq 3$) and does not impose any restrictions on the number of columns intersecting rectangles. Second, we show that *STAB* is an $\frac{e}{e-1}$-approximation algorithm for the *weighted* case of ISP_∞, i.e., the case where the columns and the rows of the grid have arbitrary positive integral weights. Third, we state here that the algorithm described by Gaur, Ibaraki, and Krishnamurti [3] can be generalized to yield a $\frac{q+1}{q}$-approximation algorithm for WRSP where the demand of each rectangle is bounded from below by an integer q. Observe that this provides a 2-approximation algorithm for the WRSP described in the introduction, where $q = 1$. Thus, this is an improvement upon the approximation ratio of the algorithm of Gaur, Ibaraki, and Krishnamurti [3] for instances with a lower bound on the rectangles' demands that is larger than 1. For the proof of this result, we refer to Kovaleva [7].

Our algorithms are based on rounding the linear programming relaxation of an integer programming formulation in an interesting way. We use the following property present in our formulation: The variables can be partitioned into two sets such that when the values of one set are fixed, one can compute the optimal values of the other variables in polynomial time, and vice versa. Next, we consider different ways of rounding one set of variables and compute each time the optimal values of the remaining variables, while keeping the best solution.

We also show that there exist instances of ISP_2 and ISP_∞ (see section 3) and WRSP (see [7]) for which the ratio between the values of a natural integer linear programming (ILP) formulation and its linear programming relaxation (LP-relaxation) is equal (or arbitrarily close) to the obtained approximation ratios. This suggests that these approximation ratios are unlikely to be improved by an LP-rounding algorithm based on the natural ILP formulation.

2. Preliminaries. Let us formalize the definition of WRSP. Let the grid in the input consist of t columns and m rows, numbered consecutively from left to right and from bottom to top, with positive weight w_c (v_r) attached to each column c (row r). Further, we are given n rectangles such that rectangle i has demand $d_i \in \mathbb{Z}_+$ and is specified by leftmost column l_i, rightmost column r_i, top row t_i, and bottom row b_i.

Let us give a natural ILP formulation of WRSP. In this paper we use notation $[a : b]$ for the set of integers $\{a, a + 1, \ldots, b\}$. The decision variables $y_c, z_r \in \mathbb{Z}_+$, $c \in [1 : t]$, $r \in [1 : m]$, denote the multiplicities of column c and row r, respectively.

$$\text{Minimize} \quad \sum_{r=1}^{m} v_r z_r + \sum_{c=1}^{t} w_c y_c \tag{1}$$

$$\text{subject to} \quad \sum_{r \in [b_i : t_i]} z_r + \sum_{c \in [l_i : r_i]} y_c \geq d_i \quad \forall i \in [1 : n], \tag{2}$$

$$z_r, y_c \in \mathbb{Z}_+^1 \qquad \forall r, c. \tag{3}$$

In a vector notation this can be represented as

$$\text{Minimize} \quad vz + wy \tag{4}$$

$$\text{subject to} \quad Bz + Cy \geq d, \tag{5}$$

$$z \in \mathbb{Z}_+^m,\ y \in \mathbb{Z}_+^t, \tag{6}$$

where $B \in \{0,1\}^{n \times m}$ and $C \in \{0,1\}^{n \times t}$ are the constraint matrices of inequalities (2). The linear programming relaxation is obtained by replacing the integrality constraints (6) by the nonnegativity constraints $z \in \mathbb{R}_+^m,\ y \in \mathbb{R}_+^t$.

For an instance $\mathcal{I}$ of WRSP and a vector $a \in \mathbb{Z}^n$, we introduce two auxiliary ILP problems:

$\text{IP}^z(\mathcal{I}, a)$: (7)

$$\begin{array}{ll} \text{Minimize} & vz \\ \text{subject to} & Bz \geq a, \\ & z \in \mathbb{Z}_+^m. \end{array}$$

$\text{IP}^y(\mathcal{I}, a)$: (8)

$$\begin{array}{ll} \text{Minimize} & wy \\ \text{subject to} & Cy \geq a, \\ & y \in \mathbb{Z}_+^t. \end{array}$$

LEMMA 2.1. *For any $a \in \mathbb{Z}^n$, the LP-relaxation of each of the problems $IP^z(\mathcal{I}, a)$ and $IP^y(\mathcal{I}, a)$ is integral.*

Proof. As was previously observed in [3], matrices B and C have a so-called consecutive 1's property. This implies that these matrices are totally unimodular (see, e.g., Schrijver [11]), which implies the lemma. □

COROLLARY 2.2. *The optimum value of $IP^z(\mathcal{I}, a)$ ($IP^y(\mathcal{I}, a)$) is smaller than or equal to the value of any feasible solution to its LP-relaxation.*

COROLLARY 2.3. *The problem $IP^z(\mathcal{I}, a)$ ($IP^y(\mathcal{I}, a)$) can be solved in polynomial time. Its optimal solution coincides with that of its LP-relaxation.*

In fact, the special structure of $\text{IP}^z(\mathcal{I}, a)$ and $\text{IP}^y(\mathcal{I}, a)$ allows us to solve it via a minimum cost flow algorithm: Let $MCF(p, q)$ denote the time needed to solve the minimum cost flow problem on a network with p nodes and q arcs. A proof of the following lemma can also be found in Veinott and Wagner [13].

LEMMA 2.4. *The problem $IP^z(\mathcal{I}, a)$ ($IP^y(\mathcal{I}, a)$) can be solved in time $O(MCF(t, n+t))$ ($O(MCF(m, n+m))$).*

Proof. Consider the LP-relaxation of formulation $\text{IP}^y(\mathcal{I}, a)$ and substitute the current variables by new variables $u_0, \ldots, u_t$ as $y_c = u_c - u_{c-1}\ \forall c \in [1:t]$. Then it transforms into

$$\begin{array}{ll} \text{Minimize} & -w_1u_0 + (w_1 - w_2)u_2 + \cdots + (w_{t-1} - w_t)u_{t-1} + w_tu_t \\ \text{subject to} & u_{r_i} - u_{l_i - 1} \geq a_i \quad \forall i \in [1:n], \\ & u_c - u_{c-1} \geq 0 \quad \forall c \in [1:t]. \end{array} \tag{9}$$

Let us denote the vector of objective coefficients, the vector of right-hand sides, and the constraint matrix by w, a, and C, respectively, and the vector of variables by u. Then (8) can be represented as $\{\text{minimize } wu |\ Cu \geq a\}$. Its dual is $\{\text{maximize } ax |\ C^Tx = w, x \geq 0\}$. Observe that this is a minimum cost flow formulation with flow conservation constraints $C^Tx = w$, since C^T has exactly one 1 and one -1 in each column. Given an optimal solution to the minimum cost flow problem, one can obtain the optimal dual solution $u_0, \ldots, u_t$ via a shortest path computation (see Ahuja, Magnanti, and Orlin [1]), and thus optimal $y_1, \ldots, y_t$ values as well. □

3. Algorithm *STAB*. Recall that the interval stabbing problem WISP refers to the restriction of WRSP, where each rectangle in the input is intersected by exactly one row. We also refer by WISP_k to WISP, where each row intersects at most k rectangles. We assume in this section that all demands are unit ($d_i = 1$, $i \in [1:n]$), thus resulting in the following formulation:

$$\text{Minimize} \quad \sum_{r=1}^{m} v_r z_r + \sum_{c=1}^{t} w_c y_c \tag{10}$$

$$\text{subject to} \quad z_{\rho_i} + \sum_{c \in [l_i : r_i]} y_c \geq 1 \quad \forall i \in [1:n], \tag{11}$$

$$z_r, y_c \in \mathbb{Z}_+^1 \qquad \forall r, c. \tag{12}$$

Here we denote by ρ_i the index of the row intersecting rectangle i.

First we describe algorithm *STAB* for WISP. In subsection 3.1 we show that it achieves a ratio of $\frac{1}{1-(1-1/k)^k}$ for the unweighted version of WISP_k: ISP_k. In subsection 3.2 we prove that *STAB* achieves a ratio of $\frac{e}{e-1}$ for WISP. Subsection 3.3 shows that the integrality gap between the values of a natural integer programming formulation of ISP_k and its LP-relaxation for $k = 2$ and $k = \infty$ coincides with the approximation ratio of the algorithm. An alternative algorithm for the case $k = 2$ yielding the same worst-case ratio (i.e., $\frac{4}{3}$) is described in Kovaleva and Spieksma [8].

Informally, algorithm *STAB* can be described as follows: Solve the LP-relaxation of (10)–(12), and denote the solution found by $(y^{\text{lp}}, z^{\text{lp}})$. Assume, without loss of generality, that the rows are sorted as $z_1^{\text{lp}} \geq z_2^{\text{lp}} \geq \cdots \geq z_m^{\text{lp}}$. At each iteration j ($j = 0, \ldots, m$) we solve the problem (10)–(12) with a fixed vector z, the first j elements of which are set to 1, and the others to 0. As shown in Lemma 2.4, this can be done in polynomial time using a minimum cost flow algorithm. Finally, we take the best of the resulting $m + 1$ solutions. A formal description of *STAB* is shown in Figure 2.

We use notation $value(y, z) \equiv \sum_{c=1}^{t} y_c + \sum_{r=1}^{m} z_r$, $value(y) \equiv \sum_{c=1}^{t} y_c$, and $value(z) \equiv \sum_{r=1}^{m} z_r$.

1. solve the LP-relaxation of (10)–(12), and obtain its optimal solution $(y^{\text{lp}}, z^{\text{lp}})$;
2. reindex the rows of the grid so that $z_1^{\text{lp}} \geq z_2^{\text{lp}} \geq \cdots \geq z_m^{\text{lp}}$;
3. $V \leftarrow \infty$;
4. for $j = 0$ to m
   for $i = 1$ to j $\bar{z}_i \leftarrow 1$,
   for $i = j + 1$ to m $\bar{z}_i \leftarrow 0$.
   solve $\text{IP}^y(\mathcal{I}, b)$, where $b_i = 1 - \bar{z}_{\rho_i}$, $\forall i \in [1:n]$, and obtain $\bar{y}$;
   if $value(\bar{y}, \bar{z}) < V$, then $V \leftarrow value(\bar{y}, \bar{z})$, $y^* \leftarrow \bar{y}$, $z^* \leftarrow \bar{z}$;
5. return (y^*, z^*).

FIG. 2. *Algorithm STAB.*

3.1. The approximation result for ISP_k. In this subsection we show that algorithm *STAB* is a $\frac{1}{1-(1-1/k)^k}$-approximation algorithm for ISP_k. Let us first adapt

the ILP formulation (10)–(12) to ISP_k with unit demands:

$$\text{Minimize} \qquad \sum_{c=1}^{t} y_c + \sum_{r=1}^{m} z_r \tag{13}$$

$$\text{subject to} \quad z_{\rho_i} + \sum_{c \in [l_i : r_i]} y_c \geq 1 \quad \forall i \in [1:n], \tag{14}$$

$$z_r, y_c \in \mathbb{Z}_+ \qquad \forall r, c. \tag{15}$$

THEOREM 3.1. *Algorithm STAB is a* $\frac{1}{1-(1-1/k)^k}$*-approximation algorithm for* ISP_k.

Proof. Consider an instance $\mathcal{I}$ of ISP_k, and let $(y^{\mathrm{lp}}, z^{\mathrm{lp}})$ and (y^*, z^*) be, respectively, an optimal LP solution and the solution returned by the algorithm for $\mathcal{I}$. We prove the theorem by establishing that

$$value(y^*, z^*) \leq \frac{1}{1-(1-1/k)^k}\, value(y^{\mathrm{lp}}, z^{\mathrm{lp}}). \tag{16}$$

It is enough to prove the result for instances satisfying the following assumption: We assume that the optimal LP solution satisfies constraints (14) at equality; i.e.,

$$z^{\mathrm{lp}}_{\rho_i} + \sum_{c \in (l_i : r_i)} y^{\mathrm{lp}}_c = 1 \quad \forall i \in [1:n]. \tag{17}$$

We now sketch why we can assume that (17) holds. Indeed, suppose that (17) does not hold for some intervals i of some instance $\mathcal{I}$. Then we modify $\mathcal{I}$ by shortening those intervals for which (17) does not hold. More precisely, by splitting the columns with y^{lp}-values we shorten the appropriate intervals so that the assumption becomes true (see Figure 3 for an example). Thus, given $\mathcal{I}$ and $(y^{\mathrm{lp}}, z^{\mathrm{lp}})$, we create an instance $\mathcal{I}'$ for which (17) holds. It is now easy to check that an optimal LP solution for $\mathcal{I}$ (with the split columns) is also an optimal LP solution for $\mathcal{I}'$. Since in $\mathcal{I}'$ the intervals have become shorter, algorithm *STAB* applied to $\mathcal{I}'$ returns a solution with a value equal to or larger than the value of the solution returned for $\mathcal{I}$. Then inequality (16) proven for $\mathcal{I}'$ implies this inequality for $\mathcal{I}$ as well.

FIG. 3. *Example of an initial instance (left) and a new instance satisfying the assumption (right).*

We order the rows of the grid in order of nonincreasing z^{lp}-values, and we denote by l ($l \geq 0$) the number of z^{lp}-values equal to 1. Then $z^{\mathrm{lp}}_1 = \cdots = z^{\mathrm{lp}}_l = 1, 1 > z^{\mathrm{lp}}_{l+1} \geq \cdots \geq z^{\mathrm{lp}}_m \geq 0$. We assume that $value(y^{\mathrm{lp}})$ is positive (otherwise all the z^{lp}-values have to be equal to 1 and the theorem obviously holds).

By construction,

$$value(y^*, z^*) = \min_{j \in [0:m]} value(y^j, z^j) \leq \min_{j \in [l:m]} value(y^j, z^j), \tag{18}$$

where (y^j, z^j) is the jth solution generated in step 4 of *STAB*.

Let us proceed by defining a number $q_j = q_j(\Delta, \beta) \in \mathbb{R}$ for each $j \in [0:m]$ that depends on a given $\Delta \in [0,1]^m$ and $\beta > 0$ as follows:

$$\sum_{k=1}^{\lfloor q_j \rfloor} (1 - \Delta_{j+k}) + (q_j - \lfloor q_j \rfloor)(1 - \Delta_{j+\lceil q_j \rceil}) = \beta, \tag{19}$$

where we put $\Delta_j = 0$ if $j > m$. Since the left-hand side is 0 at $q_j = 0$ and continuously increases to infinity as q_j grows, there always exists a unique point q_j satisfying the equality.

We will prove the following lemma.

LEMMA 3.2.

$$value(y^j, z^j) \leq j + k \cdot q_j \left(z^{\text{lp}}, \frac{value(y^{\text{lp}})}{k} \right) \quad \forall j \in [l:m].$$

Then, assuming that Lemma 3.2 holds, it follows from (18) that

$$value(y^*, z^*) \leq \min_{j \in [l:m]} \left(j + k \cdot q_j \left(z^{\text{lp}}, \frac{value(y^{\text{lp}})}{k} \right) \right). \tag{20}$$

Theorem 3.1 follows now from the following lemma, the proof of which can be found in the appendix.

LEMMA 3.3. *Given are real numbers* $1 \geq \Delta_1 \geq \Delta_2 \geq \cdots \geq \Delta_m \geq 0$, *a positive real number* Y, *an integer* $p \geq 2$, *and an integer* $l \geq 0$. *Then the following holds:*

$$\min_{i \in [l:m]} (i + p \cdot q_i(\Delta, Y/p)) \leq \frac{1}{1 - (1 - 1/p)^p} \left(Y + \sum_{r=l+1}^{m} \Delta_r \right) + l. \tag{21}$$

By applying this lemma with $p = k$, $\Delta = z^{\text{lp}}$, and $Y = value(y^{\text{lp}})$, the right-hand side of (20) can be bounded by

$$\frac{1}{1-(1-1/k)^k} \left(value(y^{\text{lp}}) + \sum_{r=l+1}^{m} z_r^{\text{lp}} \right) + l \leq \frac{1}{1-(1-1/k)^k} \left(value(y^{\text{lp}}) + \sum_{r=l+1}^{m} z_r^{\text{lp}} + l \right),$$

and since $z_1^{\text{lp}} = \cdots = z_l^{\text{lp}} = 1$, the right-hand side of this last expression is equal to

$$\frac{1}{1 - (1 - 1/k)^k} value(y^{\text{lp}}, z^{\text{lp}}).$$

The theorem is then proved.

Proof of Lemma 3.2. Consider (y^j, z^j); for some $j \in [l:m]$, let us find an upper bound for $value(y^j, z^j)$. By construction,

- $z_r^j = 1 \ \forall r \leq j$,
- $z_r^j = 0 \ \forall r \geq j+1$,
- y^j is an optimal solution to $\text{IP}^y(\mathcal{I}, b)$, where $b_i = 1 - z_{\rho_i}^j \ \forall i \in [1:n]$.

Obviously, $value(z^j) = j$. In order to bound $value(y^j)$ we introduce a solution y'^j, which is feasible to the LP-relaxation of $\text{IP}^y(\mathcal{I}, b)$. Then, Corollary 2.2 implies that $value(y^j) \leq value(y'^j)$.

First, let us define subsets $S_1, S_2, \ldots, S_m$, where $S_r \subset [1:t]$ $\forall r = 1, \ldots, m$ (i.e., each subset consists of a set of columns of the grid), in the following way:

$$S_r = \bigcup_{i:\rho_i = r} [l_i : r_i].$$

Thus, S_r is the set of columns stabbing intervals in row r.

Fix now some $j \in [l:m]$, and construct vector y'^j as follows: For each column $c \in [1:t]$,

– if $c \in S_{j+1} \cup \cdots \cup S_m$, then denote by t the minimum index such that $c \in S_t$ and let $y'^j_c = \frac{1}{(1-z^{\mathrm{lp}}_t)} y^{\mathrm{lp}}_c$ (recall that $z^{\mathrm{lp}}_r < 1$ $\forall r \in [l+1:m]$);

– otherwise, let $y'^j_c = y^{\mathrm{lp}}_c$.

Let us now establish feasibility of y'^j with respect to the LP-relaxation of $\mathrm{IP}^y(\mathcal{I}, b)$. For any interval i we show that the following inequality holds:

$$\sum_{c \in [l_i : r_i]} y'^j_c \geq 1 - z^j_{\rho_i}. \tag{22}$$

If $\rho_i < j+1$, where ρ_i is the row number of interval i, then $z^j_{\rho_i} = 1$, and the inequality holds automatically. Consider the case $\rho_i \geq j+1$. For any $c \in S_{\rho_i}$, y'^j_c is defined as $y^{\mathrm{lp}}_c/(1 - z^{\mathrm{lp}}_t)$, where $t \leq \rho_i$. Since z^{lp}_t are nonincreasing with t, we have $y'^j_c \geq y^{\mathrm{lp}}_c/(1 - z^{\mathrm{lp}}_{\rho_i})$. Then, since $[l_i : r_i] \subseteq S_{\rho_i}$, we have $y'^j_c \geq y^{\mathrm{lp}}_c/(1 - z^{\mathrm{lp}}_{\rho_i})$ for any $c \in [l_i : r_i]$. Using this, and remembering that $(y^{\mathrm{lp}}, z^{\mathrm{lp}})$ satisfies $z^{\mathrm{lp}}_{\rho_i} + \sum_{c \in [l_i:r_i]} y^{\mathrm{lp}}_c \geq 1$, we have

$$\sum_{c \in [l_i:r_i]} y'^j_c \quad \geq \quad \frac{1}{(1 - z^{\mathrm{lp}}_{\rho_i})} \sum_{c \in [l_i:r_i]} y^{\mathrm{lp}}_c \quad \geq \quad \frac{1 - z^{\mathrm{lp}}_{\rho_i}}{1 - z^{\mathrm{lp}}_{\rho_i}} = 1.$$

Thus, we have shown that inequality (22) holds for any $i \in [1:n]$, and therefore y'^j is feasible to the LP-relaxation of $\mathrm{IP}^y(\mathcal{I}, b)$. Now Corollary 2.2 implies that

$$value(y^j) \leq value(y'^j). \tag{23}$$

In what follows we show that $value(y'^j) \leq k \cdot q_j(z^{\mathrm{lp}}, \frac{value(y^{\mathrm{lp}})}{k})$ $\forall j \in [l:m]$. By construction of y'^j, using notation $Y(S) = \sum_{c \in S} y^{\mathrm{lp}}_c$,

$$\begin{aligned} value(y'^j) = {} & \tfrac{1}{1-z^{\mathrm{lp}}_{j+1}} Y(S_{j+1}) \; + \; \tfrac{1}{1-z^{\mathrm{lp}}_{j+2}} Y(S_{j+2} \backslash S_{j+1}) \\ & + \cdots + \tfrac{1}{1-z^{\mathrm{lp}}_m} Y(S_m \backslash (S_{j+1} \cup S_{j+2} \cup \cdots \cup S_{m-1})) + Y([1:t] \backslash (S_{j+1} \cup S_{j+2} \cup \cdots \cup S_m)). \end{aligned} \tag{24}$$

Observe that for the $Y(\cdot)$-terms the following equality holds:

$$\begin{aligned} & Y(S_{j+1}) + Y(S_{j+2} \backslash S_{j+1}) + \cdots + Y(S_m \backslash (S_{j+1} \cup S_{j+2} \cup \cdots \cup S_{m-1})) \\ & \quad + Y([1:t] \backslash (S_{j+1} \cup S_{j+2} \cup \cdots \cup S_m)) = \sum_{c=1}^{t} y^{\mathrm{lp}}_c = value(y^{\mathrm{lp}}). \end{aligned} \tag{25}$$

Moreover, using the definition of S_r, our assumption (17), and the fact that there are

at most k intervals per row, we have for each $r = j+1, \ldots, m$

$$Y(S_r \backslash (S_{j+1} \cup S_{j+2} \cup \cdots \cup S_{r-1})) \leq Y(S_r) = \sum_{c \in S_r} y_c^{\text{lp}}$$

$$\leq \sum_{i:\rho_i = r} \sum_{c \in [l_i : r_i]} y_c^{\text{lp}} = \sum_{i:\rho_i = r} (1 - z_{\rho_i}^{\text{lp}}) \leq k(1 - z_r^{\text{lp}}). \tag{26}$$

Now consider the following optimization problem:

$$\max_{Y_{j+1}, Y_{j+2}, \ldots} \left(\frac{1}{1 - z_{j+1}^{\text{lp}}} Y_{j+1} + \frac{1}{1 - z_{j+2}^{\text{lp}}} Y_{j+2} + \cdots + \frac{1}{1 - z_m^{\text{lp}}} Y_m + \sum_{r=m+1}^{\infty} Y_r \right)$$

$$\text{subject to} \quad Y_{j+1} + \cdots + Y_m + \sum_{r=m+1}^{\infty} Y_r \leq value(y^{\text{lp}}), \tag{27}$$

$$0 \leq Y_r \leq k(1 - z_r^{\text{lp}}) \qquad \forall r = j+1, \ldots, m, \tag{28}$$

$$0 \leq Y_r \leq k \qquad \forall r = m+1, \ldots, \infty. \tag{29}$$

Due to (25) and (26) the following solution is feasible to this optimization problem: $Y_r = Y(S_r \backslash (S_{j+1} \cup S_{j+2} \cup \cdots \cup S_{r-1}))$ for each $r = j+1, \ldots, m$, and $\sum_{r=m+1}^{\infty} Y_r = Y((1:t) \backslash (S_{j+1} \cup S_{j+2} \cup \cdots \cup S_m))$ (distributed arbitrarily among the components of the sum while satisfying (29)). Therefore the optimum value of this optimization problem is an upper bound on the right-hand side of (24).

What does the optimum solution to this optimization problem look like? Notice that the constraint matrix of (27)–(29) is a so-called *greedy* matrix (see Hoffman, Kolen, and Sakarovitch [5]). Together with the fact that the objective coefficients are nonincreasing, a result from [5] implies that successive maximization of the variables $Y_{j+1}, Y_{j+2}, \ldots$ in this order produces an optimum solution. Thus, we obtain the following optimal solution:

$$Y_{j+1} = k(1 - z_{j+1}^{\text{lp}}),\ Y_{j+2} = k(1 - z_{j+2}^{\text{lp}}), \ldots,\ Y_{j+\lfloor q \rfloor} = k(1 - z_{j+\lfloor q \rfloor}^{\text{lp}}),$$

$$Y_{j+\lfloor q \rfloor + 1} = (q - \lfloor q \rfloor) k (1 - z_{j+\lfloor q \rfloor + 1}^{\text{lp}})$$

for some number $q \in \mathbb{R}_+$, which due to (27) has to satisfy

$$k(1 - z_{j+1}^{\text{lp}}) + k(1 - z_{j+2}^{\text{lp}}) + \cdots + k(1 - z_{j+\lfloor q \rfloor}^{\text{lp}}) + k(q - \lfloor q \rfloor)(1 - z_{j+\lfloor q \rfloor + 1}^{\text{lp}}) = value(y^{\text{lp}}),$$

where we put $z_r^{\text{lp}} = 0$ for any $r > m$. Notice that $q \equiv q_j(z^{\text{lp}}, \frac{value(y^{\text{lp}})}{k})$ (see (19)), and the optimum value of the problem (27)–(29), which bounds the right-hand side of (24) from above, is $k \cdot q_j(z^{\text{lp}}, \frac{value(y^{\text{lp}})}{k})$. This proves Lemma 3.2. □

3.2. The approximation result for WISP. In this section we consider the weighted version of ISP, without any limitation on the number of rectangles sharing a row, and prove the following result.

THEOREM 3.4. *Algorithm STAB is an* $e/(e-1) \approx 1.582$*-approximation algorithm for WISP.*

Proof. Consider an instance $\mathcal{I}$ of WISP, and let $(y^{\text{lp}}, z^{\text{lp}})$ and (y^*, z^*) be, respectively, an optimal solution to the LP-relaxation of (10)–(12) and the solution returned by the algorithm for $\mathcal{I}$. We show that their values are related as follows:

$$value(y^*, z^*) \leq \frac{e}{e-1} value(y^{\text{lp}}, z^{\text{lp}}). \tag{30}$$

Since $value(y^{\text{lp}}, z^{\text{lp}})$ is a lower bound for the optimal value of WIS, the theorem follows.

Assume, without loss of generality, that the rows of the grid are sorted so that $z_1^{\text{lp}} \geq z_2^{\text{lp}} \geq \cdots \geq z_m^{\text{lp}}$. Further, suppose there are l z^{lp}-values equal to 1, i.e., $z_1^{\text{lp}} = \ldots z_l^{\text{lp}} = 1$, and $1 > z_{l+1}^{\text{lp}} \geq z_{p+2}^{\text{lp}} \geq \cdots \geq z_m^{\text{lp}} \geq 0$.

Let (y^j, z^j) be candidate solution number j constructed by *STAB* for $\mathcal{I}$ $\forall j \in [0 : m]$. From the design of *STAB* we know that

$$value(y^*, z^*) = \min_{j \in [0:m]} value(y^j, z^j) \leq \min_{j \in [l:m]} value(y^j, z^j). \tag{31}$$

Claim 1.

$$value(y^j, z^j) \equiv wy^j + vz^j \leq \sum_{r=1}^{j} v_r + \frac{wy^{\text{lp}}}{1 - z_{j+1}^{\text{lp}}} \quad \text{for any } j \in [l : m].$$

Let us prove it. Consider (y^j, z^j) for some $j \in [l : m]$. By construction,
– $z_r^j = 1$ $\forall r \leq j$,
– $z_r^j = 0$ $\forall r \geq j + 1$,
– y^j is an optimal solution to $\text{IP}^y(\mathcal{I}, b)$ with $b_i = 1 - z_{\rho_i}$ $\forall i \in [1 : n]$.
Clearly, $vz^j \equiv \sum_{r=1}^{m} v_r z_r^j = \sum_{r=1}^{j} v_r$. Let us show that

$$wy^j \leq \frac{wy^{\text{lp}}}{1 - z_{j+1}^{\text{lp}}}. \tag{32}$$

To prove this, we establish that the fractional solution

$$\frac{1}{1 - z_{j+1}^{\text{lp}}}\, y^{\text{lp}}, \tag{33}$$

where we set $z_{m+1}^{\text{lp}} = 0$, is feasible to the LP-relaxation of $\text{IP}^y(\mathcal{I}, b)$. Since y^j is optimal to $\text{IP}^y(\mathcal{I}, b)$, Corollary 2.2 implies (32). So, let us prove the following claim.

Claim 1.1. Solution (33) is feasible to the LP-relaxation of $\text{IP}^y(\mathcal{I}, b)$ with $b_i = 1 - z_{\rho_i}$ $\forall i \in [1 : n]$. We show that constraint (8) is satisfied:

$$\frac{1}{1 - z_{j+1}^{\text{lp}}} \sum_{c \in [l_i, r_i]} y_c^{\text{lp}} \geq 1 - z_{\rho_i}^j \quad \text{for any } i \in [1 : n]. \tag{34}$$

Indeed, in case $z_{\rho_i}^j = 1$, the inequality trivially holds. Otherwise, if $z_{\rho_i}^j = 0$, it follows from the construction of z^j that $\rho_i \geq j+1$. The ordering of the z^{lp}-values implies that $z_{\rho_i}^{\text{lp}} \leq z_{j+1}^{\text{lp}}$. Then, using this and the fact that solution $(y^{\text{lp}}, z^{\text{lp}})$ satisfies constraint (14), we have

$$\frac{1}{1 - z_{j+1}^{\text{lp}}} \sum_{c \in [l_i, r_i]} y_c^{\text{lp}} \geq \frac{1}{1 - z_{j+1}^{\text{lp}}}(1 - z_{\rho_i}^{\text{lp}}) \geq \frac{1}{1 - z_{j+1}^{\text{lp}}}(1 - z_{j+1}^{\text{lp}}) = 1.$$

This proves (34) and, subsequently, Claims 1.1 and 1.

From (31) and Claim 1,

$$value(y^*, z^*) \leq \min_{j \in [l:m]} \left(\sum_{r=1}^{j} v_r + \frac{wy^{\text{lp}}}{1 - z_{j+1}^{\text{lp}}} \right)$$

$$= \sum_{r=1}^{l} v_r + \min_{j \in [l:m]} \left(\sum_{r=l+1}^{j} v_r + \frac{w y^{\text{lp}}}{1 - z_{j+1}^{\text{lp}}} \right).$$

Lemma 3.5 given below implies now that the last expression can be upper bounded by

$$\sum_{r=1}^{l} v_r + \frac{e}{e-1} \left(\sum_{r=l+1}^{m} v_r z_r^{\text{lp}} + w y^{\text{lp}} \right) \leq \frac{e}{e-1} \left(\sum_{r=1}^{l} v_r + \sum_{r=l+1}^{m} v_r z_r^{\text{lp}} + w y^{\text{lp}} \right).$$

Since $z_1^{\text{lp}} = \cdots = z_l^{\text{lp}} = 1$, the last expression can be rewritten as

$$\frac{e}{e-1} \left(\sum_{r=1}^{m} v_r z_r^{\text{lp}} + w y^{\text{lp}} \right) = \frac{e}{e-1} (v z^{\text{lp}} + w y^{\text{lp}}),$$

which establishes inequality (30) and proves the theorem. □

LEMMA 3.5. *Suppose we are given numbers* $1 > \Delta_1 \geq \Delta_2 \geq \cdots \geq \Delta_m \geq 0$ $\forall i = 1, \ldots, m$, *and* $\Delta_{m+1} = 0$. *Further, given are positive numbers* $a_1, a_2, \ldots, a_m$ *and* Y. *Then we have*

$$\min_{j=0,\ldots,m} \left(\sum_{r=1}^{j} a_r + \frac{1}{1 - \Delta_{j+1}} Y \right) \leq \frac{e}{e-1} \left(\sum_{r=1}^{m} a_r \Delta_r + Y \right). \tag{35}$$

We give the proof of this lemma in the appendix.

3.3. Tightness. In this subsection we demonstrate that the ratio between the optimum values of ISP_k and the LP-relaxation of its ILP formulation (13)–(15) can be arbitrarily close to the bounds achieved by *STAB* in case $k = 2$ and $k = \infty$ (which are, respectively, 4/3 and $e/(e-1)$).

For the case $k = 2$ this is shown by the instance of ISP_2 depicted in Figure 4 (recall that all the column and row demands and rectangle weights are unit). Here the optimal value of the problem is 2, since at least two elements (columns or rows) are needed to stab the three rectangles, whereas the optimal fractional solution has the value of 3/2.

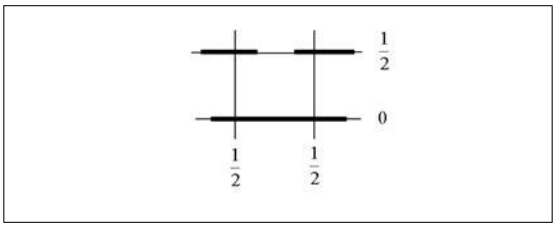


FIG. 4. *An instance of ISP*$_2$ *and an optimal fractional solution.*

In the remainder of the section we consider the problem ISP_∞, or simply ISP, without any limitation on the number of rectangles sharing a row.

THEOREM 3.6. *The integrality gap of* (13)–(15) *is arbitrarily close to* $\frac{e}{e-1}$.

Proof. For each $m \in \mathbb{N}$ we will construct an instance $\mathcal{I}_m$ of ISP and show that the value of some feasible solution to its LP-relaxation tends to be $\frac{e}{e-1}$ times its optimal value as m increases.

Let us construct $\mathcal{I}_m$ as follows. Let the grid have m rows and $t = m!$ columns. Let the rows be numbered consecutively and let each row j intersect exactly j rectangles of the instance. Let rectangles intersected by row j be numbered $j_1, \ldots, j_j$. All these rectangles are disjoint and each intersects exactly $\frac{m!}{j}$ columns (see Figure 5). So, for a rectangle j_i we have that its row number ρ_{j_i} is r, and its leftmost and rightmost columns are $l_{j_i} = \frac{m!}{j}(i-1)+1$ and $r_{j_i} = \frac{m!}{j}i$. The total number of rectangles in the instance is then $n = 1 + 2 + \cdots + m$.

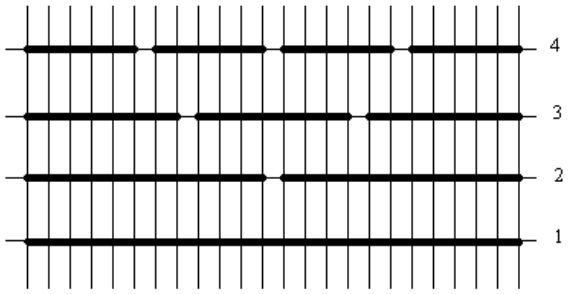


FIG. 5. *Instance* $\mathcal{I}_4$.

We claim that the following solution (y, z) is feasible to the LP-relaxation of (13)–(15) for $\mathcal{I}_m$:

$$z_j = \begin{cases} 0 & \forall j = 1, \ldots, P, \\ 1 - P/j & \forall j = P+1, \ldots, m, \end{cases} \qquad y_c = \frac{P}{m!} \quad \forall c = 1, \ldots, m!, \tag{36}$$

where $P = P(m)$ is the number satisfying

$$\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P+1} \le 1 \quad \text{and} \quad \frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P+1} + \frac{1}{P} \ge 1.$$

Denote the value of this solution by $LP(\mathcal{I}_m)$, and observe that

$$LP(\mathcal{I}_m) = \sum_{c=1}^{t} y_c + \sum_{r=1}^{m} z_r = m - P\left(\frac{1}{P+1} + \frac{1}{P+2} + \cdots + \frac{1}{m}\right).$$

Let us show feasibility of (y, z). Take any rectangle j_i and show that the constraint $z_{\rho_{j_i}} + \sum_{c \in [l_{j_i}, r_{j_i}]} y_c \ge 1$ is satisfied. Notice that the z-values of our solution also can be expressed as $z_j = \max(1 - \frac{P}{j}, 0)\ \forall j = 1, \ldots, m$. Substituting these values, and rewriting the left-hand side of constraints (14) gives

$$\begin{aligned} \max\left(1 - \frac{P}{j_i}, 0\right) + \sum_{c \in [l_{j_i}, r_{j_i}]} \frac{P}{m!} &= \max\left(1 - \frac{P}{j_i}, 0\right) + \frac{m!}{j_i}\frac{P}{m!} \\ &= \max\left(1 - \frac{P}{j_i}, 0\right) + \frac{P}{j_i}. \end{aligned}$$

Clearly, the last expression is at least equal to 1, which proves feasibility of solution (y, z) to the LP-relaxation of (13)–(15) for $\mathcal{I}_m$.

Now denote by $OPT(\mathcal{I})$ the optimum value to ISP for $\mathcal{I}$, and show that $OPT(\mathcal{I}_m) = m$. Consider any optimal integral solution, and denote by k the maximum row number, whose corresponding z-value is 0. First, this means that there are at least $m-k$ rows whose z-values are 1. Second, observe that, since there are k disjoint rectangles on row k and this row is not selected, there are at least k columns needed to stab these rectangles. Therefore, this solutions has to select at least $m-k$ rows and k columns, meaning $OPT(\mathcal{I}_m) \geq m$. Since there exists a feasible solution of value m (select all the rows, for instance), we obtain that $OPT(\mathcal{I}_m) = m$.

We use Lemma 5.3 given in the appendix to prove that the ratio

$$\frac{OPT(\mathcal{I}_m)}{LP(\mathcal{I}_m)} = \frac{m}{m - P(\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P+1})}$$

approaches $\frac{e}{e-1}$ when m increases. This establishes our tightness result. □

As mentioned in the introduction, Theorems 3.1 and 3.6 imply that it is unlikely that a better ratio for ISP_∞ can be achieved using formulation (13)–(15).

Approximation algorithms with a ratio of $\frac{e}{e-1}$ are not uncommon in the literature; integrality gaps with this ratio seem to appear less frequently. Another example of a (different) formulation with an integrality gap that equals $\frac{e}{e-1}$ is described in Hoogeveen, Skutella, and Woeginger [6].

4. Conclusion. We presented an approximation algorithm called *STAB* for two variants of the weighted rectangle stabbing problem. *STAB* achieves a ratio of $\frac{1}{1-(1-1/k)^k}$ for ISP_k, the special case where each rectangle is stabbed by a single row and by at most k columns, and where all stabbing lines have unit weight. *STAB* achieves a ratio of $\frac{e}{e-1}$ for WISP, the special case where each rectangle is stabbed by a single row. *STAB* considers different ways of rounding the LP-relaxation and outputs the best solution found in this way; it is also shown that the ratio proved equals the integrality gap when $k = 2$ and when $k = \infty$.

5. Appendix. In this appendix we give proofs of lemmas which we used in this paper.

LEMMA 3.3. *Given are real numbers* $1 \geq \Delta_1 \geq \Delta_2 \geq \cdots \geq \Delta_m \geq 0$, *a positive real number* Y, *an integer* $p \geq 2$, *and an integer* $0 \leq l < m$. *The following holds:*

$$\min_{i=l,\ldots,m} (i + p \cdot q_i(\Delta, Y/p)) \leq \frac{1}{1-(1-1/p)^p} \left(Y + \sum_{r=l+1}^{m} \Delta_r \right) + l, \tag{37}$$

where $q_i = q_i(\Delta, Y/p)$ *for each* $i \in [0:m]$ *is uniquely defined by the equality*

$$\sum_{k=1}^{\lfloor q_i \rfloor} (1 - \Delta_{i+k}) + (q_i - \lfloor q_i \rfloor)(1 - \Delta_{i+\lceil q_i \rceil}) = Y/p, \tag{38}$$

where we put $\Delta_i = 0$ *if* $i > m$.

Proof. It is enough to prove this lemma for $l = 0$. The case of other $l < m$ can be reduced to the case of $l = 0$ by changing the index to $j = i - l$ and observing that $q_{j+l}(\Delta, Y/p) = q_i(\Delta^{-l}, Y/p)$, where vector Δ^{-l} is obtained by deleting the first l elements from vector Δ. So we will prove that

$$\min_{i=0,\ldots,m} (i + p \cdot q_i(\Delta, Y/p)) \leq \frac{1}{1-(1-1/p)^p} \left(Y + \sum_{r=1}^{m} \Delta_r \right).$$

The proof consists of two lemmas. In Lemma 5.1 we show that the left-hand side of (37) is upper bounded by the following supremum:

$$\sup_{f(\cdot) \in H} G(f(\cdot)), \tag{39}$$

where

$$G(f(\cdot)) = \min_{x \in \mathbb{R}_+} \left(f(x) + p \cdot (f(x + Y/p) - f(x))\right), \tag{40}$$

and the class of functions H is defined as

$$H = \left\{ f(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+ \;\middle|\; \begin{array}{l} f(\cdot) \text{ is continuous, increasing, concave,} \\ f(0) = 0, f(x) \le x + \sum_{r=1}^{m} \Delta_r \end{array} \right\}. \tag{41}$$

In Lemma 5.2 we show that this supremum is upper bounded by the right-hand side of (37), which proves the lemma.

LEMMA 5.1.

$$\min_{i=0,\ldots,m} (i + p \cdot q_i(\Delta, Y/p)) \le \sup_{f(\cdot) \in H} G(f(\cdot)),$$

where $G(f(\cdot))$ and H are defined in (40) *and* (41).

Proof. To establish this, it is sufficient to exhibit a particular function $\hat{f}(\cdot) \in H$, such that

$$G(\hat{f}(\cdot)) = \min_{i=0,\ldots,m} (i + p \cdot q_i(\Delta, Y/p)). \tag{42}$$

Then, the supremum of $G(f(\cdot))$ over all the possible $f(\cdot) \in H$ is clearly larger than or equal to $G(\hat{f}(\cdot))$.

Before we describe the function $\hat{f}(\cdot)$, let us define an auxiliary function $F(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ as follows:

$$F(q) \equiv \sum_{r=1}^{\lfloor q \rfloor} (1 - \Delta_r) + (q - \lfloor q \rfloor)\ (1 - \Delta_{\lceil q \rceil}), \tag{43}$$

where we set $\Delta_r = 0\ \forall r \ge m + 1$.

Observe that $F(\cdot)$ is

– continuous;

– increasing, since $\Delta_r < 1$, and therefore $(1 - \Delta_r) > 0\ \forall r = 1, \ldots, \infty$;

– convex, since the coefficients Δ_r are nonincreasing with increasing r, and therefore the coefficients $(1 - \Delta_r)$ are nondecreasing with increasing r.

Furthermore,

– $F(0) = 0$;

– $F(q) \ge (q - \sum_{r=1}^{m} \Delta_r)\ \forall q \in \mathbb{R}_+$, since $F(q)$ can be also represented as

$$F(q) = q - \left(\sum_{r=1}^{\lfloor q \rfloor} \Delta_r + (q - \lfloor q \rfloor)\Delta_{\lceil q \rceil} \right),$$

and obviously $(\sum_{r=1}^{\lfloor q \rfloor} \Delta_r + (q - \lfloor q \rfloor)\Delta_{\lceil q \rceil}) \le \sum_{r=1}^{m} \Delta_r\ \forall q \in \mathbb{R}_+$;

– $F(q)$ is linear on each of the intervals $[i, i+1],\ i = 0, \ldots, m-1$, and on $[m, +\infty)$.

We are now ready to present $\hat{f}(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$. We define

$$\hat{f}(\cdot) \equiv F^{-1}(\cdot)$$

(since $F(\cdot)$ is increasing, $F^{-1}(\cdot)$ exists).

We claim that $\hat{f}(\cdot) \in H$. Indeed, $\hat{f}(\cdot)$ has the following properties:

– $\hat{f}(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ since $F(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$;

– $\hat{f}(\cdot)$ is continuous, increasing, and concave, since $F(\cdot)$ is continuous, increasing, and convex;

– $\hat{f}(0) = 0$, since $F(0) = 0$;

– $\hat{f}(x) \leq x + \sum_{r=1}^{m} \Delta_r\ \forall x \in \mathbb{R}_+$. This can be obtained from $F(q) \geq (q - \sum_{r=1}^{m} \Delta_r)\ \forall q \in \mathbb{R}_+$, using $F(q) = x,\ q = \hat{f}(x)$.

This proves that $\hat{f}(\cdot) \in H$.

To prove the lemma it remains to show that

$$G(\hat{f}(\cdot)) = \min_{i=0,\ldots,m} (i + p \cdot q_i(\Delta, Y/p)).$$

Comparing the definition of $q_i(\Delta, Y/p)$ (see (38)) and $F(\cdot)$ (see (43)), observe that for each $i \in [0 : m]$ q_i satisfies

$$F(i + q_i) - F(i) = Y/p. \tag{44}$$

Thus, $q_i = F^{-1}(F(i) + Y/p) - i$. Setting $x_i \equiv F(i)\ \forall i = 0, \ldots, m$, we find that $i = F^{-1}(x_i)$ and $q_i = F^{-1}(x_i + Y/p)) - F^{-1}(x_i)$. Replacing $F^{-1}(\cdot)$ by $\hat{f}(\cdot)$, we obtain

$$q_i = \hat{f}(x_i + Y/p) - \hat{f}(x_i)\ \forall i = 0, \ldots, m.$$

Using this together with $i = F^{-1}(x_i) = \hat{f}(x_i)$, we can rewrite

$$\min_{i=0,\ldots,m} (i + p \cdot q_i(\Delta, Y/p)) = \min_{\substack{i\,=\,0,\,\ldots\,,\,m \\ x_i\,=\,\hat{f}^{-1}(i)}} \Big(\hat{f}(x_i) + p(\hat{f}(x_i + Y/p) - \hat{f}(x_i)) \Big). \tag{45}$$

Now we need to show that the latter expression is equal to

$$G(\hat{f}(\cdot)) \equiv \min_{x\,\in\,\mathbb{R}_+} \Big(\hat{f}(x) + p(\hat{f}(x + Y/p) - \hat{f}(x)) \Big). \tag{46}$$

We do this by showing that the function $\hat{f}(x) + p(\hat{f}(x + Y/p) - \hat{f}(x))$ is continuous and concave in each of the intervals $[x_i, x_{i+1}]\ \forall i = 0, \ldots, m-1$, and is increasing in $[x_m, +\infty)$. Therefore the minimum can be achieved only at one of the endpoints $x_0, x_1, \ldots, x_m$.

Indeed, consider function $\hat{f}(x) + p(\hat{f}(x + Y/p) - \hat{f}(x))$ in $[x_i, x_{i+1}]$ for some $i \in [0 : m-1]$. It can also be written as $p\hat{f}(x + Y/p) - (p-1)\hat{f}(x)$. We know that $\hat{f}(x + Y/p)$ is concave on $[x_i, x_{i+1}]$, since it is concave everywhere in $\mathbb{R}_+$. Furthermore, $\hat{f}(x)$ is linear on each $[x_i, x_{i+1}], i \in [0 : m-1]$, since $F(\cdot)$ is linear on $[i, i+1],\ i \in [0 : m-1]$. Obviously, a concave function minus a linear function is again concave.

Now we show that $p\hat{f}(x + Y/p) - (p-1)\hat{f}(x)$ is increasing in $[x_m, +\infty)$. Since $\hat{f}(x) = F^{-1}(\cdot)$ is increasing and linear in $[x_m, +\infty)$, the growth rate of $\hat{f}(x)$ is the

same as the growth rate of $\hat{f}(x + Y/p)$ in $[x_m, +\infty)$, and thus the growth rate of $p\hat{f}(x + Y/p) - (p-1)\hat{f}(x)$ is positive. We have proved that the minimum in (46) is always achieved at one of the points $x_0, x_1, \ldots, x_m$, and therefore (46) is equal to (45). This completes the proof of Lemma 5.1. □

LEMMA 5.2.

$$\sup_{f(\cdot)\in H} G(f(\cdot)) \;\le\; \frac{1}{1-(1-1/p)^p}\, C,$$

where

$$C = Y + \sum_{r=1}^{m} \Delta_r,$$

$$G(f(\cdot)) = \min_{x\in\mathbb{R}_+} \left(f(x) + p(f(x+Y/p) - f(x))\right),$$

and the set of functions H (via notation C) is

$$H = \left\{ f(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+ \;\middle|\; \begin{array}{l} f(\cdot) \text{ is continuous, increasing, concave,} \\ f(0) = 0, f(x) \le x + C - Y \end{array} \right\}.$$

Proof. We will prove several claims and subclaims.

Claim 1.

$$\sup_{f(\cdot)\in H} G(f(\cdot)) = \sup_{g\,:\, f^g(\cdot)\in H} g,$$

where for each $g \in \mathbb{R}_+$ function $f^g(\cdot)$ is defined as follows:

– $f^g(j \cdot Y/p) = g(1-(1-1/p)^j)\ \forall j \in 0 \cup \mathbb{N}$;

– $f^g(x)$ is continuous in $[0, +\infty)$ and linear in each $[(j-1)\cdot Y/p, j\cdot Y/p],\ j \in \mathbb{N}$.

Notice that $f^g(\cdot)$ is completely defined by the above characterization.

To prove this claim it is enough to show that for any $f(\cdot) \in H$ there exists a function $f^{\hat{g}}(\cdot) \in H$, with $\hat{g} \ge 0$, such that

$$G(f(\cdot)) = G(f^{\hat{g}}(\cdot)) = \hat{g}.$$

To show that, we prove two subsidiary claims.

Claim 1.1. For any $g \ge 0$,

$$G(f^g(\cdot)) \equiv \min_{x\in\mathbb{R}_+} \left(f^g(x) + p(f^g(x+Y/p) - f^g(x))\right) = g.$$

Indeed, by construction $f^g(x)$ is linear in each of the intervals $[(j-1)\cdot Y/p,\ j\cdot Y/p]$, $j \in \mathbb{N}$. This implies that function $(f^g(x) + p(f^g(x+Y/p) - f^g(x)))$ is linear in each of these intervals as well. Therefore the minimum over all $x \ge 0$ is achieved in one of the endpoints $0,\ Y/p,\ 2Y/p, \ldots$. Consider $(f^g(x) + p\cdot(f^g(x+Y/p) - f^g(x)))$ at the point $x = j\cdot Y/p$ for some $j \in \mathbb{N} \cup 0$:

$$f^g(j\cdot Y/p) + p\cdot(f^g((j+1)\cdot Y/p) - f^g(j\cdot Y/p)).$$

Using the definition of $f^g(\cdot)$ we can rewrite it as follows:

$$g \cdot (1-(1-1/p)^j) + p \cdot (g(1-(1-1/p)^{j+1}) - g \cdot (1-(1-1/p)^j)).$$

With simple computations one can verify that the last expression is equal to g. This proves Claim 1.1.

Claim 1.2. For any $f(\cdot) \in H$ it holds that $f^{\hat{g}}(\cdot) \in H$, where $\hat{g} = G(f(\cdot))$.

Clearly, $f^{\hat{g}}(x)$ is concave. To prove that $f^{\hat{g}}(x) \leq x + C - Y \ \forall x \in \mathbb{R}_+$, it is sufficient to show that $f^{\hat{g}}(x) \leq f(x)$, since $f(\cdot) \in H$ means, e.g., $f(x) \leq x + C - Y \ \forall x \in \mathbb{R}_+$.

So, let us establish that $f^{\hat{g}}(x) \leq f(x) \ \forall x \in \mathbb{R}_+$. Recall that $f^{\hat{g}}(x)$ is linear in each of the intervals $[(j-1) \cdot Y/p, j \cdot Y/p],\ j \in \mathbb{N}$, and $f(x)$ is concave in $\mathbb{R}_+$. Then it is sufficient to show that

$$f^{\hat{g}}(x) \leq f(x) \ \forall x = j \cdot Y/p,\ j \in 0 \cup \mathbb{N}.$$

We use mathematical induction on j. For $j = 0$, $f^{\hat{g}}(0) = f(0) = 0$ and the inequality trivially holds. Suppose, for $j-1$ we have proved that $f^{\hat{g}}((j-1) \cdot Y/p) \leq f((j-1) \cdot Y/p)$, and let us show that $f^{\hat{g}}(j \cdot Y/p) \leq f(j \cdot Y/p)$.

Observe that $f^{\hat{g}}(\cdot)$ can be represented in a recursive way as follows:

$$f^{\hat{g}}(j \cdot Y/p) = \hat{g}/p + f^{\hat{g}}((j-1) \cdot Y/p)\ (1-1/p). \tag{47}$$

Since $\hat{g} = G(f(\cdot))$ we know that

$$\hat{g} \leq f((j-1) \cdot Y/p) + p \cdot (f(j \cdot Y/p) - f((j-1) \cdot Y/p)).$$

Rearranging the expression, we obtain

$$f(j \cdot Y/p) \geq \hat{g}/p + f((j-1) \cdot Y/p)\ (1-1/p).$$

By the induction hypothesis and (47) we can bound the right-hand side by

$$\hat{g}/p + f((j-1) \cdot Y/p)\ (1-1/p)\ \geq\ \hat{g}/p + f^{\hat{g}}((j-1) \cdot Y/p)\ (1-1/p) = f^{\hat{g}}(j \cdot Y/p).$$

This proves Claim 1.2.

These two claims imply that for any $f(\cdot) \in H$, there exists $f^{\hat{g}}(\cdot) \in H$, with $\hat{g} \geq 0$, such that

$$G(f(\cdot)) = G(f^{\hat{g}}(\cdot)) = \hat{g}.$$

This implies Claim 1.

Claim 2.

$$\sup_{g\,:\,f^g(\cdot) \in H} g \leq \frac{1}{1-(1-1/p)^p}\ C.$$

Indeed, $f^g(\cdot) \in H$ implies $f^g(x) \leq x + C - Y \ \forall x \in \mathbb{R}_+$ and, in particular, for $x = Y$. From this, using the definition of $f^g(\cdot)$, we obtain

$$f^g(Y) \equiv f^g(p \cdot Y/p) \equiv g(1-(1-1/p)^p) \leq Y + C - Y = C,$$

and from the last inequality, we obtain

$$g \leq \frac{1}{(1-(1-1/p)^p)}\ C,$$

which proves Claim 2 and establishes Lemma 5.2. □

Now we give a proof of Lemma 3.5. This version of the proof is due to Sgall (see the acknowledgments).

LEMMA 3.5. *Suppose we are given numbers* $1 > \Delta_1 \geq \Delta_2 \geq \cdots \geq \Delta_m \geq 0$ *and* $\Delta_{m+1} = 0$. *Further, given are positive numbers* $a_1, a_2, \ldots, a_m$ and Y. *Then we have*

$$\min_{j=0,\ldots,m} \left(\sum_{r=1}^{j} a_r + \frac{Y}{1-\Delta_{j+1}} \right) \leq \frac{e}{e-1} \left(\sum_{r=1}^{m} a_r \Delta_r + Y \right). \tag{48}$$

Proof. We use mathematical induction on the size of inequality m. For $m = 0$, the statement trivially holds. Suppose that the lemma was proved for any inequality of size smaller than m. First, consider the case $\Delta_1 \geq \frac{e-1}{e}$. We can write

$$\begin{aligned} \min_{j=0,\ldots,m} \left(\sum_{r=1}^{j} a_r + \frac{Y}{1-\Delta_{j+1}} \right) &\leq \min_{j=1,\ldots,m} \left(\sum_{r=1}^{j} a_r + \frac{Y}{1-\Delta_{j+1}} \right) \\ &= a_1 + \min_{j=1,\ldots,m} \left(\sum_{r=2}^{j} a_r + \frac{Y}{1-\Delta_{j+1}} \right). \end{aligned}$$

The latter minimum is the left-hand side of (48) for a smaller sequence: $\Delta_2, \ldots, \Delta_m$ and $a_2, \ldots, a_m$. Applying the induction hypothesis, we can bound the last expression from above as follows (we also use our bound on Δ_1):

$$\begin{aligned} a_1 + \frac{e}{e-1} \left(\sum_{r=2}^{m} a_r \Delta_r + Y \right) &\leq a_1 \cdot \Delta_1 \frac{e}{e-1} + \frac{e}{e-1} \left(\sum_{r=2}^{m} a_r \Delta_r + Y \right) \\ &= \frac{e}{e-1} \left(\sum_{r=1}^{m} a_r \Delta_r + Y \right). \end{aligned}$$

Thus, we have shown an induction step for the case $\Delta_1 \geq \frac{e-1}{e}$. For the remaining case, $\Delta_1 < \frac{e-1}{e}$, we give a direct proof below.

Suppose $\Delta_1 < \frac{e-1}{e}$. Denote the left-hand side of (48) by X, and notice that

$$\sum_{r=1}^{j} a_r \geq X - \frac{1}{1-\Delta_{j+1}} Y \quad \text{for } 0 \leq j \leq m. \tag{49}$$

The following steps are justified below:

$$\begin{aligned} \sum_{r=1}^{m} a_r \Delta_r + Y &= \sum_{j=1}^{m} \left((\Delta_j - \Delta_{j+1}) \sum_{r=1}^{j} a_r \right) + Y \\ &\geq^{(1)} \sum_{j=1}^{m} (\Delta_j - \Delta_{j+1}) X - \left(\sum_{j=1}^{m} \frac{\Delta_j - \Delta_{j+1}}{1-\Delta_{j+1}} \right) Y + Y \\ &= \Delta_1 X - \left(\sum_{j=1}^{m} \left(\frac{\Delta_j - 1}{1-\Delta_{j+1}} + 1 \right) \right) Y + Y \end{aligned}$$

$$= \Delta_1 X + \left(1 - m + \sum_{j=1}^{m} \frac{1-\Delta_j}{1-\Delta_{j+1}}\right) Y$$

$$\geq^{(2)} \Delta_1 X + \left(1 - m + m(1-\Delta_1)^{\frac{1}{m}}\right) Y$$

$$\geq^{(3)} \Delta_1 X + \left(1 - m + m(1-\Delta_1)^{\frac{1}{m}}\right)(1-\Delta_1) X$$

$$= \left(1 + m(-1 + (1-\Delta_1)^{\frac{1}{m}})(1-\Delta_1)\right) X \geq^{(4)} \left(1 - \frac{1}{e}\right) X.$$

(1) Here we use the ordering of the deltas and inequality (49).

(2) This inequality follows from the arithmetic-geometric mean inequality $\sum_{j=1}^{m} x_j \geq m(\prod_{j=1}^{m} x_j)^{1/m}$ used for positive numbers $x_j = \frac{1-\Delta_j}{1-\Delta_{j+1}}$.

(3) Here we use inequality $Y \geq (1-\Delta_1)X$, which is implied by (49) for $j = 0$ and the fact that the coefficient of Y is nonnegative, which follows from $1 - \Delta_1 \geq \frac{1}{e} \geq (1 - \frac{1}{m})^m$.

(4) This inequality is elementary calculus: The minimum of the left-hand side over all Δ_1 is achieved for $1 - \Delta_1 = (\frac{m}{m+1})^m$, and, after substituting this value, it reduces to $1 - (\frac{m}{m+1})^{m+1} \geq 1 - \frac{1}{e}$. □

LEMMA 5.3. *Let $P(m) \in \mathbb{N}$ be defined as follows:*

$$\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P(m)+1} \leq 1 \quad \textit{and} \tag{50}$$

$$\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P(m)+1} + \frac{1}{P(m)} \geq 1. \tag{51}$$

Then,

$$\lim_{m\to\infty} \frac{m}{m - P(m)(\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P(m)+1})} = \frac{e}{e-1}.$$

Proof. Let us first find $\lim_{m\to\infty} P(m)/m$. Observe that the following inequalities hold:

$$\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P(m)+1} \geq \int_{P(m)+1}^{m+1} \frac{1}{x}\, dx = \ln \frac{m+1}{P(m)+1},$$

$$\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P(m)} \leq \int_{P(m)-1}^{m} \frac{1}{x}\, dx = \ln \frac{m}{P(m)-1}$$

(the equalities follow from $\int_a^b 1/x\, dx = \ln b/a$). Then (50) and (51) imply

$$1 \geq \ln \frac{m+1}{P(m)+1}, \qquad 1 \leq \ln \frac{m}{P(m)-1}.$$

From this we have

$$\frac{m+1}{e} - 1 \le P(m) \le \frac{m}{e} + 1.$$

Dividing by m,

$$\frac{1+1/m}{e} - 1/m \le \frac{P(m)}{m} \le \frac{1}{e} + 1/m.$$

Now we see that $\lim_{m\to\infty} P(m)/m = 1/e$.

Let us now find $\lim_{m\to\infty}(\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P(m)+1})$. From (50) and (51) we have

$$1 - \frac{1}{P(m)} \le \frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P(m)+1} \le 1.$$

Since we already know that $\lim_{m\to\infty} P(m) = \infty$, we have

$$\lim_{m\to\infty}\left(\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P(m)+1}\right) = 1.$$

Now consider

$$\frac{m}{m - P(m)(\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P(m)+1})} = \frac{1}{1 - \frac{P(m)}{m}(\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P(m)+1})}.$$

Using $\lim_{m\to\infty} \frac{P(m)}{m} = 1/e$ and $\lim_{m\to\infty}(\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P(m)+1}) = 1$ we have

$$\lim_{m\to\infty} \frac{1}{1 - \frac{P(m)}{m}(\frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{P(m)+1})} = \frac{1}{1-1/e} = \frac{e}{e-1},$$

which establishes the lemma. □

Acknowledgments. We are very grateful to professor Jiří Sgall from the Mathematical Institute of the Academy of Sciences of the Czech Republic, for allowing us to include his proof of Lemma 3.5. We also thank an anonymous referee whose comments improved the paper.

REFERENCES

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[2] G. Călinescu, A. Dumitrescu, H. Karloff, and P.-J. Wan, *Separating points by axis-parallel lines*, Internat. J. Comput. Geom. Appl., 15 (2005), pp. 575–590.

[3] D. R. Gaur, T. Ibaraki, and R. Krishnamurti, *Constant ratio approximation algorithms for the rectangle stabbing problem and the rectilinear partitioning problem*, J. Algorithms, 43 (2002), pp. 138–152.

[4] R. Hassin and N. Megiddo, *Approximation algorithm for hitting objects with straight lines*, Discrete Appl. Math., 30 (1991), pp. 29–42.

[5] A. J. Hoffman, A. W. J. Kolen, and M. Sakarovitch, *Totally-balanced and greedy matrices*, SIAM J. Alg. Discrete Methods, 6 (1985), pp. 721–730.

[6] H. Hoogeveen, M. Skutella, and G. J. Woeginger, *Preemptive scheduling with rejection*, Math. Program., 94 (2003), pp. 361–374.

[7] S. Kovaleva, *Approximation of Geometric Set Packing and Hitting Set Problems*, Ph.D. thesis, Maastricht University, Maastricht, The Netherlands, 2003.

[8] S. Kovaleva and F. C. R. Spieksma, *Approximation of a geometric set covering problem*, in Proceedings of the 12th Annual International Symposium on Algorithms and Computation (ISAAC'01), Lecture Notes in Comput. Sci. 2223, Springer-Verlag, Berlin, 2001, pp. 493–501.

[9] S. Kovaleva and F. C. R. Spieksma, *Primal-dual approximation algorithms for a packing-covering pair of problems*, RAIRO Oper. Res., 36 (2002), pp. 53–72.

[10] S. Kovaleva and F. C. R. Spieksma, *Approximation of rectangle stabbing and interval stabbing problems*, in Proceedings of the 12th Annual European Symposium on Algorithms (ESA 2004), Lecture Notes in Comput. Sci. 3221, Springer-Verlag, Berlin, 2004, pp. 426–435.

[11] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley & Sons, Chichester, UK, 1986.

[12] V. V. Vazirani, *Approximation Algorithms*, Springer-Verlag, Berlin, 2001.

[13] A. F. Veinott and H. M. Wagner, *Optimal capacity scheduling: Parts* I *and* II, Oper. Res., 10 (1962), pp. 518–547.

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 769–798

THE COMPLEXITY OF GRAPH PEBBLING*

KEVIN MILANS† AND BRYAN CLARK‡

Abstract. In a graph G whose vertices contain pebbles, a pebbling move uv removes two pebbles from u and adds one pebble to a neighbor v of u. The optimal pebbling number $\widehat{\pi}(G)$ is the minimum k such that there exists a distribution of k pebbles to G so that for any target vertex r in G, there is a sequence of pebbling moves which places a pebble on r. The pebbling number $\pi(G)$ is the minimum k such that for all distributions of k pebbles to G and for any target vertex r, there is a sequence of pebbling moves which places a pebble on r. We explore the computational complexity of computing $\widehat{\pi}(G)$ and $\pi(G)$. In particular, we show that deciding whether $\widehat{\pi}(G) \leq k$ is NP-complete. Furthermore, we prove that deciding whether $\pi(G) \leq k$ is Π_2^{P}-complete and therefore both NP-hard and coNP-hard. Additionally, we provide a characterization of when an unordered set of pebbling moves can be ordered to form a valid sequence of pebbling moves.

Key words. graph pebbling, complexity, Π_2^{P}-completeness

AMS subject classifications. 05C99, 68Q17, 68R10

DOI. 10.1137/050636218

1. Introduction. Let G be a simple, undirected graph and let $p : V(G) \to \mathbb{N} \cup \{0\}$ be a distribution of pebbles to the vertices of G. We refer to the total number of pebbles $\sum_v p(v)$ as the *size* of p, denoted by $|p|$. A pebbling move uv consists of removing two pebbles from a vertex u with $p(u) \geq 2$ and adding one pebble to a neighbor v of u. After completing a pebbling move uv, we are left with a new distribution of pebbles, which we denote by p_{uv}. Similarly, if $\sigma = u_1v_1, \ldots, u_kv_k$ is a sequence of pebbling moves, denote by p_σ the distribution of pebbles that results from making the pebbling moves specified by σ. Although graph pebbling was originally developed to simplify a result in number theory (Chung provides the history [2]), it has since become an object of study in its own right. Hurlbert presents a detailed survey of recent graph pebbling results [6].

We use G and H to refer to simple, undirected graphs. We use D and E to refer to directed graphs, possibly with multiple edges. If v is a vertex in a directed multigraph D, we denote the indegree (resp., outdegree) of v by $d_D^-(v)$ (resp., $d_D^+(v)$). When D is clear from the context, we may write $d^-(v)$ for $d_D^-(v)$ or $d^+(v)$ for $d_D^+(v)$. We say that $D \subseteq E$ if, for each edge uv in D, the multiplicity of uv in E is as least as large as the multiplicity of uv in D. We denote by $D - uv$ the directed multigraph obtained from D by removing one occurrence of the edge uv; similarly, when $D \subseteq E$ we denote by $E - D$ the directed multigraph obtained from E by reducing the multiplicity of each edge uv in E by the multiplicity of uv in D. We use $V(G)$ (resp., $E(G)$) to refer to the vertex set (resp., edge set) of G, and proceed in a similar manner for directed multigraphs, except that $E(D)$ is a multiset. We define $n(G) = |V(G)|$ and $e(G) = |E(G)|$, and proceed in a similar manner for directed multigraphs. We write $d_G(u, v)$ (or $d(u, v)$ when G is clear from the context) for the length of the shortest

*Received by the editors July 17, 2005; accepted for publication (in revised form) February 22, 2006; published electronically October 24, 2006.
http://www.siam.org/journals/sidma/20-3/63621.html
†Department of Computer Science, University of Illinois, Urbana, IL 61801 (milans@uiuc.edu).
The work of this author was partially supported by NSF grant DMS-0528086.
‡Department of Physics, University of Illinois, Urbana, IL 61801 (bkclark@uiuc.edu).

uv-path in G. If p and q are pebble distributions on a graph G, we say that $p \geq q$ if $p(v) \geq q(v)$ for each vertex v in G.

Given a graph G with a pebble distribution p, we say that a vertex r in G is *reachable* if there is a sequence of pebbling moves which places a pebble on r. Note that whenever $p(r) > 0$, r is trivially reachable. The notion of reachability is fundamental to graph pebbling; most of our decision problems involve questions of reachability. We call the problem of deciding (given G, p, and r) whether r is reachable REACHABLE. The complexity of REACHABLE was first studied by Hurlbert and Kierstead, who found that REACHABLE is NP-complete via a reduction from the perfect matching problem in 4-uniform hypergraphs [7]. In section 3, we establish that REACHABLE is NP-complete, a result obtained simultaneously and independently by Watson [14].

Given a graph G and a target vertex r, the *r-pebbling number* of G, denoted $\pi(G, r)$, is the minimum k such that r is reachable under every pebble distribution of size k. Similarly, the *pebbling number* of G, denoted $\pi(G)$, is the minimum k such that every vertex in G is reachable under every pebble distribution of size k. For a connected graph G, a pigeonhole argument quickly establishes that such a k exists and thus $\pi(G)$ is well defined (see Proposition 5.1). We call the problem of deciding whether $\pi(G, r) \leq k$ (resp., $\pi(G) \leq k$) R-PEBBLING-NUMBER (resp., PEBBLING-NUMBER). In section 5, we establish that both decision problems are Π_2^{P}-complete, meaning that these problems are complete for the class of problems computable in polynomial time by a coNP machine equipped with an oracle for an NP-complete language. Consequently, these decision problems are both NP-hard and coNP-hard. It follows that R-PEBBLING-NUMBER and PEBBLING-NUMBER are in neither NP nor coNP unless NP = coNP. Watson simultaneously and independently established that R-PEBBLING-NUMBER is coNP-hard [14].

Observe that if we fix some vertex r in G and put one pebble on every other vertex, r is not reachable. It follows that $\pi(G) \geq n(G)$. It is natural to wonder which graphs achieve equality in $\pi(G) = n(G)$. Although no characterization of such graphs is known, a growing body of results provides conditions that are necessary or sufficient to imply $\pi(G) = n(G)$. Recall that G is k-connected if $n(G) \geq k + 1$ and for every set S of at most $k - 1$ vertices, $G - S$ is connected. If G has diameter 2 and is 3-connected, then $\pi(G) = n(G)$ [3]. Consequently, the probability that a random graph on n vertices satisfies $\pi(G) = n(G)$ approaches 1 as n grows. Furthermore, if G has diameter d and is (2^{2d+3})-connected, then $\pi(G) = n(G)$ [5]. On the other hand, if G contains a cut vertex, then $\pi(G) > n(G)$. Indeed, suppose v is a cut vertex in G and let u and w be vertices in separate components of $G - v$. If we put three pebbles on u, zero pebbles on v and w, and one pebble on every other vertex, then it is not possible to place a pebble on w.

The *optimal pebbling number* of G, denoted $\widehat{\pi}(G)$, is the minimum k such that there is some distribution of size k under which all vertices are reachable. We call the problem of deciding whether $\widehat{\pi}(G) \leq k$ OPTIMAL-PEBBLING-NUMBER. In section 4, we establish that OPTIMAL-PEBBLING-NUMBER is NP-complete.

It is immediate that $\widehat{\pi}(G) \leq n(G)$. If G is connected, then $\widehat{\pi}(G) \leq \lceil 2n(G)/3 \rceil$ [1]. Equality is achieved by the path [1, 9] and the cycle [1]. It is an open problem to characterize which graphs achieve equality.

Given G and distributions p and q, we say that p *covers* q if there exists a sequence of pebbling moves σ such that $p_\sigma \geq q$. The *unit distribution* assigns one pebble to each vertex in G. We call the problem of deciding whether p covers the unit distribution

TABLE 1.1
A summary of the decision problems considered in this paper.

Name	Full name	Description	Complexity
PN	PEBBLING-NUMBER	Given G,k: is $\pi(G) \leq k$?	Π_2^{P}-complete
RPN	R-PEBBLING-NUMBER	Given G, k, r: is $\pi(G, r) \leq k$?	Π_2^{P}-complete
OPN	OPTIMAL-PEBBLING-NUMBER	Given G, k: is $\widehat{\pi}(G) \leq k$?	NP-complete
PR	REACHABLE	Given G, p, r: is r reachable?	NP-complete

COVERABLE. In section 3, we establish that COVERABLE is NP-complete; this result was obtained simultaneously and independently by Watson [14].

Although most of the problems we study are computationally difficult, there are some interesting pebbling problems that are tractable. A pebble distribution q is *positive* if q assigns at least one pebble to every vertex. A distribution p is *simple* if it assigns zero pebbles to all but one vertex. The *q-cover pebbling number* of G, denoted $\gamma_q(G)$, is the minimum k such that every distribution of size k covers q. The cover pebbling theorem states that for any positive distribution q, there is a simple distribution p of size $\gamma_q(G)-1$ such that p does not cover q [13, 11]. As a consequence, given G and a positive distribution q, one can easily compute $\gamma_q(G)$ in polynomial time. In the special case that q is the unit distribution, we simply write $\gamma(G)$ for $\gamma_q(G)$.

Let us consider a simple example. Suppose we are given a graph H with a distribution of pebbles, and we wish to determine if there is a sequence of pebbling moves which ends with only one pebble left in the entire graph. We call this problem ANNIHILATION. It is not difficult to see that ANNIHILATION is NP-hard. Indeed, a reduction from HAMILTONIAN-PATH is almost immediate. Specifically, to decide if G has a Hamiltonian path, we may construct H from G by introducing a new vertex v which is adjacent to each vertex in G. We place two pebbles on v and one pebble on every other vertex in H. It is clear that G has a Hamiltonian path if and only if there is a sequence of pebbling moves which results in only one pebble in H.

What is less clear is that ANNIHILATION is in NP. If σ is a sequence of pebbling moves in G under p which results in only one pebble left in G, then the length of σ is $|p| - 1$, which may be exponentially large in the number of bits needed to represent G and p. Hence, σ may be too large to serve as a certificate for membership in ANNIHILATION. However, as we will see, the order of the moves in σ is insignificant. In fact, if we are merely told how many times σ pebbles along each direction in every edge in G, then we can quickly verify the existence of σ.

In section 2, we develop a characterization of when unordered sets of pebbling moves may be ordered in a way that yields a valid sequence of pebbling moves. In section 3, we present results on the complexity of REACHABLE and COVERABLE. We also observe that a simple greedy strategy solves REACHABLE whenever G is a tree. Section 3 uses some results from section 2. In section 4, we present our results on the complexity of computing the optimal pebbling number. Section 4 uses some results from sections 2 and 3. In section 5, we present our results on the complexity of computing the (r-)pebbling number. Section 5 uses some results from sections 2 and 3; it is generally independent of section 4. In section 6, we present our conclusions. Table 1.1 summarizes our results.

2. Pebble orderability. Many questions in graph pebbling concern the existence of a sequence of pebbling moves with certain properties. There is a natural

temptation to search for such sequences directly, by deciding which pebbling move to make first, which to make second, and so forth. In this section, we develop tools that allow us more flexibility in constructing sequences of pebbling moves. In particular, our goal is to worry only about which moves we should make and not the order in which to make them.

We define the *signature* of a sequence of pebbling moves σ in a graph G to be the directed multigraph on vertex set $V(G)$, where the multiplicity of an edge uv is the number of times σ pebbles from u to v. We say that a digraph D is *orderable* under a pebble distribution p if some ordering of $E(D)$ is a valid sequence of pebbling moves, starting at p. When p is clear from the context, we may simply write that D is orderable instead of D is orderable under p. We characterize when D is orderable under p. We call the problem of testing whether D is orderable under p ORDERABLE, or PO for short.

As it turns out, two conditions which are necessary for D to be orderable under p are also sufficient. Suppose that D is orderable and consider a vertex v. We note that v begins with $p(v)$ pebbles, D pledges that v will receive $d_D^-(v)$ pebbles from pebbling moves into v, and D requests $d_D^+(v)$ pebbling moves out of v. Because each pebbling move out of v costs two pebbles, it is clear that $p(v) + d_D^-(v)$ is at least $2d_D^+(v)$. This leads us to define the *balance* of a vertex v as

$$\text{balance}(D, p, v) = p(v) + d_D^-(v) - 2d_D^+(v).$$

The balance of v is simply the number of pebbles that remain on v after executing any sequence of pebbling moves whose signature is D; that is, for any σ whose signature is D, we have that $p_\sigma(v) = \text{balance}(D, p, v)$.

If D is orderable under p, then the balance of each vertex must be nonnegative. We call this condition the *balance condition*. The balance condition alone is not sufficient: If D is a directed cycle and each vertex has one pebble, then the balance of each vertex is zero, but we cannot make any pebbling moves, and thus D is not orderable. However, as was implicitly observed by Moews [8], if D is acyclic, then the balance condition is sufficient.

THEOREM 2.1 (acyclic orderability characterization; see [8]). *If D is an acyclic digraph with distribution p, then D is orderable under p if and only if the balance condition is satisfied.*

Proof. We have observed that the balance condition is necessary. Conversely, if the balance condition is satisfied, then we obtain a sequence of pebbling moves σ whose signature is D by iteratively selecting a source u in D, making all pebbling moves out of u, and deleting u from D. ∎

Despite the simplicity of the acyclic orderability characterization, we are already able to obtain one of our most useful corollaries. It makes precise our intuition that if we are trying to place pebbles on a target vertex r, it is never advantageous to pebble around in a cycle. Our proof is similar to that of Moews [8].

COROLLARY 2.2 (no cycle lemma; see [4, 8]). *Suppose that D is orderable under p. There exists an acyclic $D' \subseteq D$ such that D' is orderable under p and* $\text{balance}(D', p, v) \geq \text{balance}(D, p, v)$ *for all v.*

Proof. Let D' be a digraph obtained by iteratively removing cycles from D until no cycles remain. Observe that removing a cycle C does not change the balance of vertices outside of C but does increase the balance of vertices in C by one. It follows that $\text{balance}(D', p, v) \geq \text{balance}(D, p, v) \geq 0$ for all v. Hence, D' is acyclic and satisfies the balance condition. By the acyclic orderability characterization, D' is orderable. ∎

In most contexts, if a sequence σ of pebbling moves satisfies certain criteria, then so will any sequence σ' which satisfies $p_{\sigma'} \geq p_\sigma$. As we have seen, in these situations, we are able to restrict our attention to sequences of pebbling moves whose signatures are acyclic. Indeed, all of our major results fall into this category and therefore require only the orderability characterization for acyclic digraphs.

Nevertheless, one may wish to study the existence of sequences of pebbling moves which purposefully remove pebbles from the graph, as in the ANNIHILATION decision problem. Let us return to our orderability characterization for arbitrary D. As we have seen, in general the balance condition is not sufficient. However, as we show in our next lemma, a directed cycle with one pebble on each vertex is the only minimal, nontrivial situation which satisfies the balance condition and does not allow us to make any pebbling moves.

LEMMA 2.3. *Suppose that D with distribution p satisfies the balance condition, D is connected, and $e(D) \geq 1$. If we cannot make any pebbling move described by an edge in D, then D is a directed cycle and each vertex has exactly one pebble.*

Proof. Observe that D does not have any source vertices. Indeed, if v were a source, then the balance condition would imply that v has enough pebbles to make all pebbling moves out of v requested by D. Therefore v must have outdegree zero, and so v is an isolated vertex, which contradicts that D is connected and contains an edge.

Let n be the number of vertices in D, let m be the number of edges in D, let $X \subseteq V(D)$ be the set of all sinks, let $Y = V(D) - X$ be the set of all nonsinks, let k be the number of edges with heads in Y and tails in X, and let z be the number of nonsinks that have exactly one pebble. Note that $m = \sum_v d^-(v) = k + \sum_{v \in Y} d^-(v)$ and $m = \sum_v d^+(v) = \sum_{v \in Y} d^+(v)$. Furthermore, for each $v \in Y$, we have that $p(v) \leq 1$; otherwise, $p(v) \geq 2$ and v has outdegree at least one, contradicting that there are no pebbling moves available. It follows that $z = \sum_{v \in Y} p(v)$. Adding the inequality $\text{balance}(D, p, v) \geq 0$ over all $v \in Y$, we obtain

$$\sum_{v \in Y} d^-(v) + \sum_{v \in Y} p(v) \geq 2 \sum_{v \in Y} d^+(v)$$

or, equivalently, $m - k + z \geq 2m$, and thus $m + k \leq z$. Because D has no sources, every vertex has indegree at least one and thus $m \geq n$. Therefore $n \leq m \leq m + k \leq z \leq n$. It follows that $n = m = z$, so that every vertex in D is neither a sink nor a source and has exactly one pebble. Furthermore, because $n = m$, each vertex in D has indegree and outdegree exactly one. It follows that D is a directed cycle. ∎

Of course, any sequence of pebbling moves leaves a pebble somewhere in the graph; therefore if D contains an edge and D is orderable under p, then $\text{balance}(D, p, v) \geq 1$ for some vertex v. In fact, a slight generalization of this observation will serve as our second necessary condition. To develop this condition, we first recall the component digraph.

Let D be a directed multigraph. A strongly connected component A of D is *trivial* if A consists of a single vertex with indegree and outdegree zero. Define $\text{comp}(D)$, the *component digraph* of D, to be the digraph obtained by contracting each strongly connected component of D to a single vertex.

Suppose that D is orderable under p, and consider a sink A in $\text{comp}(D)$. Because A is a sink component, any pebbling move whose source is in A also has its sink in A; it follows that unless A is trivial, then there must be some vertex v in A with $\text{balance}(D, p, v) \geq 1$. We call the condition that every nontrivial sink in $\text{comp}(D)$

contains a vertex of positive balance the *sink condition*. Note that in the directed cycle example, each vertex has balance zero, and thus the sink condition fails.

As we now show, the balance condition together with the sink condition are sufficient for D to be orderable. We require a simple proposition.

PROPOSITION 2.4. *If D is a strongly connected digraph and $D-uv$ is not strongly connected, then* $\mathrm{comp}(D-uv)$ *contains a single sink A, u is in A, and v is not in A.*

THEOREM 2.5 (orderability characterization). *D is orderable under p if and only if*

(1) *(balance condition) every vertex has nonnegative balance, and*
(2) *(sink condition) every nontrivial sink A in* $\mathrm{comp}(D)$ *contains some vertex with balance at least one.*

Proof. We have observed that both conditions are necessary. Assume (1) and (2) hold. We show that D is orderable under p by induction on $e(D)$. If $e(D) = 0$, the statement is trivial. In the remaining cases, we assume that D has at least one edge.

We consider the case that there is a source v in D with outdegree at least one. Because $\mathrm{balance}(D, p, v) \geq 0$, v has enough pebbles to make all the pebbling moves that D requests out of v. Let σ be an arbitrary ordering of these moves and obtain D' from D by removing all edges whose source is v. We argue that D' is orderable under p_σ. It is clear that D' under p_σ satisfies the balance condition. Observe that every sink in $\mathrm{comp}(D')$ either consists of v (and is therefore trivial) or is a sink in $\mathrm{comp}(D)$. It follows that every nontrivial sink in $\mathrm{comp}(D')$ is a nontrivial sink in $\mathrm{comp}(D)$ and hence contains some vertex with balance at least one. By induction, D' is orderable under p_σ. In the remaining cases, we assume that every source in D is an isolated vertex.

Next, we consider the case where $\mathrm{comp}(D)$ contains a source A with outdegree at least one. Let uv be an edge from a vertex u in A to a vertex v outside of A. We check that A under p satisfies both the balance condition and the sink condition. The balance condition follows from observing that A is a source in $\mathrm{comp}(D)$. Because A is strongly connected and $\mathrm{balance}(A, p, u) \geq 2$, we have that A satisfies the sink condition. By induction, there is an ordering σ of $E(A)$ which is a valid sequence of pebbling moves. We argue that $D - E(A)$ is orderable under p_σ. It is clear that $D - E(A)$ under p_σ satisfies the balance condition. Because every nontrivial sink in $\mathrm{comp}(D-E(A))$ is a nontrivial sink in $\mathrm{comp}(D)$, $D-E(A)$ satisfies the sink condition. Because every source in D is an isolated vertex, it must be that there is some edge e in D whose tail is u; this edge e is contained in A. Therefore $D-E(A)$ contains fewer edges than D, so that the inductive hypothesis implies that $D - E(A)$ is orderable under p_σ. In the remaining cases we assume that every source in $\mathrm{comp}(D)$ is an isolated vertex in $\mathrm{comp}(D)$.

Because $\mathrm{comp}(D)$ is acyclic and every source in $\mathrm{comp}(D)$ is an isolated vertex in $\mathrm{comp}(D)$, it follows that D consists of disjoint, strongly connected components. Because D is orderable if and only if each component of D is orderable, we assume without loss of generality that D is a single, strongly connected component. If we can make a pebbling move uv which leaves $D-uv$ strongly connected, then it is clear that $D - uv$ under p_{uv} satisfies both conditions and thus D is orderable.

It remains to consider the case that every possible pebbling move results in a digraph which is no longer strongly connected. By Lemma 2.3, we have that some pebbling move uv is possible.

First, suppose that uv is the only edge out of u. Note that because D is strongly

connected, u must have indegree at least one. Furthermore, because uv is a valid pebbling move, we have $p(u) \geq 2$. It follows that balance$(D, p, u) \geq 1$. It is clear that $D - uv$ under p_{uv} satisfies the balance condition; by Proposition 2.4, we have that it also satisfies the sink condition. By induction $D - uv$ is orderable under p_{uv}.

Otherwise, let $uw \in E(D)$, $w \neq v$. Let z be a vertex in D with balance$(D, p, z) \geq 1$ (we allow $z \in \{u, w, v\}$), let P be a uz-path, and let Q be a zu-path. Observe that $uv \notin P$ or $uw \notin P$. In the former case, u and z are in the same strongly connected component in $D - uv$; in the latter case, u and z are in the same strongly connected component in $D - uw$. By Proposition 2.4 we have that either $D - uv$ under p_{uv} or $D - uw$ under p_{uw} satisfies both conditions. It then follows that D is orderable under p. □

Observe that if D is acyclic, then the sink condition is trivially satisfied, and we recover the acyclic orderability characterization. Our general orderability characterization yields a quick method for checking whether D is orderable, and thus ORDERABLE is in P. As a consequence, we see that ANNIHILATION is in NP.

Before we conclude this section, we use our tools to prove some technical lemmas which will be useful in later sections. We define a *proper sink* to be a sink with indegree at least one.

LEMMA 2.6. *Suppose D is acyclic and orderable under p. Then for any vertex w, there exists $D' \subseteq D$ such that D' is orderable under p and*

$$\begin{aligned}\text{balance}(D', p, w) &= \text{balance}(D, p, w) + 2d_D^+(w) \\ &= p(w) + d_D^-(w).\end{aligned}$$

Additionally, if $d_D^+(w) > 0$ or D has proper sinks other than w, then we may take D' to be a proper subgraph of D.

Proof. Observe that if uv is an edge in D with v a sink, then $D - uv$ satisfies the balance condition. Let D' be a digraph obtained from D by iteratively deleting edges into sinks other than w until no such edges remain. Because D' is acyclic and satisfies the balance condition, the acyclic orderability characterization implies that D' is orderable. Observe that w is a sink, or else D' would contain an edge uv with $v \neq w$ a sink. Furthermore, every edge into w in D remains in D'. It follows that balance$(D', p, w) =$ balance$(D, p, w) + 2d_D^+(w)$. □

Often, we wish to explore the consequences of the existence of a sequence of pebbling moves with certain properties. In many contexts, considering a minimum sequence of pebbling moves with the properties in question provides us with additional structure. For example, the no cycle lemma implies that a minimum sequence of pebbling moves witnessing that p covers q must be acyclic.

LEMMA 2.7 (minimum signatures lemma). *Let σ be a minimum sequence of pebbling moves in G under p which places at least k pebbles on r, where $p(r) \leq k$. If D is the signature of σ, then D is acyclic, contains no proper sinks except possibly r, the outdegree of r is 0, and the indegree of r is $k - p(r)$.*

Proof. By the no cycle lemma, D is acyclic, or else we obtain a shorter sequence of pebbling moves placing at least k pebbles on r. By Lemma 2.6, the outdegree of r is zero and no vertex except possibly r is a proper sink, or again we obtain a shorter sequence.

Because $d_D^+(r) = 0$, we have balance$(D, p, r) = p(r) + d_D^-(r)$. Together with balance$(D, p, r) \geq k$, we have that $d_D^-(r) \geq k - p(r)$. If $d_D^-(r) > k - p(r)$, then balance$(D, p, r) > k$. Obtain D' from D by deleting one edge into r. Notice that D' satisfies the balance condition and furthermore balance$(D', p, r) =$ balance$(D, p, r) -$

$1 \geq k$. It follows from the acyclic orderability characterization that we obtain a shorter sequence. ∎

If we are interested in minimum sequences of pebbling moves that place k pebbles on some vertex r in a set R of target vertices, the structure of these sequences is further constrained. Not only do their signatures obey the conditions found in the minimum signatures lemma, but the outdegree of each vertex in R is bounded.

LEMMA 2.8. *Let σ be a sequence of pebbling moves in G under p that places at least $k > 0$ pebbles on a vertex $r \in R$ which, among all sequences placing at least k pebbles on some vertex in R, minimizes the total number of pebbling moves. Let D be the signature of σ. For each $v \in R$, we have that $d_D^+(v) < k/2$.*

Proof. Observe that D is acyclic, or else we contradict the no cycle lemma. Suppose for a contradiction that there is $v \in R$ with $d_D^+(v) \geq k/2$. Because $k > 0$, we have $d_D^+(v) > 0$ and thus Lemma 2.6 yields a shorter sequence of pebbling moves placing at least k pebbles on v, a contradiction. ∎

3. Pebble reachability. Recall that the pebbling number of a graph $\pi(G)$ is the minimum k such that every vertex is reachable under every distribution of size k. It is natural, then, to explore the decision problem that results when we fix a particular distribution and target vertex; that is, given G, p, and r, is r reachable? We call this problem REACHABLE, or PR for short. As we show, PR is NP-complete, even when the inputs are restricted so that G is bipartite and has maximum degree three, and each vertex starts with at most two pebbles.

Analogously, fixing the distribution in the cover pebbling number $\gamma(G)$ yields another decision problem: Given G and p, does p cover the unit distribution? We call this problem COVERABLE, abbreviated PC. Although deciding whether $\gamma(G) \leq k$ is possible in polynomial time [13, 11], PC is NP-complete.

A sequence of pebbling moves σ is *nonrepetitive* if for every (unordered) pair of vertices $\{u, v\}$, σ contains at most one pebbling move between the vertices u and v. Similarly to PR, we may ask, given G, p, and r, whether r is reachable via a nonrepetitive sequence of pebbling moves. We call this problem NPR (nonrepetitive pebble reachability). We show that NPR is NP-complete. Our reduction is from a restricted form of 3SAT whose instances ϕ are all in a canonical form.

DEFINITION 3.1. *A 3CNF formula ϕ is in* canonical form *if*

(1) *ϕ has at least two clauses,*
(2) *each clause contains two or three variables,*
(3) *each variable appears at most three times in ϕ,*
(4) *each variable appears either once or twice in its positive form, and*
(5) *each variable appears exactly once in its negative form.*

It is well known that 3SAT remains NP-complete when (2)–(3) are required [12]. Suppose ϕ is a 3SAT formula which satisfies (2)–(3) but not necessarily (1), (4), or (5). Indeed, if a variable x always appears in its positive (negative) form in ϕ, we obtain a simpler, equivalent formula by setting x to true (false), thus removing all clauses containing x ($\overline{x}$). If x appears twice in its negative form, we simply switch all negative occurrences of x to positive occurrences and all positive occurrences of x to negative occurrences. In this way, we obtain an equivalent formula ϕ' satisfying (2)–(5). If ϕ' has at least two clauses, then ϕ' satisfies each of (1)–(5); otherwise, ϕ' contains zero clauses and is trivially satisfiable, so we may replace ϕ' with any fixed, satisfiable 3CNF formula in canonical form. We define R3SAT to be this restricted form of 3SAT.

Our reduction from R3SAT to NPR employs several simple gadgets. The AND

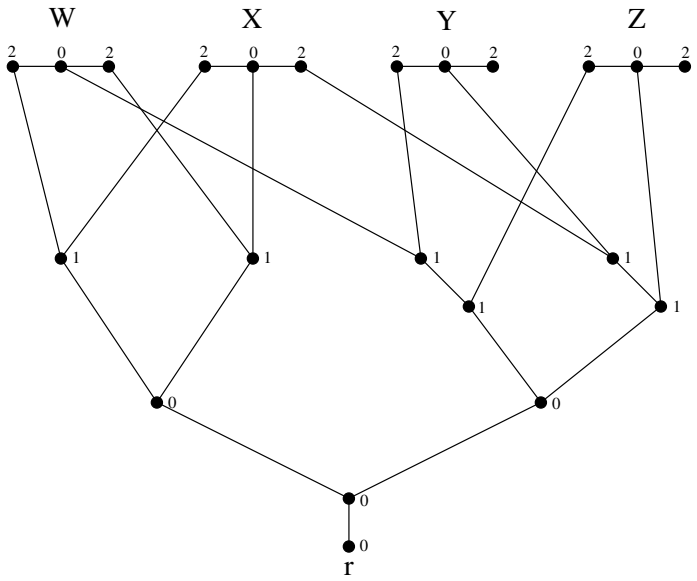


FIG. 3.1. *If* $\phi = (w \vee x) \wedge (w \vee \overline{x}) \wedge (\overline{w} \vee y \vee z) \wedge (x \vee \overline{y} \vee \overline{z})$, *then* $G^{\text{NPR}}(\phi)$ *appears above.*

gadget is a vertex v that has two input edges and one output edge; initially, v is given zero pebbles. Notice that if σ is nonrepetitive and contains a pebbling move from v along the output edge, then σ must contain pebbling moves into v along both input edges. The OR gadget is identical, except that v is initially given a single pebble. In this case, if σ is nonrepetitive and contains a pebbling move from v along the output edge, then σ must contain a pebbling move into v along one of the input edges. Using 2-ary AND (OR) gadgets, one easily constructs k-ary AND (OR) gadgets.

The variable gadget is a path $v_1v_2v_3$. The endpoint vertices $\{v_1, v_3\}$ are initially given two pebbles, and the internal vertex v_2 is initially given zero pebbles. The endpoint vertices correspond to the positive occurrence(s) of the variable in ϕ, and the internal vertex corresponds to the negative occurrence of the variable in ϕ. The variable gadget has two or three output edges, depending upon how many times the corresponding variable appears in ϕ. If x_i appears three times in ϕ, then its associated variable gadget X_i has three output edges, one incident to each v_i. If x_i appears twice in ϕ, then X_i has two output edges, one incident to each of v_1 and v_2. We say that the output edges incident to v_1 and v_3 are *positive output edges*, and the output edge incident to v_2 is the *negative output edge*.

Given an instance ϕ of R3SAT, we construct $G = G^{\text{NPR}}(\phi)$ as follows. For each variable x_i in ϕ, we introduce a variable gadget X_i in G. For each clause c_j containing $k \in \{2, 3\}$ variables, we introduce a k-ary OR gadget C_j. The output edges of the X_i's are identified with the input edges of the C_j's in the natural way: If x_i appears in c_j, a positive output edge of X_i is identified with an input edge of C_j, and if $\overline{x_i}$ appears in c_j, the negative output edge of X_i is identified with an input edge of C_j. The output edges of the C_j's are connected to the input edges of an m-ary AND gadget A, where m is the number of clauses in ϕ. Finally, the output edge of A is connected to the target vertex r.

Example. If $\phi = (w \vee x) \wedge (w \vee \overline{x}) \wedge (\overline{w} \vee y \vee z) \wedge (x \vee \overline{y} \vee \overline{z})$, then $G^{\text{NPR}}(\phi)$ appears in Figure 3.1.

PROPOSITION 3.2. *Let ϕ be an instance of* R3SAT *with n variables and m clauses. Then $G^{\text{NPR}}(\phi)$ has $O(n+m)$ vertices.*

THEOREM 3.3. NPR *is NP-complete, even when G has maximum degree three and each vertex starts with at most two pebbles.*

Proof. It is immediate that NPR is in NP. Let ϕ be an instance of R3SAT and let $G = G^{\text{NPR}}(\phi)$. Observe that each vertex in G starts with at most two pebbles and the maximum degree in G is three.

We claim that ϕ is satisfiable if and only if there is a nonrepetitive sequence of pebbling moves which ends with a pebble on r. Suppose that ϕ is satisfiable via $f : \{x_1, \ldots, x_n\} \to \{\text{true}, \text{false}\}$. We construct a nonrepetitive sequence of pebbling moves which ends with a pebble on r as follows. For each variable x_i with $f(x_i) =$ false, we make a pebbling move from each endpoint of X_i to the interior vertex of X_i. Notice that after executing these pebbling moves, for each x_i with $f(x_i) =$ true, we have two pebbles on each endpoint of X_i, and for each x_i with $f(x_i) =$ false, we have two pebbles on the interior vertex of X_i. Because f satisfies ϕ, each clause gadget C_i has some input edge which is incident to a vertex in a variable gadget with two pebbles. By construction, each vertex in a variable gadget is incident to at most one clause gadget input edge; therefore we are able to make pebbling moves into each clause gadget C_i. By the construction of our clause gadgets, we are then able to make pebbling moves out of each clause gadget and, by construction, along each of the inputs to the m-ary AND gadget. It follows that we are able to make a pebbling move along the output of our AND gadget, which places a pebble on r. It is easily observed that our sequence of pebbling moves is nonrepetitive.

Conversely, suppose that σ is a nonrepetitive sequence of pebbling moves which ends with a pebble on r. We construct a satisfying assignment f as follows. Because σ contains a pebbling move across the output of the AND gadget A, it follows that σ contains pebbling moves across the output of each clause gadget C_i. Hence, for each clause gadget C_i, σ contains a pebbling move across an input edge e_i of C_i. If e_i is incident to an endpoint of X_j, then we set $f(x_j) =$ true; otherwise, if e_i is incident to the interior vertex of X_j, we set $f(x_j) =$ false. We claim that we do not attempt to set both $f(x_j) =$ true and $f(x_j) =$ false. Indeed, if we set $f(x_j) =$ false, then σ contains a pebbling move out of the interior vertex v of X_j along an input edge to some clause gadget. Because σ is nonrepetitive, v starts with zero pebbles, and v has degree three, it must be that σ contains pebbling moves from each of the endpoints in X_j into v. Because each endpoint of X_j starts with only two pebbles and σ is nonrepetitive, the moves into v are the only pebbling moves which originate from the endpoints of X_j. Therefore σ does not contain a pebbling move out of an endpoint of X_j along an input edge of a clause gadget, and hence we never attempt to set $f(x_j) =$ true. If the truth values for any variables remain unset, we set them arbitrarily. Now f witnesses that ϕ is satisfiable. ∎

One of the major tools available to us when designing interesting graph pebbling problems is the path; on a path, the pebbling moves available to us are rather limited. If we are in a situation where we need not concern ourselves with pebbling in cycles, then our options on a path become even more limited. Furthermore, if the path is long, it may be difficult to pebble across. Before using paths to reduce NPR to PR, we explore some basic properties.

LEMMA 3.4. *Let G be a graph which contains an induced path $P = v_0, \ldots, v_{n+1}$ containing $n+2$ vertices, and suppose that each of the n internal vertices in P contains c pebbles. Let D be an acyclic signature of a sequence of pebbling moves so that*

the edge v_1v_0 has multiplicity $a_0 \geq c$. Then the multiplicity of $v_{n+1}v_n$ is at least $2^n(a_0 - c) + c$.

Proof. Observe that the claim is trivial if $a_0 = 0$; we assume that $a_0 \geq 1$. For $1 \leq i \leq n$, let a_i be the multiplicity of $v_{i+1}v_i$. We claim that for all $1 \leq i \leq n$, we have that

(1) $a_i + c \geq 2a_{i-1}$, and
(2) $a_i \geq a_0$.

Suppose for a contradiction that $i \geq 1$ is the least integer for which (1) or (2) fails, and consider the vertex v_i. By our selection of i, $a_{i-1} \geq a_0$ and therefore D requests at least a_0 pebbling moves out of v_i along edge v_iv_{i-1}. Because $a_0 \geq c$ and $a_0 \geq 1$, we have that $2a_0 > c$; hence, by the balance condition at v_i, the indegree of v_i in D is at least one. Because D is acyclic, D contains no edges of the form $v_{i-1}v_i$. Because v_i is an internal vertex in an induced path in G, the only other edge incident to v_i is v_iv_{i+1}. It follows that the indegree of v_i in D is exactly the multiplicity of $v_{i+1}v_i$, and so the indegree of v_i in D is a_i. Therefore the balance condition at v_i implies that $a_i + c \geq 2a_{i-1}$, which, together with $a_0 \geq c$ and $a_{i-1} \geq a_0$, implies $a_i \geq a_0$.

Solving our recurrence in (1), we find that $a_i \geq 2^i(a_0 - c) + c$. ∎

We use our path lemma to argue that if we can pebble across a long path several times, then we can place many pebbles on the originating endpoint of the path. Using Lemma 2.6, we obtain the following corollary.

COROLLARY 3.5. *Under the assumptions of Lemma* 3.4, *there exists $D' \subseteq D$ such that D' is orderable and* $\text{balance}(D', p, v_{n+1}) \geq 2^{n+1}(a_0 - c) + 2c$. *If in addition we have $d^+_D(v_{n+1}) > 0$, then we may take D' to be a proper subgraph of D.*

Our reduction used the notion of nonrepetitive sequences of pebbling moves. In fact, there is a natural correspondence between the nonrepetitive sequences of pebbling moves in a graph G and (arbitrary) sequences of pebbling moves in another graph $\mathcal{S}(G, \alpha)$.

DEFINITION 3.6. *We obtain $\mathcal{S}(G, \alpha)$ from G by replacing each edge in G with a path containing α internal vertices so that $d_{\mathcal{S}(G,\alpha)}(u, v) = (1 + \alpha)\, d_G(u, v)$ for each pair u, v of vertices of G. We call these paths* one use paths.

As our next lemma shows, the correspondence holds whenever α is sufficiently large with respect to the number of pebbles in G.

LEMMA 3.7. *Fix a graph G and a parameter $t \geq 0$. Suppose that*

$$\alpha \geq \max\{\lg 2t,\, 4 \lg e(G)\}$$

and let $H = \mathcal{S}(G, \alpha)$. Let p be a pebble distribution on G of size at most t and define a pebble distribution q on H so that q and p agree on $V(G)$ and q assigns one pebble each to the internal vertices of H's one use paths. We have the following claims.

(1) *If σ is a nonrepetitive sequence of pebbling moves in G, then there exists a sequence of pebbling moves σ' in H such that p_σ and $q_{\sigma'}$ agree on $V(G)$.*
(2) *Conversely, if σ is a sequence of pebbling moves in H, then there exists a nonrepetitive sequence of pebbling moves σ' in G such that $p_{\sigma'}(v) \geq q_\sigma(v)$ for all v in G.*

Proof. Claim 1 is clear. Suppose that σ is a sequence of pebbling moves in H. By the no cycle lemma, we may assume without loss of generality that the signature D of σ is acyclic. We define a digraph D' with vertex set $V(G)$ as follows. Let uv be an edge in G and let $u = w_0, \ldots, w_{\alpha+1} = v$ be the corresponding one use path in H. The multiplicity of the edge uv in D' is the multiplicity of the edge $w_\alpha w_{\alpha+1}$ in D. Because D is acyclic, the balance condition implies that if D contains the edge

$w_\alpha w_{\alpha+1}$, then D contains all edges $w_k w_{k+1}$. It follows that D' is also acyclic. It is easily seen that balance$(D', p, v) \geq$ balance(D, q, v) for each v in D'. By the acyclic orderability characterization, we obtain a sequence of pebbling moves σ' such that $p_{\sigma'}(v) \geq q_\sigma(v)$ for all v in G. It remains to show that D' has no edges of multiplicity at least two, so that σ' is nonrepetitive.

Suppose for a contradiction that uv is an edge in D' with multiplicity at least two; again, let $u = w_0, \ldots, w_{\alpha+1} = v$ be the corresponding one use path in H. It follows that $w_\alpha w_{\alpha+1}$ has multiplicity at least two in D. Recalling that q assigns each of the internal vertices w_i one pebble, Lemma 3.4 implies that the multiplicity of $w_0 w_1$ is at least $2^\alpha + 1$. Because each pebbling move reduces the total number of pebbles by one, certainly the size of q is at least $2^\alpha + 2$. But $|q| = |p| + \alpha e(G)$ and together with $t \leq 2^{\alpha-1}$ and $\alpha e(G) \leq 2^{\alpha-1}$, we obtain a contradiction. ∎

COROLLARY 3.8. REACHABLE *is NP-complete, even when G is bipartite and has maximum degree three, and each vertex starts with at most two pebbles.*

Proof. By the no cycle lemma and the acyclic orderability characterization, PR is in NP. We reduce the fragment of NPR targeted by our reduction from R3SAT to NPR as follows. Consider a graph G with maximum degree three, a distribution of pebbles p which places at most two pebbles on each vertex in G, and a target vertex r. Let α be the least odd number larger than $\max\{\lg 2\,|p|\,,\, 4 \lg e(G)\}$. Our reduction outputs $H = \mathcal{S}(G, \alpha)$ with pebble distribution q as in Lemma 3.7 and target vertex r. Observe that H is bipartite and has maximum degree three, and each vertex starts with at most two pebbles. By Lemma 3.7, r is reachable via a nonrepetitive sequence of pebbling moves in G if and only if r is reachable in H. ∎

Let ϕ be an instance of R3SAT. We define $G^{\text{PR}}(\phi) = \mathcal{S}(G^{\text{NPR}}(\phi), \alpha)$ with α chosen as in our corollary; that is, G^{PR} is the composition of our reduction from R3SAT to NPR and our reduction from NPR to PR.

COROLLARY 3.9. COVERABLE *is NP-complete, even when G is bipartite and has maximum degree three, and each vertex starts with at most three pebbles.*

Proof. By the no cycle lemma and the acyclic orderability characterization, we have that PC is in NP. We reduce PR to PC as follows. Let G be a graph with pebble distribution p and target vertex r. Define a new distribution q of pebbles so that $q(v) = p(v) + 1$ for all $v \neq r$ and $q(r) = p(r)$. We claim that r is reachable under p if and only if q covers the unit distribution. The forward direction is clear.

Suppose that σ is a minimum sequence of pebbling moves witnessing that q covers the unit distribution, and let D be the signature of σ. By the no cycle lemma, D is acyclic. Because balance$(D, q, v) \geq 1$, we have that balance$(D, p, v) \geq 0$ for all v and balance$(D, p, r) \geq 1$. It follows from the acyclic orderability characterization that D is orderable under p. Together with balance$(D, p, r) \geq 1$, we have that r is reachable under p. ∎

As we have seen, REACHABLE is NP-complete, even under some restrictions of the inputs. However, as we now observe, if we restrict G to be a tree, then we can solve REACHABLE in polynomial time using a simple greedy strategy. A *greedy pebbling move* is a pebbling move uv such that $d(v, r) < d(u, r)$. Moews established that the maximum number of pebbles that can be placed on a target vertex r in T is achievable using greedy pebbling moves [8]. We extend this slightly and argue that every maximal sequence of greedy pebbling moves places the maximum possible number of pebbles on r. The *greedy pebbling strategy* arbitrarily makes greedy pebbling moves until no greedy pebbling move is possible.

PROPOSITION 3.10 (greedy tree lemma; see [8]). *In a tree T with target r, the*

maximum number of pebbles that can be placed on r is achieved with the greedy pebbling strategy.

Proof. Suppose for a contradiction that under p, it is possible to place k pebbles on r, but if we make the greedy pebbling move uv, it is no longer possible to place at least k pebbles on r. Let σ be a minimum sequence of pebbling moves placing k pebbles on r, and let D be the signature of σ. By the no cycle lemma, D is acyclic. If D contains the edge uv, then the acyclic orderability characterization implies that $D - uv$ is orderable under p_{uv}, implying that it is possible to place k pebbles on r even after pebbling uv. Otherwise, if D does not contain the edge uv, then $d^+(u) = 0$, or else D contains a proper sink other than r, contradicting the minimum signatures lemma. Therefore σ does not contain any pebbling moves out of u, and so uv followed by σ is a legal sequence of pebbling moves placing at least k pebbles on r. ∎

4. Complexity of optimal pebbling number. Recall that the optimal pebbling number $\widehat{\pi}(G)$ of a graph G is the least number k such that there exists a distribution of size k under which every vertex is reachable. As in the introduction, we define OPTIMAL-PEBBLING-NUMBER (abbreviated OPN) to be the problem of deciding, given G and k, whether $\widehat{\pi}(G) \leq k$. In this section, we show that OPN is NP-complete. We observe that OPN is in NP; indeed, we may witness that $\widehat{\pi}(G) \leq k$ by providing a distribution p of size k and, for each r, the signature D_r of a sequence of pebbling moves showing that r is reachable. More care is needed to establish that OPN is NP-hard. As in our proof that PR is NP-hard, we establish that OPN is NP-hard through an intermediate decision problem.

Let G be a graph and let p be a distribution of pebbles to G. A vertex r is *determinative* if r being reachable under p implies that every vertex in G is reachable under p. Informally, if r is determinative, then no vertex in G is more difficult to pebble than r. Our intermediate decision problem is REACHABLE with the added restriction that r is determinative. We call this problem DPR (determinative pebble reachability).

PROPOSITION 4.1. DPR *is NP-complete, even when each vertex starts with at most two pebbles.*

Proof. Because REACHABLE is in NP, it is immediate that DPR is in NP as well. We show that our reduction G^{PR} from R3SAT to PR actually produces an instance of DPR. Let ϕ be an instance of R3SAT, and let $G = G^{\text{PR}}(\phi)$ with distribution p and target r. We show that r is determinative. Suppose that it is possible to place a pebble on r or, equivalently, that ϕ is satisfiable. Consider a vertex $v \in G$. If v is an internal vertex in a one use path introduced in our reduction from NPR to PR, then v begins with one pebble and thus v is reachable trivially.

It remains to consider the case that v is a vertex introduced in our reduction from R3SAT to NPR, so that v is either in an OR gadget, in a variable gadget, in an AND gadget, or equal to r. If v is in an OR gadget, then v begins with a pebble. If v is an endpoint of a variable gadget, then v begins with two pebbles. If v is the interior vertex of a variable gadget, then we may place a pebble on v by pebbling from either of the endpoints (which start with two pebbles) across the one use path. Otherwise, if v is in an AND gadget or $v = r$, then we use the satisfiability of ϕ to place a pebble on v. ∎

Before we are able to present our reduction from DPR to OPN, we require some technical lemmas. The following weighting argument is well known and is a fundamental tool in graph pebbling.

PROPOSITION 4.2 (standard weight equation). *Let G be a graph with distribution*

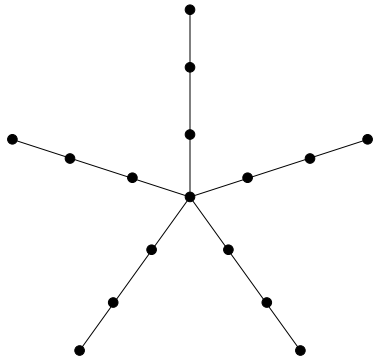


FIG. 4.1. star$(3, 5)$.

p and target vertex r, and let a_i be the number of pebbles at distance i from r. If it is possible to place s pebbles on r, then we have $\sum_{i\geq 0} 2^{-i} a_i \geq s$.

Proof. Observe that it is not possible to make a pebbling move which increases the sum $\sum_{i\geq 0} 2^{-i} a_i$. □

The following graph will be useful to us in two different contexts: first, as a gadget and second, in establishing the correctness of our reduction from DPR to OPN.

DEFINITION 4.3. *We define* star(α, β) *to be the result of replacing each edge in $K_{1,\beta}$ with a path containing α edges so that* star(α, β) *has $\alpha\beta$ edges. Equivalently,* star$(\alpha, \beta) = \mathcal{S}(K_{1,\beta}, \alpha - 1)$.

Example. star$(3, 5)$ appears in Figure 4.1.

Our reduction from DPR to OPN produces a graph whose global structure is similar to that of star$(\cdot)$. Our instance of DPR plays the role of the center vertex, and the gadgets that we add play the role of the leaves. When we argue the correctness of our reduction, we apply the following lemma to limit the pebble distributions that we must consider. The lemma shows that, despite its simplicity, the standard weight equation can yield nontrivial results.

LEMMA 4.4. *Fix $\alpha \geq 1$ and $\beta \geq 2$. Let p be a distribution of $\beta 2^\alpha$ pebbles to* star(α, β) *with the property that for each leaf l in* star(α, β)*, it is possible to place 2^α pebbles on l. If $2(\beta^2 + 1) < 2^\alpha$, then p is the distribution which places 2^α pebbles on each leaf and zero pebbles on the other vertices.*

Proof. Let v be the center vertex of star(α, β) and, for each $0 \leq i \leq \alpha$, let a_i be the number of pebbles at distance i from v. For each leaf l, it is possible to place 2^α pebbles on l, and Proposition 4.2 yields an equation; we sum these equations. Because there are β leaves, we obtain $\beta 2^\alpha$ on the right-hand side. A pebble at distance i from v is at distance $\alpha - i$ from its closest leaf and $\alpha + i$ from all other leaves. It follows that pebbles at distance i from v contribute $1/2^{\alpha-i} + (\beta - 1)/2^{\alpha+i}$ to the left-hand side of the equation. We obtain

$$\sum_{i=0}^{\alpha} \left(\frac{1}{2^{\alpha-i}} + \frac{\beta - 1}{2^{\alpha+i}} \right) a_i \geq \beta 2^\alpha$$

and, after some simplification,

$$\sum_{i=0}^{\alpha} \left(2^i + \frac{\beta - 1}{2^i} \right) a_i \geq \beta 4^\alpha.$$

Let $f(x) = 2^x + (\beta - 1)2^{-x}$ so that pebbles at distance i contribute $f(i)$ to the left-hand side. Analyzing the derivative $f'(x) = \ln 2\,(2^x - (\beta-1)2^{-x})$, we find that $f'(x) = 0$ has one solution, namely $x_0 = \log_4(\beta - 1)$. Furthermore, for $x > x_0$, we have $f'(x) > 0$ and for $x < x_0$, we have $f'(x) < 0$. It follows that $f(x)$ has a global minimum at $x = x_0$, $f(x)$ is decreasing on $(-\infty, x_0]$, and $f(x)$ is increasing on $[x_0, \infty)$.

Let $m = \sum_{i=0}^{\alpha-1} a_i$ be the number of pebbles not at distance α from v; we show that $m < 1$, implying that $m = 0$. Noting that $a_\alpha = \beta 2^\alpha - m$, we have that

$$\left(\max_{0\le i\le \alpha-1} f(i)\right) m + f(\alpha)\,(\beta 2^\alpha - m) \ge \beta 4^\alpha.$$

Because of the monotonicity properties of f, we have $\max_{0\le i\le\alpha-1} f(i) \in \{f(0), f(\alpha-1)\}$. Because $2(\beta^2+1) < 2^\alpha$, certainly $2\beta < 2^\alpha$ and therefore

$$f(0) = \beta < 2^{\alpha-1} \le 2^{\alpha-1} + (\beta-1)2^{1-\alpha} = f(\alpha-1).$$

It follows that $\max_{0\le i\le\alpha-1} f(i) = f(\alpha-1)$. Observe that $f(\alpha)-f(\alpha-1) = 2^{\alpha-1}-(\beta-1)/2^\alpha$. Because $\beta - 1 < 2(\beta^2+1) < 2^\alpha$, we have that $f(\alpha)-f(\alpha-1) > 2^{\alpha-1}-1 \ge 0$. After substitution and further simplification, we obtain

$$m \le \frac{\beta 2^\alpha\,(f(\alpha) - 2^\alpha)}{f(\alpha) - f(\alpha-1)}.$$

Substituting our formula for $f(\alpha)$ into the numerator yields

$$m \le \frac{\beta(\beta-1)}{f(\alpha)-f(\alpha-1)} \le \frac{\beta^2}{f(\alpha)-f(\alpha-1)}.$$

Because $2(\beta^2+1) < 2^\alpha$, we have that $\beta^2 < 2^{\alpha-1}-1$; recalling that $f(\alpha)-f(\alpha-1) > 2^{\alpha-1}-1$, we have $m < 1$ as required.

It follows that p places every pebble at distance α from v. It remains to show that p places 2^α pebbles on each leaf. Fix an arbitrary leaf l, and let n be the number of pebbles that p places on l. Applying the standard weight equation to l, we have that

$$n + \frac{\beta 2^\alpha - n}{2^{2\alpha}} \ge 2^\alpha.$$

After simplification, we obtain that

$$n \ge 2^\alpha - \frac{2^\alpha(\beta-1)}{4^\alpha - 1}.$$

Similarly to the previous paragraph, we show that $n > 2^\alpha - 1$. We have that

$$\frac{2^\alpha(\beta-1)}{4^\alpha-1} \le \frac{2^\alpha(\beta-1)}{4^\alpha - 2^\alpha} = \frac{\beta-1}{2^\alpha-1}.$$

Because $\beta \le 2(\beta^2+1) < 2^\alpha$, we have that $(\beta-1)/(2^\alpha-1) < 1$ and hence $n > 2^\alpha - 1$ as required. Therefore p assigns each leaf at least 2^α pebbles, and the lemma follows. □

We now have the tools necessary to present our reduction from DPR to OPN. Let G be a graph with pebble distribution p and determinative target vertex r. Let $m = |p|$,

let $\alpha = \lceil \lg\left(2(m^2+1)+1\right) \rceil$, and let $\beta = 2^\alpha m + 2$. We construct a graph H with the property that $\widehat{\pi}(H) \leq m2^\alpha$ if and only if r is reachable in G.

We construct H from G by attaching a copy of $\text{star}(\alpha, \beta)$ to each pebble in G. That is, for each pebble on a vertex u, we introduce a copy of $\text{star}(\alpha, \beta)$ and attach it to u by identifying u with one of the leaves of our copy of $\text{star}(\alpha, \beta)$.

LEMMA 4.5. *r is reachable in G under p if and only if $\widehat{\pi}(H) \leq m2^\alpha$.*

Proof. ($\implies$). Suppose r is reachable. Define a distribution q of $m2^\alpha$ pebbles to H by placing 2^α pebbles at the centers of each of the m copies of $\text{star}(\alpha, \beta)$ in H. Consider a vertex v in H. If v belongs to a copy S of $\text{star}(\alpha, \beta)$, then v is at a distance at most α from the center of S; because the center of S begins with 2^α pebbles, v is reachable. Otherwise, v must be a vertex in G. Because r is reachable and determinative under p, to show that v is reachable, it suffices to show that q covers p. But each star can contribute one pebble to the vertex it shares with G, and thus q covers p.

($\impliedby$). Let q be a distribution of $m2^\alpha$ pebbles to H witnessing that $\widehat{\pi}(H) \leq m2^\alpha$. We claim that if u is the center vertex of a copy S of $\text{star}(\alpha, \beta)$, then it is possible to place 2^α pebbles on u starting from q. Indeed, because S contains $\beta - 1 > m2^\alpha$ pendant paths with endpoint u, there is some path to which q assigns no pebbles (except possibly at u). Let $w_0 w_1, \ldots, w_\alpha$ be one such path with $w_0 = u$. Because every vertex is reachable under q, certainly w_α is reachable; let D be a signature of a minimum sequence of pebbling moves that places a pebble on w_α. Because w_α is a leaf and q assigns no pebbles to w_α, $w_{\alpha-1}w_\alpha$ is an edge in D; therefore Corollary 3.5 implies that we can place 2^α pebbles on u.

When a graph has a pebble distribution, contracting a set of vertices S changes the pebble distribution in the natural way: Pebbles on vertices in S are collected at the vertex of contraction. Construct H' and pebbling distribution q' from H and q by applying the following contractions:

(1) Contract all vertices in H that are also in G to a single vertex v.
(2) For each copy S of $\text{star}(\alpha, \beta)$, contract the vertices in S that are at distance at least α from v.

Observe that H' is exactly $\text{star}(\alpha, m)$, with center vertex v. Because the contraction operation cannot make pebbling more difficult, it is possible to place 2^α pebbles on each leaf in H' starting from q'. Because $2(m^2+1) < 2^\alpha$, applying Lemma 4.4 to $H' = \text{star}(\alpha, m)$ implies that q' must assign 2^α pebbles to each leaf of H'. It follows that q assigns 2^α pebbles to each copy of $\text{star}(\alpha, \beta)$ in H in such a way that each pebble is at a distance at least α away from the vertices in G.

Let E be the signature of a minimum sequence of pebbling moves in H starting from q which places a pebble on r. Consider a copy S of $\text{star}(\alpha, \beta)$ attached to a vertex u in G. We claim that E contains at most one edge from S into u. Indeed, if this were otherwise, then by Corollary 3.5 there would exist $E' \subseteq E$ which placed at least $2 \cdot 2^\alpha$ pebbles on the center vertex of S. However, this is impossible because E is acyclic with edges from S into G, and q assigns only 2^α pebbles to S.

Obtain E' from E by deleting all edges except those in G. Because each vertex u in G receives a pebble from p for every attached copy of $\text{star}(\alpha, \beta)$, we have that $\text{balance}(E', p, u) \geq \text{balance}(E, q, u)$ for all vertices u of G. It follows from the acyclic orderability characterization that E' is orderable under p; together with $\text{balance}(E', p, r) \geq \text{balance}(E, q, r) \geq 1$, we have that r is reachable under p. ∎

We conclude with this section's main theorem.

THEOREM 4.6. OPTIMAL-PEBBLING-NUMBER *is NP-complete.*

Proof. We have already observed that OPN is in NP and exhibited a reduction from DPR to OPN. It remains to check that our reduction is computable in polynomial time for the restricted class of DPR shown to be NP-hard in Proposition 4.1. Consider an instance G, p, r of DPR with $n = n(G)$ and $m = |p|$ at most $2n$. It suffices to show that the pair $H, m2^{\alpha}$ produced by our reduction is not too large. Our reduction uses gadgets $\mathrm{star}(\alpha, \beta)$ with $\alpha \leq \lceil \lg(2(4n^2+1)+1) \rceil$ and $\beta \leq 2^{\alpha}m+2 = O(n^3)$. It follows that each gadget $\mathrm{star}(\alpha, \beta)$ has at most $O(n^3 \log n)$ vertices. Because our reduction introduces at most $2n$ gadgets, H contains a total of at most $n + 2nO(n^3 \log n) = O(n^4 \log n)$ vertices. □

5. Complexity of pebbling number. Although the optimal pebbling number has received some study, combinatorialists have focused more attention on the pebbling number. Recall that the r-pebbling number $\pi(G, r)$ is the minimum k such that r is reachable under every distribution of size k. Similarly, the pebbling number $\pi(G)$ is the minimum k such that every vertex is reachable under every distribution of size k. It is clear from the definitions that if n is the number of vertices in G, then $\widehat{\pi}(G) \leq n \leq \pi(G)$. At first glance, it may not be clear that $\pi(G)$ is well defined. In fact, if G is not connected, then we can place arbitrarily many pebbles in a single component and we will not be able to place pebbles on vertices outside the component. However, for connected graphs, $\pi(G)$ is well defined; we implicitly assume that G is connected. Indeed, if d is the diameter of G, every vertex is reachable provided that our distribution is forced to place at least 2^d pebbles on some vertex. We record this observation as a proposition.

PROPOSITION 5.1. *Let G be a graph with diameter d. Then $\pi(G) \leq (2^d - 1)\, n+1$.*

We call the problem of deciding whether $\pi(G, r) \leq k$ R-PEBBLING-NUMBER (abbreviated RPN); similarly, we define PEBBLING-NUMBER (abbreviated PN) to be the problem of deciding whether $\pi(G) \leq k$. In this section, we establish that PN and RPN are Π_2^{P}-complete. First, note that both languages are in Π_2^{P}. Indeed, to decide if $\pi(G) \leq k$, our machine need only check that for all distributions p of size k and all target vertices r, there exists a digraph $D_{p,r}$ orderable under p that places a pebble on r. The distributions of size k, the target vertices, and the digraphs $D_{p,r}$ are all describable using $\mathrm{poly}(n, \log k)$ bits. Further, ORDERABLE is in P. It follows that PN is in Π_2^{P}. A similar argument shows that RPN is in Π_2^{P}.

The seminal Π_2^{P}-complete problem is a quantified version of 3SAT whose instances consist of a 3CNF formula ϕ over a set of universally quantified variables and a set of existentially quantified variables (see [10]). We say that ϕ is *valid* if for every setting of the universally quantified variables, there is a setting of the existentially quantified variables which satisfies ϕ. The decision problem $\forall\exists$3SAT is to determine whether ϕ is valid. Just as 3SAT remains NP-complete when ϕ is restricted to be in canonical form (recall Definition 3.1), $\forall\exists$3SAT remains Π_2^{P}-complete when ϕ is restricted to be in canonical form. We call this restriction R$\forall\exists$3SAT.

We show that RPN is Π_2^{P}-complete by a reduction from R$\forall\exists$3SAT. Whereas our reduction to OPN produces graphs H with the property that only one distribution can possibly succeed in witnessing $\widehat{\pi}(H) \leq k$, our reduction to RPN produces graphs with the property that almost all distributions succeed in being able to place a pebble on r. It is the rare "difficult" distributions—those which may not allow a pebble to be placed on r—that correspond to settings of the universally quantified variables in our R$\forall\exists$3SAT formula. Given a distribution of k pebbles to the graph we produce, either r is easily reachable or the distribution corresponds to a setting f of the universally quantified variables in ϕ, and r is reachable if and only if ϕ is satisfiable under f.

Our reduction from R∀∃3SAT to RPN involves the construction of several graphs, each building on the previous construction. We refer to the ith graph we produce as $G_i = G_i(\phi)$. We present the reduction with respect to a fixed instance ϕ of R∀∃3SAT.

5.1. The underlying graph. We obtain G_1 from ϕ by modifying $G^{\text{NPR}}(\phi)$ slightly. That is, for each universally quantified variable x_i in ϕ, we remove both edges from the variable gadget X_i in $G^{\text{NPR}}(\phi)$ associated with x_i and remove one pebble each from the endpoints of X_i so that the endpoints of X_i start with one pebble instead of two. (We leave intact variable gadgets X_j corresponding to existentially quantified variables x_j in ϕ.) Let $n_1 = n(G_1)$ be the number of vertices in G_1, let $e_1 = e(G_1)$, and let p_1 be the distribution on G_1. The following definition gives the correspondence between settings of the universally quantified variables in ϕ and distributions of pebbles in G_1.

DEFINITION 5.2. *For each setting f of the universally quantified variables in ϕ, let $p_{1,f}$ be the distribution of pebbles to G_1 given by adding the following pebbles to p_1. For each x_i with $f(x_i) = true$, add one pebble to each of the two vertices associated with positive occurrences of x_i in ϕ. For each x_i with $f(x_i) = false$, add two pebbles to the vertex associated with the negative instance of x_i.*

Observe that under any $p_{1,f}$, each vertex in G_1 contains at most two pebbles. Our interest in G_1 under the distributions $p_{1,f}$ is based on the following proposition, whose proof is similar to that of Theorem 3.3.

PROPOSITION 5.3. *There is a nonrepetitive sequence of pebbling moves which places a pebble on r in G_1 starting from $p_{1,f}$ if and only if there is a setting of the existentially quantified variables in ϕ which, together with f, satisfies ϕ.*

Note that for any two settings f, f' of the universally quantified variables in ϕ, $|p_{1,f}| = |p_{1,f'}|$; let $t = |p_{1,f}|$. Because $p_{1,f}$ assigns at most two pebbles to each vertex in G_1, $t \leq 2n_1$. We obtain G_2 from G_1 by setting $\alpha = \lceil \max\{\lg 2t, 4 \lg e_1\} \rceil$ and replacing each edge in G_1 with a path of length $\alpha + 1$; that is, $G_2 = \mathcal{S}(G_1, \alpha)$ (recall Definition 3.6). Let n_2 be the number of vertices in G_2.

Let p_2 be the distribution of pebbles to G_2 so that p_2 and p_1 agree on all vertices in G_1, and $p_2(v) = 1$ for all vertices v introduced in our construction of G_2 from G_1. Similarly, let $p_{2,f}$ be the distribution of pebbles to G_2 so that $p_{2,f}$ and $p_{1,f}$ agree on all vertices in G_1, and $p_{2,f}(v) = 1$ for all vertices v introduced in our construction of G_2 from G_1.

We call G_2 the *underlying graph* and a distribution $p_{2,f}$ an *underlying distribution*. Observe that by Lemma 3.7, there is a nonrepetitive sequence of pebbling moves which places a pebble on r in G_1 under $p_{1,f}$ if and only if there is an arbitrary sequence of pebbling moves in G_2 under $p_{2,f}$ which places a pebble on r. Together with Proposition 5.3, this results in the following proposition.

PROPOSITION 5.4. *There is a sequence of pebbling moves which places a pebble on r in G_2 starting from $p_{2,f}$ if and only if there is a setting of the existentially quantified variables in ϕ, which, together with f, satisfies ϕ.*

One useful property of the underlying graph together with an underlying distribution is that it is not possible to accumulate more than five pebbles on any vertex. This property will be instrumental in arguing that the gadgets we attach to the underlying graph behave correctly.

PROPOSITION 5.5. *It is not possible to place more than five pebbles on any vertex in G_2 starting from any $p_{2,f}$.*

Proof. Suppose for a contradiction that it is possible to place at least six peb-

bles on a vertex u in G_2. First, suppose u is a vertex introduced in our construction of G_2 from G_1 so that u is an internal vertex w_i, $1 \leq i \leq \alpha$, in a one use path $P = w_0 w_1, \ldots, w_\alpha w_{\alpha+1}$. Let D be a signature of a minimum sequence of pebbling moves which places at least six pebbles on w_i. Because $p_{2,f}(w_i) = 1$, to have balance$(D, p_{2,f}, w_i) \geq 6$ we must have that the indegree of w_i is at least five. It follows by the pigeonhole principle that the multiplicity of either $w_{i-1}w_i$ or $w_{i+1}w_i$ is at least three. If the former is true, we can apply Corollary 3.5 to obtain a sequence of pebbling moves that places $2^i(3-1) + 2 \cdot 1 \geq 6$ pebbles on w_0. Similarly, if the latter is true, we apply Corollary 3.5 to obtain a sequence of pebbling moves that places $2^{\alpha+1-i}(3-1) + 2 \cdot 1 \geq 6$ pebbles on $w_{\alpha+1}$. Because w_0 and $w_{\alpha+1}$ are vertices in G_1, it suffices to show that it is not possible to place more than five pebbles on any vertex in G_1.

Suppose that u is in G_1. Because it is possible to place at least six pebbles on u in G_2 starting from $p_{2,f}$, by Lemma 3.7, there is a nonrepetitive sequence of pebbling moves that places at least six pebbles on u in G_1 starting from $p_{1,f}$. But this is clearly impossible, because the maximum degree in G_1 is three and each vertex receives at most two pebbles from $p_{1,f}$. □

Now that we have established the important properties of the underlying graph and the underlying distributions, we attach gadgets to the vertices in the underlying graph. Just as the star gadgets we attach in our reduction from DPR to OPN force any potentially successful distribution to take a certain form, our gadgets here force any potentially unsuccessful distribution to take a form which effectively induces one of the underlying distributions on the underlying graph.

5.2. The gadgets. We introduce three classes of gadgets: the null gadget, the fork gadget, and the eye gadget. In this section, we explore the relevant properties of our gadgets as isolated graphs.

All classes of gadgets share some common properties. The gadgets have *attachment vertices*; later, we will attach gadgets to the underlying graph by identifying the attachment vertices of a gadget with vertices in the underlying graph. A *supply quota* s assigns each attachment vertex v a number $s(v)$; each gadget has one or more supply quotas. Under a particular distribution q, a gadget *satisfies* s if q covers s.

The gadgets have *overflow vertices* which are adjacent to r; we call the edges between the overflow vertices and r the *overflow edges*. We say that a gadget has an *overflow threshold* of k if r is reachable via an overflow edge under every distribution of size k.

Let q be a distribution of pebbles to a gadget. If the gadget is able to satisfy any one of its supply quotas, or if r is reachable via an overflow edge, we say that the gadget is *potent* under q. We say that a gadget has a *potency threshold* of k if the gadget is potent under every distribution of k pebbles.

Every gadget has one or more *critical distributions*, each of equal size. If q is a critical distribution and s is a supply quota, we say that q *breaches* s if there exists an attachment vertex v such that it is possible to place more than $s(v)$ pebbles on v starting from q.

Our critical distributions and supply quotas are in bijective correspondence; that is, for each critical distribution there is a corresponding supply quota and vice versa. Each critical distribution q exhibits the following *critical distribution properties*:

(1) starting from q, r is not reachable via an overflow edge,

(2) q does not breach its corresponding supply quota.

As we present the gadgets, their supply quotas, and their critical distributions, we

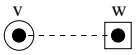


FIG. 5.1. *The null gadget. The dashed line represents a path with c edges, the circle around v indicates that v is an attachment vertex, and the box around w indicates that w is an overflow vertex.*

will establish an overflow threshold, a potency threshold, and the critical distribution properties.

To motivate the study of these parameters, we outline their use in our proof of the correctness of our reduction. Given an R∀∃3SAT instance ϕ, we compute H and k such that ϕ is valid if and only if $\pi(H, r) \leq k$. We construct H by attaching various gadgets to the underlying graph and we set k to be the sum, over all gadgets, of the size of the gadget's critical distributions.

Suppose that ϕ is valid and consider a distribution of k pebbles to H. If some gadget receives fewer pebbles than its potency threshold, the pigeonhole principle implies that some gadget receives more pebbles than its overflow threshold, and hence r is reachable. Otherwise, all gadgets are potent. If r is reachable via some gadget's overflow edge, we are done. Otherwise, every gadget is able to satisfy one of its supply quotas; this implies that our initial distribution on H covers some $p_{2,f}$. Because ϕ is satisfiable under f, we obtain from Proposition 5.4 a sequence of pebbling moves in the underlying graph which places a pebble on r.

The converse direction is somewhat trickier but proceeds roughly as follows. Suppose that $\pi(H, r) \leq k$ and consider a setting f of the universally quantified variables of ϕ. We assign pebbles to H by selecting (according to f) a critical distribution for each gadget. Because $\pi(H, r) \leq k$, we obtain a signature D of a minimum sequence of pebbling moves which places a pebble on r. Next, we argue that our critical distribution properties still apply even though the gadgets have been attached to the underlying graph. Then we show how D can be used to obtain a sequence of pebbling moves in the underlying graph starting from $p_{2,f}$ which places a pebble on r. A final application of Proposition 5.4 implies that ϕ is satisfiable under f.

As we discuss our gadgets, we will have occasion to refer to a number of absolute constants. When these constants are largely unimportant, we refer to them with the notation $O(1)$ instead of naming them individually. Our gadgets are defined in terms of two parameters, β and c. We set $c = 3$ (in fact, any constant c so that 2^c exceeds the constant obtained in Proposition 5.5 will do). We postpone fixing the precise value of β; suffice it to say we will choose $\beta = \lg n_2 + O(1)$. Our gadgets use small paths of length c to provide some separation between the underlying graph and more sensitive areas of our gadgets. We use larger paths of length β so that the number of pebbles in a gadget's critical distribution far exceeds its potency threshold.

5.2.1. The null gadget. The null gadget is a path with c edges and appears in Figure 5.1. We use the null gadget to ensure that every vertex in the underlying graph is not too far away from r, so that distributions which concentrate pebbles on the underlying graph quickly imply that r is reachable. The null gadget has a single supply quota s, with $s(v) = 0$; its corresponding critical distribution q assigns zero pebbles to each vertex in the null gadget.

Overflow threshold. Because c is a fixed constant, the null gadget is a fixed graph which does not depend upon ϕ. By Proposition 5.1, its pebbling number is a fixed constant, say a, not depending upon ϕ. Clearly, if there are $2a$ pebbles in the null gadget, then it is possible to place two pebbles on w and hence one pebble on r. It

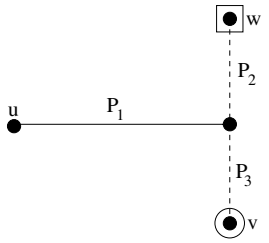


FIG. 5.2. *The fork gadget. The dashed lines represent paths P_2, P_3 with c edges, the solid line represents a path P_1 with β edges, the circle around v indicates that v is an attachment vertex, and the box around w indicates that w is an overflow vertex.*

follows that $O(1)$ is an overflow threshold for the null gadget.

Potency threshold. Because s is trivially satisfied, the null gadget has a potency threshold of 0.

Critical distribution properties. Because q assigns zero pebbles to the null gadget, it is clear that under q, the null gadget does not breach s, nor is it possible to place a pebble on r via the null gadget's overflow edge.

5.2.2. The fork gadget. The fork gadget consists of three paths P_1, P_2, P_3 which share only a common endpoint, as shown in Figure 5.2. The fork gadget is responsible for injecting one pebble in the underlying graph at the attachment location, much like the star gadgets in the previous section. It has one supply quota s with $s(v) = 1$; the corresponding critical distribution q is given by $q(u) = 2 \cdot 2^{\beta+c} - 1$ and $q(x) = 0$ for all $x \neq u$.

Overflow threshold. The fork gadget has an overflow threshold of $2 \cdot 2^{\beta+c} + O(1)$. Indeed, if the fork gadget is unable to place two pebbles on w (and hence one on r), there can be at most $O(1)$ pebbles on P_2 and P_3. Second, there can be at most $2 \cdot 2^{\beta+c} - 1$ pebbles in P_1 and P_2. It follows that the fork gadget can contain at most $2 \cdot 2^{\beta+c} + O(1)$ pebbles if r is not reachable via an overflow edge.

Potency threshold. The fork gadget has a potency threshold of $2^{\beta+c} + O(1)$. Indeed, if the fork gadget is not potent, then it must have at most $O(1)$ pebbles on P_2, or else it would be able to place a pebble on r. Similarly, it must have at most $2^{\beta+c} - 1$ pebbles on P_1 and P_3, or else it would be able to place a pebble on v and therefore satisfy s.

Critical distribution properties. Both the standard weight equation and the greedy tree lemma show that under q, the fork gadget does not breach s, nor is r reachable via an overflow edge.

5.2.3. The eye gadget. The eye gadget is the most complex of our three gadgets, and it is at the heart of our reduction. Our reduction attaches one eye gadget for each universally quantified variable in ϕ. The eye gadget is shown in Figure 5.3.

The eye gadget has two supply quota/critical distribution pairs. The pair (s^+, q^+) corresponds to a positive (true) setting of the variable x and the pair (s^-, q^-) corresponds to a negative (false) setting of x. We call s^+ the *positive supply quota* and we call s^- the *negative supply quota*. Similarly, we call q^+ the *positive critical distribution* and q^- the *negative critical distribution*.

We define the supply quotas via $s^+(v_1) = s^+(v_3) = 1$, $s^+(v_2) = 0$, and $s^-(v_1) = s^-(v_3) = 0$, $s^-(v_2) = 2$. Similarly, the critical distributions are given by $q^+(u_1) = q^+(u_3) = 2 \cdot 2^{\beta+c} - 1$, $q^+(u_0) = q^+(u_2) = 2^{\beta+c} - 1$, and $q^-(u_1) = q^-(u_3) = 2^{\beta+c} - 1$,

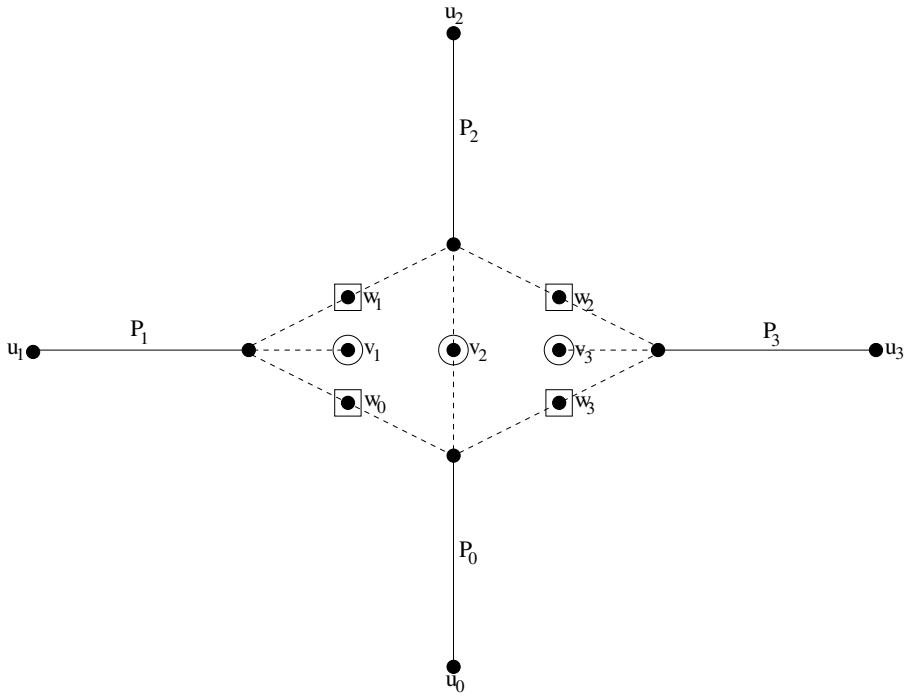


FIG. 5.3. *The eye gadget. The dashed lines represent paths with c edges, and the solid lines represent paths P_0, P_1, P_2, P_3 with β edges. The circled vertices v_i are attachment vertices; the boxed vertices w_j are overflow vertices.*

$q^-(u_0) = q^-(u_2) = 2 \cdot 2^{\beta+c} - 1$. We define $q^+(x)$and $q^-(x)$ to be zero whenever $x \notin \{u_0, u_1, u_2, u_3\}$.

Let F be the subgraph of the eye gadget obtained by removing the u_i and all interior vertices of the paths P_i. Observe that F depends only on c and therefore, like the null gadget, F is a fixed graph, not depending upon ϕ. It follows that $\pi(F) = O(1)$.

Overflow threshold. The eye gadget has an overflow threshold of $6 \cdot 2^{\beta+c} + O(1)$. Suppose the eye gadget contains k pebbles and it is not possible to place a pebble on r via one of the overflow edges. We show that $k \leq 6 \cdot 2^{\beta+c} + O(1)$. Immediately, we have that F contains at most $2\pi(F) = O(1)$ pebbles or else it would be possible to place two pebbles on w_0 and hence one pebble on r. To bound the number of pebbles in the P_i's, we consider two cases. First, suppose that each P_i contains fewer than $2^{\beta+c}$ pebbles; in this case, we have that $k \leq 4 \cdot 2^{\beta+c} + O(1)$. Otherwise, suppose that P_j has at least $2^{\beta+c}$ pebbles. Clearly, P_j has at most $2 \cdot 2^{\beta+c} - 1$ pebbles, or else we could use these pebbles to place a pebble on r via the overflow vertex w_j; similarly, the opposite path P_{j+2} contains at most $2 \cdot 2^{\beta+c} - 1$ pebbles (subscript arithmetic is understood modulo 4). Finally, the remaining paths P_{j-1}, P_{j+1} each contain at most $2^{\beta+c} - 1$ pebbles; indeed, if P_{j-1} (P_{j+1}) contained $2^{\beta+c}$ pebbles, we could use them to place one pebble on w_{j-1} (w_j) and we could use $2^{\beta+c}$ pebbles from P_j to place a second pebble on w_{j-1} (w_j). It follows that the P_i's contain at most $6 \cdot 2^{\beta+c} - 4$ pebbles, and thus $k \leq 6 \cdot 2^{\beta+c} + O(1)$.

Potency threshold. The eye gadget has a potency threshold of $5 \cdot 2^{\beta+c} + O(1)$. Suppose the eye gadget contains k pebbles, r is not reachable via an overflow edge, and it is not possible to satisfy s^+ or s^-. We show that $k \leq 5 \cdot 2^{\beta+c} + O(1)$. As before,

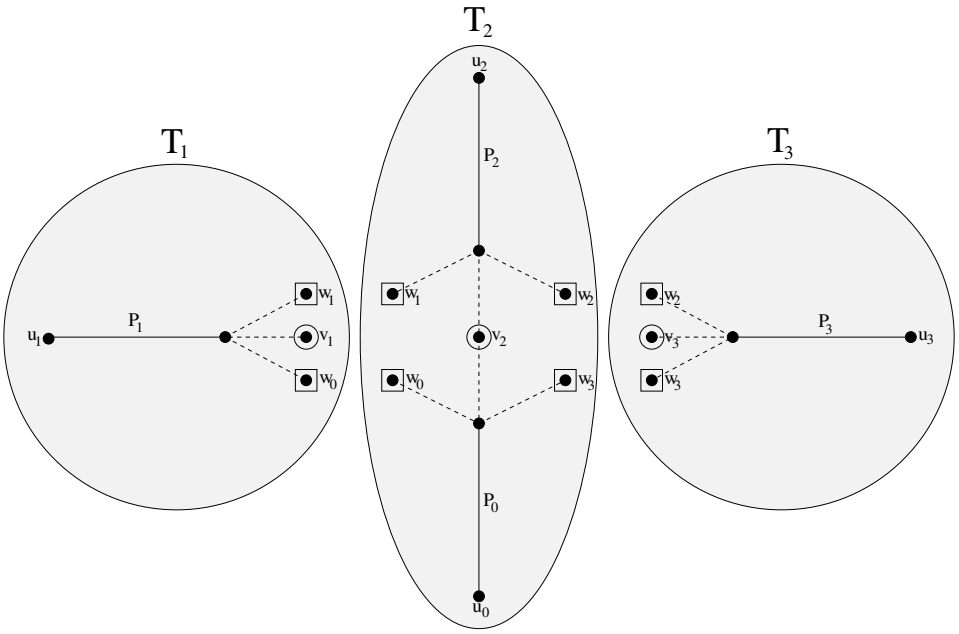


FIG. 5.4. *The overflow vertices split the eye gadget into three trees* T_1, T_2, T_3.

we have that F contains at most $O(1)$ pebbles. To bound the number of pebbles in the P_i, we consider the same two cases as before. If each path has fewer than $2^{\beta+c}$ pebbles, we immediately have $k \leq 4 \cdot 2^{\beta+c} + O(1)$ and are done. Otherwise, suppose P_j has at least $2^{\beta+c}$ pebbles. Once again, we have that P_j contains at most $2 \cdot 2^{\beta+c} - 1$ pebbles, and P_{j-1}, P_{j+1} each contain at most $2^{\beta+c} - 1$. However, now the opposite path P_{j+2} has at most $2^{\beta+c} - 1$ pebbles. Indeed, if P_j, P_{j+2} both contain at least $2^{\beta+c}$ pebbles, then we can either place one pebble each on v_1 and v_3, satisfying s^+ (as is the case if $\{j, j+2\} = \{1, 3\}$), or we can place two pebbles on v_2, satisfying s^- (as is the case if $\{j, j+2\} = \{0, 2\}$). It follows that the paths contain at most $5 \cdot 2^{\beta+c} - 4$ pebbles, implying $k \leq 5 \cdot 2^{\beta+c} + O(1)$.

Critical distribution properties. It remains to verify the critical distribution properties for q^+ and q^-. First, we show that under $q \in \{q^+, q^-\}$, r is not reachable via an overflow vertex. Let R be the set of overflow vertices in the eye gadget, and let D be the signature of a minimum sequence of pebbling moves that places two pebbles on a vertex in R. By Lemma 2.8, we have that each vertex in R has outdegree zero in D. Observe that deleting R from the eye gadget results in a graph with three components; let A_1 be the component containing P_1, let A_2 be the component containing P_0 and P_2, and let A_3 be the component containing P_3. Let T_1 be the subtree of the eye gadget induced by the set of vertices $V(A_1) \cup \{w_0, w_1\}$; similarly, let T_2 be induced by $V(A_2) \cup R$, and let T_3 be induced by $V(A_3) \cup \{w_2, w_3\}$. The trees T_i appear in Figure 5.4.

Let D_l be the digraph obtained by deleting from D all pebbling moves outside of T_l. Because D is acyclic, it is immediate that each D_l is acyclic. Observe that for all $x \in V(T_l) - R$, we have $\text{balance}(D_l, q, x) = \text{balance}(D, q, x)$. Furthermore, because $d^+_D(w_i) = 0$ for all $w_i \in R$, we have that $d^+_{D_l}(w_i) = 0$. Therefore by the acyclic orderability characterization, D_l is orderable.

Let w_i be the overflow vertex on which D places two pebbles. Because $q(w_i) = 0$, we have that the indegree of w_i in D is at least two. Suppose two edges into w_i are contained in the same tree T_l. Then D_l is the signature of a sequence of pebbling moves in T_l starting from q that places at least two pebbles on w_i. By the greedy tree lemma, the greedy pebbling strategy in T_l under q places at least two pebbles on w_i. However it is easily checked that regardless of $q \in \{q^+, q^-\}$, $T_l \in \{T_1, T_2, T_3\}$, and $w_i \in R$, the greedy strategy in T_l under q places at most one pebble on w_i. Alternatively, suppose that D contains edges into w_i from two distinct trees. Because w_i is in T_2 and one other tree, it must be that D contains an edge into w_i from T_2. Then D_2 is a signature of a sequence of pebbling moves in T_2 starting from q which places a pebble on w_i; therefore the greedy strategy in T_2 starting from q places a pebble on w_i. Because the greedy strategy in T_2 starting from q^+ is unable to place any pebbles on any overflow vertex, it follows that $q = q^-$. Suppose that D contains an edge into w_i from $T_l \in \{T_1, T_3\}$. Then D_l is the signature of a sequence of pebbling moves in T_l starting from q^- that places a pebble on w_i; therefore the greedy strategy in T_l starting from q^- places a pebble on w_i. But now a familiar contradiction is at hand: It is easily checked that regardless of $T_l \in \{T_1, T_3\}$ and $w_i \in R$, the greedy strategy in T_l starting from q^- is unable to place a pebble on w_i.

Let $(s, q) \in \{(s^+, q^+), (s^-, q^-)\}$. It remains to show that q does not breach s. Suppose for a contradiction that D is the signature of a minimum sequence of pebbling moves which witnesses that q breaches s. We have that the outdegree of each overflow vertex $w_i \in R$ is zero; indeed, if $d^+_D(w_i) \geq 1$, then by Lemma 2.6 we would obtain a sequence of pebbling moves placing two pebbles on w_i, a contradiction. As before, let D_l be the digraph obtained from D by deleting all edges outside of T_l; as before, we have that D_l is orderable in T_l. It follows that if D places more than $s(v_l)$ pebbles on v_l, then D_l witnesses that it is possible to place more than $s(v_l)$ pebbles on v_l in T_l starting from q. By the greedy tree lemma, the greedy strategy places more than $s(v_l)$ pebbles on v_l in T_l starting from q. But now we have a contradiction: We easily check that regardless of $(s, q) \in \{(s^+, q^+), (s^-, q^-)\}$ and $l \in \{1, 2, 3\}$, the greedy strategy in T_l starting from q places exactly $s(v_l)$ pebbles on v_l.

5.2.4. Summary. We summarize the various parameters of our gadgets in Table 5.1.

TABLE 5.1

Gadget	Potency threshold	Size of critical distributions	Overflow threshold
Null	0	0	$O(1)$
Fork	$2^{\beta+c} + O(1)$	$2 \cdot 2^{\beta+c} - 1$	$2 \cdot 2^{\beta+c} + O(1)$
Eye	$5 \cdot 2^{\beta+c} + O(1)$	$6 \cdot 2^{\beta+c} - 4$	$6 \cdot 2^{\beta+c} + O(1)$

From Table 5.1, we obtain the gap lemma.

LEMMA 5.6 (gap lemma). *There exists a nonnegative constant C (depending only on c) such that for each gadget, the overflow threshold exceeds the size of the critical distributions by at most C, and for the fork and eye gadgets, the size of the critical distributions exceeds the potency threshold by at least $2^{\beta+c} - C$.*

5.3. Construction of H. We set $\beta = \lceil \lg 3Cn_2 \rceil = \lg n_2 + O(1)$, with C as in Lemma 5.6.

Armed with our gadgets and our underlying graph G_2, we are able to describe the last step in our reduction from R$\forall\exists$3SAT to RPN. For each pebble in p_2 on a

vertex z in the underlying graph, we attach a fork gadget to z by identifying the attachment vertex v in the fork gadget with z. For each triplet z_1, z_2, z_3 of vertices in G_2 corresponding to a universally quantified variable x in ϕ, with z_1, z_3 corresponding to positive occurrences of x in ϕ and z_2 corresponding to the negative occurrence of x in ϕ, we attach an eye gadget by identifying the attachment vertex v_i in the eye gadget with z_i in the underlying graph. Finally, for any vertex $z \neq r$ in the underlying graph to which we did not attach a fork or eye gadget, we attach a null gadget by identifying v in the null gadget with z in the underlying graph. Let H be the resulting graph, and let k be the sum, over all gadgets in H, of the size of the gadget's critical distributions. Our reduction from R$\forall\exists$3SAT to RPN outputs H, k, and r.

Note that we attach gadgets to the underlying graph by identifying attachment vertices in gadgets with vertices in the underlying graph so that in H, each attachment vertex v is a member of the underlying graph and also a member of a gadget. Furthermore, by our construction, every vertex other than r in the underlying graph is identified with an attachment vertex; thus the vertices in the underlying graph are exactly the attachment vertices together with r.

We pause to observe two important properties about H.

PROPOSITION 5.7. *In constructing H, we attach at most two gadgets to every vertex in the underlying graph.*

Proof. Recall that p_2 assigns at most two pebbles to any vertex in the underlying graph; furthermore, p_2 assigns at most one pebble to any vertex associated with a universally quantified variable in ϕ. ∎

PROPOSITION 5.8. *The diameter of H is at most $2\beta + O(1)$.*

Proof. It suffices to show that for each z in H, the distance from z to r is at most $\beta + O(1)$. If $z \neq r$, then z is contained in some gadget. In each gadget, every vertex is at most $\beta + O(1)$ from an overflow vertex. ∎

5.4. R-PEBBLING-NUMBER is $\Pi_2^{\mathbf{P}}$-complete.

PROPOSITION 5.9. *Let $a_1, \ldots, a_n, b_1, \ldots, b_n$, and x be real numbers with $\sum_{i=1}^{n} a_i \geq \sum_{i=1}^{n} b_i$. If $a_n < b_n - x$, then there exists i such that $a_i > b_i + x/(n-1)$.*

Proof. The proof is by contradiction. Otherwise,

$$\begin{aligned}\sum_{i=1}^{n} a_i &= \sum_{i=1}^{n-1} a_i + a_n \\ &< \left(\sum_{i=1}^{n-1} b_i + \frac{x}{n-1}\right) + b_n - x \\ &= \sum_{i=1}^{n} b_i. \end{aligned}$$ ∎

We have accumulated the tools needed to show the correctness of our reduction.

THEOREM 5.10. *ϕ is valid if and only if $\pi(H, r) \leq k$.*

Proof. ($\Longrightarrow$). Suppose that ϕ is valid and let p be a pebble distribution on H of size k. We may assume $p(r) = 0$. Let l be the number of gadgets in H, label the gadgets as $Q_1, \ldots, Q_l$, let a_i be the number of pebbles that p assigns to Q_i, and let b_i be the size of Q_i's critical distributions. Because every vertex in H other than r belongs to at least one gadget, we have $\sum_{i=1}^{l} a_i \geq k = \sum_{i=1}^{l} b_i$.

We consider several cases. First, suppose there is some gadget Q_i to which p assigns fewer pebbles than Q_i's potency threshold; by the gap lemma, we have that

$a_i < b_i - (2^{\beta+c} - C)$. By Proposition 5.9, there is some Q_j to which p assigns at least $(2^{\beta+c} - C)/(l-1)$ pebbles more than the size of Q_j's critical distributions. By Proposition 5.7, $l - 1 \leq l \leq 2n_2$. It follows that Q_j contains at least

$$\begin{aligned}\frac{2^{\beta+c} - C}{2n_2} &\geq \frac{2^{\beta} - C}{2n_2} \\ &\geq \frac{3Cn_2 - C}{2n_2} \\ &\geq \frac{2Cn_2}{2n_2} \\ &\geq C\end{aligned}$$

more pebbles than the size of its critical distributions. It follows from Lemma 5.6 that Q_j contains at least as many pebbles as its overflow threshold and therefore we can place a pebble on r via one of Q_j's overflow edges. Otherwise, p assigns every gadget at least as many pebbles as its potency threshold. If there is some gadget which is able to place a pebble on r via an overflow edge, then we are done. Otherwise, for every gadget Q, there is a supply quota s such that Q under p satisfies s. Using these supply quotas, we obtain a setting f of the universally quantified variables in ϕ as follows. We set $f(x) =$ true if the eye gadget associated with x satisfies its positive supply quota s^+; otherwise, the eye gadget associated with x must meet the negative supply quota s^-, and we set $f(x) =$ false. We claim that p covers $p_{2,f}$. In each gadget, execute the pebbling moves witnessing that the gadget satisfies its supply quota. The fork gadgets alone produce a distribution that is at least as good as p_2, and the eye gadgets supply the additional pebbles prescribed by $p_{2,f}$. Because ϕ is valid, it follows from Proposition 5.4 that r is reachable.

($\Longleftarrow$). Suppose that $\pi(H, r) \leq k$ and let f be a setting of the universally quantified variables in ϕ. We obtain a setting of the existentially quantified variables in ϕ witnessing that ϕ is satisfiable under f. Naturally, we study a pebble distribution p on H of size k corresponding to f; we construct p by setting $p(r) = 0$ and choosing a critical distribution q_i for each gadget Q_i. If Q_i is not an eye gadget, then Q_i has only one critical distribution and our selection of q_i is forced. If Q_i is an eye gadget, we let q_i be the positive critical distribution q^+ if $f(x) =$ true and we let q_i be the negative critical distribution q^- otherwise. Note that p does not assign any pebbles to any vertex in the underlying graph. Let s_i be the supply quota associated with q_i.

Let H' be the graph obtained from H by removing all the overflow edges. Our first task is to establish the analog of Proposition 5.5 for H'.

Claim 1. *In H' starting from p, it is not possible to place more than five pebbles on any vertex in the underlying graph.*

Proof. Suppose for a contradiction that D is the signature of a minimum sequence of pebbling moves that places at least six pebbles on some vertex w in the underlying graph.

Subclaim 1. *D does not contain an edge whose head is inside the underlying graph and whose tail is outside the underlying graph.*

Proof. Suppose for a contradiction that uv is an edge in D from a vertex u in the underlying graph to some vertex v not in the underlying graph. Because H' does not contain any overflow edges, it must be that uv is an edge on a path of length c in some gadget; let this path be $P = x_0, \ldots, x_c$, with $u = x_c$ and $v = x_{c-1}$. It follows that D contains the edge x_1x_0, or else D contains a cycle or a proper sink other than

w, contradicting the minimum signatures lemma. Because $p(x_i) = 0$ for each internal vertex of P, we have by Corollary 3.5 that it is possible to place $2^c = 8 \geq 6$ pebbles on u using fewer pebbling moves, a contradiction. Therefore D does not contain an edge from the underlying graph to a vertex outside the underlying graph. □

SUBCLAIM 2. *For each u in the underlying graph, the number of edges in D into u with heads outside the underlying graph is at most $p_{2,f}(u)$.*

Proof. If this is not the case, then there is some gadget Q_i attached to u such that D contains more than $s_i(u)$ edges from Q_i into u. Construct D' from D by deleting all edges not contained in Q_i. Clearly, $D' \subseteq D$ is acyclic; we show that D' is orderable by verifying the balance condition. Consider a vertex v in Q_i. Recall that H' does not contain overflow edges, and therefore if v is not an attachment vertex, then the neighborhood of v is contained in Q_i. It follows that if v is not an attachment vertex, we have $\text{balance}(D', q_i, v) = \text{balance}(D, p, v)$. Alternatively, if v is an attachment vertex, we have that $d^+_{D'}(v) = 0$ or else D' (and hence D) would contain an edge from a vertex v in the underlying graph to a vertex outside the underlying graph, contradicting our previous subclaim. It follows that if v is an attachment vertex, we have $\text{balance}(D', q_i, v) \geq 0$. By the acyclic orderability characterization, we have that D' is orderable under q_i. Together with $d^-_{D'}(u) > s_i(u)$ and $d^+_{D'}(u) = 0$ (recall that u is an attachment vertex), we have that $\text{balance}(D', q_i, u) > s_i(u)$. Therefore D' witnesses that it is possible to place more than $s_i(u)$ pebbles on u in Q_i starting from q_i, contradicting Q_i's critical distribution properties. □

We return to our proof of Claim 1. Construct D' from D by removing all edges from D that are not in the underlying graph. Clearly, $D' \subseteq D$ is acyclic. We show that D' is orderable under $p_{2,f}$ by checking the balance condition. For each u in the underlying graph, we have $\text{balance}(D', p_{2,f}, u) \geq \text{balance}(D, p, u)$. Indeed, at most $p_{2,f}(u)$ edges into u are deleted from D in our construction of D'; however, $p(u) = 0$, so that $p_{2,f}$ offsets this decrease in balance. It follows that D' is orderable under $p_{2,f}$. Together with $\text{balance}(D', p_{2,f}, w) \geq \text{balance}(D, p, w) \geq 6$, we have that it is possible to place at least six pebbles on w starting from $p_{2,f}$ in the underlying graph, contradicting Proposition 5.5. This completes our proof of Claim 1. □

We return to our proof of Theorem 5.10. Let D be the signature of a minimal sequence of pebbling moves in H starting from p that places a pebble on r.

CLAIM 2 (no backflow into gadgets claim). *D does not contain an edge from a vertex inside the underlying graph to a vertex outside the underlying graph.*

Proof. By the minimum signatures lemma, we have that D contains at most one pebbling move along an overflow edge and any such pebbling move must be directed from an overflow vertex into r. Construct D' from D by removing this edge if it exists. Because r has outdegree zero in D, the acyclic orderability characterization implies that D' is orderable. Furthermore, because D' does not contain any pebbling move along overflow edges, D' yields a sequence of pebbling moves in H'.

Because D' is constructed from D by removing at most one edge into r, it suffices to show that D' does not contain an edge from a vertex inside the underlying graph to a vertex outside the underlying graph. Suppose for a contradiction that D' contains an edge uv from u inside the underlying graph to v outside the underlying graph. It must be that uv is a pebbling move along a path P of length c in some gadget. Let $P = x_0, \ldots, x_c$ with $u = x_c$ and $v = x_{c-1}$. It follows that D' contains the edge x_1x_0. Indeed, if D' does not have x_1x_0 as an edge, neither does D (after all, $x_0 \neq r$), and so D contains a cycle or a proper sink other than r, contradicting the minimum signatures lemma. Therefore D' contains the pebbling move x_1x_0.

Recalling that p assigns each internal vertex of P zero pebbles, Corollary 3.5 implies that there is an orderable $D'' \subseteq D'$ which places at least $2^c = 8$ pebbles on $x_c = u$. But now D'' is a signature witnessing that it is possible to place at least six pebbles on u in H' starting from p, contradicting Claim 1. □

Let us resume our proof of Theorem 5.10. Construct D_i from D by deleting from D all edges not contained in Q_i or in Q_i's overflow edges.

CLAIM 3. *D_i is orderable under q_i, and for each attachment vertex v, we have that* $\text{balance}(D_i, q_i, v) = d^-_{D_i}(v)$.

Proof. Because $D_i \subseteq D$, D_i is acyclic and thus it suffices to verify the balance condition. Because $d^+_D(r) = 0$, clearly $d^+_{D_i}(r) = 0$ and thus the balance condition is satisfied at r. Consider a vertex v in Q_i. Unless v is an attachment vertex, all edges incident to v in D also appear in D_i, and so $\text{balance}(D_i, q_i, v) = \text{balance}(D, p, v)$. Otherwise, if v is an attachment vertex, then $d^+_{D_i}(v) = 0$ or else D would contain an edge from a vertex in the underlying graph to a vertex outside the underlying graph, contradicting Claim 2. Together with $q_i(v) = 0$, it follows that $\text{balance}(D_i, q_i, v) = d^-_{D_i}(v)$. By the acyclic orderability characterization, D_i is orderable under q_i. □

CLAIM 4. *For each u in the underlying graph, D contains at most $p_{2,f}(u)$ edges from outside the underlying graph into u.*

Proof. Suppose that u is a counterexample to the claim. If $u = r$, then there is some gadget Q_i such that D contains an edge wr into r along one of Q_i's overflow edges. But D_i also contains wr and, by Claim 3, D_i is orderable under q_i. Clearly, $\text{balance}(D_i, q_i, r) \geq 1$ and therefore r is reachable in Q_i under q_i, contradicting the critical distribution properties of Q_i. Otherwise, if $u \neq r$, then there is some gadget Q_i such that D contains more than $s_i(u)$ edges into u from vertices in Q_i. But these edges are also in D_i, so that $d^-_{D_i}(u) > s_i(u)$. By Claim 3, D_i is the signature of a sequence of pebbling moves in Q_i under q_i placing more than $s_i(u)$ pebbles on u, contradicting Q_i's critical distribution properties. □

Let us complete our proof of Theorem 5.10. Construct E from D by deleting from D any edges outside the underlying graph. We show that E is orderable under $p_{2,f}$. Clearly, $E \subseteq D$ is acyclic and therefore it suffices to check the balance condition. Consider a vertex u in the underlying graph, and let m be the number of edges into u from outside the underlying graph. In constructing E from D, the balance of u decreases by m; by Claim 4, we have $m \leq p_{2,f}(u)$. Because $p(u) = 0$, changing distributions from p to $p_{2,f}$ increases the balance of u by $p_{2,f}(u)$. It follows that $\text{balance}(E, p_{2,f}, u) \geq \text{balance}(D, p, u)$. Therefore E is orderable under $p_{2,f}$ and so r is reachable in the underlying graph under $p_{2,f}$. A final application of Proposition 5.4 implies that ϕ is satisfiable under f. This completes our proof of Theorem 5.10. □

We are now able to complete our proof that R-PEBBLING-NUMBER is Π^{P}_2-complete.

THEOREM 5.11. R-PEBBLING-NUMBER *is Π^{P}_2-complete, even when the diameter of H is at most $O(\log n(H))$ and $k = \text{poly}(n(H))$.*

Proof. We have already observed that RPN is in Π^{P}_2 and checked the correctness of our reduction; it remains to check the diameter condition on H and that H and k are not too large relative to ϕ so that our reduction is computable in polynomial time. By Proposition 5.8, the diameter of H is at most $2\beta + O(1) = 2\lceil \lg 3Cn_2 \rceil + O(1)$. Because n_2 is the number of vertices in the underlying graph, we have $n_2 \leq n(H)$ and therefore the diameter of H is at most $2\lceil \lg 3Cn(H) \rceil + O(1) = O(\log n(H))$.

It remains to check the size condition on H and k. Because G_1 has the same number of vertices as $G^{\text{NPR}}(\phi)$, Proposition 3.2 implies that the size of G_1 is poly-

nomial in the size of ϕ. Because the underlying graph G_2 is $\mathcal{S}(G_1, \alpha)$ with $\alpha = \lceil \max\{\lg 2t,\, 4 \lg e(G_1)\} \rceil$ and $t \leq 2n(G_1)$, we have that the size of the underlying graph is polynomial in the size of G_1. Observe that each gadget has size linear in $\beta = \lg n_2 + O(1)$. Together with Proposition 5.7, we have that the size of H is polynomial in the size of G_2. It follows that the size of H is polynomial in the size of ϕ. Finally, every gadget's critical distribution size is at most $O(2^\beta) = O(n_2)$; together with Proposition 5.7, we have that k is polynomial in n_2 and hence polynomial in $n(H)$. □

5.5. PEBBLING-NUMBER is Π_2^{P}-complete. After having established Theorem 5.11, it is relatively easy to show that PN is Π_2^{P}-complete.

THEOREM 5.12. PEBBLING-NUMBER *is* Π_2^{P}*-complete.*

Proof. We have already observed that PN is in Π_2^{P}. To show that PN is Π_2^{P}-hard, we reduce RPN to PN. Let G be a graph with target vertex r, and let $k \geq 0$ be an integer. We produce H and k' so that $\pi(G, r) \leq k$ if and only if $\pi(H) \leq k'$. By Theorem 5.11, our reduction may assume that the diameter d of G is at most $c' \lg n(G)$ for an absolute constant c' and $k = \mathrm{poly}(n(G))$.

We construct H and k' as follows. Let $n = n(G)$ and set $\alpha = \lceil k n^{c'} \rceil$. We let H be the graph consisting of α copies of G that share r, so that $H - r$ is α disjoint copies of $G - r$. We set $k' = \alpha k$. Observe that k' and the size of H are polynomial in the size of G. It remains to show that $\pi(G, r) \leq k$ if and only if $\pi(H) \leq k'$.

($\Longrightarrow$). Suppose $\pi(G, r) \leq k$. Consider a distribution of $k' = \alpha k$ pebbles to H and let u be some target vertex in H. Observe that $d(u, r) \leq d$ and therefore to place a pebble on u, it suffices to show that we can place 2^d pebbles on r. Our strategy is as follows. If there is some copy of G with at least k pebbles, then we arbitrarily select a set S of k pebbles from this copy of G; because $\pi(G, r) \leq k$, we can use these pebbles to place a pebble on r. We repeat this strategy until we are unable to find a copy of G with at least k pebbles. Let s be the number of pebbles we are able to place on r via this strategy. Observe that after executing this strategy s times, at least $k\alpha - ks$ unused pebbles remain in H, and furthermore, if more than $\alpha(k-1)$ unused pebbles remain in H, then some copy of G contains at least k unused pebbles. It follows that

$$k\alpha - ks \leq \alpha(k-1),$$

and therefore $s \geq \alpha/k \geq n^{c'} = 2^{c' \lg n} \geq 2^d$.

($\Longleftarrow$). Suppose $\pi(H) \leq k'$, and let p be a distribution of k pebbles to G. If $p(r) > 0$, then r is trivially reachable. Otherwise, we define a distribution q of $k' = \alpha k$ pebbles to H by distributing k pebbles in each copy of G according to p. Let D be the signature of a minimum sequence of pebbling moves that places a pebble on r. By the minimum signatures lemma, all edges of D are contained in a single copy of G. It follows that r is reachable in G under p. □

6. Conclusions. As we have seen, many graph pebbling problems on unrestricted graphs are computationally difficult. We have seen that REACHABLE and OPTIMAL-PEBBLING-NUMBER are both NP-complete. The authors believe it more likely than not that REACHABLE remains NP-complete even when the graphs are restricted to be planar. However, we have more hope that REACHABLE may fall to P when the graphs are restricted to be outerplanar. It may be interesting to investigate the computational complexity of these problems when the inputs are restricted to be planar or outerplanar.

We have also seen that PEBBLING-NUMBER is Π_2^{P}-complete and therefore both NP-hard and coNP-hard. It follows that unless the polynomial hierarchy collapses

to the first level, PEBBLING-NUMBER is in neither NP nor coNP. Consequently, given G and k, it is unlikely that we can compute in polynomial time a collection $\mathcal{P}$ of candidate distributions of size k such that if $\pi(G) > k$, then some vertex in G is not reachable from some $p \in \mathcal{P}$ (or else PN would be in NP).

We have shown that COVERABLE and REACHABLE are both NP-complete; however, the computational complexity of these problems diverges when we introduce a universal quantifier over pebble distributions. When we add such a quantifier to COVERABLE, we obtain the problem of determining if $\gamma(G) \leq k$, which is possible in polynomial time [13, 11]. The computational difficulties in COVERABLE are smoothed out by the consideration of all pebble distributions of size k: There is a nice structure to the maximum pebble distributions from which a graph cannot be covered with pebbles. On the other hand, by adding a similar universal quantifier to REACHABLE, we obtain RPN, which asks us to decide if $\pi(G, r) \leq k$. Instead of observing a decrease in the computational complexity, we have encountered a Π_2^{P}-complete problem.

We recall that the graph pebbling community has shown a good deal of interest in developing necessary conditions and sufficient conditions for equality in $\pi(G) = n(G)$. Of course, the ultimate goal is to develop a characterization for when equality holds. We should remark that our hardness result for PEBBLING-NUMBER does not suggest that any such characterization need be complex from a computational point of view. Indeed, our PEBBLING-NUMBER hardness result produces G and k with $k > n(G)$. It may be interesting to explore the complexity of deciding whether $\pi(G) = n(G)$.

Acknowledgments. We thank David Bunde, Jeff Erickson, and Sariel Har-Peled for helpful suggestions throughout the revision process. Additionally, we are grateful for the corrections and suggestions of our anonymous referees.

REFERENCES

[1] D. P. Bunde, E. W. Chambers, D. Cranston, K. Milans, and D. B. West, *Pebbling and optimal pebbling in graphs*, J. Graph Theory, to appear.
[2] F. R. K. Chung, *Pebbling in hypercubes*, SIAM J. Discrete Math., 2 (1989), pp. 467–472.
[3] T. A. Clarke, R. A. Hochberg, and G. H. Hurlbert, *Pebbling in diameter two graphs and products of paths*, J. Graph Theory, 25 (1997), pp. 119–128.
[4] B. Crull, T. Cundiff, P. Feltman, G. H. Hurlbert, L. Pudwell, Z. Szaniszlo, and Z. Tuza, *The cover pebbling number of graphs*, Discrete Math., 296 (2005), pp. 15–23.
[5] A. Czygrinow, G. Hurlbert, H. A. Kierstead, and W. T. Trotter, *A note on graph pebbling*, Graphs Combin., 18 (2002), pp. 219–225.
[6] G. Hurlbert, *Recent progress in graph pebbling*, Graph Theory Notes N.Y., 49 (2005), pp. 25–37.
[7] G. H. Hurlbert and H. Kierstead, *private communication*, 2005.
[8] D. Moews, *Pebbling graphs*, J. Combin. Theory Ser. B, 55 (1992), pp. 244–252; also available online from http://djm.cc/dmoews/pebbling-graphs.ps.
[9] L. Pachter, H. S. Snevily, and B. Voxman, *On pebbling graphs*, Congr. Numer., 107 (1995), pp. 65–80.
[10] C. H. Papadimitriou, *Computational Complexity*, Addison–Wesley, Reading, MA, 1994.
[11] J. Sjöstrand, *The cover pebbling theorem*, Electron. J. Combin., 12 (2005), Note 22.
[12] C. A. Tovey, *A simplified NP-complete satisfiability problem*, Discrete Appl. Math., 8 (1984), pp. 85–89.
[13] A. Vuong and M. I. Wyckoff, *Conditions for Weighted Cover Pebbling of Graphs*, preprint; available online from http://www.arxiv.org/abs/math.CO/0410410 (2004).
[14] N. G. Watson, *The Complexity of Pebbling and Cover Pebbling*, preprint; available online from http://www.arxiv.org/abs/math.CO/0503511 (2005).

SIAM J. DISCRETE MATH.
Vol. 20, No. 3, pp. 799–810

ON GRAPHS HAVING NO CHROMATIC ZEROS IN $(1, 2)$*

F. M. DONG† AND K. M. KOH‡

Abstract. For a graph G of order $n \geq 2$, an ordering $(x_1, x_2, \ldots, x_n)$ of the vertices in G is called a *double-link ordering* of G if $x_1x_2 \in E(G)$ and x_i has at least two neighbors in $\{x_1, x_2, \ldots, x_{i-1}\}$ for all $i = 3, 4, \ldots, n$. This paper shows that certain graphs possessing a kind of double-link ordering have no chromatic zeros in the interval $(1, 2)$. This result implies that all graphs with a 2-tree as a spanning subgraph, certain graphs with a Hamiltonian path, all complete t-partite graphs, where $t \geq 3$, and all $(v(G) - \Delta(G) + 1)$-connected graphs G have no chromatic zeros in the interval $(1, 2)$.

Key words. chromatic polynomial, chromatic zero

AMS subject classification. 05C15

DOI. 10.1137/04061787X

1. Introduction.

For a simple graph G with vertex set $V(G)$ and edge set $E(G)$, let $P(G, \lambda)$ denote its chromatic polynomial. A zero of $P(G, \lambda)$ is also called a *chromatic zero* of G.

It is known that every graph has no chromatic zeros in the two intervals $(-\infty, 0)$ and $(0, 1)$ (see [4] or [7], for instance). Jackson [5] showed that every graph has no chromatic zeros in $(1, 32/27]$ as well. Let $v(G)$ and $b(G)$ denote the order (number of vertices) and the number of blocks of a graph G.

THEOREM 1.1 (see [5]). *For any connected graph G with $v(G) \geq 2$, the inequality $(-1)^{v(G)+b(G)-1}P(G, \lambda) > 0$ holds for all $\lambda \in (1, 32/27]$.*

Thomassen [9] further showed that for any interval (a, b) with $b > 32/27$, there exist graphs having chromatic zeros in (a, b). Hence $(-\infty, 0)$, $(0, 1)$, and $(1, 32/27]$ are the only intervals in which every graph has no chromatic zeros.

The problem we study in this paper is motivated by the following conjecture proposed by Jackson [5].

CONJECTURE 1.1. *If G is a 3-connected and nonbipartite graph, then G has no chromatic zeros in $(1, 2)$.*

Applying Theorem 1.1, it can be shown that the fact that G has no chromatic zeros in $(1, 2)$ is actually equivalent to saying that the inequality in Theorem 1.1 holds for all $\lambda \in (1, 2)$.

COROLLARY 1.1. *Let G be a connected graph with $v(G) \geq 2$. Then G has no chromatic zeros in $(1, 2)$ if and only if $(-1)^{v(G)+b(G)-1}P(G, \lambda) > 0$ for all $\lambda \in (1, 2)$.*

The condition that G is nonbipartite in the above conjecture is necessary, since, as was pointed out by Jackson [5], every 2-connected bipartite graph of odd order has a chromatic zero in $(1, 2)$. Here we give a very short proof of this observation, which will be used in the following sections.

LEMMA 1.1. *Let G be a connected bipartite graph with $v(G) \geq 2$. If $v(G) + b(G)$ is even, then G has a chromatic zero in $(1, 2)$.*

*Received by the editors October 29, 2004; accepted for publication (in revised form) May 17, 2006; published electronically November 3, 2006.
http://www.siam.org/journals/sidma/20-3/61787.html
†Mathematics and Mathematics Education, National Institute of Education, Nanyang Technological University, Singapore 637616 (fmdong@nie.edu.sg). Corresponding author.
‡Department of Mathematics, National University of Singapore, Singapore 117543 (matkohkm@nus.edu.sg).

Proof. Since G is bipartite, $P(G,2) > 0$. Since $P(G,\lambda)$ is continuous, there exists a real number δ with $0 < \delta < 1$ such that $P(G,\lambda) > 0$ for all $\lambda \in (2-\delta, 2)$. As $v(G) + b(G)$ is even, $(-1)^{v(G)+b(G)-1}P(G,\lambda) < 0$ holds for all $\lambda \in (2-\delta,2)$. By Corollary 1.1, the result holds. □

Let G be a graph with $n = v(G) \geq 2$. An ordering $(x_1, x_2, \ldots, x_n)$ of the vertices in G is called a *double-link ordering* of G if $x_1x_2 \in E(G)$ and

$$|N(x_i) \cap V_{i-1}| \geq 2 \tag{1}$$

for all $i = 3, 4, \ldots, n$, where $V_i = \{x_1, x_2, \ldots, x_i\}$ and $N(x)$ is the set of vertices in G adjacent to x. Throughout this paper, the notation V_i is fixed whenever a double-link ordering $(x_1, x_2, \ldots, x_n)$ of $V(G)$ is given.

In this paper, we shall exhibit various families of graphs which have no chromatic zeros in $(1,2)$. We first establish a general result (namely, Theorem 2.1) in section 2 that certain graphs which possess a kind of double-link orderings have no chromatic zeros in $(1,2)$. Let Γ denote the family of these graphs. In section 3, we give a sufficient condition for a graph to be in Γ. We then show in section 4 that Γ includes certain graphs having a Hamiltonian path, complete t-partite graphs, where $t \geq 3$, complete bipartite graphs with an additional edge, and graphs which are $(v(G) - \Delta(G) + 1)$-connected, where $\Delta(G)$ denotes the maximum degree of G. Finally in section 5, we give a necessary condition for a graph to be in Γ and also propose some conjectures.

2. A family of graphs with a double-link ordering. Let $\mathcal{D}$ be the family of graphs with a double-link ordering. It is clear that $\mathcal{D}$ contains only one graph of order 2 (i.e., K_2), one graph of order 3 (i.e., K_3), and two graphs of order 4 (i.e., K_4 and $K_4 - e$). For any $n \geq 2$, $K_n \in \mathcal{D}$, but for all $n \geq 3$, the path P_n does not belong to $\mathcal{D}$.

By definition, we have the following lemma.

LEMMA 2.1. *Let G be a graph with $v(G) \geq 3$. Then $G \in \mathcal{D}$ if and only if G contains a vertex x such that $d(x) \geq 2$ and $G - x \in \mathcal{D}$.*

The following is a property of the graphs in $\mathcal{D}$. The size (number of edges) of G is denoted by $e(G)$.

LEMMA 2.2. *Let $G \in \mathcal{D}$ with $v(G) \geq 3$. Then $e(G) \geq 2v(G) - 3$ and G is 2-connected.*

Proof. The result $e(G) \geq 2v(G) - 3$ follows directly from the definition.

When $v(G) = 3$, $G \cong K_3$, and so G is 2-connected. It follows by induction from Lemma 2.1 that G is 2-connected if $v(G) \geq 3$. □

Our aim in this section is to study a subfamily of $\mathcal{D}$ and show that every graph in this subfamily has no chromatic zeros in the interval $(1,2)$.

Let $G \in \mathcal{D}$ with $n = v(G) \geq 2$, and $(x_1, x_2, \ldots, x_n)$ be a double-link ordering of $V(G)$. If there exist $u_i, v_i \in N(x_i) \cap V_{i-1}$ with $u_i \neq v_i$ for $i = 3, 4, \ldots, n$ such that the inequality

$$|I| > |\{i : \{u_i, v_i\} \subseteq I, 3 \leq i \leq n\}| \tag{2}$$

holds for every nonempty independent set I of G, then $(x_1, x_2, \ldots, x_n)$ is called a *γ-ordering* of G.

Since $G \in \mathcal{D}$, by definition, we have $G[V_2] \cong K_2$ and $G[V_3] \cong K_3$ (when $n \geq 3$), where $G[A]$ denotes the subgraph of G induced by A for any $A \subseteq V(G)$. Thus, for any nonempty independent set I of G, $\{u_i, v_i\} \not\subseteq I$ for $i = 3, 4$. Hence inequality (2)

can be replaced by the following inequality:

$$|I| > |\{i : \{u_i, v_i\} \subseteq I, 5 \le i \le n\}|. \tag{3}$$

Let Γ be the family of graphs having a γ-ordering. It is clear that $\Gamma \subseteq \mathcal{D}$. By definition, if $G \in \mathcal{D}$ and $v(G) \le 5$, then $G \in \Gamma$. But, for any $n \ge 6$, there exists a graph $G \in \mathcal{D}$ with $v(G) = n$ such that $G \notin \Gamma$. For example, if G is a graph obtained from the complete bipartite graph $K_{2,n-2}$ by adding one edge which joins any two vertices in the part with $n-2$ vertices, then it can be shown that $G \in \mathcal{D}$ but $G \notin \Gamma$ if $n \ge 6$. Thus Γ is a proper subfamily of $\mathcal{D}$.

By the definition of a γ-ordering, we have the following lemma.

LEMMA 2.3. *Let $G \in \Gamma$ with a γ-ordering $(x_1, x_2, \ldots, x_n)$, where $n \ge 3$. Then $(x_1, x_2, \ldots, x_i)$ is a γ-ordering of $G[V_i]$, and thus $G[V_i] \in \Gamma$ for $i = 2, 3, \ldots, n$.*

In this section, we shall show that each graph in Γ has no chromatic zeros in $(1, 2)$. To establish this result, we need to show the following:

(i) If $G \in \Gamma$ and G is a spanning subgraph of a graph H, then $H \in \Gamma$.

(ii) If $G \in \Gamma$ and $x \in V(G)$ with $d(x) = 2$, then $G - x \in \Gamma$ and $(G - x) \cdot uv \in \Gamma$, where $\{u, v\} = N(x)$ and $(G - x) \cdot uv$ is the graph obtained from $G - x$ by contracting u and v.

LEMMA 2.4. *Let G be a spanning subgraph of a graph H. If $G \in \Gamma$, then $H \in \Gamma$.*

Proof. Since $G \in \Gamma$, G has a γ-ordering $(x_1, x_2, \ldots, x_n)$, where $n = v(G) \ge 2$. Then there exist $u_i, v_i \in N_G(x_i) \cap V_{i-1}$ with $u_i \ne v_i$ for $i = 3, 4, \ldots, n$ such that (3) holds for every nonempty independent set I of G.

Since G is a spanning subgraph of H, $(x_1, x_2, \ldots, x_n)$ is also a double-link ordering of H and $u_i, v_i \in N_H(x_i) \cap V_{i-1}$ for $i = 3, 4, \ldots, n$. Each independent set I of H is also independent in G. Thus (3) also holds for every nonempty independent set I of H. Therefore $H \in \Gamma$. □

On the other hand, if $G \in \Gamma$, what edge uv can be removed from G such that $G - uv \in \Gamma$?

LEMMA 2.5. *Let $G \in \Gamma$ and $(x_1, x_2, \ldots, x_n)$ be a γ-ordering of G. Assume that u_i, v_i are distinct vertices in $N(x_i) \cap V_{i-1}$ for $i = 3, 4, \ldots, n$ such that inequality (3) holds for every nonempty independent set I of G. Let $uv \in E(G)$. If*

$$\{u, v\} \not\subseteq \{u_i : i = 3, 4, \ldots, n\} \cup \{v_i : i = 3, 4, \ldots, n\} \tag{4}$$

and

$$uv \notin \{x_i u_i : i = 3, 4, \ldots, n\} \cup \{x_i v_i : i = 3, 4, \ldots, n\}, \tag{5}$$

then $(x_1, x_2, \ldots, x_n)$ is also a γ-ordering of $G - uv$ and so $G - uv \in \Gamma$.

Proof. By (4), $\{u, v\} \ne \{u_3, v_3\} = \{x_1, x_2\}$, and so $x_1 x_2 \in E(G - uv)$. By (5), $u_i, v_i \in N_{G-uv}(x_i) \cap V_{i-1}$ for $i = 3, 4, \ldots, n$. Suppose that there exists a nonempty independent set I of $G - uv$ such that

$$|I| \le |\{i : \{u_i, v_i\} \subseteq I, 5 \le i \le n\}|.$$

Then I is not an independent set of G. Thus $u, v \in I$. By (4), either u or v does not belong to $\{u_i, v_i : i = 3, 4, \ldots, n\}$, say $u \notin \{u_i, v_i : i = 3, 4, \ldots, n\}$. Then

$$|\{i : \{u_i, v_i\} \subseteq I \backslash \{u\}, 5 \le i \le n\}| = |\{i : \{u_i, v_i\} \subseteq I, 5 \le i \le n\}| > |I \backslash \{u\}|$$

implying that inequality (3) does not hold for the independent set $I \backslash \{u\}$ of G, a contradiction.

Therefore the result holds. □

LEMMA 2.6. *Let $G \in \Gamma$ and $x \in V(G)$ with $d(x) = 2$. Then*

(i) *$G - x \in \Gamma$; and*

(ii) *$(G - x) \cdot uv \in \Gamma$ if $N(x) = \{u, v\}$ with $uv \notin E(G)$.*

Proof. (i) Let $n = v(G)$. It is easy to verify that $G - x \in \Gamma$ if $n \leq 4$. Now let $n \geq 5$.

Let $(x_1, x_2, \ldots, x_n)$ be a γ-ordering of G. By the definition of a γ-ordering, if $3 \leq i < n-1$ and $x_i x_{i+1} \notin E(G)$, then $(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, x_i, x_{i+2}, \ldots, x_n)$ is also a γ-ordering of G.

Let $x = x_i$. We may assume that $i \geq 3$. Since $d(x) = 2$, by definition, $N(x) \subseteq V_{i-1} = \{x_1, x_2, \ldots, x_{i-1}\}$. Thus $xx_j \notin E(G)$ for all $j = i+1, \ldots, n$. By the above argument, if $i < n$, we can get a new γ-ordering of G by exchanging x_i (i.e., x) and x_{i+1}. Repeating this process, we would have a γ-ordering ending with x. Hence we may assume that $x = x_n$, and we have $G - x \in \Gamma$.

(ii) Let $(x_1, x_2, \ldots, x_n)$ be a γ-ordering of G and $x = x_n$. Since $G \in \Gamma$, by Lemma 2.5, there exist $u_i, v_i \in N(x_i) \cap V_{i-1}$ with $u_i \neq v_i$ for $i = 5, 6, \ldots, n$ such that (3) holds for every nonempty independent set I of G.

Since $N(x_n) = \{u, v\}$, we have $\{u, v\} = \{u_n, v_n\}$. Since $u_n v_n \notin E(G)$, we have $\{u_i, v_i\} \neq \{u_n, v_n\}$ for all $i = 5, 6, \ldots, n-1$; otherwise, (3) does not hold for the independent set $I = \{u_n, v_n\}$.

Let $G' = (G - x_n) \cdot u_n v_n$. Assume that $u_n = x_s$ and $v_n = x_t$, where $s < t \leq n-1$. For convenience, we still denote by x_s the new vertex in G' after contracting u_n and v_n in $G - x_n$. Then $(x_1, x_2, \ldots, x_{t-1}, x_{t+1}, \ldots, x_{n-1})$ is a double-link ordering of G' with $u_i', v_i' \in N_{G'}(x_i) \cap V_i'$, where $V_i' = V_i \backslash \{x_t\}$ and

$$u_i' = \begin{cases} u_i & \text{if } u_i \neq x_t, \\ x_s & \text{otherwise} \end{cases} \quad \text{and} \quad v_i' = \begin{cases} v_i & \text{if } v_i \neq x_t, \\ x_s & \text{otherwise.} \end{cases}$$

Since $\{u_i, v_i\} \neq \{x_s, x_t\}$, we have $u_i' \neq v_i'$ for $i \in \{3, 4, \ldots, t-1, t+1, \ldots, n-1\}$.

Suppose that there exists a nonempty independent set I' of G' such that

$$|I'| \leq |\{i : \{u_i', v_i'\} \subseteq I', i = 5, 6, \ldots, n-1, i \neq t\}|. \tag{6}$$

Then we have $x_s \in I'$; otherwise, (3) does not hold for the nonempty independent set I' of G.

Let $I_0 = (I' \backslash \{x_s\}) \cup \{u_n, v_n\}$. Then $|I_0| = |I'| + 1$ and I_0 is a nonempty independent set of G. Since $\{u_n, v_n\} \subseteq I_0$ and $\{u_i, v_i\} \subseteq I_0$ whenever $\{u_i', v_i'\} \subseteq I'$ for each $5 \leq i \leq n-1$ with $i \neq t$, we have

$$\begin{aligned} |I_0| &= 1 + |I'| \\ &\leq 1 + |\{i : \{u_i', v_i'\} \subseteq I', i = 5, 6, \ldots, n-1, i \neq t\}| \\ &\leq |\{i : \{u_i, v_i\} \subseteq I_0, i = 5, 6, \ldots, n\}|. \end{aligned}$$

Thus (3) does not hold for the independent set I_0 of G, a contradiction.

Hence $G' = (G - x_n) \cdot u_n v_n = (G - x) \cdot uv \in \Gamma$. □

We are now ready to prove our main result in this section.

THEOREM 2.1. *For any $G \in \Gamma$, $(-1)^{v(G)} P(G, \lambda) > 0$ for all $\lambda \in (1, 2)$.*

Proof. We have $v(G) \geq 2$. If $2 \leq v(G) \leq 3$, then $G = K_n$, where $n = 2$ or 3. Thus the result holds for $v(G) \leq 3$. Assume that the result holds for all $G \in \Gamma$ with $v(G) < n$, where $n \geq 4$.

Suppose on the contrary that the result does not hold for some graph of order n in Γ. Let G be such a graph with minimum $e(G)$. Let $\lambda \in (1,2)$ be such that $(-1)^{v(G)}P(G,\lambda) < 0$.

By definition, G has a γ-ordering $(x_1, x_2, \ldots, x_n)$. Then there exist $u_i, v_i \in N(x_i) \cap V_{i-1}$ with $u_i \neq v_i$ for $i = 3, 4, \ldots, n$ such that (3) holds for every nonempty independent set I of G.

It is clear that $d(x_n) \geq 2$.

Case 1. $d(x_n) = 2$.

So $N(x_n) = \{u_n, v_n\}$. By Lemma 2.3, $G - x_n \in \Gamma$. If $u_n v_n \in E(G)$, then

$$(-1)^n P(G,\lambda) = (2-\lambda)(-1)^{n-1}P(G - x_n, \lambda) > 0,$$

a contradiction. Hence $u_n v_n \notin E(G)$.

By Lemma 2.4, $(G - x_n) + u_n v_n \in \Gamma$ and, by Lemma 2.6, $(G - x_n) \cdot u_n v_n \in \Gamma$. Since

$$\begin{aligned} P(G,\lambda) &= P(G + u_n v_n, \lambda) + P(G \cdot u_n v_n, \lambda) \\ &= (\lambda - 2)P((G - x_n) + u_n v_n, \lambda) + (\lambda - 1)P((G - x_n) \cdot u_n v_n, \lambda), \end{aligned}$$

we have

$$\begin{aligned} (-1)^n P(G,\lambda) &= (2-\lambda)(-1)^{n-1}P((G - x_n) + u_n v_n, \lambda) \\ &\quad + (\lambda - 1)(-1)^{n-2}P((G - x_n) \cdot u_n v_n, \lambda) \\ &> 0, \end{aligned}$$

a contradiction.

Case 2. $d(x_n) \geq 3$.

There exists $w \in N(x_n) \backslash \{u_n, v_n\}$. By Lemma 2.5, $G - x_n w \in \Gamma$. Since $G - x_n \in \Gamma$ and $G \cdot x_n w$ can be considered as a graph obtained from $G - x_n$ by adding some edges, by Lemma 2.4, $G \cdot x_n w \in \Gamma$. Thus, by the assumption on the minimality of $e(G)$,

$$(-1)^n P(G - x_n w, \lambda) > 0$$

and

$$(-1)^{n-1} P(G \cdot x_n w, \lambda) > 0.$$

Hence

$$(-1)^n P(G,\lambda) = (-1)^n P(G - x_n w, \lambda) + (-1)^{n-1} P(G \cdot x_n w, \lambda) > 0,$$

a contradiction.

Therefore $(-1)^{v(G)}P(G,\lambda) > 0$ for all $G \in \Gamma$ with $v(G) = n$. □

3. Graphs in Γ. In the preceding section, it is shown that every graph in Γ has no chromatic zeros in $(1,2)$. However, we don't know exactly what graphs are included in Γ, although some families of graphs in Γ have been found (see section 4).

Question. *What graphs are included in* Γ*?*

Given a double-link ordering $(x_1, x_2, \ldots, x_n)$ of G, if $(x_1, x_2, \ldots, x_n)$ is a γ-ordering of G, then the following inequality follows from (3) for every nonempty independent set I of G:

$$|I| > |\{i : N(x_i) \cap V_i \subseteq I, i = 5, \ldots, n\}|. \tag{7}$$

We conjecture that the converse is also true.

CONJECTURE 3.1. *Let $(x_1, x_2, \ldots, x_n)$ be a double-link ordering of a graph G. Then $(x_1, x_2, \ldots, x_n)$ is a γ-ordering if the following inequality holds for every nonempty independent set I of G:*

$$|I| > |\{i : N(x_i) \cap V_i \subseteq I, i = 5, \ldots, n\}|. \tag{8}$$

In this section, we present a sufficient condition, which is stronger than (8), for a graph to be in Γ. We first introduce two results.

LEMMA 3.1. *Let $(x_1, x_2, \ldots, x_n)$ be a double-link ordering of a graph G, where $n = v(G) \geq 2$, and $U_0 = \{i : 5 \leq i \leq n, N(x_i) \cap V_{i-1}$ is independent in $G\}$. Then $(x_1, x_2, \ldots, x_n)$ is a γ-ordering if and only if there exist $u_i, v_i \in N(x_i) \cap V_{i-1}$ with $u_i \neq v_i$ for all $i \in U_0$ such that*

$$|I| > |\{i : \{u_i, v_i\} \subseteq I, i \in U_0\}| \tag{9}$$

holds for every nonempty independent set I of G.

Proof. The necessity is obvious by (3) and we need only to prove the sufficiency.

Assume that there exist $u_i, v_i \in N(x_i) \cap V_{i-1}$ with $u_i \neq v_i$ for all $i \in U_0$ such that (9) holds for every nonempty independent set I of G.

Let $U_1 = \{5, 6, \ldots, n\} \backslash U_0$. For each $i \in U_1$, there exist $u_i, v_i \in N(x_i) \cap V_{i-1}$ such that $u_i v_i \in E(G)$. For any independent set I of G, $\{u_i, v_i\} \not\subseteq I$ for each $i \in U_1$, and so

$$\{i : \{u_i, v_i\} \subseteq I, 5 \leq i \leq n\} = \{i : \{u_i, v_i\} \subseteq I, i \in U_0\}.$$

Thus (3) follows from (9) for every nonempty independent set I of G, and therefore $(x_1, x_2, \ldots, x_n)$ is a γ-ordering. □

LEMMA 3.2. *Let $A_1, A_2, \ldots, A_k$ be any k sets such that the following inequality holds for every nonempty set $S \subseteq \{1, 2, \ldots, k\}$:*

$$|S| < \left| \bigcup_{i \in S} A_i \right|. \tag{10}$$

Then there exist $u_i, v_i \in A_i$ with $u_i \neq v_i$ for $i = 1, 2, \ldots, k$ such that

$$|S| < \left| \bigcup_{i \in S} \{u_i, v_i\} \right| \tag{11}$$

holds for every nonempty $S \subseteq \{1, 2, \ldots, k\}$.

Proof. By Hall's theorem, there exist distinct $u_1, u_2, \ldots, u_k$ such that $u_i \in A_i$ for $i = 1, 2, \ldots, k$. We now select all v_i's by the following algorithm:

1. Let $Q = \emptyset$.
2. Choose any $i \in \{1, 2, \ldots, k\} \backslash Q$ such that $A_i \backslash \{u_j : 1 \leq j \leq k, j \notin Q\}$ is not empty. Let v_i be any member in $A_i \backslash \{u_j : 1 \leq j \leq k, j \notin Q\}$ and replace Q by $Q \cup \{i\}$.
3. If $Q = \{1, 2, \ldots, k\}$, then stop; otherwise, go to step 2.

By condition (10),

$$\left(\bigcup_{1 \leq j \leq k \, \frac{j}{\notin} Q} A_i \right) \backslash \{u_j : 1 \leq j \leq k, j \notin Q\} \neq \emptyset$$

and so step 2 is workable. Thus all v_i's can be determined by this algorithm.

Now let $r_1 r_2 \dots r_k$ be the permutation of the set $\{1, 2, \dots, k\}$ such that $v_{r_{i+1}}$ is selected after v_{r_i} in the above algorithm for $i = 1, 2, \dots, k-1$. Notice from step 2 of this algorithm that

$$v_{r_i} \notin \{u_{r_j} : i \le j \le k\}, \quad i = 1, 2, \dots, k.$$

Let S be any nonempty subset of $\{1, 2, \dots, k\} = \{r_1, r_2, \dots, r_k\}$. Let t be the minimum number in $\{1, 2, \dots, k\}$ such that $r_t \in S$. Then

$$v_{r_t} \notin \{u_{r_j} : r_j \in S, j = 1, 2, \dots, k\} = \{u_{r_j} : r_j \in S, j = t, t+1, \dots, k\}$$

and so

$$|\{u_{r_j}, v_{r_j} : r_j \in S, j = 1, 2, \dots, k\}| \ge |\{u_{r_j} : r_j \in S, j = 1, 2, \dots, k\}| + 1 > |S|.$$

Hence (11) follows. □

THEOREM 3.1. *Let $(x_1, x_2, \dots, x_n)$ be a double-link ordering of G, where $n = v(G) \ge 2$. If*

$$|U| < \left| \bigcup_{i \in U} (N(x_i) \cap V_{i-1}) \right| \tag{12}$$

holds for every nonempty $U \subseteq \{i : 5 \le i \le n, N(x_i) \cap V_{i-1}$ is independent in $G\}$, then $(x_1, x_2, \dots, x_n)$ is a γ-ordering of G and so $G \in \Gamma$.

Proof. Let $U_0 = \{i : 5 \le i \le n, N(x_i) \cap V_{i-1}$ is independent in $G\}$. By condition (12) and Lemma 3.2, there exist $u_i, v_i \in N(x_i) \cap V_{i-1}$ with $u_i \ne v_i$ for all $i \in U_0$ such that

$$|U| < \left| \bigcup_{i \in U} \{u_i, v_i\} \right| \tag{13}$$

holds for every nonempty $U \subseteq U_0$.

Let I be any nonempty independent set of G. For

$$U = \{i : \{u_i, v_i\} \subseteq I, i \in U_0\},$$

we have $\bigcup_{i \in U} \{u_i, v_i\} \subseteq I$, and by (13),

$$|\{i : \{u_i, v_i\} \subseteq I, i \in U_0\}| = |U| < \left| \bigcup_{i \in U} \{u_i, v_i\} \right| \le |I|.$$

By Lemma 3.1, $(x_1, x_2, \dots, x_n)$ is a γ-ordering of G. □

4. Some families of graphs in Γ. In this section, we shall show that Γ includes the following families of graphs:

(i) graphs containing a 2-tree as a spanning subgraph;
(ii) 2-connected plane near-triangulations;
(iii) all t-partite graphs, where $t \ge 3$;
(iv) graphs with a Hamiltonian path $x_1 x_2 \dots x_n$ such that $(x_1, x_2, \dots, x_n)$ is a double-link ordering;

(v) graphs obtained from $K_{m,n} = (A, B; E)$, where $2 \le n \le m+1$, by adding one edge joining two vertices in B, where $|A| = m$ and $|B| = n$;

(vi) all $(v(G) - \Delta(G) + 1)$-connected graphs G.

A graph G is called a *chordal graph* if either $v(G) = 1$ or G contains a vertex x such that

(i) $G - x$ is a chordal graph, and

(ii) either $d(x) = 0$ or the subgraph of G induced by $N(x)$ is complete.

By the definition of a chordal graph, it can be shown that every 2-connected chordal graph belongs to Γ, which follows from the next result.

LEMMA 4.1. *Let G be a graph with $v(G) \ge 3$ and $x \in V(G)$. If $G - x \in \Gamma$ and $N(x)$ is not an independent set of G, then $G \in \Gamma$.*

Proof. As $G - x \in \Gamma$, $G - x$ has a γ-ordering $(x_1, x_2, \ldots, x_{n-1})$, where $n = v(G)$. Let $x_n = x$. Since $N(x_n)$ is not an independent set of G,

$$\begin{aligned} &\{i : N(x_i) \cap V_{i-1} \text{ is independent, } 5 \le i \le n\} \\ &= \{i : N(x_i) \cap V_{i-1} \text{ is independent, } 5 \le i \le n-1\}. \end{aligned}$$

By Lemma 3.1, the result holds. □

COROLLARY 4.1. *If a graph G contains a double-link ordering $(x_1, x_2, \ldots, x_n)$, where $n = v(G) \ge 2$, such that $N(x_i) \cap V_{i-1}$ is not independent for all $i = 5, 6, \ldots, n$, then $G \in \Gamma$.*

Dong and Koh [3] showed that if G contains a 2-tree as a spanning subgraph, then G contains no chromatic zeros in $(1, 2)$. By Corollary 4.1, such a graph actually belongs to Γ, so their result follows from Theorem 2.1.

Birkhoff and Lewis [1] showed that every plane near-triangulation has no chromatic zeros in $(1, 2)$. By Corollary 4.1, every 2-connected plane near-triangulation belongs to $G \in \Gamma$, because if $v(G) \ge 4$, then G contains a vertex x such that $G - x$ is a 2-connected plane near-triangulation and $N(x)$ is not independent. Thus, Birkhoff and Lewis' result is a special case of Theorem 2.1.

COROLLARY 4.2. *Every complete t-partite graph G, where $t \ge 3$, contains a 2-tree as a spanning subgraph and hence belongs to Γ.*

Proof. Let $xyzx$ be a triangle in G. For every $w \in V(G) \backslash \{x, y, z\}$, we have $|N(w) \cap \{x, y, z\}| \ge 2$. Thus G contains a spanning 2-tree, and so $G \in \Gamma$. □

Thomassen [8] showed that any graph with a Hamiltonian path has no chromatic zeros in $(1, t_0]$, where

$$t_0 = \frac{2}{3} + \frac{1}{3}\sqrt[3]{26 + 6\sqrt{33}} + \frac{1}{3}\sqrt[3]{26 - 6\sqrt{33}} = 1.29559\ldots, \tag{14}$$

but for any $\epsilon > 0$, there exists a graph with a Hamiltonian path which has a chromatic zero in $(t_0, t_0 + \epsilon)$.

By Theorem 2.1, we will show that there is a large family of graphs with a Hamiltonian path which have no chromatic zeros in $(1, 2)$.

LEMMA 4.2. *Let G be a graph and $x \in V(G)$ with $d(x) \ge 2$. If $G - x$ has a γ-ordering $(x_1, x_2, \ldots, x_{n-1})$, where $n = v(G)$, such that $xx_{n-1} \in E(G)$, then $(x_1, x_2, \ldots, x_{n-1}, x)$ is a γ-ordering of G and thus $G \in \Gamma$.*

Proof. Assume that $(x_1, x_2, \ldots, x_{n-1})$ is a γ-ordering of $G - x$ and $xx_{n-1} \in E(G)$. Then there exist $u_i, v_i \in N(x_i) \cap V_{i-1}$ with $u_i \ne v_i$ for $i = 3, 4, \ldots, n-1$ such that

$$|I| > |\{i : \{u_i, v_i\} \subseteq I, i = 5, \ldots, n-1\}| \tag{15}$$

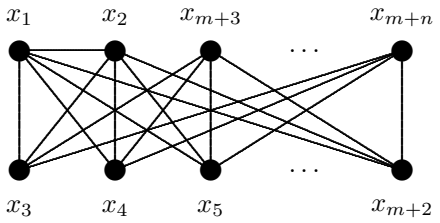


FIG. 1.

holds for every nonempty independent set I of G.

Let $u_n = x_{n-1}$ and v_n be any vertex in $N(x)\backslash\{x_{n-1}\}$. Suppose that I is a nonempty independent set I of G such that

$$|I| \leq |\{i : \{u_i, v_i\} \subseteq I, i = 5, \ldots, n\}|. \tag{16}$$

Then, by (15), $\{u_n, v_n\} \subseteq I$ and so $x_{n-1} \in I$. However, as $x_{n-1} \notin \{u_i, v_i : i = 1, 2, \ldots, n-1\}$, by (15) again,

$$\begin{aligned} |I| &= 1 + |I\backslash\{x_{n-1}\}| \\ &> 1 + |\{i : \{u_i, v_i\} \subseteq I\backslash\{x_{n-1}\}, i = 5, \ldots, n-1\}| \\ &= |\{i : \{u_i, v_i\} \subseteq I, i = 5, \ldots, n\}|, \end{aligned}$$

contradicting (16). Hence the result holds. □

The next result follows directly from Lemma 4.2.

THEOREM 4.1. *Let G be a graph with a Hamiltonian path $x_1x_2\ldots x_n$, where $n \geq 2$. If $(x_1, x_2, \ldots, x_n)$ is a double-link ordering of G, then $G \in \Gamma$.*

By Lemma 1.1, a 2-connected bipartite graph G has chromatic zeros in $(1, 2)$ as long as $v(G)$ is odd. The next result shows that for any 2-connected complete bipartite graph, adding one edge properly to this graph produces a graph having no chromatic zeros in $(1, 2)$.

THEOREM 4.2. *Let m, n be integers with $2 \leq n \leq m + 1$. Let G be the graph obtained from the complete bipartite graph $K_{m,n} = (A, B; E)$ by adding one edge joining any two vertices in B, where $|A| = m$ and $|B| = n$. Then $G \in \Gamma$.*

Proof. Let $A = \{x_3, x_4, \ldots, x_{m+2}\}$ and $B = \{x_1, x_2, x_{m+3}, \ldots, x_{m+n}\}$. Assume that G is obtained from the complete bipartite graph $K_{m,n} = (A, B; E)$ by adding one edge x_1x_2, as shown in Figure 1.

Note that $(x_1, x_2, \ldots, x_{m+n})$ is a double-link ordering of G. For $i \geq 3$, $N(x_i) \cap V_{i-1}$ is independent if and only if $i \in \{m + 3, m + 4, \ldots, m + n\}$. Since

$$N(x_i) \cap V_{i-1} = \{x_3, x_4, \ldots, x_{m+2}\}$$

for all $i = m + 3, m + 4, \ldots, m + n$, the inequality

$$|U| \leq n - 2 < m = |\{x_3, x_4, \ldots, x_{m+2}\}| = \left|\bigcup_{i \in U} (N(x_i) \cap V_{i-1})\right|$$

holds for every nonempty $U \subseteq \{m + 3, m + 4, \ldots, m + n\}$. Hence, by Theorem 3.1, $(x_1, x_2, \ldots, x_{m+n})$ is a γ-ordering of G. □

Clearly, $0 \le \Delta(G) \le v(G) - 1$ holds for each graph G. It is easy to show that if $\Delta(G) = v(G) - 1$, then $P(G, \lambda) > 0$ for all $\lambda \in (1, 2)$. Indeed, it can be shown that if $\Delta(G) = v(G) - 1$ and G is 2-connected, then G contains a 2-tree as a spanning subgraph and so $G \in \Gamma$ by Corollary 4.1. In this section, we shall generalize this result.

THEOREM 4.3. *If a graph G is $(v(G) - \Delta(G) + 1)$-connected, then $G \in \Gamma$.*

Proof. Let $x \in V(G)$ such that $d(x) = \Delta(G)$. Let

$$k = |V(G) \backslash (N(x) \cup \{x\})| = v(G) - \Delta(G) - 1.$$

Then G is $(k+2)$-connected, and thus the subgraph induced by $N(x) \cup \{x\}$, denoted by H, is 2-connected. Since $d_H(x) = v(H) - 1$ and H is 2-connected, H contains a spanning 2-tree. Thus H contains a double-link ordering $(x_1, x_2, \ldots, x_t)$ such that $N_H(x_i) \cap V_{i-1}$ is not independent in H for all $i = 3, 4, \ldots, t$, where $t = v(H) = 1 + d_G(x) = 1 + \Delta(G)$.

Let $V(G) \backslash (N(x) \cup \{x\}) = \{x_{t+1}, x_{t+2}, \ldots, y_{t+k}\}$. Since G is $(k+2)$-connected, we have $d_G(x_{t+i}) \ge k + 2$ for each $1 \le i \le k$ and thus

$$\begin{aligned} |N_G(x_{t+i}) \cap V_{t+i-1}| &= d_G(x_{t+i}) - |N_G(x_{t+i}) \cap \{x_{t+j} : j = i+1, \ldots, k\}| \\ &\ge k + 2 - (k - i) \\ &= i + 2. \end{aligned}$$

Let U be any nonempty subset of $\{t + i : i = 1, 2, \ldots, k\}$ and $t + r$ be the maximum number in U. Then

$$|U| \le |r| < |N_G(x_{t+r}) \cap V_{t+r-1}| \le \left| \bigcup_{j \in U} (N_G(x_j) \cap V_{j-1}) \right|.$$

By Theorem 3.1, $(x_1, x_2, \ldots, x_{t+k})$ is a γ-ordering of G. □

The conclusion of Theorem 4.3 is no longer true if G is not $(v(G) - \Delta(G) + 1)$-connected. Consider the complete bipartite graph $K_{m,n}$, where $2 \le n \le m$. Observe that

$$v(K_{m,n}) - \Delta(K_{m,n}) + 1 = (m + n) - m + 1 = n + 1 \tag{17}$$

and $K_{m,n}$ is n-connected but not $(n+1)$-connected. However, by Lemma 1.1, $K_{m,n}$ has chromatic zeros in $(1, 2)$ if $m + n$ is odd.

Dong and Koh [2] showed that if $\Delta(G) \ge v(G) - 2$, then G contains no chromatic zeros in the interval $(1, d)$, where

$$d = \frac{5}{3} + \frac{1}{6}\sqrt[3]{12\sqrt{69} - 44} - \frac{1}{6}\sqrt[3]{12\sqrt{69} + 44} = 1.430159709\ldots. \tag{18}$$

Furthermore this result does not hold if d is replaced by any larger number. By Theorem 4.3, however, we have the following corollary.

COROLLARY 4.3. *If G is a 3-connected graph with $\Delta(G) = v(G) - 2$, then $G \in \Gamma$, and so G contains no chromatic zeros in $(1, 2)$.*

5. A necessary condition. In this section, we first present a necessary condition for a graph G to be in Γ and then propose some conjectures related to the existence of chromatic zeros in $(1, 2)$. Let $c(H)$ denote the number of components of a graph H.

LEMMA 5.1. *Let $(x_1, x_2, \ldots, x_n)$ be any ordering of the vertices in a graph G, where $n = v(G) \geq 2$. Then for any $S \subseteq V(G)$, there exists an independent set T of G such that $|T| = c(G - S)$ and*

$$T \subseteq \{x_i \in V(G)\backslash S : N(x_i) \cap V_i \subseteq S, i = 1, 2, \ldots, n\}. \tag{19}$$

Proof. Let $c(G - S) = c$ and $G_1, G_2, \ldots, G_c$ be the components of $G - S$. Let $T = \{x_{m_k} : k = 1, 2, \ldots, c\}$, where $m_k = \min\{j : x_j \in V(G_k), j = 1, 2, \ldots, n\}$. It is clear that T is independent in G. Since $(N(x_{m_k}) \cap V_{m_k}) \cap V(G_k) = \emptyset$, we have $N(x_{m_k}) \cap V_{m_k} \subseteq S$ for each $k = 1, 2, \ldots, c$. Thus the lemma holds. □

THEOREM 5.1. *For any $G \in \Gamma$, $c(G - S) \leq |S|$ holds for every nonempty independent set S of G.*

Proof. Let $n = v(G)$. The result is obvious if $n \leq 4$. Assume that $n \geq 5$.

Let $(x_1, x_2, \ldots, x_n)$ be any γ-ordering of G. Let $u_i, v_i \in N(x_i) \cap V_{i-1}$ for $i = 5, 6, \ldots, n$ be such that inequality (3) holds for every nonempty independent set I of G. Suppose that there exists a nonempty independent set S of G with $c(G-S) \geq |S|+1$. We shall show that (3) does not hold for S, a contradiction.

By Lemma 5.1, there exists an independent set T of G such that $|T| = |S| + 1$ and

$$T \subseteq \{x_i \in V(G)\backslash S : N(x_i) \cap V_i \subseteq S, i = 1, 2, \ldots, n\}.$$

Since $x_1x_2 \in E(G)$, $\{x_1, x_2\} \not\subseteq T$. For $i = 3, 4$, $N(x_i) \cap V_i$ is not independent in G, and so $x_i \notin T$. Hence $|T \cap \{x_5, \ldots, x_n\}| \geq |T| - 1 = |S|$. For each $i = 5, 6, \ldots, n$, if $x_i \in T$, then $\{u_i, v_i\} \subseteq N(x_i) \cap V_i \subseteq S$. Thus

$$|\{i : \{u_i, v_i\} \subseteq S, 5 \leq i \leq n\}| \geq |T \cap \{x_5, \ldots, x_n\}| \geq |S|,$$

a contradiction. □

By Theorem 5.1, the condition in the following conjecture is thus weaker than that in Theorem 4.1.

CONJECTURE 5.1. *Let G be a graph with a Hamiltonian path. If $c(G - S) \leq |S|$ holds for every nonempty independent set S of G, then $(-1)^{v(G)}P(G, \lambda) > 0$ holds for all $\lambda \in (1, 2)$.*

We believe that the condition "G has a Hamiltonian path" in Conjecture 5.1 is redundant.

CONJECTURE 5.2. *If $c(G-S) \leq |S|$ holds for every nonempty independent set S of a graph G, then $(-1)^{v(G)}P(G, \lambda) > 0$ holds for all $\lambda \in (1, 2)$.*

It is obvious that the condition in the following conjecture, which was first proposed by Thomassen [10], is stronger than that in Conjectures 5.1 and 5.2, because for a Hamiltonian graph G, the inequality $c(G - S) \leq |S|$ holds for every $S \subseteq V(G)$.

CONJECTURE 5.3. *If G is a Hamiltonian graph, then $(-1)^{v(G)}P(G, \lambda) > 0$ holds for all $\lambda \in (1, 2)$.*

Acknowledgment. The authors wish to thank the referees for their very helpful suggestions and comments.

REFERENCES

[1] G. D. Birkhoff and D. C. Lewis, *Chromatic polynomials*, Trans. Amer. Math. Soc., 60 (1946), pp. 355–451.
[2] F. M. Dong and K. M. Koh, *Domination numbers and zeros of chromatic polynomials*, Discrete Math., submitted.
[3] F. M. Dong and K. M. Koh, *Two results on real zeros of chromatic polynomials*, Combin. Probab. Comput., 13 (2004), pp. 809–813.
[4] F. M. Dong, K. M. Koh, and K. L. Teo, *Chromatic Polynomials and Chromaticity of Graphs*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2005.
[5] B. Jackson, *A zero-free interval for chromatic polynomials of graphs*, Combin. Probab. Comput., 2 (1993), pp. 325–336.
[6] R. C. Read, *An introduction to chromatic polynomials*, J. Combin. Theory, 4 (1968), pp. 52–71.
[7] R. C. Read and W. T. Tutte, *Chromatic polynomials*, in Selected Topics in Graph Theory 3, L. W. Beineke and R. J. Wilson, eds., Academic Press, San Diego, 1988, pp. 15–42.
[8] C. Thomassen, *Chromatic zeros and Hamiltonian paths*, J. Combin. Theory Ser. B, 80 (2000), pp. 218–224.
[9] C. Thomassen, *The zero-free intervals for chromatic polynomials of graphs*, Combin. Probab. Comput., 6 (1997), pp. 497–506.
[10] C. Thomassen, *On the number of Hamiltonian cycles in bipartite graphs*, Combin. Probab. Comput., 5 (1996), pp. 437–442.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 811–828

UNIFORM FORMULAE FOR COEFFICIENTS OF MEROMORPHIC FUNCTIONS IN TWO VARIABLES. PART I*

MANUEL LLADSER†

Abstract. Uniform asymptotic formulae for arrays of complex numbers of the form $(f_{r,s})$, with r and s nonnegative integers, are provided as r and s converge to infinity at a comparable rate. Our analysis is restricted to the case in which the generating function $F(z,w) := \sum f_{r,s} z^r w^s$ is meromorphic in a neighborhood of the origin. We provide uniform asymptotic formulae for the coefficients $f_{r,s}$ along directions in the (r,s)-lattice determined by regular points of the singular variety of F. Our main result derives from the analysis of a one dimensional parameter-varying integral describing the asymptotic behavior of $f_{r,s}$. We specifically consider the case in which the phase term of this integral has a unique stationary point; however, we allow the possibility that one or more stationary points of the amplitude term coalesce with this. Our results find direct application in certain problems associated to the Lagrange inversion formula as well as bivariate generating functions of the form $v(z)/(1 - w \cdot u(z))$.

Key words. asymptotic enumeration, analysis of algorithms, bivariate generating functions, canonical representations, coalescing saddles, combinatorial enumeration, discrete random structures, uniform asymptotic expansions

AMS subject classifications. 05A16, 41A60

DOI. 10.1137/040620849

1. Introduction.

Suppose that $G(z,w)$ and $H(z,w)$ are analytic functions of the complex variables z and w in an open polydisk centered at the origin and assume that $H(0,0) \neq 0$. Then, the function

$$F(z,w) := \frac{G(z,w)}{H(z,w)}$$

is analytic in a neighborhood of the origin in $\mathbb{C}^2$; in particular, it has a power series expansion of the form $\sum f_{r,s} z^r w^s$, where the indices r and s are nonnegative integers. In what follows we use the notation $[z^r w^s]F$ to refer to the coefficient of $z^r w^s$ in the power series expansion of F. We also use the notation $(r,s) \to \infty$ as a shorthand for $r \to \infty$ and $s \to \infty$.

Generating functions of the above form occur frequently in the study of discrete random structures and analysis of algorithms (see [14] for a comprehensive account of examples). For a wide class of bivariate functions of this kind the coefficients $[z^r w^s]F$ are expected, up to an exponential factor, to be of order $s^{-(p+1)/n}$ as $(r,s) \to \infty$ with r/s fixed. Here, the coefficients p and n are functions of the ratio r/s. In particular, the asymptotic behavior of $[z^r w^s]F$ can be understood even if r/s varies but in such a way that p and n do not change. In this paper we show how to provide uniform asymptotic formulae for the coefficients $[z^r w^s]F$ as $(r,s) \to \infty$ when r/s is restricted to a set of values where the coefficient p may not remain constant.

From this point on we assume as given a point (ζ, ω) which is a strictly minimal simple zero of H. By simple zero we mean that $H(\zeta,\omega) = 0$; however, the complex

*Received by the editors December 15, 2004; accepted for publication (in revised form) April 5, 2006; published electronically December 5, 2006.
http://www.siam.org/journals/sidma/20-4/62084.html
†Department of Applied Mathematics, University of Colorado at Boulder, P.O. Box 526 UCB, Boulder, CO 80309-0526 (manuel.lladser@colorado.edu).

gradient $\nabla H(\zeta,\omega) \neq 0$. By strictly minimal zero we mean that $\zeta \cdot \omega \neq 0$ and that (ζ,ω) is the only zero of $H(z,w)$ in the polydisk where $|z| \leq |\zeta|$ and $|w| \leq |\omega|$.

The pioneering work of Pemantle and Wilson [13] implies that it is possible to determine an asymptotic expansion for the coefficients of F only along a certain direction in the (r,s)-lattice specified by (ζ,ω). This direction corresponds to the line

$$\operatorname{dir}(\zeta,\omega) := \{(r,s) \in \mathbb{R}^2 : r \cdot \omega\, H_w(\zeta,\omega) - s \cdot \zeta\, H_z(\zeta,\omega) = 0\}, \tag{1.1}$$

where H_w and H_z, respectively, denote the complex partial derivative of H with respect to w and z. For simplicity it will be assumed ahead that $H_w(\zeta,\omega) \neq 0$. In particular, $(r,s) \in \operatorname{dir}(\zeta,\omega)$ if and only if $r/s = d(\zeta,\omega)$, where

$$d(\zeta,\omega) := \frac{\zeta\, H_z(\zeta,\omega)}{\omega\, H_w(\zeta,\omega)}\,.$$

The strict minimality of (ζ,ω) implies that $d(\zeta,\omega) \geq 0$ (see Lemma 2.1 in [13]). Furthermore, if this quantity is a rational number, Pemantle and Wilson show that there are integers $n = n(\zeta,\omega) \geq 2$ and $p = p(\zeta,\omega) \geq 0$ and coefficients $c_j = c_j(\zeta,\omega)$, with $j \geq p$ and $c_p \neq 0$, such that

$$[z^r w^s]F \approx \frac{\zeta^{-r}\omega^{-s}}{2\pi} \sum_{j=p}^{\infty} c_j\, s^{-(j+1)/n}, \tag{1.2}$$

for all $(r,s) \in \operatorname{dir}(\zeta,\omega)$, as $(r,s) \to \infty$. The asymptotic notation used above is in the standard sense where the sequence $(s^{-(j+1)/n})_{j\geq p}$ is the so called auxiliary asymptotic sequence. This means that the difference between the left- and right-hand side terms above, with the summation truncated to the term in which $j = m$, is $O(\zeta^{-r}\,\omega^{-s}\,s^{-(m+2)/n})$, as $(r,s) \to \infty$.

The technique used to obtain the asymptotic formula in (1.2) proceeds by relating the coefficient $[z^r w^s]F$, with $(r,s) \in \operatorname{dir}(\zeta,\omega)$, to a one dimensional Fourier or Laplace like integral of the form

$$\frac{\zeta^{-r}\omega^{-s}}{2\pi} \int a(\theta;\zeta,\omega) \exp\{-s \cdot f(\theta;\zeta,\omega)\} d\theta\,. \tag{1.3}$$

We will refer loosely to $a(\theta;\zeta,\omega)$ and $f(\theta;\zeta,\omega)$, respectively, as the derived amplitude and phase term. Roughly speaking, the expansion in (1.2) is in powers of $s^{-1/n}$ because the derived phase term vanishes to degree n in the variable θ about $\theta = 0$, which turns out to be the dominant critical point of the integral. Furthermore, $[z^r w^s]F$ is of order $\zeta^{-r}\,\omega^{-s}\,s^{-(p+1)/n}$ because the derived amplitude term vanishes to degree p at $\theta = 0$. (See [6] and [3] for a compelling introduction to the main techniques used to study the asymptotic behavior of Fourier–Laplace integrals.)

The asymptotic expansion in (1.2) holds usually along a wider set of directions in the (r,s)-lattice. Indeed, suppose that K is a compact set of strictly minimal simple zeros of H and consider the set

$$\Lambda := \bigcup_{(\zeta,\omega)\in K} \left\{(r,s) \in \mathbb{R}^2 :\, \frac{r}{s} = d(\zeta,\omega)\right\}. \tag{1.4}$$

Observe that Λ is a cone if K is connected. Theorem 3.3 in [13] implies that (1.2) holds uniformly as $(r,s) \to \infty$, with $(r,s) \in \Lambda$, provided that the derived amplitude

and phase term in (1.3) do not change their degree of vanishing about $\theta = 0$ as (ζ, ω) varies over K. In particular, if for each $(r, s) \in \Lambda$, $(\zeta, \omega) = (\zeta(r, s), \omega(r, s)) \in K$ is such that $d(\zeta, \omega) = r/s$, then

$$[z^r w^s]F \sim c_p(\zeta, \omega) \cdot \frac{\zeta^{-r}\omega^{-s}}{2\pi} \cdot s^{-(p+1)/n}, \tag{1.5}$$

uniformly for all $(r, s) \in \Lambda$, as $(r, s) \to \infty$. The notation used above means that the ratio of the two sides tends to 1 as $(r, s) \to \infty$.

However, when there is a change of degree of the derived amplitude or phase term in (1.3) the hypotheses of [13] are not met and therefore no conclusion may be drawn from it. When the change of degree is in the phase term one needs to build a bridge between differently scaled regions. This is hard work and will be presented in the forthcoming paper [11]. (For further details on this case, see Theorem 6.6 in Chapter 6 in [12].)

The main contribution of the present paper is to settle the case in which only the derived amplitude term may undergo a change of degree. Although it is not mentioned in [13] the asymptotic formula in (1.2) is still valid for $(r, s) \in \Lambda$ but it requires a more careful interpretation. To amplify on this consider the case in which at a particular point $(\zeta_c, \omega_c) \in K$, the derived amplitude term in (1.3) vanishes to degree q yet, for all $(\zeta, \omega) \in K$ nearby (ζ_c, ω_c), the derived amplitudes vanish to some degree $p < q$. Then (1.5) implies that up to the exponential factor $\zeta^{-r}\omega^{-s}$,

$$[z^r w^s]F \text{ is of order } \begin{cases} s^{-(q+1)/n} \text{ if } r/s = d(\zeta_c, \omega_c); \\ s^{-(p+1)/n} \text{ otherwise,} \end{cases} \tag{1.6}$$

as $(r, s) \in \Lambda$ goes to infinity, provided that r/s remains constant.

A problem of interest is how to bridge the gap between the asymptotic orders in (1.6) as r/s approaches $d(\zeta_c, \omega_c)$, as $(r, s) \to \infty$. As we shall see in the coming section, we resolve this problem with great generality, and can provide a uniform asymptotic expansion for the coefficients $[z^r w^s]F$ as long as $(r, s) \in \Lambda$ and $|r/s - d(\zeta_c, \omega_c)|$ is sufficiently small. Our main result builds upon the asymptotic analysis of an integral such as the one in (1.3), which does not rely on having the term $a(\theta; \zeta, \omega)$ vanish to constant degree as (ζ, ω) varies over K. The technique we propose to analyze integrals of this kind draws on the techniques of Chester, Friedman, and Ursell [5], the results of Levinson on polynomial canonical representations [8], [9], and the work of Pemantle and Wilson [13]. All of these techniques are founded on complex variable methods. For a compelling introduction to function theory of one or several complex variables, see [4], [15], or [17].

Under appropriate hypotheses, our main result implies that $[z^r w^s]F$ has (up to an exponentially decreasing factor) an asymptotic expansion of the form

$$[z^r w^s]F = \sum_{j=p}^{q} c_j(r/s) \cdot s^{-(j+1)/n} + o(s^{-(q+1)/n}), \tag{1.7}$$

where the coefficients $c_j(r/s)$ are analytic functions of r/s and, except for $j = q$, they all vanish when $r/s = d(\zeta_c, \omega_c)$. Furthermore, the above expansion is uniform as $(r, s) \to \infty$ provided that r/s is sufficiently close to $d(\zeta_c, \omega_c)$. Observe how the condition of having $c_j(d(\zeta_c, \omega_c)) = 0$, for $j \neq q$, and $c_q(d(\zeta_c, \omega_c)) \neq 0$ explains the asymptotic behavior described in (1.6).

In well-behaved situations one finds that the coefficients $c_j(r/s)$ in (1.7) are all nonnegative. This sign constraint prevents cancellation between the terms participating in the summation in (1.7). As a result, one obtains that

$$[z^r w^s]F = (1 + o(1)) \cdot \sum_{j=p}^{q} c_j(r/s) \cdot s^{-(j+1)/n},$$

if $r/s \to d(\zeta_c, \omega_c)$ as $(r, s) \to \infty$. On the contrary, when the coefficients $c_j(r/s)$ have mixed signs, and depending on the rate at which r/s approaches to $d(\zeta_c, \omega_c)$, it is possible that terms in the summation in (1.7) cancel one another and therefore

$$\sum_{j=p}^{q} c_j(r/s) \cdot s^{-(j+1)/n} = o(s^{-(q+1)/n}).$$

Thus, in situations where $r/s \to d(\zeta_c, \omega_c)$ in such away that the above asymptotic formula holds, our main result allows us to conclude only that $[z^r w^s]F$ is $o(s^{-(q+1)/n})$ as $(r, s) \to \infty$. The effect of cancellation to determine asymptotic formulae for the coefficients $[z^r w^s]F$ is illustrated in Example 2.5 in the next section.

2. Main definitions and results with applications. To state our main definition we recall that if $U(z, w)$ is analytic in an open neighborhood of a point (z_0, w_0) in $\mathbb{C} \times \mathbb{C}$, then it is possible to represent U in the form

$$U(z, w) = \sum_{k=0}^{\infty} U_k(z) \cdot (w - w_0)^k,$$

where $U_k(z) := \frac{1}{k!} \frac{\partial^k U}{\partial w^k}(z, w_0)$. The above series is usually referred to as the Hartogs series of U in powers of $(w - w_0)$ about the point (z_0, w_0). This series is uniformly convergent for all (z, w) in polydisks of the form $\{(z, w) : |z - z_0| \leq \epsilon, |w - w_0| \leq \epsilon\}$ provided that the polydisk is completely contained in the domain where U is analytic (see section 4.5 in [12]).

To state our main result the following definition will be used.

DEFINITION 2.1. *Given nonnegative integers $p < q$ and a function $U(z, w)$ analytic in an open neighborhood of a point (z_0, w_0) in $\mathbb{C} \times \mathbb{C}$, we say that U has a p-to-q change of degree about $w = w_0$ as $z \to z_0$ provided that the Hartogs series of U in powers of $(w - w_0)$ about the point (z_0, w_0) is of the form $U(z, w) = U_p(z) \cdot (w - w_0)^p + \cdots + U_q(z) \cdot (w - w_0)^q + \cdots$, where $U_j(z_0) = 0$ for all $p \leq j < q$; however, $U_q(z_0) \neq 0$. On the contrary, if $U(z, w) = U_p(z) \cdot (w - w_0)^p + \cdots$ with $U_p(z_0) \neq 0$, we say that U vanishes to constant degree p about $w = w_0$ as $z \to z_0$. Alternatively, we will sometimes say that U has a p-to-p change of degree about $w = w_0$ as $z \to z_0$.*

In what follows, $G(z, w)$ and $H(z, w)$ are given analytic functions in some open polydisk D centered at the origin in $\mathbb{C}^2$ and it is assumed that $H(0, 0) \neq 0$. We also assume as given a compact set $K \subset D$ of strictly minimal simple zeros of H containing a particular point (ζ_c, ω_c) such that $H_w(\zeta_c, \omega_c) \neq 0$. The implicit function theorem (see IV.5.6 in [4]) lets us then parametrize the zero set of H near (ζ_c, ω_c) in the form $\omega = g(\zeta)$, where g is a certain analytic function of ζ near $\zeta = \zeta_c$.

For each $(\zeta, \omega) \in K$, $\mathrm{dir}(\zeta, \omega)$ is the line defined as in (1.1) and Λ is the cone defined in (1.4). For each $(r, s) \in \Lambda$ such that $r/s = d(\zeta_c, \omega_c)$ we define $(\zeta(r, s), \omega(r, s)) := (\zeta_c, \omega_c)$. Furthermore, for each $(r, s) \in \Lambda$ we let $(\zeta, \omega) = (\zeta(r, s), \omega(r, s)) \in K$ be such

that $(r,s) \in \mathrm{dir}(\zeta,\omega)$. For the validity of our main result, we require the continuity condition

$$(\zeta(r,s),\omega(r,s)) \to (\zeta_c,\omega_c),$$

as $r/s \to d(\zeta_c,\omega_c)$. Indeed, since

$$(r,s) \in \mathrm{dir}(\zeta,\omega) \iff \frac{r}{s} = -\frac{\zeta g'(\zeta)}{g(\zeta)},$$

to satisfy the continuity condition it is enough to select $\zeta(r,s) = \zeta$ and $\omega(r,s) = g(\zeta)$, where ζ is the closest solution to $\zeta = \zeta_c$ (among a finite number of solutions) to the equation above. In particular, ζ and ω can be thought of as homogeneous functions of degree zero in the variable (r,s).

We define

$$a(\zeta,\theta) := \frac{-G(\zeta e^{i\theta}, g(\zeta e^{i\theta}))}{g(\zeta e^{i\theta})\, H_w(\zeta e^{i\theta}, g(\zeta e^{i\theta}))}, \tag{2.1}$$

$$f(\zeta,\theta) := \ln\left\{\frac{g(\zeta e^{i\theta})}{g(\zeta)}\right\} - i\theta\frac{\zeta g'(\zeta)}{g(\zeta)}, \tag{2.2}$$

which are analytic for all θ sufficiently small and ζ sufficiently close to ζ_c.

Our main result is as follows.

THEOREM 2.2. *Let $G(z,w)$, $H(z,w)$, K, (ζ_c,ω_c), etc., be as above. Define $F(z,w) := G(z,w)/H(z,w)$. If there are nonnegative integers $p \le q$ such that $a(\zeta,\theta)$ has a p-to-q change of degree about $\theta = 0$ as $\zeta \to \zeta_c$, while $f(\zeta,\theta)$ vanishes to constant degree n about $\theta = 0$ as $\zeta \to \zeta_c$, then there is a constant $C > 0$ and functions $A_k(\zeta)$ and $B_k(\zeta;s)$, with $p \le k \le q$, analytic in ζ near $\zeta = \zeta_c$, such that*

$$A_k(\zeta_c) = 0,\ p \le k < q, \tag{2.3}$$

$$A_q(\zeta_c) \neq 0, \tag{2.4}$$

and

$$[z^r w^s]F = \frac{\zeta^{-r}\omega^{-s}}{2\pi}\left\{\sum_{k=p}^{q} A_k(\zeta)\cdot B_k(\zeta;s) + O(e^{-s\cdot C})\right\}, \tag{2.5}$$

uniformly for all $(r,s) \in \Lambda$ such that r/s is sufficiently close to $d(\zeta_c,\omega_c)$. Furthermore, there are coefficients $c_k(\zeta;j)$, with $j \ge k$, which are analytic in ζ near $\zeta = \zeta_c$ such that each coefficient B_k above admits an asymptotic expansion of the form

$$B_k(\zeta;s) \approx \sum_{j=k}^{\infty} c_k(\zeta;j)\cdot\left(1 + (-1)^j\cdot D(j,n)\right)\cdot\frac{1}{n}\Gamma\left(\frac{j+1}{n}\right)\cdot s^{-(j+1)/n}, \tag{2.6}$$

as $s \to \infty$, uniformly for all $(r,s) \in \Lambda$ such that r/s is sufficiently close to $d(\zeta_c,\omega_c)$, where we have defined

$$D(j,n) := \begin{cases} 1, & n \text{ even}; \\ \exp\left(-\frac{i\pi(j+1)}{n}\cdot sign\{i\cdot[\theta^n]f(\zeta_c,\theta)\}\right), & n \text{ odd}. \end{cases} \tag{2.7}$$

Remark 2.1. The analytic coefficients A_k in (2.3) and (2.4) together with an auxiliary function $\alpha = \alpha(\zeta, \theta)$ are the unique analytic solutions (near $\zeta = \zeta_c$ and $\theta = 0$) to the system of equations

$$\begin{aligned} \int_0^\theta a(\zeta, w)dw &= \sum_{k=p}^{q} \frac{A_k(\zeta)}{k+1}\alpha^{k+1}, \\ A_k(\zeta_c) &= 0\,, p \le k < q, \\ A_q(\zeta_c) &\neq 0, \\ \alpha = \alpha(\zeta, \theta) &= \theta + \cdots. \end{aligned}$$

In particular, it follows that

$$A_p(\zeta) = [\theta^p]a(\zeta, \theta), \tag{2.8}$$

$$A_q(\zeta_c) = [\theta^q]a(\zeta_c, \theta)\,. \tag{2.9}$$

Remark 2.2. In (2.6) one has that

$$c_k(\zeta; k) = \left([\theta^n]f(\zeta, \theta)\right)^{-(k+1)/n}. \tag{2.10}$$

More generally, the coefficients $c_k(\zeta; j)$ are characterized by the identity $c_k(\zeta; j) = [\beta^j]\alpha^k \frac{\partial \alpha}{\partial \beta}$ where, for all ζ sufficiently close to ζ_c, the variables α and β are related to each other through the variable θ via the relations

$$\begin{aligned} &\alpha = \alpha(\zeta, \theta), \\ &\beta = \alpha \cdot ([\theta^n]f(\zeta, \theta))^{1/n} \cdot \left(1 + \frac{f(\zeta, \theta) - ([\theta^n]f(\zeta, \theta))\ \alpha^n}{([\theta^n]f(\zeta, \theta))\ \alpha^n}\right)^{1/n}. \end{aligned}$$

Theorem 2.2 is essentially equivalent to Theorem 3.3 in [13] when the amplitude term $a(\zeta, \theta)$ in (2.1) vanishes to constant degree in the variable θ about $\theta = 0$ (the case $p = q$). The first application we show is concerned with precisely this case. The generating function in the following example is analyzed in [13]. However, here we perform a similar analysis but from the perspective of Theorem 2.2.

Example 2.3 (lattice paths). The Delannoy numbers (see [16, p. 185]) are the coefficients $f_{r,s}$ that count the number of paths in the $\mathbb{Z} \times \mathbb{Z}$-lattice that join $(0, 0)$ with (r, s) with steps of the form $(0, 1)$, $(1, 1)$, and $(1, 0)$. With the understanding that $f_{0,0} = 1$ and $f_{r,s} = 0$ whenever $r < 0$ or $s < 0$, it follows that $f_{r,s} = f_{r-1,s} + f_{r-1,s-1} + f_{r,s-1}$, for all integers $r, s \ge 0$ except when $(r, s) = (0, 0)$. Using this recursion it is almost direct to see that

$$F(z, w) := \sum_{r,s\ge 0} f_{r,s}\, z^r\, w^s = \frac{1}{1 - z - w - zw}\,.$$

The strictly minimal simple zeros of the denominator of F are all of the form (ζ, ω), with $\zeta \in (0, 1)$ and $\omega = g(\zeta) := \frac{1-\zeta}{1+\zeta}$. Furthermore, one finds that

$$(r, s) \in \operatorname{dir}(\zeta, \omega) \iff \frac{r}{s} = \frac{2\zeta}{1 - \zeta^2}\,.$$

This allows an asymptotic analysis for $[z^r w^s]F$ as $(r, s) \to \infty$, uniformly for (r, s) in any cone of the form $\Lambda = \{(r, s) : d_1 \le r/s \le d_2\}$, with $d_1 > 0$ and $d_2 > 0$ arbitrary constants. On the other hand, as shown in [13], one finds for $(r, s) \in \Lambda$ that

$$(r, s) \in \operatorname{dir}(\zeta, \omega) \iff \zeta = \frac{\sqrt{r^2 + s^2} - s}{r}\,, \omega = \frac{\sqrt{r^2 + s^2} - r}{s}\,.$$

Using definitions (2.1) and (2.2) it follows that

$$\begin{aligned} a(\zeta,\theta) &= \frac{1}{1-\zeta e^{i\theta}} \\ &= \frac{1}{1-\zeta} + \frac{i\zeta}{(1-\zeta)^2}\theta + \cdots, \\ f(\zeta,\theta) &= \ln\left\{\frac{(1-\zeta e^{i\theta})(1+\zeta)}{(1+\zeta e^{i\theta})(1-\zeta)}\right\} + \frac{2i\zeta}{1-\zeta^2}\theta \\ &= \frac{\zeta(1+\zeta^2)}{(1-\zeta^2)^2}\theta^2 + \frac{i\zeta(1+6\zeta^2+\zeta^4)}{3(1-\zeta^2)^3}\theta^3 + \cdots. \end{aligned}$$

Since $a(\zeta,\theta)$ and $f(\zeta,\theta)$, respectively, vanish to constant degree 0 and 2 at $\theta = 0$, for all $\zeta \in (0,1)$, Theorem 2.2 implies that there is a constant $c > 0$ and coefficients $B(r,s)$ such that

$$\begin{aligned} [z^r w^s]F &= \frac{1}{2\pi}\left(\frac{\sqrt{r^2+s^2}-s}{r}\right)^{-r}\left(\frac{\sqrt{r^2+s^2}-r}{s}\right)^{-s} \\ &\quad \cdot \left\{\frac{r\cdot B(r,s)}{r+s-\sqrt{r^2+s^2}} + O(e^{-s\cdot c})\right\}, \\ B(r,s) &= 2\sqrt{\pi}\frac{s}{r}\left(\frac{\sqrt{r^2+s^2}-s}{r} + \frac{r}{\sqrt{r^2+s^2}-s}\right)^{-1/2}\cdot s^{-1/2} + O(s^{-3/2}), \end{aligned}$$

uniformly for $(r,s) \in \Lambda$ as $(r,s) \to \infty$. In particular, it follows that

$$[z^r w^s]F \sim \left(\frac{\sqrt{r^2+s^2}-s}{r}\right)^{-r}\left(\frac{\sqrt{r^2+s^2}-r}{s}\right)^{-s}\cdot\sqrt{\frac{rs}{2\pi(r+s-\sqrt{r^2+s^2})^2\sqrt{r^2+s^2}}},$$

whenever $(r,s) \to \infty$ at a comparable rate.

Although the computations in Theorem 2.2 can be involved, it gives a precise and unified understanding of the elements that are important to take into consideration when analyzing the asymptotic behavior of the coefficients of meromorphic functions in two variables. Furthermore, the calculations greatly simplify in situations where the coefficients $A_k(\zeta)$ are easily available. This is the main point of the following result which is a direct consequence of Remark 2.1.

COROLLARY 2.4. *Under the hypothesis of Theorem* 2.2 *but for the special case in which* $q = p+1$, *if for all* ζ *sufficiently close to* ζ_c, $\theta(\zeta)$ *is the only nontrivial solution of the equation* $a(\zeta,\theta) = 0$, *with* θ *in some open neighborhood of* $\theta = 0$, *then*

$$A_p(\zeta) = [\theta^p]a(\zeta,\theta), \tag{2.11}$$

$$A_{p+1}(\zeta) = \left(\frac{(-1)^{p+1}}{(p+1)(p+2)}\cdot\{[\theta^p]a(\zeta,\theta)\}^{p+2}\cdot\left\{\int_0^{\theta(\zeta)} a(\zeta,\xi)d\,\xi\right\}^{-1}\right)^{1/(p+1)}, \tag{2.12}$$

where the branch of the $(p+1)$*-root above is to be selected so as to have* $\lim_{\zeta\to\zeta_c} A_{p+1}(\zeta) = [\theta^{p+1}]a(\zeta_c,\theta)$.

Example 2.5 (Lagrange inversion formula). If $t(x)$ is an analytic function of x near $x = 0$ such that $t(x) = x \cdot u(t(x))$, for a certain analytic function $u(x)$ with $u(0) \neq 0$, then $[x^r]t(x) = [x^{r-1}](u(x))^r/r$ (see section 5.4 in [16]). More generally, many problems related to the Lagrange inversion formula naturally lead to the study of the asymptotic behavior of coefficients of the form $[x^r](u(x))^s v(x)$, as $(r, s) \to \infty$ (see [7] and [1]). These coefficients are related to those of a bivariate generating function via the identity

$$[x^r](u(x))^s v(x) = [z^r w^s] \frac{v(z)}{1 - wu(z)}. \tag{2.13}$$

(See the final remark in section 2 in [2] and Remark 5.22 in [12] for the uses of multivariate generating functions in problems associated with the Lagrange inversion formula. See [18] for a discussion in the context of Riordan arrays.)

In what follows we assume that the radius of convergence of $v(z)$ is greater than or equal to that of $u(z)$. In the context of Theorem 2.2, a point of the form $(\zeta, 1/u(\zeta))$ is a strictly minimal simple zero of the denominator in the right-hand side of (2.13) provided that $\zeta \cdot u(\zeta) \neq 0$ and that $|u(x)|$ is maximized on the circumference $|x| = |\zeta|$ solely at $x = \zeta$. We emphasize that this condition is easily satisfied for $\zeta > 0$ and within the radius of convergence of $u(z)$ whenever $u(z)$ is aperiodic and has nonnegative Taylor coefficients. Asymptotic formulae for the coefficients in (2.13) are then available along the directions in the (r, s)-lattice where $r/s = \zeta u'(\zeta)/u(\zeta)$. Furthermore, if K is a compact set of strictly minimal simple zeros and $(\zeta_c, 1/u(\zeta_c))$ is an interior point of K, then Theorem 2.2 can be used to provide asymptotic formulae in an open cone of directions in the (r, s)-lattice containing the line $r/s = \zeta_c u'(\zeta_c)/u(\zeta_c)$, provided that there are nonnegative integers $p \leq q$ and n such that

$$a(\zeta, \theta) := v(\zeta e^{i\theta}), \tag{2.14}$$

$$f(\zeta, \theta) := \ln\left\{\frac{u(\zeta)}{u(\zeta e^{i\theta})}\right\} + i\theta \frac{\zeta u'(\zeta)}{u(\zeta)}, \tag{2.15}$$

respectively, have a p-to-q and n-to-n change of degree about $\theta = 0$ as $\zeta \to \zeta_c$.

To fix these ideas consider the case in which $u(x) := (1-x)^{-1}$ and $v(x) := (1-2x)$. Then every point of the form $(\zeta, 1-\zeta)$, with $\zeta \in (0, 1)$, is a strictly minimal simple zero of the denominator in the right-hand side of (2.13). Furthermore, $(r, s) \in \mathrm{dir}(\zeta, 1-\zeta)$ if and only if $r/s = \zeta/(1-\zeta)$; in particular,

$$(r, s) \in \mathrm{dir}(\zeta, 1-\zeta) \iff \zeta = \frac{r}{r+s}.$$

This motivates us to define $\zeta(r, s) := r/(r+s)$ for all (r, s) such that $r \cdot s > 0$.

Observe that back in (2.14) and (2.15) one finds that

$$a(\zeta, \theta) = (1 - 2\zeta) - 2i\zeta\theta + \cdots,$$
$$f(\zeta, \theta) = \frac{\zeta}{2(1-\zeta)^2}\theta^2 + \cdots.$$

While $f(\zeta, \theta)$ vanishes to constant degree 2 about $\theta = 0$, for all $\zeta \in (0, 1)$, $a(\zeta, \theta)$ has a 0-to-1 change of degree about $\theta = 0$, as $\zeta \to 1/2$. As a result, using Theorem 2.2, we can determine the asymptotic behavior of $[x^r](1-x)^{-s}(1-2x)$ as $(r, s) \to \infty$ so long as r and s grow at a comparable rate.

Theorem 2.2 implies almost immediately that

$$[x^r](1-x)^{-s}(1-2x) = \frac{1}{\sqrt{2\pi}}\left(\frac{r}{r+s}\right)^{-r}\left(\frac{s}{r+s}\right)^{-s}\cdot\left\{\left(1-\frac{r}{s}\right)\left(1+\frac{s}{r}\right)^{1/2}s^{-1/2} + O(s^{-1})\right\}, \tag{2.16}$$

as $(r,s)\to\infty$, uniformly for r/s restricted to a compact subset of $(0,1)\cup(1,\infty)$.

On the other hand, Corollary 2.4 implies that there is an $\epsilon>0$ such that

$$[x^r](1-x)^{-s}(1-2x) = \frac{1}{\sqrt{2\pi}}\left(\frac{r}{r+s}\right)^{-r}\left(\frac{s}{r+s}\right)^{-s} \cdot\left\{A_0\left(\frac{r}{r+s}\right)\cdot B_0(r,s) + A_1\left(\frac{r}{r+s}\right)\cdot B_1(r,s)\right\},$$

as $(r,s)\to\infty$, uniformly for $(1-\epsilon)\le r/s\le(1+\epsilon)$, where

$$\begin{aligned}
A_0(\zeta) &:= 1-2\zeta,\\
B_0(r,s) &= \alpha_0\left(\frac{r}{r+s}\right)s^{-1/2}+O(s^{-3/2}),\\
\alpha_0(\zeta) &:= \frac{1-\zeta}{\sqrt{\zeta}},\\
A_1(\zeta) &:= \frac{i(1-2\zeta)^2}{2(1-2\zeta+\ln(2\zeta))},\\
B_1(r,s) &= \alpha_1\left(\frac{r}{r+s}\right)s^{-3/2}+O(s^{-5/2}),\\
\alpha_1(\zeta) &:= -i(1-\zeta)^2\cdot\frac{5-9\zeta-12\zeta^2+20\zeta^3+2(1+5\zeta-8\zeta^2)\ln(2\zeta)}{4\zeta\sqrt{\zeta}(1-2\zeta)(1-2\zeta+\ln(2\zeta))}.
\end{aligned}$$

Observe that Theorem 2.2 asserts that $A_1(\zeta)$ and $\alpha_1(\zeta)$ are analytic about any $\zeta\in(0,1)$. The apparent singularity of $A_1(\zeta)$ at $\zeta=1/2$ is not so because its denominator vanishes to degree 2 about $\zeta=1/2$. On the other hand, the numerator and denominator of $\alpha_1(\zeta)$ vanish to degree 3 at about $\zeta=1/2$. Indeed, the first few terms of the Taylor series of $A_1(\zeta)$ and $\alpha_1(\zeta)$ at about $\zeta=1/2$ are found to be

$$\begin{aligned}
A_1(\zeta) &= -i-\frac{4i}{3}\left(\zeta-\frac12\right)+\frac{2i}{9}\left(\zeta-\frac12\right)^2+\cdots,\\
\alpha_1(\zeta) &= -\frac{i\sqrt{2}}{4}+\frac{31i\sqrt{2}}{24}\left(\zeta-\frac12\right)-\frac{503i\sqrt{2}}{180}\left(\zeta-\frac12\right)^2+\cdots.
\end{aligned}$$

Since $A_0(r/(r+s))=0$ whenever $r=s$, the above expansion for $[x^r](1-x)^{-s}(1-2x)$ implies that

$$[x^r](1-x)^{-s}(1-2x) = -\frac{4^{(s-1)}}{\sqrt{\pi}}\left\{s^{-3/2}+O(s^{-5/2})\right\}, \tag{2.17}$$

as $(r,s)\to\infty$ with $r=s$. This corresponds to the asymptotic expansion one would obtain after using Stirling's formula to find the leading asymptotic order of the factorial terms in the identity

$$[x^r](1-x)^{-r}(1-2x) = -\frac{(2r-2)!}{r((r-1)!)^2}.$$

Formulae (2.16) and (2.17) characterize the asymptotic behavior of the coefficients $[x^r](1-x)^{-s}(1-2x)$ as $(r,s)\to\infty$ along the diagonal line $r=s$ or along directions completely away from it. More explicit asymptotic formulae for these coefficients, as $r/s\to 1$, can be obtained looking at the Taylor coefficients of the functions $A_0(\zeta)\cdot\alpha_0(\zeta)$ and $A_1(\zeta)\cdot\alpha_1(\zeta)$ at about $\zeta=1/2$. Indeed, it follows for all constant $\delta>0$ that

$$[x^r](1-x)^{-s}(1-2x) = \frac{-1}{2\sqrt{\pi}}\left(\frac{r}{r+s}\right)^{-r}\left(\frac{s}{r+s}\right)^{-s} \tag{2.18}$$

$$\cdot\left\{\frac{r-s}{r+s}\cdot s^{-1/2}+\frac{s^{-3/2}}{2}+O(s^{-5/2})\right\}, \tag{2.19}$$

as $(r,s)\to\infty$, uniformly for (r,s) in the region $1-\delta/s\le r/s\le 1+\delta/s$.

If r/s approaches 1 from above, then cancellation between the first two terms in the curly bracket in (2.18) is ruled out. As a result, if $r/s=1+|O(s^{-1})|$, then

$$[x^r](1-x)^{-s}(1-2x)\sim\frac{-1}{2\sqrt{\pi}}\left(\frac{r}{r+s}\right)^{-r}\left(\frac{s}{r+s}\right)^{-s}\cdot\left\{\frac{r/s-1}{r/s+1}\cdot s+\frac{1}{2}\right\}\cdot s^{-3/2}. \tag{2.20}$$

This means that in the (r,s)-lattice a bandwidth of size s^{-1} from above the line $r=s$ is what separates the behavior of $[x^r](1-x)^{-s}(1-2x)$ as prescribed in (2.16) from the one in (2.17).

On the other hand, if r/s approaches 1 from below, then a cascade effect of cancellation in (2.18) may reduce the size of $[x^r](1-x)^{-s}(1-2x)$ to arbitrarily small orders. Refined estimates in this case depend on the precise rate of convergence of r/s toward 1. To amplify this, consider coefficients $\alpha>0$, $\beta\ge 1$, $\gamma\neq 0$, and $\delta>0$, and suppose that

$$\frac{r}{s}=1-\alpha s^{-\beta}+\gamma s^{-(\beta+\delta)}+o(s^{-(\beta+\delta)}).$$

In particular, $(r-s)/(r+s)=-\alpha s^{-\beta}(1+\alpha s^{-\beta}/2)/2+\gamma s^{-(\beta+\delta)}/2+o(s^{-2\beta}+s^{-(\beta+\delta)})$. Using this in (2.18) we obtain that

$$[x^r](1-x)^{-s}(1-2x)\sim\frac{-1}{4\sqrt{\pi}}\left(\frac{r}{r+s}\right)^{-r}\left(\frac{s}{r+s}\right)^{-s}\cdot\begin{cases}(1-\alpha)s^{-3/2} & \text{if } \alpha\neq 1 \text{ and } \beta=1;\\ \gamma s^{-(3/2+\delta)} & \text{if } \alpha=1,\beta=1, \text{ and } 0<\delta<1;\\ s^{-3/2} & \text{if } \beta>1.\end{cases}$$

As a result and unlike the asymptotic description in (2.20), we see that if $(r,s)\to\infty$ with $r/s\uparrow 1$, then there is no well-defined bandwidth that separates the asymptotic behavior of $[x^r](1-x)^{-s}(1-2x)$ as prescribed in (2.16) from the one in (2.17). Furthermore, if $\alpha=\beta=1$ and $0<\delta<1$, then $[x^r](1-x)^{-s}(1-2x)$ is of an asymptotic order smaller than anyone observed as $(r,s)\to\infty$ along any diagonal line in the (r,s)-lattice. This finding is consistent with the identity

$$[x^r](1-x)^{-s}(1-2x)=\frac{(s-r-1)\cdot(r+s-2)!}{r!\cdot(s-1)!},$$

from which we see that $[x^r](1-x)^{s}(1-2x)=0$ whenever $r/s=1-s^{-1}$.

Remark 2.3. The determination of the coefficients $A_k(\zeta)$ in Theorem 2.2 becomes more difficult the bigger the change of degree of the amplitude term $a(\zeta,\theta)$ in (2.1). However, the linear dependence between the asymptotic expansion of the coefficients of F and of $a(\zeta,\theta)$ can be exploited to overcome this problem. Indeed, if $a(\zeta,\theta)$ has a p-to-q change of degree in the variable θ, with $p<q$, then one can rewrite $a(\zeta,\theta)=a_0(\zeta,\theta)+a_1(\zeta,\theta)$, where $a_0(\zeta,\theta)$ is a polynomial in the variable θ (of degree less than q) and $a_1(\zeta,\theta)$ vanishes regardless of ζ to constant degree q in θ. Theorem 2.2 can now be used to obtain an asymptotic expansion for each of the terms in $a_0(\zeta,\theta)$ as well as for $a_1(\zeta,\theta)$. Combining these linearly, one obtains an asymptotic expansion for $[z^r w^s]F$ that resembles the one in (2.5).

3. Proof of main results.

3.1. Associating a parameter-varying integral. In this section we show some preliminary results that are required to prove Theorem 2.2. We assume that there are functions $G(z,w)$ and $H(z,w)$ analytic in an open polydisk D centered at $(0,0)$ on which $F(z,w)$, the generating function associated to the coefficients $(f_{r,s})$, satisfies the identity $F(z,w)=G(z,w)/H(z,w)$. In addition, we assume as given a compact set $K\subset D$ of strictly minimal simple zeros of H containing a particular point (ζ_c,ω_c). It is assumed that $H_w(\zeta_c,\omega_c)\neq 0$. In particular, the implicit function theorem implies that (ζ_c,ω_c) has an open neighborhood of the form $Z\times W\subset D$ and there is an analytic map $g: Z\to W$ such that for all $(z,w)\in Z\times W$, $H(z,w)=0$ if and only if $w=g(z)$. Without loss of generality, we may assume that $0\notin W$.

We now adopt the following notation. For all $0<\epsilon<\pi/2$, the notation $|\arg\{z\}|\leq\epsilon$ signifies that $z=|z|e^{i\theta}$, for some $\theta\in[-\epsilon,\epsilon]$. Accordingly, the notation $|\arg\{z\}|\geq\epsilon$ is used to mean that $z=|z|e^{i\theta}$, for some $\theta\in[\epsilon,\pi]\cup[-\epsilon,-\pi]$.

LEMMA 3.1. *For all $\epsilon_1>0$ sufficiently small there is a $\delta_1>0$ such that for all $(\zeta,\omega)\in K$, H is zero-free on the set $\{(z,w): |z|=|\zeta|, |\arg(z/\zeta)|\geq\epsilon_1, |w|\leq(1+\delta_1)|\omega|\}$.*

Proof. Without loss of generality, assume that $0<\epsilon_1<\pi/2$. If K consisted of only one point, the lemma would follow directly from the continuity of H together with the strict minimality of its only element. More generally, define for each $(\zeta,\omega)\in K$ the quantity $\delta_1(\zeta,\omega)$ to be the supremum of those $\delta>0$ such that H is zero-free on the set $\{(z,w): |z|=|\zeta|, |\arg(z/\zeta)|\geq\epsilon_1, |w|\leq(1+\delta)|\omega|\}$. To prove the lemma, it is enough to show that $\inf\{\delta_1(\zeta,\omega):(\zeta,\omega)\in K\}>0$. We prove this by contradiction. Assuming otherwise there would be a sequence of points $(\zeta_j,\omega_j)\in K$ such that $\delta_1(\zeta_j,\omega_j)\to 0$, as $j\to\infty$. In particular, for all j sufficiently large, there would be a (z_j,w_j) such that $|z_j|=|\zeta_j|$, $|\arg\{z_j/\zeta_j\}|\geq\epsilon_1$, $|w_j|=(1+\delta_1(\zeta_j,\omega_j))|\omega_j|$, and $H(z_j,w_j)=0$. But, since K is a compact set, there is no loss of generality in assuming that $(\zeta_j,\omega_j)\to(\zeta,\omega)\in K$ and $(z_j,w_j)\to(z,w)$, as $j\to\infty$. In particular, $|z|=|\zeta|$, $|w|=|\omega|$, $z\neq\zeta$; however, $H(z,w)=0$. This contradicts the fact that (ζ,ω) is a strictly minimal zero of H and therefore we conclude that $\inf\{\delta_1(\zeta,\omega):(\zeta,\omega)\in K\}>0$. This completes the proof of the lemma. □

LEMMA 3.2. *For all $\epsilon_2>0$ sufficiently small there is a $\delta_2>0$ such that all zeros of H in the set $\{(z,w): |z-\zeta_c|<\epsilon_2, |w|\leq(1+\delta_2)|g(z)|\}$ are of the form $w=g(z)$.*

Proof. The strict minimality of (ζ_c,ω_c) together with the analyticity of H imply that there is $\eta>0$ such that $w=\omega_c$ is the only zero of $H(\zeta_c,w)$ in the disk $\{w: |w|\leq(1+\eta)|\omega_c|\}$ (see Theorem 10.18 in [15]). Without loss of generality, we may assume that $\{w: |w-\omega_c|\leq\eta|\omega_c|\}\subset W$. Observe that $H(\zeta_c,w)$ is zero-free on the set $\{w:\eta|\omega_c|\leq|w-\omega_c|$ and $|w|\leq(1+\eta)|\omega_c|\}$. Thus, since H is uniformly continuous, it follows for all $\epsilon_2>0$ sufficiently small that H is zero-free in the set

$\{(z,w) : |z-\zeta_c| \leq \epsilon_2\,, \eta|\omega_c| \leq |w-\omega_c| \text{ and } |w| \leq (1+\eta)|\omega_c|\}$. In this case, the condition that $\{w : |w-\omega_c| \leq \eta|\omega_c|\} \subset W$ implies that all zeros of H in the polydisk $\{(z,w) : |z-\zeta_c| \leq \epsilon_2, |w| \leq (1+\eta)|\omega_c|\}$ are of the form $w = g(z)$. The lemma follows after selecting $\epsilon_2 > 0$ small enough and defining $\delta_2 > 0$ so as to have

$$(1+\delta_2) = (1+\eta) \inf_{z:|z-\zeta_c|\leq\epsilon_2} \left|\frac{\omega_c}{g(z)}\right| > 1\,.$$

The above inequality is always possible because $g(\zeta_c) = \omega_c$. This completes the proof of the lemma. □

The next result pretty much follows the lines of Lemma 4.1 in [13]. It is included here for the sake of completeness.

LEMMA 3.3. *For a sufficiently small choice of $\epsilon > 0$ and for all $|\theta| \leq \epsilon$ and ζ sufficiently close to ζ_c, consider the functions $a(\zeta,\theta)$ and $f(\zeta,\theta)$ as defined in* (2.1) *and* (2.2)*, respectively. Then $f(\zeta,0) = \frac{\partial f}{\partial\theta}(\zeta,0) = 0$ and, for all $(\zeta,\omega) \in K$ sufficiently close to (ζ_c,ω_c), and all nonzero θ such that $-\epsilon \leq \theta \leq \epsilon$, $\Re\{f(\zeta,\theta)\} > 0$). Furthermore, if*

$$\Sigma(\zeta;s) := \int_{-\epsilon}^{\epsilon} e^{-s\cdot f(\zeta,\theta)} a(\zeta,\theta)\, d\theta, \tag{3.1}$$

then there is a constant $c > 0$ such that

$$[z^r w^s]F = \frac{\zeta^{-r}\omega^{-s}}{2\pi}\left\{\Sigma(\zeta;s) + O(e^{-sc})\right\}, \tag{3.2}$$

uniformly for all $(r,s) \in dir(\zeta,\omega)$ and $(\zeta,\omega) \in K$ sufficiently close to (ζ_c,ω_c).

Proof. Let $\epsilon_2 > 0$ and $\delta_2 > 0$ be as in Lemma 3.2. Consider $\epsilon_3 > 0$ such that the functions $a(\zeta,\theta)$ and $f(\zeta,\theta)$ are analytic for $|\zeta-\zeta_c| \leq \epsilon_3$ and $|\theta| \leq \epsilon_3$. In addition, consider for $\epsilon_1 > 0$ the sets

$$\begin{aligned} K_c &:= \{(\zeta,\omega) \in K : |\zeta-\zeta_c| \leq \epsilon_1\},\\ \gamma_1(\zeta) &:= \{z : |z| = |\zeta| \quad \text{and} \quad |\arg\{z/\zeta\}| \geq \epsilon_1\},\\ \gamma_2(\zeta) &:= \{z : |z| = |\zeta| \quad \text{and} \quad |\arg\{z/\zeta\}| \leq \epsilon_1\}\,. \end{aligned}$$

Select $\epsilon_1 > 0$ small enough so as to have $\gamma_2(\zeta) \subset \{z : |z-\zeta_c| \leq \min\{\epsilon_2,\epsilon_3\}\}$, whenever $(\zeta,\omega) \in K_c$. Furthermore, choose $\epsilon_1 > 0$ sufficiently small so that the conclusion of Lemma 3.1 applies with some $\delta_1 > 0$. Select δ so as to satisfy $0 < \delta < \min\{\delta_1,\delta_2,1\}$. The strict minimality of $(\zeta,\omega) \in K_c$ implies that H is zero-free on the polydisk $\{z : |z| \leq |\zeta|\} \times \{w : |w| \leq (1-\delta)|\omega|\}$. Cauchy's formula [15] can then be used to represent the coefficients of F by the integrals

$$[z^r w^s]F = \frac{1}{2\pi}\left\{\int_{z\in\gamma_1(\zeta)} + \int_{z\in\gamma_2(\zeta)}\right\}\frac{1}{z^r}\left(\frac{1}{2\pi i}\int_{|w|=(1-\delta)|\omega|}\frac{G(z,w)}{H(z,w)\cdot w^{s+1}}dw\right)\frac{dz}{iz}, \tag{3.3}$$

where all contour integrals are in the standard counterclockwise orientation.

Lemma 3.1 implies for all $(\zeta,\omega) \in K_c$ that H is zero-free on the set $\gamma_1(\zeta) \times \{w : |w| \leq (1+\delta)|\omega|\}$. As a result,

$$\left|\int_{z\in\gamma_1(\zeta)}\frac{1}{z^r}\int_{w=(1-\delta)|\omega|}\frac{G(z,w)}{H(z,w)\cdot w^{s+1}}dw\frac{dz}{iz}\right|$$

$$= \left| \int_{z\in\gamma_1(\zeta)} \frac{1}{z^r} \int_{w=(1+\delta)|\omega|} \frac{G(z,w)}{H(z,w)\cdot w^s} \frac{dw}{iw} \frac{dz}{iz} \right|$$
$$\leq (2\pi)^2 |\zeta|^{-r} \{(1+\delta)|\omega|\}^{-s} \cdot \sup_{\Gamma_1} |F|,$$

where for convenience we have defined Γ_1 to be the set of all those points of the form (z,w) such that there exists a $(\zeta,\omega)\in K_c$ such that $z\in\gamma_1(\zeta)$ and $|w|\leq(1+\delta)|\omega|$. Since Γ_1 is compact and H is zero-free over it, then $\sup_{\Gamma_1}|F|$ must be finite. Back in (3.3), this implies that

$$[z^r w^s]F = \frac{1}{2\pi} \int_{z\in\gamma_2(\zeta)} \frac{1}{z^r} \left(\frac{1}{2\pi i} \int_{|w|=(1-\delta)|\omega|} \frac{G(z,w)}{H(z,w)\cdot w^{s+1}} dw \right) \frac{dz}{iz}$$
$$+ O(|\zeta|^{-r}|\omega|^{-s}(1+\delta)^{-s}), \tag{3.4}$$

uniformly for all $r,s\geq 0$ and all $(\zeta,\omega)\in K_c$. However, Lemma 3.2 implies that for each $(\zeta,\omega)\in K_c$ and $z\in\gamma_2(\zeta)$, $w=g(z)$ is the only singularity of the integrand above within the disk $\{w : |w|\leq(1+\delta)|g(z)|\}$. The residue theorem in one variable [15] lets us conclude that

$$\frac{1}{2\pi i}\int_{|w|=(1-\delta)|\omega|} \frac{G(z,w)}{H(z,w)\cdot w^{s+1}} dw = \frac{-G(z,g(z))}{H_w(z,g(z))\cdot\{g(z)\}^{s+1}}$$
$$+ \frac{1}{2\pi i}\int_{|w|=(1+\delta)|g(z)|} \frac{G(z,w)}{H(z,w)\cdot w^{s+1}} dw\,.$$

But, observe that if $|w|=(1+\delta)|g(z)|$ and $z\in\gamma_2(\zeta)$, then the strict minimality of $(\zeta,\omega)\in K_c$ implies that $|g(z)|\geq|g(\zeta)|=|\omega|$. In particular,

$$\left| \frac{1}{2\pi} \int_{z\in\gamma_2(\zeta)} \frac{1}{z^r} \left(\frac{1}{2\pi i} \int_{|w|=(1+\delta)|g(z)|} \frac{G(z,w)}{H(z,w)\cdot w^{s+1}} dz \right) \frac{dz}{iz} \right|$$
$$\leq |\zeta|^{-r}\{(1+\delta)|\omega|\}^{-s} \cdot \sup_{\Gamma_2}|F|,$$

where we have defined Γ_2 to be the set of points (z,w) for which there exists a $(\zeta,\omega)\in K_c$ such that $z\in\gamma_2(\zeta)$ and $|w|=(1+\delta)|g(z)|$. Since Γ_2 is a compact set and H is zero-free over it, from (3.4) we can conclude that

$$[z^r w^s]F = \frac{1}{2\pi}\int_{z\in\gamma_2(\zeta)} \frac{1}{z^r} \frac{-G(z,g(z))}{H_w(z,g(z))\cdot\{g(z)\}^{s+1}} \frac{dz}{iz} + O(|\zeta|^{-r}|\omega|^{-s}(1+\delta)^{-s}), \tag{3.5}$$

uniformly for all $r,s\geq 0$ and all $(\zeta,\omega)\in K_c$.

The integral on the right-hand side in (3.5) can be parametrized using polar coordinates. Indeed, substituting $z=\zeta e^{i\theta}$, with $-\epsilon\leq\theta\leq\epsilon$, one obtains that

$$[z^r w^s]F = \frac{\zeta^{-r}\omega^{-s}}{2\pi}\int_{-\epsilon}^{\epsilon} e^{-s\cdot f(\theta;\zeta,r/s)} a(\zeta,\theta)d\theta + O(|\zeta|^{-r}|\omega|^{-s}(1+\delta)^{-s}),$$

uniformly for all $r,s\geq 0$ and all $(\zeta,\omega)\in K_c$, where $a(\zeta,\theta)$ is defined as in (2.1) and $f(\theta;\zeta,\lambda) := \ln\left\{\frac{g(\zeta e^{i\theta})}{g(\zeta)}\right\} + i\lambda\theta$. Observe that $f\left(0;\zeta,\frac{r}{s}\right)=0$ and

$$\frac{\partial f}{\partial\theta}\left(0;\zeta,\frac{r}{s}\right) = i\left(\frac{\zeta g'(\zeta)}{g(\zeta)} + \frac{r}{s}\right)$$
$$= i\left(\frac{r}{s} - \frac{\zeta H_z(\zeta,\omega)}{\omega H_w(\zeta,\omega)}\right).$$

In particular, we see that $\frac{\partial f}{\partial \theta}\left(0;\zeta,\frac{r}{s}\right)=0$ for all $(r,s)\in \operatorname{dir}(\zeta,\omega)$. Furthermore, the strict minimality of $(\zeta,\omega)\in K_c$ implies that $|g(\zeta e^{i\theta})|>|g(\zeta)|$, for all nonzero θ such that $-\epsilon\leq\theta\leq\epsilon$ and, as a result, $\Re\{f(\zeta;\theta)\}>0$ for all such θ. Lemma 3.3 follows by noticing that whenever $(r,s)\in\operatorname{dir}(\zeta,\omega)$, then $f\left(0;\zeta,r/s\right)=f(\zeta;\theta)$, with $f(\zeta;\theta)$ as defined in (2.2). □

3.2. Polynomial canonical representations. A result of Levinson [10] implies that if a function $H(u,v)$ is analytic in a neighborhood of the origin in $\mathbb{C}^2$ and its Hartogs series vanishes to degree $q\geq 1$ in the variable v about the origin, then H admits a near $(0,0)$ a representation of the form

$$H(u,v)=\sum_{j=0}^{q}H_j(u)w^j. \tag{3.6}$$

Above the coefficient functions H_j are analytic near the origin and such that $H_j(0)=0$ for all $0\leq j<q$; however, $H_q(0)\neq 0$. In addition, $w=w(u,v)$ is a certain analytic function near the origin such that $w(u,0)=0$ and $\frac{\partial w}{\partial v}(u,0)=1$. In [12] it is proved using one complex variable methods that this representation is indeed unique. The following more precise representation will be more suitable to prove our main result.

LEMMA 3.4. *Let $0\leq p\leq q$ with $q\geq 1$ be nonnegative integers. Suppose that $H(u,v)$ is analytic in a neighborhood of the origin and has a p-to-q change of degree about $v=0$ as $u\to 0$. Then, H admits near the origin a unique representation of the form*

$$H(u,v)=\sum_{k=p}^{q}H_k(u)\cdot w^k, \tag{3.7}$$

where $H_k(0)=0$, for $p\leq k<q$, $H_q(0)\neq 0$, and $w=w(u,v)$ is such that $w(u,0)=0$ and $\frac{\partial w}{\partial v}(u,0)=1$. Furthermore,

$$H_p(u)=\frac{1}{p!}\frac{\partial^p H}{\partial v^p}(u,0)\,. \tag{3.8}$$

Proof. The uniqueness of the representation in (3.7) is immediate from the uniqueness of the representation in (3.6). Suppose that the representation in (3.6) applies for all (u,v) in an open neighborhood of the polydisk $\{(u,v):|u|\leq\epsilon \text{ and } |v|\leq\epsilon\}$, for some $\epsilon>0$. Since $H(u,v)$ has a p-to-q change of degree about $v=0$ as $u\to 0$, H has a Hartogs series of the form

$$H(u,v)=\sum_{k=p}^{\infty}h_k(u)\,v^k,$$

where the coefficients h_k are analytic for $|u|\leq\epsilon$, $h_p(u)$ is not identically zero in any neighborhood of $u=0$, and $h_q(0)\neq 0$.

Consider the map $\Phi(u,v)=(u,w(u,v))$. The conditions imposed over w in (3.6) imply that the Jacobian matrix $\frac{\partial\Phi}{\partial(u,v)}(0,0)$ is triangular with all entries equal to 1 along the diagonal. Since $\Phi(0,0)=(0,0)$, the inverse mapping theorem lets us assume without loss of generality that Φ is holomorphic and 1-to-1 over the polydisk $\{(u,v):|u|\leq\epsilon,|v|\leq\epsilon\}$. In particular, for all u such that $|u|\leq\epsilon$, $w(u,\cdot)$ is 1-to-1 for $|v|\leq\epsilon$. Furthermore, since $w(u,0)=0$, the open mapping theorem implies that

there are $\rho_1, \rho_2 > 0$ such that $\{w : |w| \le \rho_1\} \subset w(u, \{v : |v| < \epsilon\})$ and the preimage of $\{v : |v| < \rho_1\}$ under $w(u, \cdot)$ contains the disk $\{v : |v| \le \rho_2\}$. As a result, using Cauchy's formula in (3.6) and then the substitution $w = w(u, v)$, it follows for all $0 \le j \le q$ that

$$\begin{aligned} H_j(u) &= \frac{1}{2\pi i} \int_{|w|=\rho_1} \frac{1}{w^{j+1}} \left(\sum_{k=0}^{q} H_k(u) \cdot w^k \right) dw \\ &= \frac{1}{2\pi i} \int_{|v|=\rho_2} \frac{H(u,v)}{\{w(u,v)\}^{j+1}} \frac{\partial w}{\partial v}(u,v)\, dv \\ &= \frac{1}{2\pi i} \sum_{k=p}^{\infty} h_k(u) \cdot \int_{|v|=\rho_2} \frac{v^k}{\{w(u,v)\}^{j+1}} \frac{\partial w}{\partial v}(u,v)\, dv, \end{aligned}$$

where for the last identity we have used that Hartogs series of H converges uniformly over compact subsets of $\{(u, v) : |u| \le \epsilon \text{ and } |v| \le \epsilon\}$. However, observe that the conditions imposed over w in (3.6) imply that, for all $j < p \le k$, the function $\frac{v^k}{\{w(u,v)\}^{j+1}} \frac{\partial w}{\partial v}(u, v)$ is analytic in v in an open neighborhood of $\{v : |v| \le \rho_2\}$. Consequently, for $j < p$, all the terms in the above summation vanish and therefore $H_j(u) = 0$. This shows (3.7). Furthermore, if $j = p$, then the residue theorem implies that

$$\begin{aligned} H_p(u) &= \frac{h_p(u)}{2\pi i} \cdot \int_{|v|=\rho_2} \frac{v^p}{\{w(u,v)\}^{p+1}} \frac{\partial w}{\partial v}(u,v)\, dv \\ &= h_p(u) \cdot \operatorname{Res}\left(\frac{v^p}{\{w(u,v)\}^{p+1}} \frac{\partial w}{\partial v}(u,v); v = 0 \right) \\ &= h_p(u)\,. \end{aligned}$$

This shows (3.8) and completes the proof of the lemma. □

3.3. Asymptotic analysis. In this section we prove Theorem 2.2. This is accomplished by analyzing the asymptotic behavior of the integral $\Sigma(\zeta; s)$ in (3.2), as $s \to \infty$. Observe that $\Sigma(\zeta; s) = \Sigma_1(\zeta; s) + \Sigma_2(\zeta; s)$, where we have defined

$$\Sigma_i(\zeta; s) := \int_0^{\epsilon} e^{-s \cdot f(\zeta, (-1)^{i+1}\theta)} a(\zeta, (-1)^{i+1}\theta) d\theta, \quad i = 1, 2.$$

Because of the similarity of $\Sigma_1(\zeta; s)$ and $\Sigma_2(\zeta; s)$, we analyze only the asymptotic behavior of $\Sigma_1(\zeta; s)$ under the hypotheses that $f(\zeta, \theta)$ and $a(\zeta, \theta)$ have, respectively, an n-to-n and p-to-q change of degree about $\theta = 0$ as $\zeta \to \zeta_c$, and that $f(\zeta, \theta)$ has the properties stated in Lemma 3.3. A similar analysis of the asymptotic behavior of $\Sigma_2(\zeta; s)$ is summarized at the end of this section.

Lemma 3.3 implies that $n \ge 2$. In particular, we may write

$$f(\zeta, \theta) = u(\zeta) \cdot \theta^n + \cdots, \tag{3.9}$$

where u is a certain analytic function near ζ_c such that $u(\zeta_c) \ne 0$. Since for all nonzero $\theta \in [-\epsilon, \epsilon]$, $\Re\{f(\zeta_c, \theta)\} > 0$, we must have $\Re\{u(\zeta_c)\} \ge 0$.

On the other hand, Lemma 3.4 implies that there is a unique representation of the form

$$\int_0^{\theta} a(\zeta, w) dw = \sum_{k=p}^{q} \frac{A_k(\zeta)}{k+1} \alpha^{k+1}, \tag{3.10}$$

where $A_k(\zeta_c) = 0$ for all $p \leq k < q$, $A_q(\zeta_c) \neq 0$, and $\alpha = \alpha(\zeta, \theta)$ is such that $\alpha(\zeta, 0) = 0$ and $\frac{\partial \alpha}{\partial \theta}(\zeta, 0) = 1$. The coefficients A_k, $p \leq k \leq q$, correspond to those appearing in Remark 2.1. The inverse mapping theorem implies that $\Psi_1(\zeta, \theta) := (\zeta, \alpha(\zeta, \theta))$ is a biholomorphic map from an open neighborhood of $(\zeta, \theta) = (\zeta_c, 0)$ to an open neighborhood of $(\zeta, \alpha) = (\zeta_c, 0)$. In particular, assuming that $\epsilon > 0$ is sufficiently small, we can perform in $\Sigma_1(\zeta; s)$ the change of variables $\alpha = \alpha(\zeta, \theta)$ to obtain that

$$\Sigma_1(\zeta; s) = \sum_{k=p}^{q} A_k(\zeta) \int_0^{\alpha(\zeta,\epsilon)} e^{-s \cdot g(\zeta,\alpha)} \alpha^k d\alpha,$$

where $g(\zeta, \alpha) := f(\Psi_1^{-1}(\zeta, \alpha))$. This last function is analytic in a neighborhood of the origin. Furthermore, it's Hartogs series about $(\zeta, \alpha) = (\zeta_c, 0)$ in powers of α is of the form $g(\zeta, \alpha) = u(\zeta)\alpha^n + \cdots$ with $u(\zeta)$ as in (3.9). This motivates us to consider the map

$$\Psi_2(\zeta, \alpha) := \left(\zeta, \alpha \cdot (u(\zeta))^{1/n} \cdot \left(1 + \frac{g(\zeta, \alpha) - u(\zeta)\alpha^n}{u(\zeta)\alpha^n} \right)^{1/n} \right),$$

where the principal branch of the nth root function is to be used in both cases. Since $u(\zeta_c) \neq 0$ and $\Re\{u(\zeta_c)\} \geq 0$, it follows that Ψ_2 is well defined and holomorphic near $(\zeta_c, 0)$. Furthermore, if $\beta = \beta(\zeta, \alpha)$ is such that $\Psi_2(\zeta, \alpha) = (\zeta, \beta(\zeta, \alpha))$, the inverse mapping theorem implies that $\Psi_2(\zeta, \alpha)$ is biholomorphic between open neighborhoods of $(\zeta, \alpha) = (\zeta_c, 0)$ and $(\zeta, \beta) = (\zeta_c, 0)$. In particular, it follows that $g(\Psi_2^{-1}(\zeta, \beta)) = \beta^n$ and therefore

$$\Sigma_1(\zeta; s) = \sum_{k=p}^{q} A_k(\zeta) \int_0^{\beta(\zeta,\alpha(\zeta,\epsilon))} e^{-s \cdot \beta^n} (\alpha(\zeta, \beta))^k \frac{\partial \alpha}{\partial \beta}(\zeta, \beta) d\beta,$$

provided that $\epsilon > 0$ is chosen sufficiently small to start with. We claim that the domain of integration of the integrals participating in the summation above can be replaced by a real interval of the form $[0, \delta]$, for some $\delta > 0$. For this observe that the condition $\Re\{f(\zeta_c, \epsilon)\} > 0$ implies that $\Re\{(\beta(\zeta_c, \alpha(\zeta_c, \epsilon)))^n\} > 0$. On the other hand, since $\beta(\zeta_c, \alpha(\zeta_c, \epsilon)) = (u(\zeta_c))^{1/n}\epsilon + O(\epsilon^2)$, with $\Re\{u(\zeta_c)\} \geq 0$, we conclude that $|\arg\{\beta(\zeta_c, \alpha(\zeta_c, \epsilon))\}| < \pi/(2n)$. Since $\beta(\zeta, \alpha(\zeta, \epsilon)) \to \beta(\zeta_c, \alpha(\zeta_c, \epsilon))$, as $\zeta \to \zeta_c$, we conclude that $|\arg\{\beta(\zeta, \alpha(\zeta, \epsilon))\}| < \pi/(2n)$ for all ζ sufficiently close to ζ_c. Choosing $\delta := \Re\{\beta(\zeta_c, \alpha(\zeta_c, \epsilon))\}$, it follows that there is a constant $c > 0$ such that

$$\int_\delta^{\beta(\zeta,\alpha(\zeta,\epsilon))} e^{-s \cdot \beta^n} (\alpha(\zeta, \beta))^k \frac{\partial \alpha}{\partial \beta}(\zeta, \beta) d\beta = O(e^{-sc}),$$

as $s \to \infty$, uniformly for all ζ sufficiently close to ζ_c and for all $p \leq k \leq q$. This implies that

$$\Sigma_1(\zeta; s) = \sum_{k=p}^{q} A_k(\zeta) \int_0^{\delta} e^{-s \cdot \beta^n} (\alpha(\zeta, \beta))^k \frac{\partial \alpha}{\partial \beta}(\zeta, \beta) d\beta + O(e^{-sc}), \tag{3.11}$$

as $s \to \infty$, uniformly for all ζ sufficiently close to ζ_c. An asymptotic expansion for the integrals participating in the summation above is easily obtained using the standard stationary phase method (see Chapter 6 in [3]). Indeed, since Hartogs series

of $(\alpha(\zeta,\beta))^k \frac{\partial \alpha}{\partial \beta}(\zeta,\beta)$ in powers of β about $(\zeta,\beta)=(\zeta_c,0)$ must be of the form

$$(\alpha(\zeta,\beta))^k \frac{\partial \alpha}{\partial \beta}(\zeta,\beta) = \sum_{j=k}^{\infty} c_k(\zeta;j)\beta^j,$$

with $c_k(\zeta;k) = (u(\zeta))^{-(k+1)/n}$, then from (3.11) it follows that

$$\Sigma_1(\zeta;s) = \sum_{k=p}^{q} A_k(\zeta)\cdot B_k(\zeta;s) + O(e^{-sc}), \tag{3.12}$$

$$B_k(\zeta;s) \approx \sum_{j=k}^{\infty} \frac{c_k(\zeta;j)}{n}\Gamma\left(\frac{j+1}{n}\right)\cdot s^{-(j+1)/n}, \tag{3.13}$$

uniformly for all ζ sufficiently close to ζ_c as $s\to\infty$. The coefficients $c_k(\zeta;j)$ correspond to those appearing in Remark 2.2. Equations (3.12) and (3.13) provide a complete asymptotic description for $\Sigma_1(\zeta;s)$ which is uniform for all ζ sufficiently close to ζ_c as $s\to\infty$.

To obtain an asymptotic expansion for the term $\Sigma_2(\zeta;s)$, the uniqueness of the decomposition in (3.10) is relevant to relate the coefficients appearing in the expansion of $\Sigma_2(\zeta;s)$ with those in (3.12) and (3.13). Without delving into details it follows that

$$\Sigma_2(\zeta;s) = \sum_{k=p}^{q} A_k(\zeta)\cdot \tilde{B}_k(\zeta;s) + O(e^{-sc}), \tag{3.14}$$

where for the case in which n is even it applies that

$$\tilde{B}_k(\zeta;s) \approx \sum_{j=k}^{\infty} \frac{(-1)^j c_k(\zeta;j)}{n}\Gamma\left(\frac{j+1}{n}\right)\cdot s^{-(j+1)/n}; \tag{3.15}$$

however, for the case in which n is odd,

$$\tilde{B}_k(\zeta;s) \approx \sum_{j=k}^{\infty} \frac{(-1)^j D(j,n) c_k(\zeta;j)}{n}\Gamma\left(\frac{j+1}{n}\right)\cdot s^{-(j+1)/n}, \tag{3.16}$$

where $D(j,n) := \exp(-\frac{i\pi(j+1)}{n}\cdot \operatorname{sign}\{i[\theta^n]f(\zeta_c,\theta)\})$. Equations (2.5) and (2.6) in Theorem 2.2 are now a direct consequence of (3.12)–(3.16). This completes the proof of Theorem 2.2. ∎

Acknowledgments. I would like to thank my graduate advisor, Robin Pemantle, and his collaborator, Mark Wilson, for their insights and support in my work on asymptotic analysis and generating functions. Special thanks also to Jean-Pierre Rosay and Saleh Tanveer for their helpful inputs in the preliminary version of my graduate dissertation which in one way or another has been reflected in here.

REFERENCES

[1] C. Banderier, P. Flajolet, G. Schaeffer, and M. Soria, *Planar maps and Airy phenomena*, in Proceedings of the 27th International Colloquium on Automata, Languages and Programming, Geneva, Switzerland, Lecture Notes in Comput. Sci. 1853, Springer, Berlin, 2000, pp. 388–402.

[2] C. Banderier, P. Flajolet, G. Schaeffer, and M. Soria, *Random maps, coalescing saddles, singularity analysis, and Airy phenomena*, Random Structures Algorithms, 19 (2001), pp. 194–246.
[3] N. Bleistein and R. Handelsman, *Asymptotic Expansion of Integrals*, Dover Publications, Inc., New York, 1986.
[4] H. Cartan, *Elementary Theory of Analytic Functions of One or Several Complex Variables*, Addison-Wesley, Reading, MA, 1973.
[5] C. Chester, B. Friedman, and F. Ursell, *An extension of the method of steepest descents*, Proc. Cambridge Philos. Soc., 53 (1957), pp. 599–611.
[6] N. DeBruijn, *Asymptotic Methods in Analysis*, Dover Publications, Inc., New York, 1981.
[7] M. Drmota, *A bivariate asymptotic expansion of coefficients of powers of generating functions*, European J. Combin., 15 (1994), pp. 139–152.
[8] N. Levinson, *A canonical form for an analytic function of several variables at a critical point*, Bull. Amer. Math. Soc., 66 (1960), pp. 68–69.
[9] N. Levinson, *A polynomial canonical form for certain analytic functions of two variables at a critical point*, Bull. Amer. Math. Soc., 66 (1960), pp. 366–368.
[10] N. Levinson, *Transformation of an analytic function of several variables to a canonical form*, Duke Math. J., 28 (1961), pp. 345–353.
[11] M. Lladser, *Uniform Formulae for the Coefficients of Meromorphic Functions in Two Variables, Part* II*: The Airy Phenomena*, in preparation.
[12] M. Lladser, *Asymptotic Enumeration via Singularity Analysis*, Ph.D. thesis, The Ohio State University, Columbus, OH, 2003.
[13] R. Pemantle and M. C. Wilson, *Asymptotics of multivariate sequences.* I*. Smooth points of the singular variety*, J. Combin. Theory Ser. A, 97 (2002), pp. 129–161.
[14] R. Pemantle and M. Wilson, *Twenty Combinatorial Examples of Asymptotics Derived from Multivariate Generating Functions*, preprint, 2005.
[15] W. Rudin, *Real and Complex Analysis,* 3rd ed., Higher Mathematics Series, McGraw-Hill, New York, 1987.
[16] R. Stanley, *Enumerative Combinatorics*, vol. I and II of Cambridge Stud. Adv. Math., Cambridge University Press, Cambridge, UK, 1999.
[17] J. Taylor, *Several Complex Variables with Connections to Algebraic Geometry and Lie Groups*, Graduate Studies in Mathematics 46, AMS, Providence, RI, 2002.
[18] M. Wilson, *Asymptotics of Riordan arrays*, in 2005 International Conference on Analysis of Algorithms, Discrete Math. Theor. Comput. Sci. Proc. AD, 2005, pp. 323–333.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 829–840

ON MINIMUM DEGREE IMPLYING THAT A GRAPH IS H-LINKED*

RONALD J. GOULD†, ALEXANDR KOSTOCHKA‡, AND GEXIN YU§

Abstract. Given a fixed multigraph H, possibly containing loops, with $V(H) = \{h_1, \ldots, h_m\}$, we say that a graph G is H-linked if for every choice of m vertices $v_1, \ldots, v_m$ in G, there exists a subdivision of H in G such that v_i is the branch vertex representing h_i (for all i). This generalizes the concept of k-linked graphs (as well as a number of other well-known path or cycle properties). In this paper we determine a sharp lower bound on $\delta(G)$ (which depends upon H) such that each graph G on at least $10(|V(H)| + |E(H)|)$ vertices satisfying this bound is H-linked.

Key words. minimum degree, connectivity, k-linked, H-linked

AMS subject classifications. 05C40, 05C38

DOI. 10.1137/050624662

1. Introduction. For terms not defined here, see [9]. A graph is *k-linked* if for every sequence of $2k$ vertices, $v_1, \ldots, v_k, w_1, \ldots, w_k$, there are internally disjoint paths $P_1, \ldots, P_k$ such that P_i joins v_i and w_i. The literature contains numerous results and important open problems dealing with k-linked graphs. In this paper we are concerned with the following generalization of k-linked graphs.

Let H be a multigraph. An *H-subdivision* in a graph G is a pair of mappings $f : V(H) \to V(G)$ and g: $E(H)$ into the set of paths in G such that:

(a) $f(u) \neq f(v)$ for all distinct $u, v \in V(H)$;
(b) for every $uv \in E(H)$, $g(uv)$ is an $f(u), f(v)$-path in G, and distinct edges map to internally disjoint paths in G.

A graph G is *H-linked* if every injective mapping $f : V(H) \to V(G)$ can be extended to an H-subdivision in G. In other words, G is H-linked if G contains an H-subdivision with prescribed branching vertices for any such prescription. This notion is a common generalization of the notions of k-linked, k-ordered, and k-connected graphs. In particular, if G is k-linked, then G is H-linked for every H with k edges and no isolated vertices. Since every $10k$-connected graph is k-linked (see [8]), every $10(|E(H)| + |V(H)|)$-connected graph is H-linked.

The idea of H-linked graphs originated with Jung [3], but had not been considered in full generality until recently, when the concept was first considered independently in [5] and [10].

In [6] and [7], H-linkage was considered for loopless multigraphs H with k edges and minimum degree at least two. The following was shown in [7].

THEOREM 1. *Let H be a loopless graph with $|E(H)| = k$ and $\delta(H) \geq 2$. Every simple graph G of order $n \geq 5k + 6$ with $\delta(G) \geq \lceil \frac{n+k}{2} \rceil - 1$ is H-linked. If $H = C_k$,*

*Received by the editors February 17, 2005; accepted for publication (in revised form) March 6, 2006; published electronically December 5, 2006.
http://www.siam.org/journals/sidma/20-4/62466.html
†Department of Mathematics and Computer Science, Emory University, Atlanta, GA 30322 (rg@mathcs.emory.edu).
‡Department of Mathematics, University of Illinois, Urbana, IL 61801 and Institute of Mathematics, Novosibirsk, 630090, Russia (kostochk@math.uiuc.edu). This author was partially supported by NSF grant DMS-0400498 and grant 03-01-00796 of the Russian Foundation for Basic Research.
§Department of Mathematics, University of Illinois, Urbana, IL 61801 (gexinyu@uiuc.edu).

then every graph G of order $n \geq 5k+6$ with $\delta(G) \geq \lceil \frac{n}{2} \rceil + \lfloor \frac{k}{2} \rfloor - 1$ is H-linked. The minimum degree conditions are sharp.

It was also verified in [7] that under the conditions of Theorem 1, any H-subdivision in G can be extended to an H-subdivision that spans $V(G)$. This extended an earlier result of Kierstead, Sárközy, and Selkow [4] on k-ordered graphs. Extension results like this and that of [2] provide a framework for generalizing both linkage and strong Hamiltonian-type results, as both involve questions on spanning subgraphs.

In [6], the work in [7] was sharpened. At the same time a result similar to that of [6] was shown in [1].

Let $B(H)$ denote the maximum number of edges in an edge-cut of H. In terms of $B(H)$, the main results in [6] and [1] can be summarized as follows.

THEOREM 2. *Every simple graph G of order n with $\delta(G) \geq \lceil \frac{n+B(H)}{2} \rceil - 1$ is H-linked provided*

(i) *See* [6]. *H is a loopless connected multigraph with k edges and $\delta(H) \geq 2$ and G is of order $n \geq 7.5k$.*

(ii) *See* [1]. *H is a connected multigraph, possibly containing loops, and G is of sufficiently large order n.*

The purpose of this paper is to provide a merging of ideas from [6] and [1] and to prove a more general result describing the situations also for disconnected graphs H. That is, we wish to show for all multigraphs H, possibly containing loops, a sharp lower bound on $\delta(G)$ sufficient to ensure that each graph G on at least $10(|V(H)|+|E(H)|)$ vertices will be H-linked. It turns out that for disconnected H, the bound is more sophisticated.

We will say that a multigraph H is *uneven* if it does not contain even cycles. Denote by $c(H)$ the number of uneven components of H. Let

$$b(H) = \begin{cases} |V(H)| - 1 & \text{if } H \text{ is uneven,} \\ B(H) + c(H) & \text{otherwise.} \end{cases}$$

Note that for uneven graphs, the value $b(H) = |V(H)| - 1$ is exactly one less than that from the second part of the formula.

Our proof is based on the proof in [6], modified to handle the more general conditions on H. Our main result is the following theorem.

THEOREM 3. *Let H be a multigraph with $e(H)$ edges (loops or nonloops) and let $k_1 = k_1(H) = e(H) + c(H)$. Let G be a simple graph of order $n \geq 9.5(k_1+1)$. If*

$$\delta(G) \geq \left\lceil \frac{n + b(H)}{2} \right\rceil - 1, \tag{1.1}$$

then G is H-linked. Moreover, every injective mapping $f : V(H) \to V(G)$ can be extended to an H-subdivision in G containing at most $5k_1 + 2$ vertices.

Restriction (1.1) cannot be weakened. In the next section we will prove this and derive some simple facts on edge cuts in connected graphs. In the subsequent three sections we prove Theorem 3 for the case of loopless H, and in the final section we prove the theorem in full generality. We briefly discuss the ideas of the proof at the end of section 3.

2. On edge cuts and constructions. It is well known (see, e.g., [9, p. 51]) that

$$B(H) \geq (k+1)/2 \tag{2.1}$$

for every H with $k > 0$ edges.

The following property makes uneven components special for our theorem.

LEMMA 4. *A connected graph H is uneven if and only if $B(H) = |V(H)| - 1$.*

Proof. Suppose first that H is a connected uneven graph with m cycles. Then no two cycles in H share an edge and hence $|E(H)| = |V(H)| - 1 + m$. Furthermore, any edge cut in H misses at least one edge in each (odd) cycle of H, and hence $B(H) \leq |E(H)| - m$. Therefore $B(H) \leq |V(H)| - 1$. On the other hand, if we delete one edge from each (odd) cycle of H, then we obtain a bipartite graph. Hence $B(H) = |V(H)| - 1$.

Suppose now that a connected graph H contains an even cycle $C = (w_1, \ldots, w_{2m})$. Let $V_1 = \{w_1, w_3, \ldots, w_{2m-1}\}$ and $V_2 = \{w_2, w_4, \ldots, w_{2m}\}$. Then at least $|V_1| + |V_2|$ edges in H connect $|V_1|$ with $|V_2|$. If $V_1 \cup V_2 = V(G)$, then we have $B(H) \geq |V(H)|$. Otherwise, since H is connected, there is a vertex w_{2m+1} adjacent to $V_1 \cup V_2$. If w_{2m+1} is adjacent to V_1, then we add it to V_2, otherwise add it to V_1. In any case, the number of edges between the new V_1 and V_2 is greater than between the old ones. We continue adding vertices to $V_1 \cup V_2$ so that with each added vertex, the number of edges between V_1 and V_2 grows by at least one. When we add the last vertex of H, we get a partition (V_1, V_2) of $V(H)$ such that the number of edges between V_1 and V_2 is at least $|V(H)|$. □

Now we show that restriction (1.1) in Theorem 3 cannot be weakened.

Suppose first that the multigraph H has no uneven components. In this case, by definition, $b(H) = B(H) = \max_{X \subset V(H)} e(X, V(H) - X)$. Let this maximum be achieved at the set $X_0 \subset V(H)$ and let $Y_0 = V(H) - X_0$. Let G be formed from two complete graphs G_1 and G_2 of order l that intersect on $b(H) - 1$ vertices. If the set S chosen as the image of $V(H)$ under f is such that the vertices of X_0 lie in $G_1 - G_2$ and the vertices of Y_0 lie in $G_2 - G_1$, then $G_1 \cap G_2$ is not large enough to allow an embedding of H. Further, $\delta(G) = l - 1$. Since $|V(G)| = 2l - b(H) + 1$, we see that $\delta(G) = \frac{n + b(H) - 3}{2}$. Thus, (1.1) is necessary in this case.

Suppose now that H has both uneven components and components containing even cycles. Let H_0 be the subgraph of H induced by all uneven components of H and H_1 be the subgraph of H induced by all other components. By our definition and Lemma 4,

$$b(H) = B(H_1) + |V(H_0)|. \tag{2.2}$$

Let X_1 be a subset of $V(H_1)$ such that $(X_1, V(H_1) - X_1)$ is a maximum edge cut and let $Y_1 = V(H_1) - X_1$. Then $B(H_1) = e(X_1, Y_1)$. Consider the same graph G as the previous paragraph. Let the mapping f be such that the image of X_1 is completely in $G_1 - G_2$, the image of Y_1 is completely in $G_2 - G_1$, and the image of $V(H_0)$ is completely in $G_1 \cap G_2$. Then only $b(H) - 1 - |V(H_0)|$ vertices of $G_1 \cap G_2$ are not occupied by vertices of H_0. By (2.2), this is not enough to embed all paths from the image of X_1 to the image of Y_1.

If every component of H is uneven, we will map the vertices of all but one component, say C_0, from H to $G_1 \cap G_2$ and then place the vertices of C_0 into $G_1 - G_2$ and $G_2 - G_1$ so that we need $|V(C_0)| - 1$ paths to connect $G_1 - G_2$ with $G_2 - G_1$.

3. Preliminaries. In this and the next two sections we consider only loopless H. First, for the purposes of our proof, we wish to show that it suffices to consider only H with no uneven components, or H that are connected and contain an odd cycle, or $H = K_2$. Note that if H' is obtained from H by adding an edge e' and if $k_1(H') \leq k_1(H)$ and $b(H') \leq b(H)$, then, since $H' \supset H$, the fact that a graph G is

H'-linked implies that G is H-linked. Now, if H has at least two components and a component H_1 of H is uneven, then by adding an edge connecting H_1 with another component, we decrease $c(H)$. This means that $b(H)$ and $k_1(H)$ do not change. Thus, in this case it is enough to consider only the cases when H is connected or has no uneven components. Furthermore, if H is a tree on at least 3 vertices, then adding to H an edge connecting two vertices at distance two does not change $c(H)$ or $b(H)$, but now H contains an odd cycle. If H is a tree on 2 vertices, then $H = K_2$ and hence $b(H) = 1$. Thus, it suffices to consider the case when H has no uneven components, or H is a connected graph containing an odd cycle, or $H = K_2$, and the reduction we desired is possible.

Suppose that $e(H) = k$. Let $f : V(H) \to V(G)$ be an injective mapping and $W = f(V(H))$. Let $E(H) = \{e_j = u_j^0 v_j^0 : 1 \le j \le k\}$. Let $u_j = f(u_j^0)$ and $v_j = f(v_j^0)$.

If $H = K_2$, then $k = 1$ and $b(H) = 1$. In this case, if an n-vertex graph G satisfies the conditions of the theorem, then $\delta(G) \ge (n-1)/2$. Therefore u_1 and v_1 are either adjacent or have a common neighbor. This settles the case of $H = K_2$, and from now on we assume that either H is connected and has a cycle or has no uneven components. In this case, $|W| = |V(H)| \le k$.

For each edge $e_j = u_j^0 v_j^0 \in E(H)$, we define functions $\beta(e_j, u_j^0)$, $\beta(e_j, v_j^0)$ inductively as follows:

(1) If H has no vertices of degree one, then for every j, let $\beta(e_j, u_j^0) = 1/deg_H(u_j^0)$ and $\beta(e_j, v_j^0) = 1/deg_H(v_j^0)$.
(2) If H has a pendant vertex u_s^0 (which is incident with the edge $e_s = u_s^0 v_s^0$), let $H' = H - u_s^0$. Since H' is a smaller graph without acyclic components, we can define $\beta(e_j, u_j^0), \beta(e_j, v_j^0)$ for every $j \ne s$ and then let $\beta(e_s, u_s^0) = 1$ and $\beta(e_s, v_s^0) = 0$.

For simplicity, we denote $\beta(e_j, u_j^0)$ by β_j, and $\beta(e_j, v_j^0)$ by γ_j. By construction, for every $j = 1, \dots, k$,

$$0 \le \beta_j, \gamma_j \le 1 \text{ and } \beta_j + \gamma_j \le 1. \tag{3.1}$$

Also, for every $u^0 \in V(H)$,

$$\sum_{\{e \in E(H) : u^0 \in e\}} \beta(e, u^0) = 1, \quad \text{and hence} \quad \sum_{j=1}^{k} (\beta_j + \gamma_j) = |V(H)| = |W|. \tag{3.2}$$

Say that a family $\mathcal{C}$ of the form $\{P_1, \dots, P_k\}$ is a *partial H-linkage* if each P_j is either the set $\{u_j, v_j\}$ or a u_j, v_j-path and the following conditions hold:

(I) $|X| \le |W| + 3k - 2b(H) + 2\alpha + 3$, where $X = \bigcup_{j=1}^{k} V(P_j)$ and α is the number of P_j-s that are paths;
(II) The internal vertices of the paths P_j's are pairwise disjoint and disjoint from W.

Consider $\mathcal{C}_0 = \{\{u_1, v_1\}, \dots, \{u_k, v_k\}\}$. This family satisfies the properties (I) and (II) above with $X = \bigcup_{j=1}^{k} \{u_j, v_j\} = W$ and $\alpha = 0$. Therefore, $\mathcal{C}_0$ is a partial H-linkage.

A partial H-linkage $\mathcal{C} = \{P_1, \dots, P_k\}$ is *optimal*, if as many P_j-s as possible are paths and, subject to this, the set $X = \bigcup_{j=1}^{k} V(P_j)$ is as small as possible. We will prove that an optimal partial H-linkage is an H-subdivision. This will imply our theorem (for loopless H).

Suppose, to the contrary, that $\mathcal{C} = \{P_1, \dots, P_k\}$ is an optimal partial H-linkage but is not an H-subdivision. Let, for definiteness, $P_k = \{u_k, v_k\}$ and $u_k v_k \notin E(G)$.

Denote $X = \bigcup_{j=1}^{k} V(P_j)$, $x = u_k$, and $y = v_k$. Let $A = N(x) - X$, $B = N(y) - X$, and $R = V(G) - (X \cup A \cup B)$.

By (1.1) and (2.1), each of A and B has size at least

$$\delta(G) - (|X| - 2) \geq \frac{n + b(H) - 2}{2} - (|W| + 3k - 2b(H) + 2(k-1) + 3 - 2)$$
$$\geq \frac{9.5k + b(H) - 2}{2} - 6k + 1 + 2b(H) = 2.5b(H) - 1.25k \geq 1.25.$$

It follows that we may choose distinct $a_1, a_2 \in A$ and $b_1, b_2 \in B$.

For $v \in V(G)$, let $d_j(v)$ denote the number of neighbors of v in the interior of P_j plus β_j if $u_j \in N_G(v)$ and plus γ_j if $v_j \in N_G(v)$ (β_j and γ_j are defined above (3.1)). By (3.2), we have

$$\sum_{j=1}^{k} d_j(v) = |N_G(v) \cap X| \qquad \forall v \in V(G). \tag{3.3}$$

Let l_p be the number of P_j's of length p for $p \geq 1$, and l_0 be the number of P_j's that are not paths. Then

$$|X| = |W| + \sum_{p \geq 1} (p-1) l_p = \sum_{j=1}^{k} (\beta_j + \gamma_j) + \sum_{p \geq 1} (p-1) l_p \tag{3.4}$$

and

$$k = \sum_{p \geq 0} l_p = \alpha + l_0. \tag{3.5}$$

We will assume that every path P_j is of the form $P_j = u_j, w_{1,j}, \ldots, w_{p_j-1,j}, v_j$. Sometimes, for simplicity we will write p instead of p_j and w_i instead of $w_{i,j}$ if j is clear from the context. In the rest of the paper, for every $j = 1, \ldots, k$ and fixed $a_1, a_2 \in A$, $b_1, b_2 \in B$, we denote $M_j = d_j(x) + d_j(y)$ and $L_j = d_j(a_1) + d_j(a_2) + d_j(b_1) + d_j(b_2)$.

In order to add an x, y-path to $\mathcal{C}$ and still satisfy condition (I), we are allowed to use only two additional vertices. In the next section, we prove that, for an optimal $\mathcal{C}$, the set X satisfies an inequality stronger than (I) and this allows us to use five additional vertices when constructing an x, y-path. We will eventually show that if even with the help of that many vertices we are not able to create an x, y-path, possibly changing already constructed paths, then either x or y has a low degree.

4. Main lemma. We begin with a lemma needed in the proof of Lemma 6.

LEMMA 5. *Let $a_1, a_2 \in A$, $b_1, b_2 \in B$. For a $P_j = u_j, w_1, \ldots, w_{p-1}, v_j$, let $s_j = M_j + 0.5L_j$, $\beta = \beta_j$, and $\gamma = \gamma_j$. Define*

$$D_1(p, \beta, \gamma) = \begin{cases} p + 2 + 2\beta + 2\gamma & \text{for } p \leq 1, \\ p + 4 + 2\beta + 2\gamma & \text{for } p \geq 2. \end{cases}$$

Then

(a) $s_j \leq D_1(p, \beta, \gamma)$.

(b) $s_k \leq 2(\beta_k + \gamma_k)$. *Furthermore, if $xy = u_k v_k \notin E(G)$, then $s_k = \beta_k + \gamma_k$.*

Proof. Let $\lambda = \max\{\beta, \gamma\}$. By definition (see (3.1)), $\lambda \le 1$, $\min\{\beta, \gamma\} \le 0.5$, and $L_k = 2\beta_k + 2\gamma_k$. If $xy \in E(G)$, then $M_k = \beta_k + \gamma_k$; otherwise, $M_k = 0$. This proves (b).

CLAIM 1. *Let* $Z = \{a_1, a_2, b_1, b_2\}$.

(i) *For each* $z \in Z$*, the distance in* P_j *between any two neighbors of* z *is at most two. In particular, each* $z \in Z$ *has at most* 3 *neighbors in* P_j*.*

(ii) *If* $p \ge 3$*, then no* $z \in Z$ *is a common neighbor of* u_j *and* v_j*.*

(iii) *If* $p \ge 3$*, then* x *and* y *have no interior neighbors of distance at most* $p - 3$ *in* P_j*.*

(iv) *If* $p \ge 3$*, then* x *(respectively,* y*) has no interior neighbors at distance at most* $p - 4$ *in* P_j *from interior neighbors of* b_1 *and* b_2 *(respectively, of* a_1 *and* a_2*).*

Proof. If some $z \in Z$ is adjacent to w_i and w_{i+m} for some $m \ge 3$ (we treat u_j as w_0 and v_j as w_p), then we can replace P_j by a shorter u_j, v_j-path, a contradiction to the optimality of $\mathcal{C}$. This proves (i), and (ii) is a partial case of (i).

If x and y have interior neighbors at distance at most $p - 3$ in P_j, then we can delete P_j from $\mathcal{C}$ and add a shorter x, y-path. This proves (iii). The same trick proves (iv), completing the proof of the claim. □

In order to prove (a), we consider several cases (depending on p).

Case 1. $p = 0$. By (3.1), $L_j \le 4(\beta + \gamma) \le 4$. Therefore $s_j = M_j + 0.5L_j \le 2(\beta + \gamma) + 2 = D_1(0, \beta, \gamma)$.

Case 2. $p = 1$. Trivially,

$$s_j \le 2(\beta + \gamma) + 0.5(4(\beta + \gamma)) \le 2(\beta + \gamma) + 2 < D_1(1, \beta, \gamma).$$

Case 3. $p = 2$. If each of x and y is adjacent to w_1 and some $z \in Z$ is adjacent to both u_j and v_j, then $\mathcal{C}$ is not optimal: we can replace P_j by the path u_j, z, v_j and add the path xw_1y. Otherwise, either $M_j \le 2(\beta + \gamma) + 1$ and hence

$$s_j \le 2(\beta + \gamma) + 1 + 0.5(4(\beta + \gamma + 1)) \le 2(\beta + \gamma) + 6 = D_1(2, \beta, \gamma),$$

or $L_j \le 4(\lambda + 1)$ and hence

$$s_j \le 2(\beta + \gamma + 1) + 0.5(4(\lambda + 1)) \le 2(\beta + \gamma) + 6 = D_1(2, \beta, \gamma).$$

Case 4. $p = 3$. By (iii), $M_j \le 2(\beta + \gamma) + 2$. If $L_j \le 10$, then $s_j \le D_1(3, \beta, \gamma)$. Otherwise, because of the symmetry between A and B, we may assume that $d_j(a_1) + d_j(a_2) > 5$ and that $d_j(a_1) > 2.5$. Then by (ii), we may assume that a_1 is adjacent to w_1, w_2, and v_j and that a_2 is adjacent to w_1 and w_2 (and maybe to one more vertex). If $yw_2 \in E(G)$, then we can replace P_j with u_j, w_1, a_1, v_j and add the path x, a_2, w_2, y, a contradiction to the optimality of $\mathcal{C}$. If neither x nor y is adjacent to w_2, then by (iii), $M_j \le 2(\beta + \gamma) + 1$, by (ii), $L_j \le 4(2 + \lambda) \le 12$, and therefore $s_j \le 2(\beta + \gamma) + 7 = D_1(3, \beta, \gamma)$. If $xw_2 \in E(G)$ and some $b \in \{b_1, b_2\}$ is adjacent to w_2, then we can replace P_j with u_j, w_1, a_1, v_j and add the path x, w_2, b, y. Finally, if neither b_1w_2 nor b_2w_2 is in $E(G)$, then by (i), $d_j(b_1) + d_j(b_2) \le 2(1 + \lambda) \le 4$, and hence by (ii) $L_j \le 6 + 4 = 10$.

Case 5. $p \ge 4$. If x has r interior neighbors and $r \ge 2$, then by (iii), $d_j(y) \le \beta + \gamma$ and by (iv), $d_j(b_i) \le \max\{0, 3 - r\} + \lambda$. Together with (i) this shows that in this case,

$$s_j \le 2\beta + 2\gamma + r + 3 + \max\{0, 3 - r\} + \lambda.$$

If $r \ge 3$, then $s_j \le 2\beta + 2\gamma + p - 1 + 3 + \lambda \le p + 3 + 2\beta + 2\gamma \le D_1(p, \beta, \gamma)$. If $r = 2$, then $s_j \le 2\beta + 2\gamma + r + 4 + \lambda \le 2\beta + 2\gamma + p + 3 \le D_1(p, \beta, \gamma)$, again.

Thus, we can assume that each of x and y has at most one interior neighbor in P_j. By (iv) $d_j(a_i) + d_j(y) \le \beta + \gamma + \lambda + 3$ and $d_j(b_i) + d_j(x) \le \beta + \gamma + \lambda + 3$ for $i = 1, 2$. Therefore, $s_j \le 2\lambda + 6 + 2\beta + 2\gamma \le 2\beta + 2\gamma + p + 2 + 2 = D_1(p, \beta, \gamma)$. This completes the proof of (a) and hence, of Lemma 5. □

LEMMA 6. *Let $a_1, a_2 \in A$, $b_1, b_2 \in B$, $Z = \{a_1, a_2, b_1, b_2\}$, and $V_0 = (A \cup B) - Z - N_G(Z)$. Then $|X| \le |W| + 3k - 2b(H) + 2\alpha - |R| - |V_0|$.*

Proof. Let

$$\Sigma' = \deg_G(x) + \deg_G(y) + \frac{1}{2}(\deg_G(a_1) + \deg_G(a_2) + \deg_G(b_1) + \deg_G(b_2)). \tag{4.1}$$

Observe that every vertex $w \notin X$ contributes to Σ' at most 2: if $w \in R$, then it is not adjacent to x and y, and if $w \in A$ (respectively, $w \in B$), then it is not adjacent to y, b_1, and b_2 (respectively, to x, a_1, and a_2). By definition, every vertex in V_0 is not adjacent to any vertex in Z, and therefore contributes at most 1 to Σ'. Furthermore, every $z \in Z$ contributes at most 1.5 to Σ', since it is not adjacent to itself. Therefore,

$$\Sigma' \le 4 \cdot 1.5 + 2(|A \cup B| - 4) + 2|R| + \sum_{j=1}^{k} s_j - |V_0|. \tag{4.2}$$

By Lemma 5, (3.2), and (3.5),

$$\begin{aligned} \sum_{j=1}^{k} s_j &\le k + l_0 + 2l_1 + \sum_{p \ge 2} (p+3) l_p + 2 \sum_{j=1}^{k} (\beta_j + \gamma_j) - 1 \\ &= k + l_0 + 2l_1 + \sum_{p \ge 2} (p+3) l_p + 2|W| - 1. \end{aligned} \tag{4.3}$$

Therefore,

$$\begin{aligned} \Sigma' \le 2(|A \cup B| + |R|) - 2 - |V_0| + 2 \left(|W| + l_0 + \sum_{p \ge 1} p l_p \right) \\ - 1 - l_0 + \sum_{p \ge 2} (3 - p) l_p + k. \end{aligned}$$

By (3.4) and (3.5), the last expression is equal to $2n + 3k - |V_0| - 3 - l_0 - \sum_{p \ge 2} (p-3) l_p$. Combining this again with (3.4) and (3.5), we get

$$|X| + \Sigma' \le 2n + |W| + 3k + 2\alpha - 3 - l_0 - 2l_1 - |V_0|.$$

By the assumption of Theorem 3, $\delta(G) \ge \frac{n + b(H)}{2} - 1$ and hence $\Sigma' \ge 2n + 2b(H) - 4$. Thus,

$$\begin{aligned} |X| &\le |W| + 3k - 2b(H) + 2\alpha - l_0 - 2l_1 - |V_0| + 1 \\ &\le |W| + 3k - 2b(H) + 2\alpha - |V_0|. \end{aligned} \tag{4.4}$$

If an $r \in R$ has a neighbor $a_0 \in A$ and a neighbor $b_0 \in B$, then one can add to $\mathcal{C}$ the path $P_k = x, a_0, r, b_0, y$. The new set of paths will be a better partial linkage, since the new X would have size at most $|W| + 3k - 2b(H) + 2(\alpha + 1) + 1$. Since this

contradicts the choice of $\mathcal{C}$, no $r \in R$ has both a neighbor in A and a neighbor in B. Thus every $r \in R$ contributes at most 1 to Σ', and (4.2) becomes

$$\Sigma' \le 4 \cdot 1.5 + 2(|A \cup B| - 4) + |R| + \sum_{j=1}^{k} s_j - |V_0|.$$

Correspondingly, (4.4) transforms into

$$|X| \le |W| + 3k - 2b(H) + 2\alpha - |V_0| - |R|. \qquad (4.5)$$ □

5. Completion of the case of loopless H. Lemma 6 has the following two immediate consequences.

LEMMA 7. $|A| + |B| > 2k$.

Proof. By Lemma 6 and (2.1), $|A| + |B| = n - (|X| + |R|) \ge n - (|W| + 3k - 2b(H) + 2\alpha) \ge 9.5k - (k + 3k - 2\frac{k+1}{2} + 2(k-1)) = 4.5k + 3 > 2k$. □

LEMMA 8. *Each $v \in V(G)$ is adjacent to at least 3 vertices in $A \cup B - V_0$. In particular, either v has 2 neighbors in A that belong to or are adjacent to the set $\{a_1, a_2\}$, or 2 neighbors in B that belong to or are adjacent to the set $\{b_1, b_2\}$.*

Proof. Recall that by the definition of V_0, $A \cup B - V_0 = Z \cup (N_G(Z) \cap (A \cup B))$. Hence, by Lemma 6,

$$\begin{aligned}\delta(G) - (|X| + |R| + |V_0|) &\ge 0.5(9.5k + b(H) - 2) - |W| - 3k + 2b(H) - 2\alpha \\ &\ge 4.75k + 0.5b(H) - 1 - k - 3k + 2b(H) - 2(k-1) \\ &= 2.5b(H) - 1.25k + 1 \ge 2.25 > 2.\end{aligned}$$

Thus each vertex has at least 3 neighbors in $V(G) - X - R - V_0 = A \cup B - V_0$. □

For given $a_1, a_2 \in A$, $b_1, b_2 \in B$, let $A'' = A''(a_1, a_2)$ (respectively, $B'' = B''(b_1, b_2)$) denote the set of vertices in X having at least 2 neighbors in A (respectively, in B) that belong to or are adjacent to the set $\{a_1, a_2\}$ (respectively, $\{b_1, b_2\}$). The above lemma yields that for every choice of a_1, a_2, b_1, and b_2,

$$A'' \cup B'' = X. \qquad (5.1)$$

LEMMA 9. *For every nonadjacent $s, t \in A$ (or B), $|N(s) \cap N(t) - X| \ge 3$.*

Proof. Suppose to the contrary that $a_1, a_2 \in A$, $a_1a_2 \notin E(G)$ and the cardinality of the set T of common neighbors of a_1 and a_2 outside of X is at most two. Consider arbitrary $b_1, b_2 \in B$ and let $Z = \{a_1, a_2, b_1, b_2\}$. Then the contribution of every $a \in A - Z - T$ to the sum Σ' defined in (4.1) is at most 1.5. Thus, repeating the proof of Lemma 6, the right-hand side of the inequality corresponding to (4.5) will be less by $0.5|A - Z - T|$. Hence, since $|(Z \cap A) \cup T| \le 4$, instead of (4.5), we will get $|X| \le |W| - |R| + 3k - 2b(H) + 2\alpha - |V_0| - 0.5(|A - V_0| - 4)$. In other words,

$$|X| + 0.5|A| + |R| \le |W| + 3k - 2b(H) + 2\alpha + 2 \le 6k - 2b(H). \qquad (5.2)$$

On the other hand, $\deg_{G-X}(a_1) + \deg_{G-X}(a_2) \le |A| + |T| + |R| - 2$ (the -2 arises because neither a_1 nor a_2 is adjacent to a_1 or a_2). It follows that

$$2\frac{n + b(H)}{2} - 2 \le 2\delta(G) \le 2|X| + |A| + |R|,$$

which together with (5.2) yields $n + b(H) - 2 \le 2(6k - 2b(H))$. Thus, $n \le 12k - 5b(H) + 2 \le 12k - 5\frac{k+1}{2} + 2 = 9.5k - 0.5$, a contradiction. ∎

For the rest of the section, we fix some distinct $a_1, a_2 \in A$ and $b_1, b_2 \in B$, and let $A'' = A''(a_1, a_2)$ and $B'' = B''(b_1, b_2)$.

LEMMA 10. *Let $\mathcal{C}$ be optimal, $1 \le j \le k-1$, and either $\{u_j, v_j\} \subset A''$ or $\{u_j, v_j\} \subset B''$. Then for each $a \in A$ and $b \in B$,*

$$(N(a) \cap N(b) \cap P_j) \backslash \{u_j, v_j\} = \emptyset.$$

Proof. Assume to the contrary that $r \in N(a) \cap N(b) \cap P_j \backslash \{u_j, v_j\}$. Let $P'_k = (x, a, r, b, y)$. Without loss of generality, assume that $\{u_j, v_j\} \subset A''$. Then there exist $s \in N(u_j) \cap A \backslash \{a\}$ and $t \in N(v_j) \cap A \backslash \{a\}$. If $s = t$ or s is adjacent to t, then let $P'_j = (u_j, s, t, v_j)$.

If s and t are nonadjacent, then by Lemma 9, we have $|(N(s) \cap N(t)) \backslash X| \ge 3$, and therefore there exists $q \in N(s) \cap N(t) \backslash (X \cup \{a, b\})$. In this case, let $P'_j = (u_j, s, q, t, v_j)$. In both cases, P'_j is a path disjoint from P'_k. Thus, in both cases we increase the number of P_j-s that are paths by one and, by (4.5), maintain $|X| \le |W| + 3k - 2b(H) + 2(\alpha + 1) + 3$. This is a contradiction which completes the proof. ∎

LEMMA 11. *Let $\mathcal{C}$ be optimal, $1 \le j \le k-1$, $P_j = (w_0, w_1, \ldots, w_p)$, where $w_0 = u_j \in A''$, and $w_p = v_j \in B''$. If some w_i, $1 \le i \le p-1$, has a neighbor $a_0 \in A \cup \{x\}$ and a neighbor $b_0 \in B \cup \{y\}$, then each $w_{i'}$ for $i < i' \le p$ has no neighbors in $A - a_0$ and each $w_{i''}$ for $0 \le i'' < i$ has no neighbors in $B - b_0$.*

Proof. Suppose some $w_{i'}$ for $i < i' \le p$ has a neighbor $a' \in A - a_0$. By the definition of A'', u_j has a neighbor $a'' \in A - a_0$. By Lemma 9, the length of a shortest path P' from a'' to a' in $G[A - a_0]$ is at most two. Thus, we can replace P_j by the path $(u_j, a'', P', a', w_{i'}, P'_j, v_j)$ (where P'_j is the part of P_j connecting $w_{i'}$ with v_j) and add the path $P_k = (x, a_0, w_i, b_0, y)$. The new set of $\alpha + 1$ paths has at most $|X| + 5$ vertices, which by (4.5) is at most $|W| + 3k - 2b(H) + 2(\alpha + 1) + 3$, a contradiction to the choice of $\mathcal{C}$. Note that a similar argument works for $w_{i''}$. ∎

Similarly to $d_j(v)$, let $d_j(u, v)$ denote the number of common neighbors of u and v "inside" P_j plus $\beta_j \cdot |N(u) \cap N(v) \cap \{u_j\}|$ plus $\gamma_j \cdot |N(u) \cap N(v) \cap \{v_j\}|$.

LEMMA 12. *Let $\mathcal{C}$ be optimal, $a \in A$, $b \in B$. Then there exists some $j = j(a, b)$ such that $d_j(a, b) > 1$.*

Proof. Since $N(a) \cap N(b) \cap (V(G) - X + x + y) = \emptyset$ (otherwise we can find a path $xazby$ not using any vertex of X), we have

$$\sum_{j=1}^{k-1} d_j(a, b) = |N(a) \cap N(b)| \ge 2\delta(G) - (n-2) \ge b(H). \tag{5.3}$$

Suppose that $d_j(a, b) \le 1$ for each $1 \le j \le k-1$. Then we will find an edge cut in H with more than $\sum_{j=1}^{k-1} d_j(a, b)$ edges, a contradiction to (5.3). Let E' be the set of edges e_j in H such that an internal vertex of P_j contains a vertex of $N(a) \cap N(b)$. Let V' be the set of vertices u^0 in H such that the vertex $f(u^0)$ (i.e., the branching vertex in G corresponding to u^0) is in $N(a) \cap N(b)$. Recall that $x, y \notin N(a) \cap N(b)$. By our assumption, no vertex in V' is incident to an edge in E', and for each $e_j \in E'$, the path P_j contains exactly one vertex of $N(a) \cap N(b)$. Thus, it is enough to find in H an edge cut of size greater than $|E'| + |V'|$.

Let V_0 denote the set of vertices in all components of H containing at least one edge of $E' \cup \{e_k\}$ and let H_0 be the subgraph of H induced by V_0. Again by Lemma 10, for each $e_j \in E'$, either $u_j \in A'' - B''$ and $v_j \in B'' - A''$ or $v_j \in A'' - B''$

and $u_j \in B'' - A''$. Recall that $x = f(u_k^0)$, $y = f(v_k^0)$, $x \in A'' - B''$, and $y \in B'' - A''$. It follows that the set $E' \cup \{e_k\}$ is contained in an edge-cut in H. Let V_1 and V_2 be the disjoint subsets of $V(H_0)$ such that

(a) each edge in $E' \cup \{e_k\}$ is incident to a vertex in V_1 and a vertex in V_2, and
(b) each vertex in $V_1 \cup V_2$ is incident to an edge in $E' \cup \{e_k\}$.

By the above, $V' \cap (V_1 \cup V_2) = \emptyset$ and hence $|V(H) - (V_1 \cup V_2)| \geq |V'|$. If $V_1 \cup V_2 \neq V_0$, then there is a vertex $u^0 \in V_0 - (V_1 \cup V_2)$ adjacent to $V_1 \cup V_2$. If u^0 is adjacent to V_1, then we add u^0 to V_2, otherwise add it to V_1. In any case the number of edges between the new V_1 and V_2 is greater than between the old ones. We continue adding vertices to $V_1 \cup V_2$ so that with each added vertex, the number of edges between V_1 and V_2 grows by at least one until we add all vertices of $V_0 - (V_1 \cup V_2)$. When we add the last vertex of H_0, we get a partition (V_1, V_2) of V_0 such that the number of edges between V_1 and V_2 is at least

$$|E' \cup \{e_k\}| + |V_0 - (V_1 \cup V_2)| \geq |E'| + 1 + |V' \cap V_0|.$$

If $H_0 = H$, then we get a contradiction to (5.3). If $H_0 \neq H$, then every component H_i of $H - V_0$ has an even cycle and by Lemma 4, H_i has an edge cut with at least $|V(H_i)|$ edges. This together with the partition (V_1, V_2) of V_0 will give an edge cut of H with at least $|E'|+1+|V' \cap V_0|+|V(H)-V_0| \geq |E'|+1+|V'|$ edges, a contradiction to (5.3). □

LEMMA 13. *Let $\mathcal{C}$ be optimal, $1 \leq j \leq k-1$. Then there is at most one $a \in A$, such that there is more than one $b \in B$ with $j = j(a, b)$.*

Proof. Let $P_j = (w_0, w_1, \ldots, w_p)$, where $w_0 = u_j$ and $w_p = v_j$. Assume to the contrary that there are $a_1, a_2 \in A$ and $b_1, b_2, b_3, b_4 \in B$ such that $j(a_1, b_1) = j(a_1, b_2) = j(a_2, b_3) = j(a_2, b_4) = j$, where $a_1 \neq a_2$, $b_1 \neq b_2$, $b_3 \neq b_4$. By Lemma 10, we may assume that $u_j \in A'' \backslash B''$ and $v_j \in B'' \backslash A''$.

Since $\beta_j + \gamma_j \leq 1$, there exists i, $1 \leq i \leq p-1$, such that $w_i \in N(a_1) \cap N(b_1)$. Since $b_3 \neq b_4$, we may assume that $b_3 \neq b_1$. By Lemma 11, no vertex in $V(P_j) - w_i$ can belong to $N(a_2) \cap N(b_3)$. However, this contradicts the fact that $d_j(a_2, b_3) > 1$. □

By Lemma 7, $|A| + |B| > 2k$. We may assume that $|A| \leq |B|$. Thus $|B| \geq k$. If $|A| \geq k$, then since $|B| \geq k$, for each $a \in A$ there is some $j(a)$ and $b_1(a)$ and $b_2(a)$ such that $j(a) = j(a, b_1(a)) = j(a, b_2(a))$. Furthermore, since $|A| \geq k$, for some $a_1, a_2 \in A$, the indices $j(a_1)$ and $j(a_2)$ are the same. This contradicts Lemma 13.

Thus we may assume that $|A| < k$. Since $|B| \geq k$, for each $a \in A$ there is some $j(a)$ and $b_1(a)$ and $b_2(a)$ such that $j(a) = j(a, b_1(a)) = j(a, b_2(a))$. Let $J = \{j(a) \mid a \in A\}$. By Lemma 13, the indices $j(a)$ are distinct for distinct $a \in A$ and hence $|J| = |A|$.

LEMMA 14. *Suppose that $j \in J$. Then x is not adjacent to some interior vertex of P_j.*

Proof. Let $P_j = (w_0, w_1, \ldots, w_p)$, where $w_0 = u_j$ and $w_p = v_j$. By the definition of J, there exists $a \in A$ and $b_1, b_2 \in B$ such that $d_j(a, b_1), d_j(a, b_2) > 1$. Since $\beta_j + \gamma_j \leq 1$, this implies that $p \geq 2$. Assume that $u_j \in A'' - B''$ and $v_j \in B'' - A''$.

Since $u_j \notin B''$, we may assume that $u_j b_1 \notin E(G)$. Let $w_{i'}, w_{i''} \in N(a) \cap N(b_1)$ and $i' < i''$. By our choice of $w_{i'}$, $1 \leq i' \leq p-1$. If $xw_{i'} \in E(G)$, then we get a contradiction to Lemma 11 with $a_0 = x$, since $w_{i''}a \in E(G)$. Thus, $xw_{i'} \notin E(G)$. □

By Lemma 14, x is not adjacent to at least $|J|$ vertices in $X - W$. It also is not adjacent to itself. Thus, $|N(x) \cap X| \leq |X| - |J| - 1 \leq |W| + 3k - 2b(H) + 2(k-1) -$

$|J| - 1 \le 6k - 2b(H) - 3 - |J|$. Since $|J| = |A| = |N(x) - X|$, we get

$$\frac{n + b(H)}{2} - 1 \le \deg(x) \le 6k - 2b(H) - 3,$$

which yields $n \le 12k - 5b(H) - 4 < 9.5k - 6.5$, a contradiction. This contradiction proves that an optimal partial H-linkage is an H-linkage in the case of loopless H.

By condition (I) in the definition of a partial H-linkage, $|X| \le |W| - 2b(H) + 5k + 3 \le 5k + 2$.

6. Proof of the general case. As in section 2, it is enough to consider H that either has no uneven components or is connected and has an odd cycle other than a loop, or has at most two vertices. Let H have k' nonloop edges and k'' loops, in total $k = k' + k''$ edges. Recall that $n \ge 9.5(k_1 + 1)$, where $k_1 = k + c(H)$. Note that $b(H)$ does not depend on k'', thus $b(H) \ge 0.5k'$.

Let $f : V(H) \to V(G)$ be an injective mapping and $W = f(V(H))$. Let $E(H) = \{e_j = u_j^0 v_j^0 : 1 \le j \le k\}$. We may assume that the first k' edges are not loops. Let $u_j = f(u_j^0)$ and $v_j = f(v_j^0)$.

Let H' be the multigraph obtained from H by deleting all loops and let $k_1' = k' + c(H')$. Since H' is loopless, our theorem is proved for it, and thus f can be extended to an H'-subdivision in G on at most $5k_1' + 2$ vertices. If H' has an acyclic component, then so does H, and hence by the above, $|V(H')| \le 2$. It was observed in section 3 that in this case G has a subdivision of H' on at most 3 vertices. Thus, in either case, f can be extended to an H'-subdivision in G on at most $5k' + 2$ vertices. Among such H'-subdivisions choose one, say, F_1, with the fewest vertices and let $X_1 = V(F_1)$. We will extend F_1 to a partial H-subdivision F such that:

(I′) as many loops as possible are mapped to internally disjoint cycles of length at most 4, and

(II′) among partial H-subdivisions satisfying (I′), the set $X = V(F)$ has the smallest size.

We claim that such a partial H-subdivision is actually an H-subdivision. Suppose not, then we may assume that F represents the images $g(e_j)$ for $1 \le j \le q$, where $k' \le q \le k - 1$.

First we observe that by the minimality of F_1 and F, every vertex outside X has at most 3 neighbors in $g(e_j)$ for each $1 \le j \le q$.

Let e_{q+1} be a loop at vertex u_{q+1}^0 and $u_{q+1} = f(u_{q+1}^0)$. Consider graph $G' = G - (X - u_{q+1})$.

If H is not an isolated vertex, then every $x \in W$ is in X_1 (in fact, x belongs to $g(e_j)$ for some $1 \le j \le k'$), therefore, u_{q+1} has at most $3(q - k')$ neighbors in $X - X_1$ by (I′). If H is an isolated vertex, then $k' = 0$, $V(H) = \{u_{q+1}\}$, and u_{q+1} has at most $2q$ neighbors in X. It follows that

$$\deg_{G'}(u_{q+1}) \ge \deg_G(u_{q+1}) - 5k' - 2 - 3(q - k') \ge \frac{n + k'/2}{2} - 1 - 5k' - 2 - 3(q - k')$$

$$\ge \frac{n}{2} - 4.75q - 3 \ge \frac{9.5(k+1)}{2} - 4.75(k-1) - 3 \ge 6.5.$$

Let $S = N_{G'}(u_{q+1})$. If some vertices of S are adjacent or have a common neighbor in G' other than u_{q+1}, then we extend our partial H-linkage. If this is not the case,

then all neighbors in G' of vertices in S, apart from u_{q+1}, are distinct. Thus,

$$\sum_{s\in S}(\deg_{G'}(s)-1)+|S|+1\le n-(|X|-1). \tag{6.1}$$

Since $S\cap X=\emptyset$, by the above, $\deg_{G'}(s)\ge \deg_G(s)-\min\{|X|,3q\}$ for every $s\in S$. Thus, (6.1) yields $|S|(\delta(G)-\min\{|X|,3q\})+1\le n-|X|+1$. Since $|S|>6$, we have

$$6\frac{n}{2}\le 6\min\{|X|,3q\}+n-|X|\le 15q+n\le 15(k-1)+n.$$

It follows that $2n<15k$, a contradiction.

Acknowledgment. We thank the referees for their very helpful comments.

REFERENCES

[1] M. Ferrara, R. J. Gould, G. Tansey, and T. Whalen, *On H-linked graphs*, Graphs & Combinatorics, 22 (2006), pp. 217–224.
[2] R. J. Gould and T. Whalen, *Subdivision extendability*, Graphs & Combinatorics, to appear.
[3] H. A. Jung, *Eine Verallgemeinerung des n-fachen Zusammenhangs für Graphen*, Math. Ann., 187 (1970), pp. 95–103.
[4] H. Kierstead, G. Sárközy, and S. Selkow, *On k-ordered Hamiltonian graphs*, J. Graph Theory, 32 (1999), pp. 17–25.
[5] A. Kostochka and G. Yu, *On H-linked graphs*, Oberwolfach Report, no. 1, (2004), pp. 42–45.
[6] A. Kostochka and G. Yu, *Minimum degree conditions for H-linked graphs*, Discrete Applied Mathematics, to appear.
[7] A. Kostochka and G. Yu, *An extremal problem for H-linked graphs*, J. Graph Theory, 50 (2005), pp. 321–339.
[8] R. Thomas and P. Wollan, *An improved linear edge bound for graph linkages*, European J. Combin., 26 (2005), pp. 309–324.
[9] D. B. West, *Introduction to Graph Theory*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 2001.
[10] T. Whalen, *Degree Conditions and Relations to Distance, Extendability, and Levels of Connectivity in Graphs*, Ph.D. thesis, Department of Mathematics and Computer Science, Emory University, Atlanta, GA, 2003.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 841–879

OPTIMAL INTERLEAVING ON TORI*

ANXIAO (ANDREW) JIANG†, MATTHEW COOK‡, AND JEHOSHUA BRUCK§

Abstract. This paper studies t-interleaving on two-dimensional tori. Interleaving has applications in distributed data storage and burst error correction, and is closely related to Lee metric codes. A *t-interleaving* of a graph is defined as a vertex coloring in which any connected subgraph of t or fewer vertices has a distinct color at every vertex. We say that a torus can be *perfectly t-interleaved* if its *t-interleaving number* (the minimum number of colors needed for a t-interleaving) meets the sphere-packing lower bound, $\lceil t^2/2 \rceil$. We show that a torus is perfectly t-interleavable if and only if its dimensions are both multiples of $\frac{t^2+1}{2}$ (if t is odd) or t (if t is even). The next natural question is how much bigger the t-interleaving number is for those tori that are not perfectly t-interleavable, and the most important contribution of this paper is to find an optimal interleaving for all sufficiently large tori, proving that when a torus is large enough in both dimensions, its t-interleaving number is at most just one more than the sphere-packing lower bound. We also obtain bounds on t-interleaving numbers for the cases where one or both dimensions are not large, thus completing a general characterization of t-interleaving numbers for two-dimensional tori. Each of our upper bounds is accompanied by an efficient t-interleaving scheme that constructively achieves the bound.

Key words. bursts, chromatic number, cluster, error-correcting code, Lee distance, multidimensional interleaving, t-interleaving, torus

AMS subject classifications. 05C15, 05C70, 94B20

DOI. 10.1137/040618655

1. Introduction.

Interleaving is an important technique used for error burst correction and network data storage. In communications, interleaving the bits of consecutive codewords guarantees that error bursts will get distributed over many codewords, thus allowing the use of conventional error-correcting codes to correct bursts of errors [16]. The concept of a one-dimensional error burst was generalized to higher dimensions by Blaum, Bruck, and Vardy in [8], where an error burst of size t is defined as a set of errors confined to a connected subgraph of t vertices in a multidimensional array. It is there that the notion of t-interleaving was introduced, the purpose being to color the vertices of a multidimensional array so that every connected subgraph of t vertices receives t distinct colors, and two- and three-dimensional t-interleaving schemes were presented. Such schemes have applications in combatting error bursts in two-dimensional magnetic media and in three-dimensional holographic storage systems and optical recording systems.

Subsequent work on t-interleaving includes [21], where t-interleaving on circulant graphs with two offsets was studied, and [24], where a dual problem of t-interleaving on two-dimensional arrays was explored. The problem of two-dimensional interleaving

*Received by the editors November 10, 2004; accepted for publication (in revised form) May 15, 2006; published electronically December 5, 2006. This work was supported in part by the Lee Center for Advanced Networking at the California Institute of Technology, and by NSF grant CCR-TC-0208975. A one-page abstract presenting part of the results in this paper appeared in the Proceedings of the IEEE International Symposium on Information Theory, held in Chicago, 2004.
http://www.siam.org/journals/sidma/20-4/61865.html

†Computer Science Department, Texas A&M University, College Station, TX, 77843-3112 (ajiang@cs.tamu.edu).

‡Computation and Neural Systems Department, California Institute of Technology, MC 136-93, Pasadena, CA, 91125 (cook@paradise.caltech.edu).

§Electrical Engineering Department, California Institute of Technology, MC 136-93, Pasadena, CA, 91125 (bruck@paradise.caltech.edu).

with repetitions was introduced in [7] by Blaum, Bruck, and Farrell, and was extensively studied in [10] by Etzion and Vardy. That problem is to interleave colors on a two-dimensional mesh (array or its variation) in such a way that in every connected subgraph of t vertices, each color appears at most r times. Here t and r are given parameters, and the concept of interleaving with repetitions is a generalization of t-interleaving. More work on interleaving with repetitions includes [17] and [19]. Interleaving schemes on two-dimensional arrays achieving the Reiger bound were studied by Abdel-Ghaffar in [1], where error bursts of both rectangular shapes and arbitrary connected shapes were considered. More examples of interleaving for coping with shaped error bursts include [3] and [6], where the error bursts considered are respectively circular and rectangular.

Interleaving schemes have also been used for network data storage. In [12], an algorithm was presented to interleave N colors on a tree whose edges have lengths, in such a way that for every point of the tree (including a vertex or a point part way along an edge), the smallest ball centered at the point that contains at least N vertices will contain all N colors. That algorithm is useful for minimizing data retrieval delay in distributed data storage systems in hierarchical or tree-like networks. A related interleaving algorithm aimed at the graceful degradation of data-storage performance in faulty environments was presented in [14]. In [13], a scheme called *multicluster interleaving* was studied, which is a scheme to interleave colors on a path or a cycle such that every m disjoint intervals of length L in the path or cycle together contain at least K distinct colors, where $K > L$. Multicluster interleaving can be used for data storage on array-networks, ring-networks, or disks where data gets accessed through multiple access points.

This paper is the first to study t-interleaving on two-dimensional tori. Tori provide an important network structure for parallel and distributed systems [9], [18], [20], [22]. The use of t-interleaving on tori has applications in both burst error correction and distributed data storage, similar to [8], [21], [24], [12] and [14]. Specifically, for distributed data storage, a t-interleaving on a two-dimensional torus ensures that for every vertex, the colors assigned within $\lfloor \frac{t-1}{2} \rfloor$ hops are all distinct. The topic of t-interleaving on tori is closely related to a research topic in coding theory called *Lee metric codes* [2], [4], [5], [11], [15]. In a t-interleaved n-dimensional torus, the set of vertices having any given color is a Lee metric code of length n whose minimum distance is t, and the set of Lee metric codes corresponding to different colors partitions the whole code space.

Here we present some definitions so that we can state our claims precisely. These definitions are straightforward generalizations of the definition of t-interleaving originally given in [8] for arrays.

DEFINITION 1.1. *Let G be a graph. By an* interleaving*, we will mean a vertex coloring, as follows. We say that G is* interleaved (*or there is an* interleaving *on G*) *if each vertex of G is assigned one of a finite number of distinct colors. We say that G is* t-interleaved (*or there is a* t-interleaving *on G*) *if every set of t vertices, forming a connected subgraph of G, is colored by t distinct colors.*

The classic vertex coloring problem is clearly also a t-interleaving problem, where $t = 2$. On the other hand, t-interleaving a graph G is the same as vertex-coloring the power graph G^t, when the power graph G^t is defined as adding an edge to G between each pair of vertices connected by a path of t or fewer vertices. Determining the chromatic number of this kind of power graph is difficult in general. To the best of our knowledge, no result on the type of graphs we are interested in has appeared in the literature.

DEFINITION 1.2. *A two-dimensional $l_1 \times l_2$ torus is a graph containing $l_1 l_2$ vertices and $2l_1 l_2$ edges. We denote its vertices by (i, j) for $0 \le i \le l_1 - 1$ and $0 \le j \le l_2 - 1$.*

$(0,0)$	$(0,1)$	$\cdots$	$(0, l_2-1)$
$(1,0)$	$(1,1)$	$\cdots$	$(1, l_2-1)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$(l_1-1,0)$	$(l_1-1,1)$	$\cdots$	(l_1-1, l_2-1)

Each vertex (i, j) is incident to four edges, which connect it to its four neighbors according to the arrangement shown, wrapping around at the boundaries: $((i-1) \bmod l_1, j)$, $((i+1) \bmod l_1, j)$, $(i, (j-1) \bmod l_2)$, *and* $(i, (j+1) \bmod l_2)$.

Now we can define the problem of t-interleaving on tori.

DEFINITION 1.3. *The minimum number of colors used by any t-interleaving for G is called the* t-interleaving number *of G. A t-interleaving on a torus whose number of colors equals the torus' t-interleaving number is called an* optimal t-interleaving*, as it uses as few colors as possible.*

Example 1.1. The following 5×5 torus is 3-interleaved with 6 colors. The colors are shown as integers from 0 to 5. Each vertex is shown as a square cell in the grid, which is understood to have its left and right edges identified, and its top and bottom edges identified, thus forming a torus.

0	3	1	4	2
1	4	2	0	3
2	0	3	1	5
3	1	5	2	0
4	2	0	3	1

However, the 3-interleaving number of this torus is not 6, since a 3-interleaving does not require 6 colors: If we replace the two instances of color 5 with color 4, we can achieve a 3-interleaving with 5 colors. Thus the 3-interleaving number of this torus is at most 5.

To see that we need 5 colors, consider the vertex $(1, 1)$ and its four neighbors $(0, 1)$, $(2, 1)$, $(1, 0)$, and $(1, 2)$, and notice that any two of them are contained in a connected subgraph of order 3. Therefore, any 3-interleaving has to assign those 5 vertices 5 distinct colors. Thus the 3-interleaving number of this torus is 5.

Note that a torus that does not have at least t rows and t columns will have the property that there is a path of length less than t which wraps around the torus, going from a vertex to itself. While the definitions can still be understood for such small tori, often the practical application of interleaving results breaks down when this happens, and we will not consider such small tori in this paper.

Assumption 1.1. When discussing t-interleaving for a torus, we will assume that the torus has at least t rows and t columns when t is odd, and at least $t - 1$ rows and t columns when t is even.

Our objective in this paper is to find optimal t-interleavings. The t-interleaving number of a torus is by definition the number of colors of an optimal t-interleaving, one which uses the smallest number of colors. A lower bound, which we call the *sphere-packing lower bound*, can be obtained as follows. Figure 1.1 shows six graphs (subgraphs of a torus, assuming they fit on the torus) which we call *spheres* S_1, $S_2, \ldots, S_6$, respectively. In general, for any $t \ge 3$, the sphere S_t is obtained by attaching to the sphere S_{t-2} all the vertices adjacent to it. Any two vertices in S_t are connected by a path of at most $t - 1$ edges, so a t-interleaving needs to color

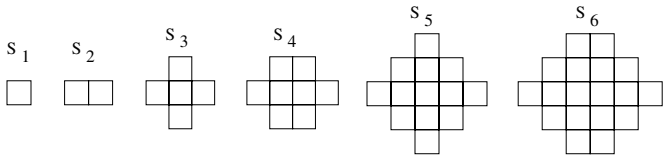


FIG. 1.1. *Six examples of spheres.*

them with different colors. So the number of vertices in S_t, which we shall denote by $|S_t|$, sets a universal lower bound for the t-interleaving number. This argument was originally proposed in [8] for studying t-interleaving on arrays. A direct calculation tells us that $|S_t| = \frac{t^2+1}{2}$ when t is odd, and $|S_t| = \frac{t^2}{2}$ when t is even. We refer to this as the *sphere-packing lower bound.*

We define *perfect t-interleaving* to be a t-interleaving using just $|S_t|$ colors, thus achieving the sphere-packing lower bound, on a torus that has at least t rows and t columns. Clearly any perfect t-interleaving is an optimal t-interleaving.

We will show that a torus can be perfectly interleaved if and only if its sizes in both dimensions are multiples of a certain function of t. Then what about tori of other sizes? Our main result will show that when a torus is sufficiently large in both dimensions, its t-interleaving number exceeds the lower bound $|S_t|$ by at most one.

A more detailed description of our results is as follows:

- We prove that an $l_1 \times l_2$ torus can be perfectly t-interleaved if and only if the following condition is satisfied: when t is odd (respectively, even), both l_1 and l_2 are multiples of $\frac{t^2+1}{2}$ (respectively, t). We reveal the close relationship between perfect t-interleaving and perfect sphere-packing, and present the *complete* set of perfect sphere-packing constructions. Based on that, we obtain a set of efficient perfect t-interleaving constructions, which includes the lattice interleaver scheme presented in [8] as a special case.
- We prove that for any torus that is sufficiently large in both dimensions, its t-interleaving number is either $|S_t|$ or $|S_t|+1$. In other words, any large torus needs at most one more color than a perfect t-interleaving would use if it were possible. More specifically, there exist integer pairs (θ_1, θ_2) such that whenever $l_1 \geq \theta_1$ and $l_2 \geq \theta_2$, the t-interleaving number of an $l_1 \times l_2$ torus is at most $|S_t|+1$. Here θ_1 and θ_2 depend on t, and naturally there is a tradeoff between them: If θ_1 takes a greater value, then the minimum value that θ_2 can take decreases or remains the same, and vice versa. We find a sequence of valid values for θ_1 and θ_2, which are shown in Theorems 4.7 and 4.8. We present optimal t-interleaving constructions for tori whose sizes exceed the found pairs (θ_1, θ_2), and we comment that those constructions, as a general interleaving method, can also be used to optimally t-interleave tori of many other sizes.
- We study upper bounds for t-interleaving numbers, and show that every $l_1 \times l_2$ torus' t-interleaving number is $|S_t| + O(t^2)$. That upper bound is tight, even if $l_1 \to +\infty$ or $l_2 \to +\infty$, meaning that having just one large dimension is not enough to guarantee any significant reduction in the t-interleaving number. When both l_1 and l_2 are of the order $\Omega(t^2)$, the t-interleaving number of an $l_1 \times l_2$ torus is $|S_t| + O(t)$.

The results can be illustrated qualitatively as Figure 1.2, but the figure is not quantitative: The coordinates of points and the shape of the curve are not exact.

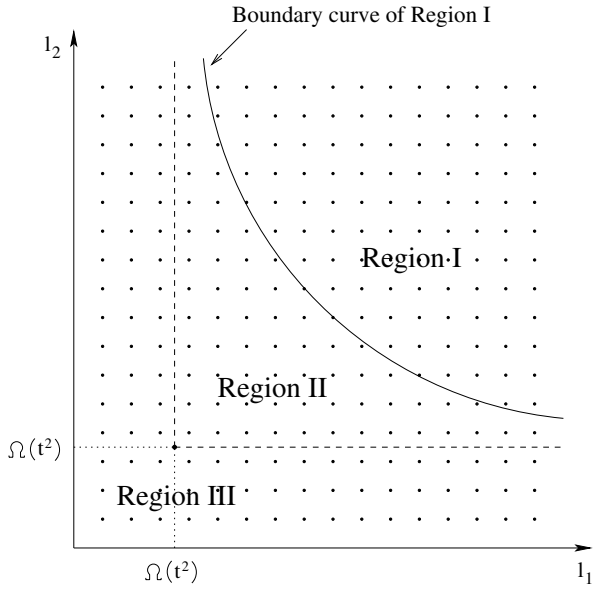


FIG. 1.2. *A qualitative illustration of the t-interleaving numbers.*

Figure 1.2 shows for any given t how the $l_1 \times l_2$ tori can be divided into different classes based on their t-interleaving numbers.

The uniform lattice of dots in Figure 1.2 represents the sizes of all the tori that can be perfectly t-interleaved. The region labeled as Region I consists of all the integer pairs (θ_1, θ_2). The boundary curve of Region I is nonincreasing and symmetric with respect to the line $l_2 = l_1$. We know the exact t-interleaving number of every torus in this region: $|S_t|$ if it is one of the lattice dots, and $|S_t| + 1$ otherwise. The most important contribution of this paper is to prove the existence of Region I and present the corresponding optimal interleaving constructions. Region II is the region where $l_1 = \Omega(t^2)$ and $l_2 = \Omega(t^2)$, in which the tori's t-interleaving numbers are upper-bounded by $|S_t| + O(t)$. Region III includes every torus, where the t-interleaving number is upper-bounded by $|S_t| + O(t^2)$. That upper bound for Region III is tight, even if l_1 or l_2 approaches $+\infty$. Thus, increasing a torus' size in only one dimension does not help reduce the t-interleaving number very effectively in general.

The rest of the paper is organized as follows. In section 2, we show the necessary and sufficient conditions for tori that can be perfectly t-interleaved, and present perfect t-interleaving constructions based on perfect sphere packing. In section 3, we present a t-interleaving method, with which we can t-interleave large tori using just one more than the optimal number of colors. In section 4, we improve upon the t-interleaving method shown in section 3, and present optimal t-interleaving constructions for tori whose sizes are large in both dimensions. As a parallel result, the existence of Region I is proved. In section 5, we prove some general bounds for the t-interleaving numbers. In section 6, we conclude this paper.

2. Perfect *t*-interleaving. In this section, we show the close relationship between *perfect t-interleaving* and *perfect sphere-packing*, and use it to prove the necessary and sufficient condition for tori to have perfect t-interleaving. We present the complete set of perfect sphere-packing constructions. Based on them, we derive

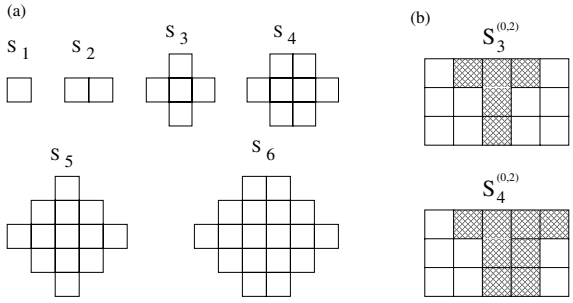


FIG. 2.1. *Examples of the sphere S_t.*

efficient perfect t-interleaving constructions.

2.1. Perfect t-interleaving and sphere-packing. The following is the definition of Lee distance in tori.

DEFINITION 2.1. *The* Lee distance *between two vertices in a torus is the number of edges in the shortest path connecting those two vertices. For two vertices in an $l_1 \times l_2$ torus G, (a_1, b_1) and (a_2, b_2), the* Lee distance *between them is denoted by $d((a_1, b_1), (a_2, b_2))$. Note that therefore, $d((a_1, b_1), (a_2, b_2)) = \min\{(a_1 - a_2) \bmod l_1, (a_2 - a_1) \bmod l_1\} + \min\{(b_1 - b_2) \bmod l_2, (b_2 - b_1) \bmod l_2\}$. Occasionally, in order to emphasize that the two vertices are in G, we also denote it by $d_G((a_1, b_1), (a_2, b_2))$.*

Clearly, an interleaving on a torus is a t-interleaving if and only if the Lee distance between any two vertices of the same color is at least t.

The following is a more detailed definition of spheres than that in the Introduction.

DEFINITION 2.2. *Let G be an $l_1 \times l_2$ torus, where $l_1 \geq 2\lfloor \frac{t-1}{2} \rfloor + 1$ and $l_2 \geq t$, and let (a, b) be a vertex in G. When t is odd, the* sphere centered at (a, b), *$S_t^{(a,b)}$ is defined to be the subgraph induced by all those vertices whose Lee distance to (a, b) is less than or equal to $\frac{t-1}{2}$. When t is even, the* sphere left-centered at (a, b), *$S_t^{(a,b)}$ is defined to be the subgraph induced by all those vertices whose Lee distance to either (a, b) or $(a, (b + 1) \bmod l_2)$ is less than or equal to $\frac{t}{2} - 1$. (a, b) is called the* center *of $S_t^{(a,b)}$ if t is odd, or the* left-center *of $S_t^{(a,b)}$ if t is even. If we do not care where the sphere is centered or left-centered, then the sphere is simply denoted by S_t. The number of vertices in the sphere is denoted by $|S_t|$.*

Example 2.1. Figure 2.1(a) shows the spheres S_1 to S_6. Figure 2.1(b) shows two spheres, $S_3^{(0,2)}$ and $S_4^{(0,2)}$, in a 3×5 torus.

For any $l_1 \times l_2$ torus, where $l_1 \geq 2\lfloor \frac{t-1}{2} \rfloor + 1$ and $l_2 \geq t$, its t-interleaving number is at least $|S_t|$. We call $|S_t|$ the *sphere-packing lower bound.* The relationship between this bound and sphere-packing will become clearer soon.

DEFINITION 2.3. *A torus G is said to have a* perfect packing of spheres *S_t if spheres S_t are packed in G in such a way that every vertex of G lies in exactly one of the spheres.*

LEMMA 2.4. (1) *Let t be odd. An interleaving on an $l_1 \times l_2$ torus (where $l_1 \geq t$ and $l_2 \geq t$) is a t-interleaving if and only if for any two vertices (a_1, b_1) and (a_2, b_2) of the same color, the two spheres centered at them, $S_t^{(a_1,b_1)}$ and $S_t^{(a_2,b_2)}$, do not share any common vertex.*

(2) *Let t be even. An interleaving on an $l_1 \times l_2$ torus (where $l_1 \geq t - 1$ and $l_2 \geq t$) is a t-interleaving if and only if for any two vertices (a_1, b_1) and (a_2, b_2) of the*

same color, the two spheres left-centered there, $S_t^{(a_1,b_1)}$ and $S_t^{(a_2,b_2)}$, do not share any common vertex and, what is more, $b_1 \neq b_2$ or $(a_1 - a_2) \neq \pm(t-1) \bmod l_1$.

Proof. (1) Let t be odd. Both $S_t^{(a_1,b_1)}$ and $S_t^{(a_2,b_2)}$ are classic spheres with radius $\frac{t-1}{2}$. If the interleaving is a t-interleaving, then the Lee distance between (a_1, b_1) and (a_2, b_2) is at least $t = 2 \cdot \frac{t-1}{2} + 1$, so $S_t^{(a_1,b_1)}$ and $S_t^{(a_2,b_2)}$ must have no intersection. The converse is also true.

(2) Let t be even. We consider two cases: $b_1 = b_2$ and $b_1 \neq b_2$.

First consider the case $b_1 = b_2$. In this case, $S_t^{(a_1,b_1)}$ and $S_t^{(a_2,b_2)}$ have no intersection if and only if $d((a_1, b_1), (a_2, b_2)) \geq 2 \cdot (\frac{t}{2} - 1) + 1 = t - 1$. Further, $d((a_1, b_1), (a_2, b_2)) = t - 1$ if and only if $(a_1 - a_2) \equiv \pm(t-1) \bmod l_1$. So the Lee distance between (a_1, b_1) and (a_2, b_2) is at least t if and only if $S_t^{(a_1,b_1)}$ and $S_t^{(a_2,b_2)}$ have no intersection and $(a_1 - a_2) \neq \pm(t-1) \bmod l_1$, which is the conclusion we want.

Now consider the case $b_1 \neq b_2$. In this case, the Lee distance between (a_1, b_1) and (a_2, b_2) is at least $t \Leftrightarrow$ both the Lee distance between $(a_1, (b_1 + 1) \bmod l_2)$ and (a_2, b_2) and the Lee distance between $(a_2, (b_2 + 1) \bmod l_2)$ and (a_1, b_1) are at least $t - 1 \Leftrightarrow S_{t-1}^{(a_1,(b_1+1) \bmod l_2)}$ does not intersect $S_{t-1}^{(a_2,b_2)}$ and $S_{t-1}^{(a_2,(b_2+1) \bmod l_2)}$ does not intersect $S_{t-1}^{(a_1,b_1)} \Leftrightarrow S_t^{(a_1,b_1)}$ and $S_t^{(a_2,b_2)}$ have no intersection. Note that $S_t^{(a_1,b_1)}$ is the union of $S_{t-1}^{(a_1,b_1)}$ and $S_{t-1}^{(a_1,(b_1+1) \bmod l_2)}$, and $S_t^{(a_2,b_2)}$ is the union of $S_{t-1}^{(a_2,b_2)}$ and $S_{t-1}^{(a_2,(b_2+1) \bmod l_2)}$. So we get the conclusion. ∎

THEOREM 2.5. *For an $l_1 \times l_2$ torus, where $l_1 \geq t$ and $l_2 \geq t$, if an interleaving on it is a perfect t-interleaving, then for every color the spheres S_t centered or left-centered at the vertices of that color form a perfect sphere-packing in the torus. The converse is also true when $t \neq 2$.*

Proof. Let us say that the torus is interleaved. We used I to denote the set of distinct colors used by the interleaving. For any color $i \in I$, we use N_i to denote the number of vertices of color i.

Let us first prove one direction. Assume that the interleaving is a perfect t-interleaving. Then $|I| = |S_t|$. By Lemma 2.4, for any $i \in I$, the spheres S_t centered or left-centered at vertices of color i do not overlap. By counting the number of vertices in the torus and in each sphere S_t, we get $N_i \leq \frac{l_1 l_2}{|S_t|}$ for any $i \in I$. Since $\sum_{i \in I} N_i = l_1 l_2$, we get $N_i = \frac{l_1 l_2}{|S_t|}$ for any $i \in I$. So for any color $i \in I$, the spheres S_t centered or left-centered at the vertices of color i form a perfect sphere-packing in the torus.

Now let us prove the converse direction. Assume $t \neq 2$. Also assume for every color that the spheres S_t centered or left-centered at the vertices of that color form a perfect sphere packing in the torus. Then $N_i = \frac{l_1 l_2}{|S_t|}$ for any $i \in I$. Since $\sum_{i \in I} N_i = l_1 l_2$, we get $|I| = |S_t|$. What is left to prove is that the interleaving is a t-interleaving. By Lemma 2.4, the interleaving can fail to be a t-interleaving only if the following situation becomes true: "t is even, and there exist two vertices (a_1, b_1) and (a_2, b_2) of the same color such that $b_1 = b_2$ and $a_1 - a_2 \equiv t - 1 \bmod l_1$." We will now show that such a situation cannot happen.

Assume that situation happens. Then it is straightforward to verify that the four vertices $(a_1 - (\frac{t}{2} - 1) \bmod l_1, b_1)$, $(a_2 + (\frac{t}{2} - 1) \bmod l_1, b_1)$, $(a_1 - (\frac{t}{2} - 2) \bmod l_1, b_1 - 1 \bmod l_2)$, and $(a_2 + (\frac{t}{2} - 2) \bmod l_1, b_1 - 1 \bmod l_2)$ are contained in either $S_t^{(a_1,b_1)}$ or $S_t^{(a_2,b_2)}$, while the two vertices $(a_1 - (\frac{t}{2} - 1) \bmod l_1, b_1 - 1 \bmod l_2)$ and $(a_2 + (\frac{t}{2} - 1) \bmod l_1, b_1 - 1 \bmod l_2)$ are neither contained in $S_t^{(a_1,b_1)}$ nor in $S_t^{(a_2,b_2)}$. The two vertices $(a_1 - (\frac{t}{2} - 1) \bmod l_1, b_1 - 1 \bmod l_2)$ and $(a_2 + (\frac{t}{2} - 1) \bmod l_1, b_1 - 1 \bmod l_2)$ cannot

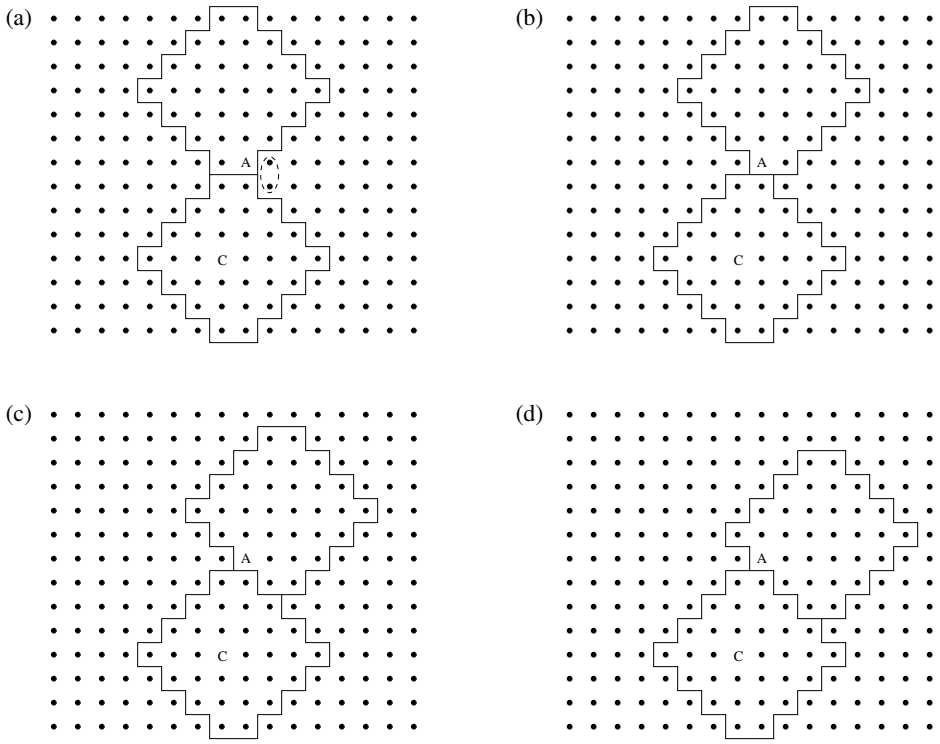


FIG. 2.2. *Relative positions of spheres and vertices.*

both be contained in spheres S_t that are left-centered at vertices having the color of (a_1, b_1) and (a_2, b_2), because they are vertically adjacent, and the vertices directly above them, below them, and to the right of them are all contained in two spheres that do not contain them, due to the shape of the sphere, as seen in Figure 2.2(a). This contradicts that fact that all the spheres S_t, left-centered at the vertices having the same color as (a_1, b_1), form a perfect sphere-packing in the torus. So the assumed situation cannot happen. Summarizing the above results, we see that the interleaving must be a perfect t-interleaving. □

THEOREM 2.6. *For an $l_1 \times l_2$ torus, where $l_1 \geq t$ and $l_2 \geq t$, if it can be perfectly t-interleaved, then the spheres S_t can be perfectly packed in it. The converse is also true when $t \neq 2$.*

Proof. Let G be an $l_1 \times l_2$ torus. For any t, Theorem 2.5 has shown that if G can be perfectly t-interleaved, then the spheres S_t can be perfectly packed in it. Now we prove the other direction. Assume $t \neq 2$, and that the spheres S_t can be perfectly packed in G. Let (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$ be a set of vertices such that the spheres S_t centered or left-centered at them form a perfect packing in G. The proof of Theorem 2.5 has essentially shown that for any i and j $(i \neq j)$, the Lee distance between (x_i, y_i) and (x_j, y_j) is at least t. Now we can interleave G is this way: Color each sphere S_t with $|S_t|$ distinct colors in the same way, so that every color is used in exactly the same position in every sphere. Clearly, for any two colors a and b, the two sets of vertices colored by a and b are translates of each other in the torus, and therefore the Lee distance between any two vertices of the same color is at least t. Thus G has a perfect t-interleaving. □

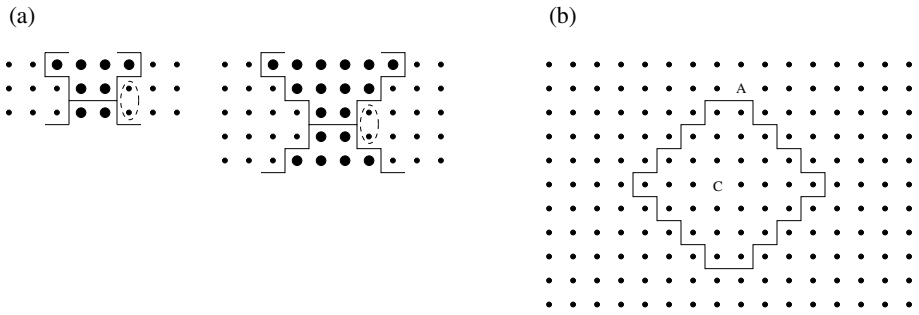


FIG. 2.3. *A sphere in a torus.*

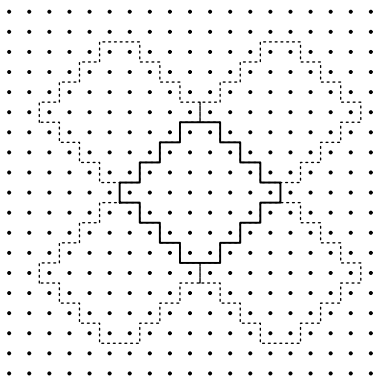


FIG. 2.4. *Four positions that a neighbor sphere might be in.*

2.2. Perfect t-interleaving and its construction. The following lemma is an important property of perfect sphere-packing. It will help us derive the necessary and sufficient condition for perfect t-interleaving.

LEMMA 2.7. *Let t be even and $t \geq 4$. When spheres S_t are perfectly packed in an $l_1 \times l_2$ torus, there exists an integer $a \in \{+1, -1\}$ such that if there is a sphere left-centered at the vertex (x, y), then there are two spheres respectively left-centered at $((x - \frac{t}{2}) \bmod l_1, (y - a \cdot \frac{t}{2}) \bmod l_2)$ and $((x + \frac{t}{2}) \bmod l_1, (y + a \cdot \frac{t}{2}) \bmod l_2)$.*

Proof. Assume that spheres S_t are perfectly packed in an $l_1 \times l_2$ torus, where $t \geq 4$ and t is even. First we observe that $l_1 \geq t$: Since a sphere S_t spans $t - 1$ rows when t is even, l_1 must be at least $t - 1$, but l_1 cannot be exactly $t - 1$ either, because then, as shown in Figure 2.3(a), the sphere will just touch itself, and it is clearly impossible to cover the two adjacent positions marked by dashed circles in Figure 2.3(a) using nonoverlapping spheres. Thus $l_1 \geq t$.

Clearly, one of the following two cases must be true, concerning the presence or absence of any of the four possible neighbor spheres shown in Figure 2.4:

- Case 1. Whenever there is a sphere left-centered at a vertex (x, y), there are four spheres respectively left-centered at the four vertices $((x - \frac{t}{2}) \bmod l_1$, $(y - \frac{t}{2}) \bmod l_2)$, $((x - \frac{t}{2}) \bmod l_1, (y + \frac{t}{2}) \bmod l_2)$, $((x + \frac{t}{2}) \bmod l_1, (y - \frac{t}{2}) \bmod l_2)$, and $((x + \frac{t}{2}) \bmod l_1, (y + \frac{t}{2}) \bmod l_2)$.
- Case 2. There exists a sphere left-centered at a vertex (x_0, y_0) such that there is no sphere left-centered at at least one of the following four vertices:

$((x_0 - \frac{t}{2}) \bmod l_1, (y_0 - \frac{t}{2}) \bmod l_2)$, $((x_0 - \frac{t}{2}) \bmod l_1, (y_0 + \frac{t}{2}) \bmod l_2)$, $((x_0 + \frac{t}{2}) \bmod l_1, (y_0 - \frac{t}{2}) \bmod l_2)$, and $((x_0 + \frac{t}{2}) \bmod l_1, (y_0 + \frac{t}{2}) \bmod l_2)$.

If Case 1 is true, then the conclusion of this lemma obviously holds. From now on, let us assume that Case 2 is true. Without loss of generality, we assume that there is one sphere left-centered at (x_0, y_0), but there is no sphere left-centered at $((x_0 - \frac{t}{2}) \bmod l_1, (y_0 + \frac{t}{2}) \bmod l_2)$.

Since $l_1 \geq t$, the vertex $((x_0 - \frac{t}{2}) \bmod l_1, (y_0 + 1) \bmod l_2)$, which we shall call vertex A, is not contained in the sphere left-centered at (x_0, y_0). An example is shown in Figure 2.3(b), where the sphere in consideration is an S_8, whose left-center (x_0, y_0) is labeled by C. The vertex A is contained in one of the perfectly packed spheres, which we shall call sphere B. The relative position of vertex A in sphere B can only be one of the following two possibilities:

- Possibility 1. The vertex A is the right-most vertex in the bottom row of the sphere B, as in Figure 2.2(a).
- Possibility 2. The vertex A is in the lower-left diagonal of the border of the sphere B, as in Figure 2.2(b), (c), and (d). Note that it cannot be the left-most vertex of the sphere B, because that is the location where we are assuming there is not a sphere.

Possibility 1, however, as we saw in Figure 2.2(a), is impossible. So we are left with Possibility 2. In the following proof we use the example of $t = 8$ for illustration, and assume that the relative position of the sphere B is as shown in Figure 2.2(b). We comment that when t takes other values or when the sphere B takes one of the three other positions, it is easy to see that the argument still holds.

Let the sphere left-centered at (x_0, y_0) be the sphere denoted by L_1 in Figure 2.5, and let sphere B be the sphere denoted by R_1 in Figure 2.5. We immediately see that the vertex denoted by E must be the right-most vertex of a sphere, so the sphere containing the vertex E must be the sphere denoted by L_2. Then we immediately see that the vertex denoted by F must be the right-most vertex in the bottom row of a sphere, so the sphere containing the vertex F must be the sphere denoted by R_2. With the same method we can fix the positions of a series of spheres L_1, L_2, L_3, L_4, ... and a series of spheres R_1, R_2, R_3, R_4, ... Since the torus is finite, we will get a series of spheres L_1, L_2, L_3, L_4, ..., L_n such that the relative position of L_n to L_1 is the same as the relative position of L_1 to L_2 (see Figure 2.5 for an illustration). Such a series of spheres forms a cycle in the torus. Since the spheres are perfectly packed in the torus, no two spheres in this cycle overlap. Similarly, the spheres R_1, R_2, ..., R_n also form a cycle in the torus. Note that we do not make any assumption about whether these two cycles overlap or not.

If those two cycles do not already contain all the spheres in the torus, then there must be some spheres outside the two cycles that are directly attached to the lower-left side of the cycle formed by $L_1, L_2, \ldots, L_n$. This is due to the very regular way the cycle is formed and the resulting shape of the cycle, which is invariant to horizontal and vertical shifts. Let D_1 be a sphere directly attached to the cycle formed by $L_1, L_2, \ldots, L_n$, as shown in Figure 2.5. Note that we do not care about the exact position of D_1, as long as it is directly attached to the lower-left side of the cycle. Then the vertex I immediately determines that the sphere containing it must be D_2, and similarly the vertex J determines the position of the sphere D_3, and so on. So we will get a series of spheres $D_1, D_2, D_3, \ldots, D_n$, which will again form a cycle. It is easy to see that this cycle does not overlap the previous two cycles. Continuing in this way, we can keep finding cycles until they cover the torus.

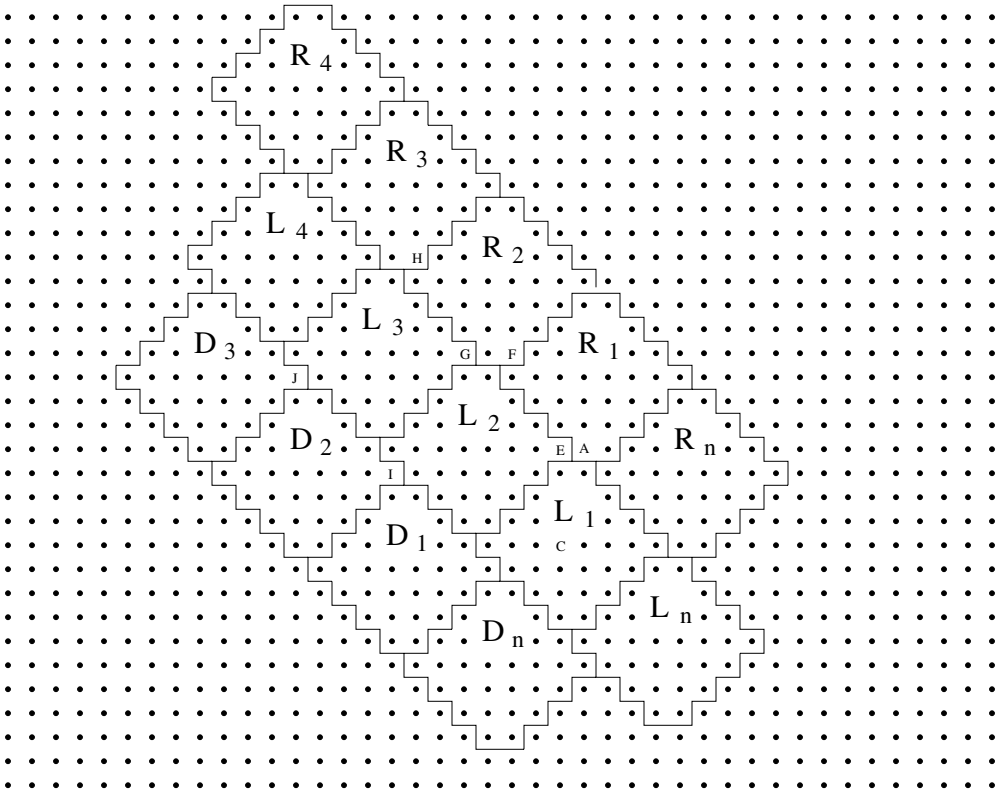


FIG. 2.5. *The packing of spheres in a torus.*

We can easily see that in each of the cycles here, if there is a sphere left-centered at a vertex (x, y), then there are two spheres respectively left-centered at $((x - \frac{t}{2}) \bmod l_1, (y - \frac{t}{2}) \bmod l_2)$ and $((x + \frac{t}{2}) \bmod l_1, (y + \frac{t}{2}) \bmod l_2)$. In the other instances of Case 2, we either find the same pattern of cycles or else we find the flipped pattern, in which whenever there is a sphere left-centered at a vertex (x, y), there are two spheres respectively left-centered at $((x - \frac{t}{2}) \bmod l_1, (y + \frac{t}{2}) \bmod l_2)$ and $((x + \frac{t}{2}) \bmod l_1, (y - \frac{t}{2}) \bmod l_2)$. The parameter a in the statement of the lemma represents which of the two patterns is being used. □

DEFINITION 2.8. *Let t be an even positive integer, let a be either $+1$ or -1, and let G be an $l_1 \times l_2$ torus. Let (x, y) be an arbitrary vertex in G. We define the* cycle containing (x, y) (*corresponding to the parameter a*) *to be the set of spheres S_t that are respectively left-centered at the vertices (x, y), $((x + \frac{t}{2}) \bmod l_1, (y + a \cdot \frac{t}{2}) \bmod l_2)$, $((x + 2 \cdot \frac{t}{2}) \bmod l_1, (y + 2a \cdot \frac{t}{2}) \bmod l_2)$, $((x + 3 \cdot \frac{t}{2}) \bmod l_1, (y + 3a \cdot \frac{t}{2}) \bmod l_2)$, ...*

The proof of the following lemma is omitted due to its simplicity.

LEMMA 2.9. *Let t be an even positive integer, let a be either $+1$ or -1, and let G be an $l_1 \times l_2$ torus. For any vertex (x, y) in G, the cycle containing it* (*corresponding to the parameter a*) *consists of $\frac{lcm(l_1, l_2, \frac{t}{2})}{\frac{t}{2}}$ distinct spheres S_t.*

The following theorem shows the necessary and sufficient condition for tori that can be perfectly t-interleaved.

THEOREM 2.10. *Let G be an $l_1 \times l_2$ torus, where $l_1 \geq t$ and $l_2 \geq t$. If t is odd, then G can be perfectly t-interleaved if and only if both l_1 and l_2 are multiples of $\frac{t^2+1}{2}$.*

If t is even, then G can be perfectly t-interleaved if and only if both l_1 and l_2 are multiples of t.

Proof. We consider the following three cases separately.

Case 1: $t = 2$. In this case, 2-interleaving is equivalent to vertex coloring, so the 2-interleaving number of G equals G's chromatic number $\chi(G)$. Let R_1 and R_2 each be a graph consisting of a single cycle, having l_1 and l_2 vertices, respectively. Then G is the Cartesian product of those two cycles, namely, $G = R_1 \otimes R_2$. It is well known [23] that for any two graphs H_1 and H_2, $\chi(H_1 \otimes H_2) = \max\{\chi(H_1), \chi(H_2)\}$. Since $l_1 \geq t = 2$ (respectively, $l_2 \geq t = 2$), we get that $\chi(R_1) \geq 2$ (respectively, $\chi(R_2) \geq 2$), and $\chi(R_1) = 2$ (respectively, $\chi(R_2) = 2$) if and only if l_1 (respectively, l_2) is a multiple of 2. So $\chi(G) = 2$ if and only if both l_1 and l_2 are multiples of 2. Since $|S_2| = 2$, we get the conclusion in this lemma.

Case 2: t is even but $t \neq 2$. First, we prove one direction. Assume that G can be perfectly t-interleaved. We will show that both l_1 and l_2 are multiples of t. Let i be a color used by a perfect t-interleaving on G. Then by Theorem 2.5, the spheres S_t left-centered at the vertices of color i form a perfect sphere-packing in G. By Lemma 2.7, there exists an integer $a \in \{+1, -1\}$ such that for any cycle containing a vertex of color i (corresponding to the parameter a), the spheres S_t in the cycle are all left-centered at vertices of color i, and therefore they do not overlap. By Lemma 2.9, the cycle containing a vertex of color i consists of $\frac{lcm(l_1, l_2, \frac{t}{2})}{\frac{t}{2}}$ distinct spheres S_t. So such a cycle consists of

$$\frac{lcm(l_1, l_2, \frac{t}{2})}{\frac{t}{2}} \cdot |S_t| = \frac{lcm(l_1, l_2, \frac{t}{2})}{\frac{t}{2}} \cdot \frac{t^2}{2} = lcm\left(l_1, l_2, \frac{t}{2}\right) \cdot t$$

vertices. Let (x_1, y_1) and (x_2, y_2) be any two vertices of color i. We can see that for the cycle containing (x_1, y_1) and the cycle containing (x_2, y_2), either they do not overlap or they are the same cycle. Therefore, the vertices in G can be partitioned into several such cycles, so $l_1 \cdot l_2$ is a multiple of $lcm(l_1, l_2, \frac{t}{2}) \cdot t$. Since $lcm(l_1, l_2, \frac{t}{2})$ is a multiple of l_1, l_2 must be a multiple of t. Similarly, l_1 must be a multiple of t, too. So if G can be perfectly t-interleaved, then both l_1 and l_2 are multiples of t.

Now we prove the other direction. Assume both l_1 and l_2 are multiples of t. Let W be such a set of vertices in G: $W = \{(x, y) | x \equiv 0 \bmod \frac{t}{2}, y \equiv 0 \bmod \frac{t}{2}, x + y \equiv 0 \bmod t\}$. It is easy to verify that the Lee distance between any two vertices in W is at least t. Now for $i = 0, 1, \ldots, \frac{t}{2} - 1$ and for $j = 0, 1, \ldots, t - 1$, define $W^{i,j}$ to be $W^{i,j} = \{((x + i) \bmod l_1, (y + j) \bmod l_2) | (x, y) \in W\}$. Clearly those $\frac{t}{2} \cdot t = |S_t|$ sets, $W^{0,0}$, $W^{0,1}$, $\ldots$, $W^{\frac{t}{2}-1, t-1}$, are a partition of the vertices in G. For each $W^{i,j}$, we color the vertices in it with the same distinct color. Clearly such an interleaving is a perfect t-interleaving. So if both l_1 and l_2 are multiples of t, then G can be perfectly t-interleaved.

Case 3: t is odd. First, we prove one direction. Assume that both l_1 and l_2 are multiples of $\frac{t^2+1}{2}$. Golomb and Welch have shown in [11] that a $\frac{t^2+1}{2} \times \frac{t^2+1}{2}$ torus can be perfectly packed by the spheres S_t for odd t. Therefore, G can also be perfectly packed by S_t because a torus has a toroidal topology and G can be folded onto itself into an $\frac{t^2+1}{2} \times \frac{t^2+1}{2}$ torus. Let C be a set of vertices in G such that the spheres S_t centered at the vertices in C form a perfect sphere-packing. Then the Lee distance between any two vertices in C is at least t. We call a set of vertices D a *translate* of C when the following condition is satisfied: "There exist integers a and b such that a vertex $(x, y) \in C$ if and only if $((x + a) \bmod l_1, (y + b) \bmod l_2) \in D$." C has $|S_t|$

different translates in total (including C itself), and those translates partition the vertices of G. For each translate, we color its vertices with one distinct color, and we get a perfect t-interleaving. So if both l_1 and l_2 are multiples of $\frac{t^2+1}{2}$, then G can be perfectly t-interleaved.

Now we prove the other direction. Assume that G can be perfectly t-interleaved. Let i be a color used by a perfect t-interleaving on G. Then by Theorem 2.5, the spheres S_t centered at the vertices of color i form a perfect sphere-packing in G. Golomb and Welch presented in [11] a way to perfectly pack spheres S_t in a torus when t is odd, which can be described as "either of the following two conditions is true: (1) Whenever there is a sphere S_t centered at a vertex (x, y), there are two spheres respectively centered at $((x+\frac{t+1}{2}) \bmod l_1, (y+\frac{t-1}{2}) \bmod l_2)$ and $((x-\frac{t-1}{2}) \bmod l_1, (y+\frac{t+1}{2}) \bmod l_2)$; (2) whenever there is a sphere S_t centered at a vertex (x, y), there are two spheres respectively centered at $((x+\frac{t-1}{2}) \bmod l_1, (y+\frac{t+1}{2}) \bmod l_2)$ and $((x-\frac{t+1}{2}) \bmod l_1, (y+\frac{t-1}{2}) \bmod l_2)$". It is easy to see that that way of packing is in fact the only way to perfectly pack S_t for odd t, whose feasibility requires both l_1 and l_2 to be multiples of $\frac{t^2+1}{2}$. Thus if G can be perfectly t-interleaved, then both l_1 and l_2 are multiples of $\frac{t^2+1}{2}$. □

Below we present the complete set of perfect sphere-packing constructions. But first let us explain a few concepts. Let G be an $l_1 \times l_2$ torus that is perfectly packed by spheres S_t, so there are $\frac{l_1 l_2}{|S_t|}$ such spheres. Define $e = \frac{l_1 l_2}{|S_t|}$, and let us say that those spheres are centered (or left-centered) at the vertices (x_1, y_1), (x_2, y_2), ..., (x_e, y_e). By *vertically* (respectively, *horizontally*) *shifting the spheres in* G, we mean to select some integer s, and get a new set of perfectly packed spheres that are centered (or left-centered) at $(x_1 + s \bmod l_1, y_1)$, $(x_2 + s \bmod l_1, y_2)$, ..., $(x_e + s \bmod l_1, y_e)$ (respectively, at $(x_1, y_1 + s \bmod l_2)$, $(x_2, y_2 + s \bmod l_2)$, ..., $(x_e, y_e + s \bmod l_2)$). By *vertically reversing the spheres in* G, we mean to get a new set of perfectly packed spheres that are centered (or left-centered) at $(-x_1 \bmod l_1, y_1)$, $(-x_2 \bmod l_1, y_2)$, ..., $(-x_e \bmod l_1, y_e)$. After such a shift or reverse operation, technically speaking, the way the spheres are perfectly packed in G is changed. However, the pattern of the sphere-packing essentially remains the same.

Construction 2.1. The complete set of perfect sphere-packing constructions.

Input: A positive integer t. An $l_1 \times l_2$ torus G, where (1) both l_1 and l_2 are multiples of t if t is even and $t \neq 2$, (2) l_2 is even if $t = 2$, and (3) both l_1 and l_2 are multiples of $\frac{t^2+1}{2}$ if t is odd.

Output: A perfect packing of the spheres S_t in G.

Construction:

1. If t is even and $t \neq 2$, then do the following:
   - Let $A_1, A_2, \ldots, A_{gcd(\frac{l_1}{t}, \frac{l_2}{t})-1}$ be $gcd(\frac{l_1}{t}, \frac{l_2}{t}) - 1$ integers, where A_i can be any integer in the set $\{0, 1, \ldots, \frac{t}{2} - 1\}$ for $i = 1, 2, \ldots, gcd(\frac{l_1}{t}, \frac{l_2}{t}) - 1$.
   - Find the $gcd(\frac{l_1}{t}, \frac{l_2}{t})$ cycles in G respectively containing the vertices $(0, 0)$, $(A_1, t+A_1)$, $(A_1+A_2, 2t+A_1+A_2)$, ..., and $(\sum_{i=1}^{gcd(\frac{l_1}{t}, \frac{l_2}{t})-1} A_i, \sum_{i=1}^{gcd(\frac{l_1}{t}, \frac{l_2}{t})-1} (t + A_i))$. The spheres S_t in those $gcd(\frac{l_1}{t}, \frac{l_2}{t})$ cycles form a perfect sphere-packing in the torus.
2. If $t = 2$, the do the following:
   - The $l_1 \times l_2$ torus G has l_1 rows, each of which can be seen as a ring of l_2 vertices. When $t = 2$, the sphere S_t simply consists of two horizontally adjacent vertices. Split each row of G into $\frac{l_2}{2}$ spheres in any way. The resulting $\frac{l_1 l_2}{2}$ spheres form a perfect sphere-packing in the torus.

3. If t is odd, then do the following:
   - Find a set of $\frac{l_1 l_2}{|S_t|}$ spheres S_t such that each of the spheres is centered at a vertex $(i \cdot \frac{t+1}{2} + j \cdot \frac{1-t}{2} \bmod l_1, i \cdot \frac{t-1}{2} + j \cdot \frac{t+1}{2} \bmod l_2)$ for some integers i and j. Those spheres form a perfect sphere-packing in the torus.

4. Horizontally shift, vertically shift, and/or vertically reverse the spheres in G in any way.

THEOREM 2.11. *Construction* 2.1 *is the* complete *set of perfect sphere-packing constructions.*

Proof. We consider the following three cases. For each case, we need to prove two things: First, the *input* part of Construction 2.1 sets the necessary and sufficient condition for a torus to have a perfect sphere-packing; second, the *construction* part of Construction 2.1 generates perfect sphere-packing correctly, and every perfect sphere-packing that exists is a possible output of it.

Case 1: t is even and $t \neq 2$. Lemma 2.7 and its proof have shown that when spheres are perfectly packed in a torus, those spheres can be partitioned into cycles. By observing the shape of the border of a cycle, we see that two adjacent cycles can freely slide along each other's border, and there are $\frac{t}{2}$ possible relative positions for two adjacent cycles. In Construction 2.1, the $\frac{t}{2}$ possible relative positions are determined by A_i, a variable that can take $\frac{t}{2}$ possible values. Now it is easy to see that step 1 of Construction 2.1 provides a perfect sphere-packing (which takes one of many possible forms, depending on the value of A_i), and step 4 changes the positions of the spheres to furthermore cover all the possible cases of perfect sphere-packing.

Case 2: $t = 2$. We skip the proof for this case due to its simplicity.

Case 3: t is odd. In this case, Construction 2.1 reproduces the sphere-packing method presented in [11], which is commonly known as the unique way to pack spheres for odd t (see the final paragraph of the proof of Theorem 2.10 for a more detailed introduction). □

Now we present perfect t-interleaving constructions that are based on perfect sphere-packing.

Construction 2.2. Perfect t-interleaving constructions

Input: A positive integer t. An $l_1 \times l_2$ torus G, where both l_1 and l_2 are multiples of t if t is even, and both l_1 and l_2 are multiples of $\frac{t^2+1}{2}$ if t is odd.

Output: A perfect t-interleaving on G.

Construction:

1 If $t \neq 2$, then do the following:
   - Use Construction 2.1 to get a perfect sphere-packing in G. Color each sphere in the same way, using $|S_t|$ distinct colors, so that each color is used exactly once in each sphere.

2 If $t = 2$, then do the following:
   - For every vertex (i, j) of G, if $i + j$ is even, color it with color 0; otherwise color it with color 1.

The following example illustrates how to use Construction 2.1 to obtain perfect sphere-packing, and how to use Construction 2.2 to obtain perfect t-interleaving.

Example 2.2. Let $t = 4$, and let G be a 12×24 torus. First, we use Construction 2.1 to find a perfect sphere-packing in G. Since t is even, step 1 of Construction 2.1 is executed. We choose $A_1, A_2, \ldots, A_{gcd(\frac{l_1}{t}, \frac{l_2}{t})-1}$ to be $A_1 = 0$, $A_2 = 1$. Note that here $gcd(\frac{l_1}{t}, \frac{l_2}{t}) - 1 = 2$. Then the $gcd(\frac{l_1}{t}, \frac{l_2}{t}) = 3$ cycles in G are as shown in Figure 2.6(a), which are three sets of spheres S_t respectively of three different background shades. The spheres in those three cycles form a perfect packing in G.

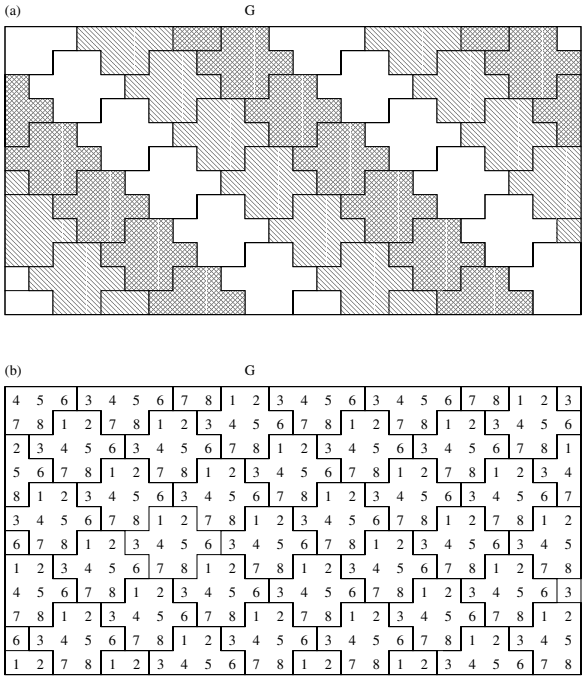


FIG. 2.6. *Example of perfect sphere-packing using Construction* 2.1*, and perfect t-interleaving using Construction* 2.2.

Next, we use Construction 2.2 to perfectly t-interleave G. Let the perfect sphere-packing remain as it is, and color all the spheres with the same pattern, using $|S_t| = 8$ distinct colors. The resulting perfect t-interleaving on G is shown in Figure 2.6(b).

We comment that Construction 2.2 provides the *complete* set of perfect t-interleaving constructions that have the following property: For any two colors, the two sets of vertices respectively colored by those two colors are translates of each other in the torus. Observing the constructions, we note that every such interleaving pattern has at least one translational periodicity other than the identity. In the previous work of [8], three t-interleaving constructions for two-dimensional arrays were presented, all based on lattice interleavers. Those three constructions can also be applied to tori because of their periodic patterns. Our Construction 2.2 generalizes the results in [8] in two ways: First, it covers more constructions based on lattice interleavers, with the results of [8] included as special cases; secondly, when t is even, it also covers constructions that do not use lattice interleavers, which we can make happen by simply letting any A_i and A_j take different values.

3. Achieving an interleaving degree within one of the optimal. Recall that an optimal interleaving need not be a perfect interleaving. A perfect interleaving uses $|S_t|$ colors, which is possible only when the dimensions satisfy the divisibility conditions of Construction 2.2. Most dimensions do not satisfy these divisibility conditions, and thus most tori do not admit a perfect interleaving—any interleaving must use more than $|S_t|$ colors. Recall that an optimal interleaving uses the minimal number of necessary colors.

In this section, we present a novel t-interleaving construction, with which we can t-interleave any large enough torus with at most one more than the optimal number

0	2	4	0	3	5	1	4
1	3	5	1	4	0	2	5
2	4	0	2	5	1	3	0
3	5	1	3	0	2	4	1
4	0	2	4	1	3	5	2
5	1	3	5	2	4	0	3

FIG. 3.1. *An example of t-interleaving with the three features.*

of colors. The construction presented here will also be used as a building block in section 4 for optimal t-interleaving.

3.1. Interleaving construction. The following definition defines several types of integer strings that are crucial to the interleaving constructions to be presented.

DEFINITION 3.1.

- *Given a positive integer t, if t is odd, then P is defined to be a string of integers, "a_1, a_2, ..., $a_{\frac{t-1}{2}}$," where $a_{\frac{t-1}{2}} = t+1$ and $a_i = t$ for $1 \le i < \frac{t-1}{2}$; if t is even, then P is defined to be a string of integers, "a_1, a_2, ..., $a_{\frac{t}{2}}$," where $a_{\frac{t}{2}} = t$ and $a_i = t-1$ for $1 \le i < \frac{t}{2}$. For example, if $t = 3$, then $P =$ "4"; if $t = 4$, then $P =$ "3,4"; if $t = 5$, then $P =$ "5,6."*
- *Given a positive integer t, if t is odd, then Q is defined to be a string of integers "b_1, b_2, ..., $b_{\frac{t+1}{2}}$," where $b_{\frac{t+1}{2}} = t+1$ and $b_i = t$ for $1 \le i < \frac{t+1}{2}$; if t is even, then Q is defined to be a string of integers "b_1, b_2, ..., $b_{\frac{t}{2}+1}$," where $b_{\frac{t}{2}+1} = t$ and $b_i = t-1$ for $1 \le i < \frac{t}{2}+1$.*
- *Given a positive integer t, an* offset sequence *is a string of P's and Q's. For example, an offset sequence consisting of one P and two Q's can be "PQQ," "QPQ" or "QQP." The offset sequence is also naturally seen as a string of integers which is the concatenation of the integer strings in its P's and Q's. For example, when $t = 3$, if an offset sequence consisting of one P and two Q's is "PQQ," then the offset sequence is also seen as "4,3,4,3,4"; when $t = 4$, if an offset sequence consisting of three P's and two Q's is "PQPPQ," then the offset sequence is also seen as "3,4,3,3,4,3,4,3,4,3,3,4." The number of integers in an offset sequence is called its* length.

In this section, we are particularly interested in one kind of t-interleaving on an $l_1 \times l_2$ torus, which has the following features:

- Feature 1: $l_1 = |S_t| + 1$. In other words, if t is odd, then $l_1 = \frac{t^2+1}{2} + 1$; if t is even, then $l_1 = \frac{t^2}{2} + 1$.
- Feature 2: The number of colors in the t-interleaving equals l_1. Also, in every column of the torus, each of the l_1 colors is assigned to exactly one vertex.
- Feature 3: If the vertex (a_1, b_1) and the vertex (a_2, b_2) have the same color, then for $i = 1, 2, \ldots, l_1 - 1$, the vertex $((a_1 + i) \bmod l_1, b_1)$ and the vertex $((a_2 + i) \bmod l_1, b_2)$ have the same color.

Example 3.1. Figure 3.1 shows a t-interleaving on an $l_1 \times l_2$ torus which has the above three features. There $t = 3$, $l_1 = |S_t| + 1 = 6$, and $l_2 = 8$.

Now let us choose a color i, where $0 \le i \le 5$, and say that the set of vertices of

color i is $\{(x_0,0),(x_1,1),\ldots,(x_{l_2-1},l_2-1)\}$. Then the string of integers "(x_1-x_0) mod $l_1,(x_2-x_1)$ mod $l_1,\ldots,(x_7-x_6)$ mod $l_1,(x_0-x_7)$ mod l_1" equals "4,4,4,3,4,4,3,4." Since when $t=3$, $P=$ "4" and $Q=$ "3,4," the above string of integers actually equals "$PPPQPQ$," which is an offset sequence of length l_2. We comment that this phenomenon is not a pure coincidence: Offset sequences do help us find t-interleavings that have the above three features. In fact, we can prove that in many cases (e.g., when $t=5$ or 7), for *any* t-interleaving on a torus that has the above three features, after horizontally shifting and/or vertically reversing the interleaving pattern, the resulting interleaving will exhibit the same phenomenon as the example shown here.

The following construction outputs a t-interleaving that has the three features.

Construction 3.1.

Input: A positive integer t. An $l_1\times l_2$ torus, where $l_1=|S_t|+1$. An integer m that equals $\lfloor\frac{t}{2}\rfloor$. Two integers p and q that satisfy the following equation set if t is odd,

$$\text{(3.1)}\qquad \begin{cases} pm+q(m+1)=l_2, \\ p(2m^2+m+1)+q(2m^2+3m+2)\equiv 0 \bmod (2m^2+2m+2), \\ p \text{ and } q \text{ are nonnegative integers, } p+q>0, \end{cases}$$

and satisfy the following equation set if t is even:

$$\text{(3.2)}\qquad \begin{cases} pm+q(m+1)=l_2, \\ p(2m^2-m+1)+q(2m^2+m)\equiv 0 \bmod (2m^2+1), \\ p \text{ and } q \text{ are nonnegative integers, } p+q>0. \end{cases}$$

Output: A t-interleaving on the $l_1\times l_2$ torus that satisfies Features 1, 2, and 3.

Construction: Let $S=$"$s_0,s_1,\ldots,s_{l_2-1}$" be an arbitrary offset sequence consisting of p P's and q Q's. For $j=1,2,\ldots,l_2$ and for $i=0,1,\ldots,l_1-1$, color the vertex $((\sum_{k=0}^{j-1}s_k+i) \bmod l_1, j \bmod l_2)$ with color i.

Example 3.2. Let $t=3$, $l_1=6$, $l_2=8$, $m=1$, $p=4$, and $q=2$. We use Construction 3.1 to t-interleave an $l_1\times l_2$ torus. Say the offset sequence S is chosen to be "$PPPQPQ$." Then Construction 3.1 outputs the t-interleaving shown in Figure 3.1.

We explain Construction 3.1 a little further. The equation set (3.1) (for odd t) and the equation set (3.2) (for even t) ensure that the offset sequence S, which consists of p P's and q Q's, exists. Furthermore, for any integer j $(0\le j\le l_2-1)$, if (a,j) and $(b,(j+1) \bmod l_2)$ are two vertices of the same color, then $b-a\equiv s_j \bmod l_1$. That is, the offset sequence S indicates the *vertical offsets* of any two vertices of the same color in adjacent columns. It is simple to verify that the t-interleaving output by Construction 3.1 satisfies all the three features listed earlier in this subsection.

The following lemma will be used to prove the correctness of Construction 3.1 and also in future analysis.

LEMMA 3.2. *Let $i\in\{0,1,\ldots,|S_t|\}$ be any of the colors used by Construction* 3.1 *to interleave the $l_1\times l_2$ torus. Let $\{(b_0,0),(b_1,1),\ldots,(b_{l_2-1},l_2-1)\}$ be the set of vertices of color i in the torus. Let m and S have the same meaning as in Construction* 3.1 *(namely, $m=\lfloor\frac{t}{2}\rfloor$, and $S=$"$s_0,s_1,\ldots,s_{l_2-1}$" is the offset sequence consisting of p P's and q Q's utilized by Construction* 3.1*). For any two integers j_1 and j_2 $(0\le j_1\ne j_2\le l_2-1)$, we define $L_{j_1\to j_2}$ as $L_{j_1\to j_2}=[(j_2-j_1) \bmod l_2]+\min\{(b_{j_2}-b_{j_1}) \bmod l_1,(b_{j_1}-b_{j_2}) \bmod l_1\}$. Then we have the following conclusions:*

- *Case* 1. *t is odd, $j_2-j_1\equiv m \bmod l_2$, and s_{j_1}, $s_{(j_1+1) \bmod l_2}$, $s_{(j_1+2) \bmod l_2}$, $\ldots$, $s_{(j_2-1) \bmod l_2}$ do not all equal t. In this case, $b_{j_2}-b_{j_1}\equiv -(m+1) \bmod l_1$ and $L_{j_1\to j_2}=t$.*

- *Case* 2. *t is odd, $j_2 - j_1 \equiv m+1 \bmod l_2$, and exactly one of s_{j_1}, $s_{(j_1+1) \bmod l_2}$, $s_{(j_1+2) \bmod l_2}$, $\ldots$, $s_{(j_2-1) \bmod l_2}$ equals $t+1$. In this case, $b_{j_2} - b_{j_1} \equiv m \bmod l_1$ and $L_{j_1 \to j_2} = t$.*
- *Case* 3. *t is even, $j_2 - j_1 \equiv 1 \bmod l_2$, and $s_{j_1} = t-1$. In this case, $b_{j_2} - b_{j_1} \equiv t-1 \bmod l_1$ and $L_{j_1 \to j_2} = t$.*
- *Case* 4. *t is even, $j_2 - j_1 \equiv m \bmod l_2$, and s_{j_1}, $s_{(j_1+1) \bmod l_2}$, $s_{(j_1+2) \bmod l_2}$, $\ldots$, $s_{(j_2-1) \bmod l_2}$ do not all equal $t-1$. In this case, $b_{j_2} - b_{j_1} \equiv -m \bmod l_1$ and $L_{j_1 \to j_2} = t$.*
- *Case* 5. *t is even, $j_2 - j_1 \equiv m+1 \bmod l_2$, and exactly one of s_{j_1}, $s_{(j_1+1) \bmod l_2}$, $s_{(j_1+2) \bmod l_2}$, $\ldots$, $s_{(j_2-1) \bmod l_2}$ equals t. In this case, $b_{j_2} - b_{j_1} \equiv m-1 \bmod l_1$ and $L_{j_1 \to j_2} = t$.*
- *If none of the above five cases is true and $j_2 - j_1 \neq t \bmod l_2$, then $L_{j_1 \to j_2} > t$. If none of the above five cases is true and $j_2 - j_1 \equiv t \bmod l_2$, then $L_{j_1 \to j_2} \geq t$.*

Proof. Let $\Delta = t+1$ if t is odd, and let $\Delta = t$ if t is even. The offset sequence S consists of P's and Q's, so it has the following property: For any $k \in \{0, 1, \ldots, l_2 - 1\}$ such that $s_k = \Delta$, the $m-1$ integers $s_{(k+1) \bmod l_2}$, $s_{(k+2) \bmod l_2}$, $\ldots$, $s_{(k+m-1) \bmod l_2}$ are all equal to $\Delta - 1$, and either $s_{(k+m) \bmod l_2}$ or $s_{(k+m+1) \bmod l_2}$ equals Δ. Also note that $b_{j_2} - b_{j_1} \equiv s_{j_1} + s_{(j_1+1) \bmod l_2} + s_{(j_1+2) \bmod l_2} + \cdots + s_{(j_2-1) \bmod l_2} \bmod l_1$. Based on those two observations, this lemma can be proved with straightforward computation. ∎

THEOREM 3.3. *Construction* 3.1 *is correct.*

Proof. Let (b_{j_1}, j_1) and (b_{j_2}, j_2) be any two vertices of the same color in the $l_1 \times l_2$ torus that was interleaved by Construction 3.1. The Lee distance between them is $d((b_{j_1}, j_1), (b_{j_2}, j_2)) = \min\{(j_2 - j_1) \bmod l_2, (j_1 - j_2) \bmod l_2\} + \min\{(b_{j_2} - b_{j_1}) \bmod l_1, (b_{j_1} - b_{j_2}) \bmod l_1\} = \min\{L_{j_1 \to j_2}, L_{j_2 \to j_1}\}$. From Lemma 3.2, it is clear that neither $L_{j_1 \to j_2}$ nor $L_{j_2 \to j_1}$ is less than t. Therefore $d((b_{j_1}, j_1), (b_{j_2}, j_2)) \geq t$. So Construction 3.1 t-interleaved the torus. And as mentioned before, this t-interleaving satisfies Features 1, 2, and 3. ∎

3.2. Existence of offset sequences. The feasibility of Construction 3.1 depends only on one thing: whether the two input parameters p and q exist or not. The following theorem shows that when the width of the torus, l_2, exceeds a threshold, p and q are guaranteed to exist.

THEOREM 3.4. *Let t be an odd (respectively, even) positive integer. When $l_2 \geq \lfloor \frac{t}{2} \rfloor (\lfloor \frac{t}{2} \rfloor + 1)(|S_t| + 1)$, there exists at least one solution (p, q) to the equation set* (3.1) (*respectively, equation set* (3.2))*, which is shown in the* input *part of Construction* 3.1.

Proof. Firstly, let us assume that t is odd. The equation set (3.1) is as follows:

$$\begin{cases} pm + q(m+1) = l_2, \\ p(2m^2 + m + 1) + q(2m^2 + 3m + 2) \equiv 0 \bmod (2m^2 + 2m + 2), \\ p \text{ and } q \text{ are nonnegative integers, } p + q > 0, \end{cases}$$

where $m = \lfloor \frac{t}{2} \rfloor$. We introduce a new variable z, and transform the above equation set equivalently to be

$$\begin{cases} \begin{pmatrix} m & m+1 \\ 2m^2 + m + 1 & 2m^2 + 3m + 2 \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} l_2 \\ z(2m^2 + 2m + 2) \end{pmatrix}, \\ p \text{ and } q \text{ are nonnegative integers; } z \text{ is a positive integer,} \end{cases}$$

which is the same as

$$\begin{cases} \begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} m & m+1 \\ 2m^2+m+1 & 2m^2+3m+2 \end{pmatrix}^{-1} \begin{pmatrix} l_2 \\ z(2m^2+2m+2) \end{pmatrix}, \\ p \text{ and } q \text{ are nonnegative integers; } z \text{ is a positive integer,} \end{cases}$$

which equals

$$\begin{cases} p = 2(m+1)(m^2+m+1)z - (2m^2+3m+2)l_2, \\ q = (2m^2+m+1)l_2 - 2m(m^2+m+1)z, \\ p \text{ and } q \text{ are nonnegative integers; } z \text{ is a positive integer.} \end{cases}$$

There exists a solution for the variables p, q, and z in the above equation set if and only if the following conditions can be satisfied:

$$\begin{cases} 2(m+1)(m^2+m+1)z - (2m^2+3m+2)l_2 \geq 0, \\ (2m^2+m+1)l_2 - 2m(m^2+m+1)z \geq 0, \\ z \text{ is a positive integer,} \end{cases}$$

which is equivalent to

$$\begin{cases} \frac{(2m^2+3m+2)l_2}{2(m+1)(m^2+m+1)} \leq z \leq \frac{(2m^2+m+1)l_2}{2m(m^2+m+1)}, \\ z \text{ is a positive integer.} \end{cases}$$

To enable a value for z to exist that satisfies the above conditions, it is sufficient to make $\frac{(2m^2+m+1)l_2}{2m(m^2+m+1)} - \frac{(2m^2+3m+2)l_2}{2(m+1)(m^2+m+1)} \geq 1$, that is, to make $l_2 \geq 2m(m+1)(m^2+m+1) = \lfloor \frac{t}{2} \rfloor (\lfloor \frac{t}{2} \rfloor + 1)(|S_t|+1)$. Therefore when $l_2 \geq \lfloor \frac{t}{2} \rfloor (\lfloor \frac{t}{2} \rfloor + 1)(|S_t|+1)$, there exists at least one solution (p, q) to the equation set (3.1).

When t is even, the conclusion can be proved in a very similar way. We skip its details. ∎

COROLLARY 3.5. *When* $l_2 \geq \lfloor \frac{t}{2} \rfloor (\lfloor \frac{t}{2} \rfloor + 1)(|S_t|+1)$, *Construction* 3.1 *can be used to output a* t*-interleaving on an* $(|S_t|+1) \times l_2$ *torus.*

Proof. When $l_2 \geq \lfloor \frac{t}{2} \rfloor (\lfloor \frac{t}{2} \rfloor + 1)(|S_t|+1)$, all the parameters in the *input* part of Construction 3.1 exist, including p and q. ∎

3.3. Interleaving with degree within one of the optimal. In this subsection, we will show how to interleave a large enough torus with at most one more than the optimal number of colors.

We define the simple term of *tiling tori* here. By tiling several interleaved tori vertically or horizontally, we get a larger torus, whose interleaving is the straightforward combination of the interleaving on the smaller tori. It is best explained with an example.

Example 3.3. Three interleaved tori, A, B, and C, are shown in Figure 3.2. The torus D is a 5×4 torus, obtained by *tiling* A *and* B *vertically* in the form of $\begin{bmatrix} A \\ B \end{bmatrix}$. The torus E is a 2×8 torus, obtained by *tiling one copy of* A *and two copies of* C *horizontally* in the form of $[\ C \quad A \quad C\ ]$.

The following construction t-interleaves a large enough torus with at most $|S_t|+2$ distinct integers.

Construction 3.2. t-interleave an $l_1 \times l_2$ torus G, where $l_1 \geq |S_t|(|S_t|+1)$ and $l_2 \geq \lfloor \frac{t}{2} \rfloor (\lfloor \frac{t}{2} \rfloor + 1)(|S_t|+1)$, using at most $|S_t|+2$ distinct integers.

1. Let G_1 be an $(|S_t|+1) \times l_2$ torus that is t-interleaved by Construction 3.1, using colors $0, 1, \ldots, |S_t|$. Let $\{(c_0, 0), (c_1, 1), \ldots, (c_{l_2-1}, l_2-1)\}$ be the set of vertices in G_1 having color 0.

(a)

A

0	1	2	3
3	2	1	0

B

0	0	0	0
1	1	1	1
2	2	2	2

C

1	2
3	4

(b)

D

0	1	2	3
3	2	1	0
0	0	0	0
1	1	1	1
2	2	2	2

E

1	2	0	1	2	3	1	2
3	4	3	2	1	0	3	4

FIG. 3.2. *Examples of tiling tori.*

G_1

0	2	0	2	0	2
1	0	1	0	1	0
2	1	2	1	2	1

G_2

3	2	3	2	3	2
0	3	0	3	0	3
1	0	1	0	1	0
2	1	2	1	2	1

G

0	2	0	2	0	2
1	0	1	0	1	0
2	1	2	1	2	1
3	2	3	2	3	2
0	3	0	3	0	3
1	0	1	0	1	0
2	1	2	1	2	1

FIG. 3.3. *Examples of Construction* 3.2.

2. Let G_2 be an $(|S_t|+2) \times l_2$ torus. Color the vertices $\{(c_0, 0), (c_1, 1), \ldots, (c_{l_2-1}, l_2 - 1)\}$ in G_2 with color $|S_t| + 1$.

3. For $j = 0, 1, \ldots, l_2 - 1$ and for $i = 1, 2, \ldots, |S_t| + 1$, color vertex $((c_j + i) \bmod (|S_t| + 2), j)$ in G_2 with color $i - 1$.

4. Let x and y be two nonnegative integers such that $l_1 = x(|S_t|+1)+y(|S_t|+2)$. Tile x copies of G_1 and y copies of G_2 vertically to get an $l_1 \times l_2$ torus G. Note that then G has been t-interleaved using at most $|S_t| + 2$ distinct integers.

Example 3.4. We use Construction 3.2 to t-interleave a 7×6 torus G, where $t = 2$. The first step is to use Construction 3.1 to t-interleave a 3×6 torus G_1. Say the offset sequence selected in Construction 3.1 is $S =$ *"QQQ"* ="1,2,1,2,1,2"; then G_1 is as shown in Figure 3.3. Then the 4×6 torus G_2 is as shown in the figure. By tiling one copy of G_1 and one copy of G_2 vertically, we get the t-interleaved torus G. $|S_t| + 2 = 4$ distinct integers are used to interleave G.

THEOREM 3.6. *Construction* 3.2 *is correct.*

Proof. It is a known fact that for any two relatively prime positive integers A and B, any integer C no less than $(A-1)(B-1)$ can be expressed as $C = xA+yB$, where x and y are nonnegative integers. Therefore in Construction 3.2, since $l_1 \geq |S_t|(|S_t|+1)$, l_1 indeed can be expressed as $l_1 = x(|S_t| + 1) + y(|S_t| + 2)$, as shown in the last step of Construction 3.2. Thus the construction can be executed from beginning to end successfully. Now we prove that the construction does t-interleave G; that is, for any two vertices (a_1, b_1) and (a_2, b_2) both of color i in G, the Lee distance between them is at least t. We consider three cases.

Case 1: $b_1 = b_2$, which means that (a_1, b_1) and (a_2, b_2) are in the same column of G. We see every column of G as a ring of length l_1 (because it is toroidal). Then, observe the colors in a column of G, and we can see that on the column, the color following color $|S_t|+1$ and before the next color $|S_t|+1$ must be the following, where

the pattern $0, 1, \ldots, |S_t|$ appears at least once:

$$0, 1, \ldots, |S_t|, 0, 1, \ldots, |S_t|, \ldots\ldots, 0, 1, \ldots, |S_t|.$$

Therefore since (a_1, b_1) and (a_2, b_2) have the same color, the Lee distance between them must be at least $|S_t| + 1 > t$.

Case 2: $b_1 \neq b_2$, and $i \neq |S_t|+1$. In this case, let us first observe two conclusions:

- The interleaving on G_2 (defined in Construction 3.2) is a t-interleaving. This can be proved as follows: Any two vertices of the same color in G_2 can be expressed as $((c_{j_1} + i_0) \bmod (|S_t| + 2), j_1)$ and $((c_{j_2} + i_0) \bmod (|S_t| + 2), j_2)$ (see steps 2 and 3 of Construction 3.2); then, $d_{G_2}(((c_{j_1} + i_0) \bmod (|S_t| + 2), j_1), ((c_{j_2} + i_0) \bmod (|S_t| + 2), j_2)) = d_{G_2}((c_{j_1}, j_1), (c_{j_2}, j_2)) \geq d_{G_1}((c_{j_1}, j_1), (c_{j_2}, j_2)) \geq t$.
- Let (α, j) and (β, j) be two vertices respectively in G_1 and G_2, which both have the same color. Then it is simple to see that $\beta = \alpha$ or $\beta = \alpha + 1$. Since G_1 has $|S_t| + 1$ rows and G_2 has $|S_t| + 2$ rows, we have $d_{G_2}((\beta, j), (0, j)) \geq d_{G_1}((\alpha, j), (0, j))$ and $d_{G_2}((\beta, j), (|S_t| + 1, j)) \geq d_{G_1}((\alpha, j), (|S_t|, j))$. That is, if u and v are two vertices respectively in G_1 and G_2, both of which are in the jth column and have the same color, then the vertical distance from v to either the top or bottom of G_2 is no less than the vertical distance from u to the top or bottom of G_1.

According to Construction 3.2, G is obtained by vertically tiling x copies of G_1 and y copies of G_2. Let us call each of those $x + y$ tori a *component torus* of G. Now, if (a_1, b_1) and (a_2, b_2) are in the same component torus of G, we know that the Lee distance between them *in* G is no less than the Lee distance between them *in that component torus*, which is at least t because that component torus is t-interleaved. If (a_1, b_1) and (a_2, b_2) are not in the same component torus of G, we do the following. We first construct a torus G', which is obtained by vertically tiling $x+y$ copies of G_1. It is simple to see that G' is t-interleaved. We call each of the $x+y$ copies of G_1 in G' a *component torus* of G'. Let us say that (a_1, b_1) and (a_2, b_2) are respectively in the k_1th and k_2th component torus *of* G. Let (c_1, b_1) and (c_2, b_2) be the two vertices of color i that are respectively in the k_1th and k_2th component torus *of* G'. Observe the shortest path between (a_1, b_1) and (a_2, b_2) *in* G, and we see that it can be split into such three intervals: from (a_1, b_1) to a border of the k_1th component torus, from the border of the k_1th component torus to the border of the k_2th component torus, and from the border of the k_2th component torus to (a_2, b_2). There is a corresponding (not necessarily shortest) path connecting (c_1, b_1) and (c_2, b_2) in G', which can be split into such three intervals similarly. Furthermore, each of the three intervals of the first path is at least as long as the corresponding interval of the second path. G' is t-interleaved, and so the second path's length is at least t. Thus the Lee distance between (a_1, b_1) and (a_2, b_2) *in* G is at least t.

Case 3: $b_1 \neq b_2$ and $i = |S_t| + 1$. In this case, it is simple to see that the two vertices in G, $(a_1 + 1 \bmod l_1, b_1)$ and $(a_2 + 1 \bmod l_1, b_2)$, both have color 0. Based on the conclusion of Case 2, $d_G((a_1 + 1 \bmod l_1, b_1), (a_2 + 1 \bmod l_1, b_2)) \geq t$. Thus $d_G((a_1, b_1), (a_2, b_2)) = d_G((a_1 + 1 \bmod l_1, b_1), (a_2 + 1 \bmod l_1, b_2)) \geq t$.

Thus Construction 3.2 correctly t-interleaved G. □

As a result of Construction 3.2, we get the following theorem.

THEOREM 3.7. *When* $l_1 \geq |S_t|(|S_t| + 1)$ *and* $l_2 \geq \lfloor \frac{t}{2} \rfloor (\lfloor \frac{t}{2} \rfloor + 1)(|S_t| + 1)$, *the* t*-interleaving number of an* $l_1 \times l_2$ *(or* $l_2 \times l_1$*) torus is at most* $|S_t| + 2$.

By combining Construction 2.2 (the construction for perfect t-interleaving) and

Construction 3.2, we can t-interleave any sufficiently large torus with at most one more than the optimal number of colors.

4. Optimal interleaving on large tori. In the previous section, it is shown that when l_2 is large enough, an $(|S_t|+1) \times l_2$ torus can be t-interleaved using $|S_t|+1$ integers. In this section, we will construct a $[k(|S_t|+1)-1] \times l_2$ torus (for some integer k) which is also t-interleaved using $|S_t|+1$ integers, by using an operation we call *removing a zigzag row*. Those two tori have a special property: When they (or multiple copies of them) are tiled vertically to get a larger torus, the larger torus is also t-interleaved with $|S_t|+1$ colors. Since $|S_t|+1$ and $k(|S_t|+1)-1$ are relatively prime, a large enough l_1 must be a linear combination of those two numbers with nonnegative integral coefficients, and therefore an $l_1 \times l_2$ torus can be t-interleaved using $|S_t|+1$ integers in this way. We present constructions to optimally t-interleave such tori, and as a parallel result, the existence of Region I (see the Introduction) is proved.

All the results of this section can be split into two parts: one for the case when t is odd, and the other for the case when t is even. Those two cases can be analyzed with very similar methods; however, their analysis and results differ in details. For succinctness, in this section, we only analyze in detail the case when t is odd, which should suffice for illustrating all the ideas. So in the first three subsections here (subsections 4.1, 4.2, and 4.3), we always assume that t *is odd*. In subsection 4.4, we present just the final result for the case when t is even. We list the major intermediate results for the case when t is even in Appendix II (section 8).

4.1. Removing a zigzag row in a torus. Below we define zigzag rows and the concept of removing a zigzag row in a torus.

DEFINITION 4.1. *A* zigzag row *in an $l_1 \times l_2$ torus is a set of l_2 vertices of the torus: $\{(a_0,0),(a_1,1),\ldots,(a_{l_2-1},l_2-1)\}$, where $0 \le a_i \le l_1-1$ for $i=0,1,\ldots,l_2-1$.*

For example, $\{(2,0),(3,1),(0,2),(0,3),(3,4)\}$ is a zigzag row in a 4×5 torus.

DEFINITION 4.2. *Let T be an $l_1 \times l_2$ torus. Let $\{(a_0,0),(a_1,1),\ldots,(a_{l_2-1},l_2-1)\}$ be a zigzag row in T. Let there be an interleaving on T, which colors T's vertex (b,c) with color $I(b,c)$, for $b=0,1,\ldots,l_1-1$ and $c=0,1,\ldots,l_2-1$. Then a torus G is said to be obtained by* removing *the zigzag row $\{(a_0,0),(a_1,1),\ldots,(a_{l_2-1},l_2-1)\}$ in T if and only if these two conditions are satisfied:*

- *G is an $(l_1-1) \times l_2$ torus.*
- *For $i=0,1,\ldots,l_1-2$ and $j=0,1,\ldots,l_2-1$, the vertex (i,j) in G has color $I(i,j)$ if $i<a_j$, and color $I(i+1,j)$ if $i \ge a_j$.*

Example 4.1. In Figure 4.1, a 6×5 torus T is shown. A zigzag row $\{(3,0), (2,1), (1,2), (3,3), (1,4)\}$ in T is circled in the figure. Figure 4.1 shows a torus G obtained by removing the zigzag row $\{(3,0), (2,1), (1,2), (3,3), (1,4)\}$ in T.

It can be readily observed that G can be seen as being derived from T in the following way: First, delete the zigzag row in T that is circled in Figure 4.1; then in each column of T, move the vertices below the circled vertex upward.

In order to get our final results, we present three rules to follow for devising a zigzag row. Let B be an $l_0 \times l_2$ torus which is t-interleaved by Construction 3.1. Note that this means $l_0 = |S_t|+1$. Let $S =$"$s_0, s_1, \ldots, s_{l_2-1}$" be the offset sequence utilized by Construction 3.1 when it was t-interleaving B. Let H be an $l_1 \times l_2$ torus obtained by tiling several copies of B vertically. Let $m = \lfloor \frac{t}{2} \rfloor$. Then the three rules for devising a zigzag row *in* H, $\{(a_0,0),(a_1,1),\ldots,(a_{l_2-1},l_2-1)\}$, are the following:

- Rule 1. For any j such that $0 \le j \le l_2-1$, if the integers s_j, $s_{(j+1) \bmod l_2}$, $\ldots$, $s_{(j+m-1) \bmod l_2}$ do not all equal t, then $a_j \ge a_{(j+m) \bmod l_2} + m$.

T

1	3	5	2	4
2	4	(6)	3	(5)
3	(5)	1	4	6
(4)	6	2	(5)	1
5	1	3	6	2
6	2	4	1	3

G

1	3	5	2	4
2	4	1	3	6
3	6	2	4	1
5	1	3	6	2
6	2	4	1	3

FIG. 4.1. *Removing a zigzag row* $\{(3,0),(2,1),(1,2),(3,3),(1,4)\}$ *in* T.

- Rule 2. For any j such that $0 \le j \le l_2 - 1$, if exactly one of the integers $s_j, s_{(j+1) \bmod l_2}, \ldots, s_{(j+m) \bmod l_2}$ equals $t+1$, then $a_j \le a_{(j+m+1) \bmod l_2} - (m-1)$.
- Rule 3. For any j such that $0 \le j \le l_2 - 1$, $m \le a_j \le l_1 - m - 1$.

LEMMA 4.3. *Let B be a torus t-interleaved by Construction* 3.1. *Let H be a torus obtained by tiling copies of B vertically, and let T be a torus obtained by removing a zigzag row in H, where the zigzag row in H follows the three rules listed above. Let G be a torus obtained by tiling copies of B and T vertically. Then, both T and G are t-interleaved.*

Proof. When $t = 1$, the proof is trivial. So we assume $t \ge 3$ in the rest of the proof. It is simple to see that H is t-interleaved, because H is obtained by tiling B, a t-interleaved torus. We assume that B is an $l_0 \times l_2$ torus (where $l_0 = |S_t| + 1$), H is an $l_1 \times l_2$ torus (where l_1 is a multiple of l_0), T is an $l_T \times l_2$ torus (where $l_T = l_1 - 1$), and G is an $l_G \times l_2$ torus. Let $m = \lfloor \frac{t}{2} \rfloor$. Let $S =$"$s_0, s_1, \ldots, s_{l_2-1}$" be the offset sequence utilized by Construction 3.1 when it was t-interleaving B.

(1) In this part, we will prove that T is t-interleaved. Let (x_1, y_1) and (x_2, y_2) be two vertices in T both of color r. We need to prove that $d_T((x_1, y_1), (x_2, y_2)) \ge t$.

Let $\{(a_0, 0), (a_1, 1), \ldots, (a_{l_2-1}, l_2 - 1)\}$ denote the zigzag row removed in H to get T. If $a_{y_1} \le x_1$, then let $z_1 = x_1 + 1$; otherwise let $z_1 = x_1$. Similarly, if $a_{y_2} \le x_2$, then let $z_2 = x_2 + 1$; otherwise let $z_2 = x_2$. Clearly, the two vertices in H, (z_1, y_1) and (z_2, y_2), also have color r.

We need to consider only the following three cases.

Case 1: $y_1 = y_2$. In this case, $d_H((z_1, y_1), (z_2, y_2))$ is a multiple of $|S_t| + 1$ (the number of rows in B), and $d_T((x_1, y_1), (x_2, y_2)) \ge d_H((z_1, y_1), (z_2, y_2)) - 1 \ge |S_t| = \frac{t^2+1}{2} > t$.

Case 2: $y_1 \ne y_2$ and $d_T((x_1, y_1), (x_2, y_2)) \le d_H((z_1, y_1), (z_2, y_2)) - 2$. Without loss of generality, we assume $x_1 \ge x_2$. Then, based on the definition of removing a zigzag row, it is simple to verify that the following must be true: $d_T((x_1, y_1), (x_2, y_2)) = d_H((z_1, y_1), (z_2, y_2)) - 2$, $a_{y_2} < z_2 < z_1 < a_{y_1}$, $(z_2 - z_1 \bmod l_1) \le (z_1 - z_2 \bmod l_1)$. By Rule 3, any vertex in the removed zigzag row is neither in the first m rows nor in the last m rows of H, so $(z_2 - z_1 \bmod l_1) \ge 2m + 3$. Thus $d_T((x_1, y_1), (x_2, y_2)) = d_H((z_1, y_1), (z_2, y_2)) - 2 > (z_2 - z_1 \bmod l_1) - 2 \ge 2m + 1 = t$.

Case 3: $y_1 \ne y_2$ and $d_T((x_1, y_1), (x_2, y_2)) \ge d_H((z_1, y_1), (z_2, y_2)) - 1$. We know that $d_H((z_1, y_1), (z_2, y_2)) \ge t$. So to show that $d_T((x_1, y_1), (x_2, y_2)) \ge t$, we just need to prove that if $d_H((z_1, y_1), (z_2, y_2)) = t$, then $d_T((x_1, y_1), (x_2, y_2)) \ge d_H((z_1, y_1), (z_2,$

$y_2)$). By Lemma 3.2, there are only two nontrivial subcases to consider, without loss of generality, as follows.

Subcase 3.1: $y_2 - y_1 \equiv m \bmod l_2$, $z_2 - z_1 \equiv -(m+1) \bmod l_1$, $d_H((z_1, y_1), (z_2, y_2)) = (y_2 - y_1 \bmod l_2) + (z_1 - z_2 \bmod l_1) = t$, and s_{y_1}, $s_{(y_1+1) \bmod l_2}$, $s_{(y_1+2) \bmod l_2}$, $\ldots$, $s_{(y_1+m-1) \bmod l_2}$ do not all equal t. If $z_1 > z_2$ (which means $z_1 = z_2 + (m+1)$), then from Rule 1, it is simple to see that $x_1 - x_2 = z_1 - z_2$, and so $d_T((x_1, y_1), (x_2, y_2)) = d_H((z_1, y_1), (z_2, y_2)) = t$. If $z_1 < z_2$ (which means that (z_1, y_1) and (z_2, y_2) are respectively in the first and last $m+1$ rows of H), since the first and last m rows of H and T must be the same, we get that $(x_1 - x_2 \bmod l_T) = (z_1 - z_2 \bmod l_1) = m+1$, and so $d_T((x_1, y_1), (x_2, y_2)) = d_H((z_1, y_1), (z_2, y_2)) = t$.

Subcase 3.2: $y_2 - y_1 \equiv m+1 \bmod l_2$, $z_2 - z_1 \equiv m \bmod l_1$, $d_H((z_1, y_1), (z_2, y_2)) = (y_2 - y_1 \bmod l_2) + (z_2 - z_1 \bmod l_1) = t$, and exactly one of s_{y_1}, $s_{(y_1+1) \bmod l_2}$, $s_{(y_1+2) \bmod l_2}$, $\ldots$, $s_{(y_1+m) \bmod l_2}$ equals $t+1$. If $z_1 < z_2$ (which means $z_1 = z_2 - m$), then from Rule 2, it is simple to see that $x_2 - x_1 = z_2 - z_1$, and so $d_T((x_1, y_1), (x_2, y_2)) = d_H((z_1, y_1), (z_2, y_2)) = t$. If $z_1 > z_2$ (which means that (z_1, y_1) and (z_2, y_2) are respectively in the last and first m rows of H), since the first and last m rows of H and T must be the same, we get that $(x_2 - x_1 \bmod l_T) = (z_2 - z_1 \bmod l_1) = m$, and so $d_T((x_1, y_1), (x_2, y_2)) = d_H((z_1, y_1), (z_2, y_2)) = t$.

Thus T is t-interleaved.

(2) In this part, we will prove that G is t-interleaved. First let us make an observation: When a t-interleaved torus K is tiled with other tori vertically to get a larger torus $\hat{G}$, for any two vertices μ and ν in K (which are now also in $\hat{G}$) of the same color, the Lee distance between them *in* $\hat{G}$, $d_{\hat{G}}(\mu, \nu)$, is clearly no less than t. Let us also notice that the torus obtained by tiling one copy of B and one copy of T vertically is t-interleaved, which can be proved with exactly the same proof as in part (1).

G is obtained by tiling multiple copies of B and T. Let us call each copy of B or T in G a *component torus*. Let (x_1, y_1) and (x_2, y_2) be two vertices in G of the same color. Assume $d_G((x_1, y_1), (x_2, y_2)) \leq t$. Then since both B and T have more than t rows, (x_1, y_1) and (x_2, y_2) must be either in the same component torus or in two adjacent component tori. Now if (x_1, y_1) and (x_2, y_2) are in the same component torus, let K denote that component torus; if (x_1, y_1) and (x_2, y_2) are in two adjacent component tori, let K be the torus obtained by vertically tiling those two component tori; let $\hat{G}$ be the same as G. By using the observation in the previous paragraph, we can readily prove that $d_G((x_1, y_1), (x_2, y_2)) \geq t$. Thus G is t-interleaved. □

4.2. Constructing the zigzag row. We presented three rules on devising a zigzag row in the previous subsection. But specifically, how can one construct a zigzag row that follows all those rules? In this subsection, we present such constructions.

Before the formal presentation, let us go over a few concepts. An offset sequence is a string of P's and Q's, where P and Q are strings of integers depending on t. For example, when $t = 5$, $P =$"$5, 6$" and $Q =$"$5, 5, 6$." Then an offset sequence "PPQ" can also be written as "$5, 6, 5, 6, 5, 5, 6$." Let us also express the offset sequence "PPQ" as "$s_0, s_1, s_2, s_3, s_4, s_5, s_6$," where $s_0 = 5$, $s_1 = 6$, $\ldots$, $s_6 = 6$. Then for $i = 0, 1, \ldots, 6$ we will call s_i the $(i+1)$th element of the offset sequence. Also, we will say that s_2 is *the first element of a* P, because it is the first element of the second P in the offset sequence. For the same reason, s_0 is the first element of a P (this time, the first P in the offset sequence), s_1 is the second (or last) element of a P (the first P in the offset sequence), s_4 is the first element of a Q, and so on.

Now we begin the formal presentation of the constructions. Let B be an $l_0 \times l_2$ torus that is t-interleaved by Construction 3.1, so $l_0 = |S_t| + 1$. Let H be an $l_1 \times l_2$

torus obtained by tiling z copies of B vertically, so $l_1 = zl_0 = z(|S_t| + 1)$. Let $S =$"$s_0, s_1, \ldots, s_{l_2-1}$" be the offset sequence utilized by Construction 3.1 when it was t-interleaving B. We say that the offset sequence S consists of p P's and q Q's, where we require $p > 0$ and $q > 0$. We require that in the offset sequence the P's and Q's be interleaved very evenly. To be specific, in the offset sequence, between any two nearby P's (including between the last P and the first P, because we see the offset sequence as being toroidal), there must be either $\lceil \frac{q}{p} \rceil$ or $\lfloor \frac{q}{p} \rfloor$ consecutive Q's; and between any two nearby Q's (including between the last Q and the first Q), there must be either $\lceil \frac{p}{q} \rceil$ or $\lfloor \frac{p}{q} \rfloor$ consecutive P's. Also, we require the offset sequence to start with a P and to end with a Q. For example, an offset sequence consisting of three P's and five Q's that satisfies the above requirements is "$PQQPQQPQ$." Let $m = \frac{t-1}{2}$. Let $L = m + m\lceil \frac{p}{q} \rceil$ if $p \geq q$, and let $L = m + (m-1)\lceil \frac{q}{p} \rceil$ if $p < q$. Below we present two constructions—Constructions 4.1 and 4.2—for constructing a zigzag row *in* H, applicable respectively when $p \geq q$ and when $p < q$. If l_1 is too small, there may not exist a zigzag row *in* H that follows the three rules. To make our constructions work, we require that

$$l_1 \geq \left(\left\lceil \frac{p}{q} \right\rceil + 1 \right) m^2 + 2m + 1$$

if $p \geq q$, and also that

$$l_1 \geq \left(\left\lceil \frac{q}{p} \right\rceil + 1 \right) m^2 + m + \left(2 - \left\lceil \frac{q}{p} \right\rceil \right)$$

if $p < q$. Note that the constructed zigzag row is denoted by $\{(a_0, 0), (a_1, 1), \ldots, (a_{l_2-1}, l_2 - 1)\}$. Also note that both constructions require $t > 3$. The analysis for the case $t = 3$, a somewhat special case, is presented in Appendix I (section 7).

Construction 4.1. Constructing a zigzag row in H, when t is odd, $t > 3$, and $p \geq q > 0$.

1. Let $s_{x_1}, s_{x_2}, \ldots, s_{x_{p+q}}$ be the integers such that $0 = x_1 < x_2 < \cdots < x_{p+q} = l_2 - m - 1$ and each s_{x_i} $(1 \leq i \leq p + q)$ is the first element of a P or Q in the offset sequence S.

Let $a_{x_1} = L$. For $i = 2$ to $p+q$, if $s_{x_{i-1}}$ is the first element of a Q, let $a_{x_i} = L$.

For $i = 2$ to $p+q$, if $s_{x_{i-1}}$ is the first element of a P, then let $a_{x_i} = a_{x_{i-1}} - m$.

2. For $i = 2$ to m and for $j = 1$ to $p + q$, let $a_{x_j+i-1} = a_{x_j+i-2} + L$.

3. Let $s_{y_1}, s_{y_2}, \ldots, s_{y_q}$ be the integers such that $y_1 < y_2 < \cdots < y_q = l_2 - 1$ and each s_{y_i} $(1 \leq i \leq q)$ is the last element of a Q in the offset sequence S.

For $i = 1$ to q, let $a_{y_i} = mL + m$.

Now we have fully determined the zigzag row, $\{(a_0, 0), (a_1, 1), \ldots, (a_{l_2-1}, l_2 - 1)\}$, in the torus H.

The zigzag row constructed by Construction 4.1 has a quite regular structure. We show it with an example.

Example 4.2. We use this example to illustrate Construction 4.1. In this example, $t = 5$, and B is an 14×18 torus as shown in Figure 4.2(a). B is t-interleaved by Construction 3.1 by using the offset sequence $S =$"$PPPQPPPQ$"="5, 6, 5, 6, 5, 6, 5, 5, 6, 5, 6, 5, 6, 5, 6, 5, 5, 6." The torus H is shown in Figure 4.2(b). H is an 28×18 torus obtained by tiling two copies of B vertically. The rest of the parameters used by Construction 4.1 are $p = 6$, $q = 2$, $m = 2$, and $L = 8$. It is not difficult to verify that the zigzag row in H constructed by Construction 4.1 is $\{(8, 0), (16, 1), (6, 2), (14, 3), (4, 4), (12, 5), (2, 6), (10, 7), (18, 8), (8, 9), (16, 10), (6, 11), (14, 12), (4, 13), (12, 14),$

(a)

B

0	9	3	12	6	1	9	4	13	7	2	10	5	13	8	2	11	6
1	10	4	13	7	2	10	5	0	8	3	11	6	0	9	3	12	7
2	11	5	0	8	3	11	6	1	9	4	12	7	1	10	4	13	8
3	12	6	1	9	4	12	7	2	10	5	13	8	2	11	5	0	9
4	13	7	2	10	5	13	8	3	11	6	0	9	3	12	6	1	10
5	0	8	3	11	6	0	9	4	12	7	1	10	4	13	7	2	11
6	1	9	4	12	7	1	10	5	13	8	2	11	5	0	8	3	12
7	2	10	5	13	8	2	11	6	0	9	3	12	6	1	9	4	13
8	3	11	6	0	9	3	12	7	1	10	4	13	7	2	10	5	0
9	4	12	7	1	10	4	13	8	2	11	5	0	8	3	11	6	1
10	5	13	8	2	11	5	0	9	3	12	6	1	9	4	12	7	2
11	6	0	9	3	12	6	1	10	4	13	7	2	10	5	13	8	3
12	7	1	10	4	13	7	2	11	5	0	8	3	11	6	0	9	4
13	8	2	11	5	0	8	3	12	6	1	9	4	12	7	1	10	5

(b)

H

P P P Q P P P Q

0	9	3	12	6	1	9	4	13	7	2	10	5	13	8	2	11	6
1	10	4	13	7	2	10	5	0	8	3	11	6	0	9	3	12	7
2	11	5	0	8	3	(11)	6	1	9	4	12	7	1	10	(4)	13	8
3	12	6	1	9	4	12	7	2	10	5	13	8	2	11	5	0	9
4	13	7	2	(10)	5	13	8	3	11	6	0	9	(3)	12	6	1	10
5	0	8	3	11	6	0	9	4	12	7	1	10	4	13	7	2	11
6	1	(9)	4	12	7	1	10	5	13	8	(2)	11	5	0	8	3	12
7	2	10	5	13	8	2	11	6	0	9	3	12	6	1	9	4	13
(8)	3	11	6	0	9	3	12	7	(1)	10	4	13	7	2	10	5	0
9	4	12	7	1	10	4	13	8	2	11	5	0	8	3	11	6	1
10	5	13	8	2	11	5	⟨0⟩	9	3	12	6	1	9	4	12	⟨7⟩	2
11	6	0	9	3	12	6	1	10	4	13	7	2	10	5	13	8	3
12	7	1	10	4	⟨13⟩	7	2	11	5	0	8	3	11	⟨6⟩	0	9	4
13	8	2	11	5	0	8	3	12	6	1	9	4	12	7	1	10	5
0	9	3	⟨12⟩	6	1	9	4	13	7	2	10	⟨5⟩	13	8	2	11	6
1	10	4	13	7	2	10	5	0	8	3	11	6	0	9	3	12	7
2	⟨11⟩	5	0	8	3	11	6	1	9	⟨4⟩	12	7	1	10	4	13	8
3	12	6	1	9	4	12	7	2	10	5	13	8	2	11	5	0	9
4	13	7	2	10	5	13	8	⟨3̄⟩	11	6	0	9	3	12	6	1	⟨10̄⟩
5	0	8	3	11	6	0	9	4	12	7	1	10	4	13	7	2	11
6	1	9	4	12	7	1	10	5	13	8	2	11	5	0	8	3	12
7	2	10	5	13	8	2	11	6	0	9	3	12	6	1	9	4	13
8	3	11	6	0	9	3	12	7	1	10	4	13	7	2	10	5	0
9	4	12	7	1	10	4	13	8	2	11	5	0	8	3	11	6	1
10	5	13	8	2	11	5	0	9	3	12	6	1	9	4	12	7	2
11	6	0	9	3	12	6	1	10	4	13	7	2	10	5	13	8	3
12	7	1	10	4	13	7	2	11	5	0	8	3	11	6	0	9	4
13	8	2	11	5	0	8	3	12	6	1	9	4	12	7	1	10	5

FIG. 4.2. *An example of Construction* 4.1.

$(2, 15)$, $(10, 16)$, $(18, 17)\}$. In Figure 4.2(b), the vertices in the zigzag row are shown in solid circles, solid hexagons, or dashed circles.

Now we briefly analyze the structure of the zigzag row in H. Let us write the offset sequence S as $S =$"$s_0, s_1, \ldots, s_{17}$." Then for $i = 0, 1, \ldots, 17$, we can see that s_i actually shows the *offset* between the ith column and the $(i + 1)$th column of H. In other words, if we shift the integers in the ith column of H down (toroidally) by s_i units, we get the $(i + 1)$th column of H, so we can think of s_i as spanning from the ith column to the $(i + 1)$th column of H. And let us say that a P or Q in the offset sequence spans the columns that all its elements span. Then, since the offset sequence here is "$PPPQPPPQ$," the range spanned by each is as indicated in Figure 4.2(b).

Let us observe the vertices in the zigzag row that are in solid circles. If we indicate them by (a_{x_1}, x_1), $(a_{x_2}, x_2), \ldots, (a_{x_{p+q}}, x_{p+q})$, where $x_1 < x_2 \cdots < x_{p+q}$, then we can see that $s_{x_1}, s_{x_2}, \ldots, s_{x_{p+q}}$ are the first elements of the P's and Q's in the offset sequence (namely, each of them is the first element of a P or a Q in the offset sequence). And we can see that the vertices in solid circles have a regular structure: The vertical position climbs up by $m = 2$ units from one vertex to the next, and drops to a base-position if it is between the spanned ranges of a Q and a P. Now let us observe the vertices in solid hexagons. We can see that they correspond to the second elements of the P's and Q's in the offset sequence, and they also have a regular structure. To be specific, the positions of the vertices in solid hexagons can be obtained by shifting the positions of the vertices in solid circles horizontally by one unit and then down by $L = 8$ units. In general, those vertices in a zigzag row that correspond to the $(i + 1)$th elements of P's and Q's can be obtained by shifting the positions of the vertices that correspond to the ith elements of P's and Q's horizontally by one unit and down by L unit (here $0 \le i < m$). As for the vertices in dashed circles, they correspond to the last elements of the Q's in the offset sequence, and they are all in the same row. The above observations can be extended in an obvious way to the general outputs of Construction 4.1.

Now we present the second construction.

Construction 4.2. Constructing a zigzag row in H, when t is odd, $t > 3$, and $0 < p < q$.

1. Let $s_{x_1}, s_{x_2}, \ldots, s_{x_{p+q}}$ be the integers such that $0 = x_1 < x_2 < \cdots < x_{p+q} = l_2 - m - 1$, and each s_{x_i} $(1 \le i \le p + q)$ is the first element of a P or Q in the offset sequence S.

Let $a_{x_1} = L$.

For $i = 2$ to $p + q$, if s_{x_i} is the first element of a P, let $a_{x_i} = L$; if $s_{x_{i-1}}$ is the first element of a P, let $a_{x_i} = L - \lceil \frac{q}{p} \rceil (m - 1)$; otherwise, let $a_{x_i} = a_{x_{i-1}} + (m - 1)$.

2. For $i = 2$ to m and for $j = 1$ to $p + q$, let $a_{x_j+i-1} = a_{x_j+i-2} + L$.

3. Let $s_{y_1}, s_{y_2}, \ldots, s_{y_q}$ be the integers such that $y_1 < y_2 < \cdots < y_q = l_2 - 1$ and each s_{y_i} $(1 \le i \le q)$ is the last element of a Q in the offset sequence S.

For $i = 1$ to q, let $a_{y_i} = a_{y_i-1} + L$.

Now we have fully determined the zigzag row, $\{(a_0, 0), (a_1, 1), \ldots, (a_{l_2-1}, l_2 - 1)\}$, in the torus H.

Like Construction 4.1, the zigzag row constructed by Construction 4.2 also has a regular (and similar) structure.

THEOREM 4.4. *The zigzag rows constructed by Constructions* 4.1 *and* 4.2 *follow all the three rules listed above* (*Rules* 1, 2, *and* 3).

The above theorem can be proved with straightforward verification. So we skip its proof.

4.3. Optimal interleaving when t is odd. In this subsection, we prove that when t is odd, for a torus whose size is large enough in both dimensions, its t-interleaving number is at most one more than the sphere packing lower bound, $|S_t|$. We also present the corresponding optimal t-interleaving construction.

LEMMA 4.5. *In equation set* (3.1) (*the equation set in Construction* 3.1)*, let the values of* t*,* m*, and* l_2 *be fixed. Let* $p = p_0, q = q_0$ *be a solution that satisfies the equation set* (3.1)*. Then, another solution,* $p = p_1, q = q_1$*, also satisfies the equation set* (3.1) *if and only if there exists an integer* c *such that* $p_1 = p_0 + c(m+1)(2m^2 + 2m + 2) \geq 0$ *and* $q_1 = q_0 - cm(2m^2 + 2m + 2) \geq 0$.

Proof. We can easily prove that "$p = p_1, q = q_1$ is a solution that satisfies the equation set (3.1) if $p_1 = p_0 + c(m+1)(2m^2 + 2m + 2) \geq 0$ and $q_1 = q_0 - cm(2m^2 + 2m + 2) \geq 0$ for some integer c," by plugging $p = p_1, q = q_1$ into the equation set (3.1). Now let us prove the other direction.

Assume that $p = p_1, q = q_1$ is a solution that satisfies the equation set (3.1). Let $x = p_1 - p_0$ and $y = q_1 - q_0$. By the first equation in (3.1), $p_1 m + q_1(m+1) = l_2 = p_0 m + q_0(m+1)$, and therefore $(p_1 - p_0)m = -(q_1 - q_0)(m+1)$, which is $xm = -y(m+1)$. So x is a multiple of $m+1$, and y is a multiple of m. Thus there exists an integer a such that $x = a(m+1)$ and $y = -am$.

Now let us look at the second equation in (3.1), $p_1(2m^2+m+1)+q_1(2m^2+3m+2) \equiv 0 \bmod (2m^2+2m+2)$. Note that $2m^2+m+1 \equiv -(m+1) \bmod (2m^2+2m+2)$ and $2m^2+3m+2 \equiv m \bmod (2m^2+2m+2)$. So $-p_1(m+1)+q_1 m \equiv 0 \bmod (2m^2+2m+2)$. Since $p_1 = p_0 + x = p_0 + a(m+1)$ and $q_1 = q_0 + y = q_0 - am$, we get $-[p_0 + a(m+1)](m+1) + (q_0 - am)m \equiv [-p_0(m+1) + q_0 m] - [a(m+1)^2 + am^2] \equiv -a(2m^2+2m+1) \equiv 0 \bmod (2m^2+2m+2)$. Since $2m^2+2m+1$ and $2m^2+2m+2$ must be relatively prime, we get $2m^2+2m+2|a$. So there exists an integer c such that $a = c(2m^2+2m+2)$. Then $p_1 = p_0+x = p_0+a(m+1) = p_0+c(m+1)(2m^2+2m+2) \geq 0$ and $q_1 = q_0+y = q_0-am = q_0-cm(2m^2+2m+2) \geq 0$, these two inequalities coming from the last condition in (3.1). That completes the proof of the other direction of this lemma. ∎

LEMMA 4.6. *In equation set* (3.1) (*the equation set in Construction* 3.1)*, let the values of* t*,* m*, and* l_2 *be fixed. Let* $\Delta_P = (m+1)(2m^2+2m+2)$ *and* $\Delta_Q = m(2m^2+2m+2)$*. If there exists a solution of* p *and* q *that satisfies the equation set* (3.1)*, then there exists a solution* $p = p^*, q = q^*$ *that satisfies not only* (3.1) *but also one of the following two inequalities:*

$$\frac{l_2}{2m+1} - \frac{\Delta_Q}{2} < q^* \leq p^* < \frac{l_2}{2m+1} + \frac{\Delta_P}{2}, \tag{4.1}$$

$$\frac{l_2}{2m+1} - \frac{\Delta_P}{2} \leq p^* < q^* \leq \frac{l_2}{2m+1} + \frac{\Delta_Q}{2}. \tag{4.2}$$

Proof. Assume that there is a solution $p = p_0, q = q_0$ that satisfies equation set (3.1). Trivially, either $p_0 \geq q_0$ or $p_0 < q_0$. First, let us assume that $p_0 \geq q_0$. If $p_0 \geq \frac{l_2}{2m+1} + \Delta_P$, then $q_0 = \frac{l_2 - p_0 m}{m+1} \leq \frac{l_2 - [l_2/(2m+1) + \Delta_P]m}{m+1} = \frac{l_2 - [l_2/(2m+1) + (m+1)(2m^2+2m+2)]m}{m+1} = \frac{l_2}{2m+1} - \Delta_Q$ (and vice versa), so then by Lemma 4.5, $p = p_0 - \Delta_P, q = q_0 + \Delta_Q$ is also a solution to (3.1), and, what is more, $p_0 - \Delta_P \geq \frac{l_1}{2m+1} \geq q_0 + \Delta_Q$. Based on the above observation, we can see that there must exist a solution $p = p_1, q = q_1$ such that $\frac{l_2}{2m+1} - \Delta_Q < q_1 \leq p_1 < \frac{l_2}{2m+1} + \Delta_P$. If $p_1 < \frac{l_2}{2m+1} + \frac{\Delta_P}{2}$, then $q_1 > \frac{l_2}{2m+1} - \frac{\Delta_Q}{2}$, so then we can simply let $p^* = p_1$ and let $q^* = q_1$. If $p_1 \geq \frac{l_2}{2m+1} + \frac{\Delta_P}{2}$, then $q_1 \leq \frac{l_2}{2m+1} - \frac{\Delta_Q}{2}$,

so then we will let $p^* = p_1 - \Delta_P$ and let $q^* = q_1 + \Delta_Q$, in which case we will have $\frac{l_2}{2m+1} - \frac{\Delta_P}{2} \le p^* < \frac{l_2}{2m+1} < q^* \le \frac{l_2}{2m+1} + \frac{\Delta_Q}{2}$. So when $p_0 \ge q_0$, this lemma holds. The case that $p_0 < q_0$ can be analyzed similarly. ∎

THEOREM 4.7. *Let t be a positive odd integer. Let $m = \frac{t-1}{2}$. Define A as*

$$\max\Big\{ \Big(\Big\lceil \tfrac{l_2+(m+1)(2m+1)(m^2+m+1)}{l_2-m(2m+1)(m^2+m+1)} \Big\rceil + 1\Big)m^2 + 2m + 1,$$
$$\Big(\Big\lceil \tfrac{l_2+m(2m+1)(m^2+m+1)}{l_2-(m+1)(2m+1)(m^2+m+1)} \Big\rceil + 1\Big)m^2 + m + 2 - \Big\lceil \tfrac{l_2+m(2m+1)(m^2+m+1)}{l_2-(m+1)(2m+1)(m^2+m+1)} \Big\rceil \Big\}.$$

Then when

$$l_2 \ge (m+1)(2m+1)(m^2+m+1) + 1$$

and

$$l_1 \ge (2m^2+2m+1)\left(\left\lceil \frac{A}{2m^2+2m+2} \right\rceil (2m^2+2m+2) - 2\right),$$

the t-interleaving number of an $l_1 \times l_2$ (or $l_2 \times l_1$) torus is either $|S_t|$ or $|S_t|+1$.

Proof. This theorem is trivially correct when $t = 1$. When $t = 3$, by the result of Appendix I (Theorem 7.1), we can also easily verify that this theorem is correct. Thus in the following analysis, we assume that $t > 3$.

Let us first define a few variables for the ease of expression. Let $\Delta_P = (m+1)(2m^2+2m+2)$, $\Delta_Q = m(2m^2+2m+2)$, $B = \frac{l_2+(m+1)(2m+1)(m^2+m+1)}{l_2-m(2m+1)(m^2+m+1)}$, $C = \frac{l_2+m(2m+1)(m^2+m+1)}{l_2-(m+1)(2m+1)(m^2+m+1)}$, $D = (\lceil B \rceil + 1)m^2 + 2m + 1$, and $E = (\lceil C \rceil + 1)m^2 + m + 2 - \lceil C \rceil$. Then clearly $A = \max\{D, E\}$.

When $l_2 \ge (m+1)(2m+1)(m^2+m+1)+1 = (m+\frac{1}{2})(m+1)(2m^2+2m+2)+1 > m(m+1)(2m^2+2m+2) = \lfloor \frac{t}{2} \rfloor (\lfloor \frac{t}{2} \rfloor + 1)(|S_t|+1)$, by Theorem 3.4, there exists at least one solution of p and q that satisfies equation set (3.1). Then by Lemma 4.6, there exists a solution $p = p^*, q = q^*$ to (3.1) that satisfies either the condition $\frac{l_2}{2m+1} - \frac{\Delta_Q}{2} < q^* \le p^* < \frac{l_2}{2m+1} + \frac{\Delta_P}{2}$ or the condition $\frac{l_2}{2m+1} - \frac{\Delta_P}{2} \le p^* < q^* \le \frac{l_2}{2m+1} + \frac{\Delta_Q}{2}$. We analyze the two cases below.

- Case 1. There is a solution $p = p^*, q = q^*$ to equation set (3.1) that satisfies the condition $\frac{l_2}{2m+1} - \frac{\Delta_Q}{2} < q^* \le p^* < \frac{l_2}{2m+1} + \frac{\Delta_P}{2}$. We use Construction 3.1 to t-interleave an $(|S_t|+1) \times l_2$ torus G_1. Note that when $l_2 \ge (m+1)(2m+1)(m^2+m+1)+1$, $\frac{l_2}{2m+1} - \frac{\Delta_Q}{2} > 0$, so $q^* > 0$. Also note that $\frac{p^*}{q^*} < \frac{l_2/(2m+1)+\Delta_P/2}{l_2/(2m+1)-\Delta_Q/2} = B$, so $D \ge (\lceil \frac{p^*}{q^*} \rceil + 1)m^2 + 2m + 1$. Let G_2 be a $[\lceil \frac{D}{|S_t|+1} \rceil (|S_t|+1)] \times l_2$ torus obtained by tiling $\lceil \frac{D}{|S_t|+1} \rceil$ copies of G_1 vertically. We use Construction 4.1 to find a zigzag row in G_2; then by removing the zigzag row in G_2, we get a torus G_3 whose size is $[\lceil \frac{D}{|S_t|+1} \rceil (|S_t|+1) - 1] \times l_2$. Clearly the number of rows in G_1, $|S_t|+1$, and the number of rows in G_3, $\lceil \frac{D}{|S_t|+1} \rceil (|S_t|+1) - 1$, are relatively prime. So for any $l_0 \times l_2$ torus G where $l_0 \ge (|S_t|+1-1)(\lceil \frac{D}{|S_t|+1} \rceil (|S_t|+1) - 1 - 1) = |S_t|(\lceil \frac{D}{|S_t|+1} \rceil (|S_t|+1) - 2)$, it can be obtained by tiling copies of G_1 and G_3 vertically, and so by Lemma 4.3, G is t-interleaved, using $|S_t|+1$ colors.
- Case 2. There is a solution $p = p^*, q = q^*$ to equation set (3.1) that satisfies the condition $\frac{l_2}{2m+1} - \frac{\Delta_P}{2} \le p^* < q^* \le \frac{l_2}{2m+1} + \frac{\Delta_Q}{2}$. We use Construction 3.1 to t-interleave an $(|S_t|+1) \times l_2$ torus G_1. Note that when $l_2 \ge (m+1) \cdot$

$(2m+1)(m^2+m+1)+1$, $\frac{l_2}{2m+1} - \frac{\Delta_P}{2} > 0$, so $p^* > 0$. Also note that $\frac{q^*}{p^*} \le \frac{l_2/(2m+1)+\Delta_Q/2}{l_2/(2m+1)-\Delta_P/2} = C$, so $E \ge (\lceil \frac{q^*}{p^*} \rceil + 1)m^2 + m + (2 - \lceil \frac{q^*}{p^*} \rceil)$. Let G_2 be an $[\lceil \frac{E}{|S_t|+1} \rceil (|S_t|+1)] \times l_2$ torus obtained by tiling $\lceil \frac{E}{|S_t|+1} \rceil$ copies of G_1 vertically. We use Construction 4.2 to find a zigzag row in G_2; then by removing the zigzag row in G_2, we get a torus G_3 whose size is $[\lceil \frac{E}{|S_t|+1} \rceil (|S_t|+1) - 1] \times l_2$. Clearly the number of rows in G_1, $|S_t|+1$, and the number of rows in G_3, $\lceil \frac{E}{|S_t|+1} \rceil (|S_t|+1) - 1$, are relatively prime. So for any $l_0 \times l_2$ torus G where $l_0 \ge (|S_t|+1-1)(\lceil \frac{E}{|S_t|+1} \rceil (|S_t|+1) - 1 - 1) = |S_t|(\lceil \frac{E}{|S_t|+1} \rceil (|S_t|+1) - 2)$, it can be obtained by tiling copies of G_1 and G_3 vertically, and so by Lemma 4.3, G is t-interleaved, using $|S_t|+1$ colors.

Now let G be an $l_1 \times l_2$ torus, where $l_2 \ge (m+1)(2m+1)(m^2+m+1)+1$ and $l_1 \ge (2m^2+2m+1)(\lceil \frac{A}{2m^2+2m+2} \rceil (2m^2+2m+2) - 2) = |S_t|(\lceil \frac{\max\{D,E\}}{|S_t|+1} \rceil (|S_t|+1) - 2)$. Based on the analysis for Cases 1 and 2, we know that G's t-interleaving number is at most $|S_t|+1$. By the sphere-packing lower bound, G's t-interleaving number is at least $|S_t|$. So G's t-interleaving number is either $|S_t|$ or $|S_t|+1$. ∎

For easy reference, we show the method for optimally t-interleaving a large torus as a construction below. Note that the construction below is applicable only when $t \ge 5$ (and, by default, t is odd). When $t = 1$, any torus can be t-interleaved with 1 integer in a trivial way. When $t = 3$, the torus can be t-interleaved with the construction to be presented in Appendix I.

Construction 4.3. Optimal t-interleaving on a large torus.

Input: An odd integer t such that $t \ge 5$. An integer m such that $m = \frac{t-1}{2}$. An $l_1 \times l_2$ torus, where

$$l_2 \ge (m+1)(2m+1)(m^2+m+1)+1$$

and

$$l_1 \ge (2m^2+2m+1)\left(\left\lceil \frac{A}{2m^2+2m+2} \right\rceil (2m^2+2m+2) - 2\right).$$

The parameter A is as defined in Theorem 4.7.

Output: An optimal t-interleaving on the $l_1 \times l_2$ torus.

Construction:

1. If both l_1 and l_2 are multiples of $|S_t|$, then the $l_1 \times l_2$ torus' t-interleaving number is $|S_t|$. In this case, we use Construction 2.2 to t-interleave the $l_1 \times l_2$ torus with $|S_t|$ distinct integers.

2. If either l_1 or l_2 is not a multiple of $|S_t|$, then the $l_1 \times l_2$ torus' t-interleaving number is $|S_t|+1$. In this case, we t-interleave the torus with $|S_t|+1$ integers in the following way: First, we t-interleave an $(|S_t|+1) \times l_2$ torus, B, by using Construction 3.1 (note that $|S_t|+1 = 2m^2+2m+2$); second, we let H be an $[\lceil \frac{A}{|S_t|+1} \rceil (|S_t|+1)] \times l_2$ torus, which is obtained by tiling $\lceil \frac{A}{|S_t|+1} \rceil$ copies of B vertically, and use Construction 4.1 or Construction 4.2 (depending on which is applicable) to find a zigzag row in H; third, we remove the zigzag row in H to get a $[\lceil \frac{A}{|S_t|+1} \rceil (|S_t|+1) - 1] \times l_2$ torus T; and finally, we find nonnegative integers x and y such that $l_1 = x(|S_t|+1) + y[\lceil \frac{A}{|S_t|+1} \rceil (|S_t|+1) - 1]$ and get an $l_1 \times l_2$ torus by tiling x copies of B and y copies of T vertically. The resulting interleaving on the $l_1 \times l_2$ torus is a t-interleaving.

4.4. Optimal interleaving when t is even. When t is even, the optimal t-interleaving on large tori can be analyzed in a very similar way as in the case of odd

t. The main result for even t is shown in the following theorem. For succinctness, we leave the major steps and intermediate results of the corresponding analysis to Appendix II.

THEOREM 4.8. *Let t be a positive even integer. Let $m = \frac{t}{2}$. Define A as*

$$\max\Big\{ \Big(\Big\lceil \tfrac{2l_2+(m+1)(2m+1)(2m^2+1)}{2l_2-m(2m+1)(2m^2+1)} \Big\rceil + 1\Big)m^2 + \Big(3 - \Big\lceil \tfrac{2l_2+(m+1)(2m+1)(2m^2+1)}{2l_2-m(2m+1)(2m^2+1)} \Big\rceil\Big)m - 3,$$
$$\Big(\Big\lceil \tfrac{2l_2+m(2m+1)(2m^2+1)}{2l_2-(m+1)(2m+1)(2m^2+1)} \Big\rceil + 1\Big)m^2 + \Big(3 - \Big\lceil \tfrac{2l_2+m(2m+1)(2m^2+1)}{2l_2-(m+1)(2m+1)(2m^2+1)} \Big\rceil\Big)m - 1$$
$$- 2\Big\lceil \tfrac{2l_2+m(2m+1)(2m^2+1)}{2l_2-(m+1)(2m+1)(2m^2+1)} \Big\rceil \Big\}.$$

Then when

$$l_2 > \frac{(m+1)(2m+1)(2m^2+1)}{2}$$

and

$$l_1 \geq 2m^2 \left(\left\lceil \frac{A}{2m^2+1} \right\rceil (2m^2+1) - 2 \right),$$

the t-interleaving number of an $l_1 \times l_2$ (or $l_2 \times l_1$) torus is either $|S_t|$ or $|S_t|+1$.

5. General bounds on interleaving numbers. We have shown that for a torus whose size is large enough in both dimensions (Theorems 4.7 and 4.8), its t-interleaving number is at most $|S_t|+1$. If the requirement on the torus' size is loosened to some extent (Theorem 3.7), then its t-interleaving number is at most $|S_t|+2$. Does that mean that for a torus of any size its t-interleaving number is always at most $|S_t|$ plus a small constant? The answer is no. The following theorem shows bounds on t-interleaving numbers.

THEOREM 5.1. (1) *The t-interleaving numbers of two-dimensional tori are $|S_t| + O(t^2)$ in general. And that upper bound is tight, even if the number of rows or the number of columns of the torus approaches infinity.* (2) *When both l_1 and l_2 are of the order $\Omega(t^2)$, the t-interleaving number of an $l_1 \times l_2$ torus is $|S_t| + O(t)$.*

Proof. (1) First, let us show that the t-interleaving numbers of two-dimensional tori are $|S_t| + O(t^2)$ in general. Let G be an $l_1 \times l_2$ torus. First we assume that t is even and $l_1 \geq t$, $l_2 \geq t$. Let $K_1 = \lfloor \frac{l_1}{t} \rfloor$, $K_2 = \lfloor \frac{l_2}{t} \rfloor$. We see G as being tiled by small blocks in the way shown in Figure 5.1, where the blocks are labeled by A or B. Note

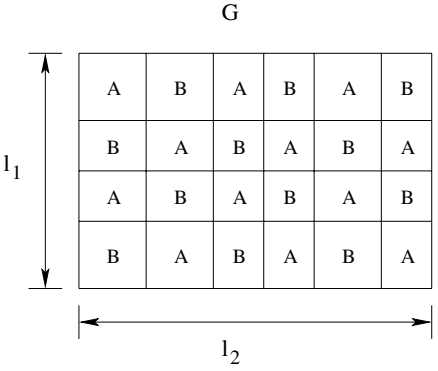


FIG. 5.1. *See G as being tiled by small blocks.*

that two blocks both labeled A are not necessarily of the same size, nor are two blocks both labeled B necessarily of the same size. For every block labeled as A (respectively, B), the four blocks around it (to its left, right, above, and below) are all labeled as B (respectively, A). Each block consists of either $\lceil \frac{l_1}{2K_1} \rceil$ or $\lfloor \frac{l_1}{2K_1} \rfloor$ rows and either $\lceil \frac{l_2}{2K_2} \rceil$ or $\lfloor \frac{l_2}{2K_2} \rfloor$ columns. Note that $\lceil \frac{l_1}{2K_1} \rceil = \lceil \frac{K_1 t + (l_1 \bmod t)}{2K_1} \rceil = \frac{t}{2} + \lceil \frac{l_1 \bmod t}{2K_1} \rceil$, $\lfloor \frac{l_1}{2K_1} \rfloor = \frac{t}{2} + \lfloor \frac{l_1 \bmod t}{2K_1} \rfloor$, $\lceil \frac{l_2}{2K_2} \rceil = \frac{t}{2} + \lceil \frac{l_2 \bmod t}{2K_2} \rceil$, and $\lfloor \frac{l_2}{2K_2} \rfloor = \frac{t}{2} + \lfloor \frac{l_2 \bmod t}{2K_2} \rfloor$. We see each block as a torus of its corresponding size. Thus for a block whose size is $\alpha \times \beta$, its vertices are denoted by (i, j) for $i = 0, 1, \ldots, \alpha - 1$ and $j = 0, 1, \ldots, \beta - 1$, just as a torus' vertices are normally denoted. Now we interleave all the blocks following these two rules: (i) only integers in the set $\{1, 2, \ldots, \lceil \frac{l_1}{2K_1} \rceil \cdot \lceil \frac{l_2}{2K_2} \rceil\}$ are used to interleave any block A, and only integers in the set $\{\lceil \frac{l_1}{2K_1} \rceil \cdot \lceil \frac{l_2}{2K_2} \rceil + 1, \lceil \frac{l_1}{2K_1} \rceil \cdot \lceil \frac{l_2}{2K_2} \rceil + 2, \ldots, 2 \cdot \lceil \frac{l_1}{2K_1} \rceil \cdot \lceil \frac{l_2}{2K_2} \rceil\}$ are used to interleave any block B; (ii) for all the blocks labeled by A (respectively, B) and for any i and j, the vertices denoted by (i, j) in them (provided they exist) all have the same color. It is very easy to see that G is t-interleaved in this way, using $2 \cdot \lceil \frac{l_1}{2K_1} \rceil \cdot \lceil \frac{l_2}{2K_2} \rceil = 2(\frac{t}{2} + \lceil \frac{l_1 \bmod t}{2K_1} \rceil)(\frac{t}{2} + \lceil \frac{l_2 \bmod t}{2K_2} \rceil) \leq 2(\frac{t}{2} + \lceil \frac{t-1}{2} \rceil)(\frac{t}{2} + \lceil \frac{t-1}{2} \rceil) = 2t^2 = |S_t| + \frac{3}{2}t^2$ distinct colors. So G's t-interleaving number is $|S_t| + O(t^2)$.

Now we assume that t is even and $l_1 < t$ or $l_2 < t$. Without loss of generality, let us say $l_1 < t$. Then we see G as being tiled horizontally by smaller tori $A_1, A_2, \ldots, A_n$, where each A_i (for $i = 1, 2, \ldots, n-1$) is an $l_1 \times t$ torus, and A_n is an $l_1 \times (l_2 \bmod t)$ torus. We interleave $A_1, A_2, \ldots, A_{n-1}$ in exactly the same way and assign $l_1 \times t$ distinct colors to each of them. We interleave A_n with a disjoint set of $l_1 \times (l_2 \bmod t)$ colors. Clearly G is t-interleaved in this way, using $l_1 \cdot t + l_1 \cdot (l_2 \bmod t) = |S_t| + O(t^2)$ distinct colors. So again, G's t-interleaving number is $|S_t| + O(t^2)$.

Finally we assume that t is odd. We can $(t+1)$-interleave G using $|S_{t+1}| + O((t+1)^2) = \frac{(t+1)^2}{2} + O((t+1)^2) = \frac{t^2+1}{2} + O(t^2) = |S_t| + O(t^2)$ distinct colors. $t+1$ is even, and a $(t+1)$-interleaving is also a t-interleaving. So G's t-interleaving number is still $|S_t| + O(t^2)$.

Now let us show that the above bound on t-interleaving numbers, $|S_t| + O(t^2)$, is tight, no matter whether t is even or odd. Consider an $l_1 \times l_2$ torus, where l_1 is the largest even integer that is no greater than $\lfloor \frac{3}{2}t \rfloor$ and l_2 is any integer greater than or equal to $\lfloor \frac{3}{4}t \rfloor$. We are first going to show that a t-interleaving can place a color at most twice in any $\lfloor \frac{3}{4}t \rfloor$ consecutive columns of the torus.

Assume that a t-interleaving places the same color on three vertices in $\lfloor \frac{3}{4}t \rfloor$ consecutive columns of the torus. Without loss of generality, let us say that those three vertices are $(0, 0)$, (a, b), and (c, d), where $0 \leq b \leq \lfloor \frac{3}{4}t \rfloor - 1$ and $0 \leq d \leq \lfloor \frac{3}{4}t \rfloor - 1$; see Figure 5.2. Since the interleaving is a t-interleaving, the Lee distance between any two of those three vertices is at least t. Let $e = \frac{l_1}{2}$ and $f = \lfloor \frac{3}{4}t \rfloor - 1$. It is not difficult to see that the Lee distance between (a, b) and (e, f) is at most $\min\{(e-a) \bmod l_1, (a-e) \bmod l_1\} + (f-b) = \frac{l_1}{2} - \min\{(0-a) \bmod l_1, (a-0) \bmod l_1\} + (f-b) = \frac{l_1}{2} + f - [\min\{(0-a) \bmod l_1, (a-0) \bmod l_1\} + b]$. Since the Lee distance between $(0, 0)$ and (a, b) is at most $\min\{(0-a) \bmod l_1, (a-0) \bmod l_1\} + b$, we know that $\min\{(0-a) \bmod l_1, (a-0) \bmod l_1\} + b \geq t$. Therefore the Lee distance between (a, b) and (e, f) is at most $\frac{l_1}{2} + f - t \leq \lfloor \frac{3}{2}t \rfloor / 2 + \lfloor \frac{3}{4}t \rfloor - 1 - t < \frac{t}{2}$. Similarly, the Lee distance between (c, d) and (e, f) is also less than $\frac{t}{2}$. Therefore the Lee distance between (a, b) and (c, d) is less than t, which is a contradiction. So a t-interleaving can place each color on at most two vertices in $\lfloor \frac{3}{4}t \rfloor$ consecutive columns of the torus.

Any $\lfloor \frac{3}{4}t \rfloor$ consecutive columns of the $l_1 \times l_2$ torus contain $l_1 \times \lfloor \frac{3}{4}t \rfloor \geq (\frac{3}{2}t - 2) \times$

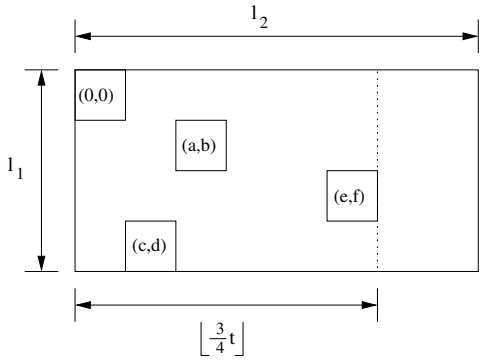


FIG. 5.2. *Four vertices in an $l_1 \times l_2$ torus.*

$(\frac{3}{4}t-1) = \frac{9}{8}t^2 - 3t + 2$ vertices, where each color is placed at most twice by a t-interleaving. Therefore the t-interleaving number of the torus is at least $\frac{\frac{9}{8}t^2-3t+2}{2} = \frac{9}{16}t^2 - \frac{3}{2}t + 1 = \frac{t^2+1}{2} + \frac{1}{16}t^2 - \frac{3}{2}t + \frac{1}{2} \geq |S_t| + \frac{1}{16}t^2 - \frac{3}{2}t + \frac{1}{2} = |S_t| + \Theta(t^2)$, which matches the upper bound $|S_t| + O(t^2)$. Since here l_2 can be *any* integer that is no less than $\lfloor \frac{3}{4}t \rfloor$, the upper bound is tight even if the number of columns (or equivalently, the number of rows) of the torus approaches infinity. The first part of this theorem has been proved by now.

(2) Let us prove the second part of this theorem. In the previous part of this proof, a method for t-interleaving an $l_1 \times l_2$ torus has been proposed for the case when t is even and $l_1 \geq t$, $l_2 \geq t$. That method uses $2(\frac{t}{2} + \lceil \frac{l_1 \bmod t}{2K_1} \rceil)(\frac{t}{2} + \lceil \frac{l_2 \bmod t}{2K_2} \rceil)$ colors. Note that $K_1 = \lfloor \frac{l_1}{t} \rfloor$ and $K_2 = \lfloor \frac{l_2}{t} \rfloor$. When both l_1 and l_2 are of the order $\Omega(t^2)$, both K_1 and K_2 are of the order of $\Omega(t)$, and then $2(\frac{t}{2} + \lceil \frac{l_1 \bmod t}{2K_1} \rceil)(\frac{t}{2} + \lceil \frac{l_2 \bmod t}{2K_2} \rceil) = 2(\frac{t}{2} + O(1))(\frac{t}{2} + O(1)) = \frac{t^2}{2} + O(t) = |S_t| + O(t)$. When t is odd, we can t-interleave an $l_1 \times l_2$ torus, where $l_1 = \Omega(t^2) = \Omega((t+1)^2)$ and $l_2 = \Omega(t^2) = \Omega((t+1)^2)$, by $(t+1)$-interleaving it using $|S_{t+1}| + O(t+1) = \frac{(t+1)^2}{2} + O(t) = \frac{t^2+1}{2} + O(t) = |S_t| + O(t)$ colors. So no matter whether t is even or odd, when both l_1 and l_2 are of the order $\Omega(t^2)$, the t-interleaving number of an $l_1 \times l_2$ torus is $|S_t| + O(t)$. ∎

6. Discussion. In this paper, we have studied the t-interleaving problem for two-dimensional tori. It has applications in both distributed data storage and burst error correction. This is the first time that the t-interleaving problem has been studied for graphs with modular structures, and consequently, novel interleaving methods different from traditional techniques (e.g., the widely used lattice-interleaver schemes in early works [8], [10], [17]) have been developed for optimal t-interleaving. The necessary and sufficient condition for tori that can be perfectly t-interleaved was proved, and the corresponding perfect t-interleaving construction was presented, based on the method of sphere-packing. The most important contribution of this paper is to prove that for tori whose sizes are large in both dimensions, which constitute by far the majority of all existing cases, their t-interleaving numbers are at most one more than the sphere-packing lower bound. Optimal t-interleaving constructions for such tori were presented, based on the method of removing-a-zigzag-row and tori-tiling. Then, some additional bounds on the t-interleaving numbers were shown. Those results together give a general characterization of the t-interleaving problem for two-dimensional tori.

The importance of the t-interleaving method based on removing-a-zigzag-row and tori-tiling is not limited to the results in Theorems 4.7 and 4.8. Those two theorems should be seen as a lower bound for the performance of the t-interleaving method. By analyzing the performance of the corresponding t-interleaving constructions more carefully, and furthermore, by keeping the main idea of the t-interleaving method but tuning its specific parameters on a case-by-case basis, we can improve the bounds derived in Theorems 4.7 and 4.8. The content of Appendix I can serve as an example in this regard. What is more, the t-interleaving method can be used to optimally t-interleave some tori whose sizes do not fall within the derived bounds.

We are interested in studying the t-interleaving problem for higher-dimensional tori, as well as finding more t-interleaving constructions. Those remain as our future research.

7. Appendix I. The optimal t-interleaving construction for odd t, Construction 4.3, if applicable only when $t \geq 5$. In this appendix, we present the optimal t-interleaving construction when $t = 3$, thus completing the result for t-interleaving on large tori while t is odd. We also use this case, $t = 3$, as an example to show how previous results can be improved if the t-interleaving problem is analyzed case by case and more carefully.

We will show that when $l_1 \geq 20$ and $l_2 \geq 15$ (or equivalently, when $l_1 \geq 15$ and $l_2 \geq 20$), an $l_1 \times l_2$ torus' 3-interleaving number is either 5 or 6. Note that $|S_3| = 5$. Below we present an construction that can optimally 3-interleave any $l_1 \times l_2$ torus where $l_1 \geq 20$ and $l_2 \geq 15$, except when $l_2 = 19$.

Construction 7.1. Optimally 3-interleave an $l_1 \times l_2$ torus, where $l_1 \geq 20$, $l_2 \geq 15$, and $l_2 \neq 19$.

1. If both l_1 and l_2 are multiples of 5, then the $l_1 \times l_2$ torus' 3-interleaving number is $|S_t| = 5$. In this case, 3-interleave the $l_1 \times l_2$ torus with five colors by using Construction 2.2.

If l_1 or l_2 is not a multiple of 5, then use steps 2–4 below to 3-interleave the $l_1 \times l_2$ torus with six colors.

2. Find nonnegative integers x_1 and x_2 such that $l_1 = 5x_1+6x_2$. Find nonnegative integers y_1, y_2, and y_3 such that $l_2 = 5y_1 + 8y_2 + 12y_3$.

3. There are six tori shown in Figure 7.1(a): a 5×5 torus A, a 5×8 torus B, a 5×12 torus C, a 6×5 torus A', a 6×8 torus B', and a 6×12 torus C'.

Get a $5 \times l_2$ torus M_1 by tiling horizontally y_1 copies of A, y_2 copies of B, and y_3 copies of C (whose order can be arbitrary).

Get a $6 \times l_2$ torus M_2 by tiling horizontally y_1 copies of A', y_2 copies of B', and y_3 copies of C', whose order needs to satisfy this rule: for $i = 1$ to $y_1 + y_2 + y_3$, if the ith module-torus in M_1 is an A (respectively, a B or a C), then the ith module in M_2 is an A' (respectively, a B' or a C').

4. Get an $l_1 \times l_2$ torus by tiling x_1 copies of M_1 and x_2 copies of M_2 (whose order can be arbitrary) vertically. The interleaving on the $l_1 \times l_2$ torus is a 3-interleaving.

Example 7.1. We use Construction 7.1 to 3-interleave an $l_1 \times l_2$ torus, where $l_1 = 11$ and $l_2 = 25$. l_1 is not a multiple of $|S_t|$, so the torus' 3-interleaving number is greater than 5. Since $l_1 = 5 + 6$ and $l_2 = 5 + 8 + 12$, the variables in Construction 7.1 can be set as follows: $x_1 = 1$, $x_2 = 1$, $y_1 = 1$, $y_2 = 1$, and $y_3 = 1$. Furthermore, we can let the torus M_1 have the form of $[ABC]$ and let the torus M_2 have the form of $[A'B'C']$. We then tile M_1 and M_2 to get the $l_1 \times l_2$ torus, which is of the form $\left[\begin{smallmatrix} A & B & C \\ A' & B' & C' \end{smallmatrix}\right]$. This 3-interleaved torus is shown in Figure 7.1(b). The interleaving used $6 = |S_3| + 1$ colors.

(a) Modules

A

0	2	4	1	3
1	3	0	2	5
2	5	1	3	0
4	0	2	5	1
5	1	3	0	2

B

0	2	4	0	3	5	1	3
1	3	5	2	4	0	2	5
2	4	1	3	5	1	4	0
3	0	2	4	0	3	5	1
5	1	3	5	2	4	0	2

C

0	2	5	1	4	0	3	5	2	4	1	3
1	3	0	2	5	1	4	0	3	5	2	4
2	5	1	4	0	3	5	2	4	1	3	0
4	0	3	5	2	4	1	3	0	2	5	1
5	1	4	0	3	5	2	4	1	3	0	2

A'

0	2	4	1	3
1	3	⑤	2	④
2	④	0	3	5
③	5	1	④	0
4	0	2	5	1
5	1	3	0	2

B'

0	2	4	0	3	5	1	3
1	3	5	①	4	0	2	④
2	4	⓪	2	5	1	③	5
3	⑤	1	3	0	②	4	0
④	0	2	4	①	3	5	1
5	1	3	5	2	4	0	2

C'

0	2	5	1	4	0	3	5	2	4	1	3
1	3	0	2	5	1	4	0	3	5	2	4
2	④	1	③	0	②	5	①	4	⓪	3	⑤
③	5	②	4	①	3	⓪	2	⑤	1	④	0
4	0	3	5	2	4	1	3	0	2	5	1
5	1	4	0	3	5	2	4	1	3	0	2

(b) Tiling of modules

0	2	4	1	3	0	2	4	0	3	5	1	3	0	2	5	1	4	0	3	5	2	4	1	3
1	3	0	2	5	1	3	5	2	4	0	2	5	1	3	0	2	5	1	4	0	3	5	2	4
2	5	1	3	0	2	4	1	3	5	1	4	0	2	5	1	4	0	3	5	2	4	1	3	0
4	0	2	5	1	3	0	2	4	0	3	5	1	4	0	3	5	2	4	1	3	0	2	5	1
5	1	3	0	2	5	1	3	5	2	4	0	2	5	1	4	0	3	5	2	4	1	3	0	2
0	2	4	1	3	0	2	4	0	3	5	1	3	0	2	5	1	4	0	3	5	2	4	1	3
1	3	5	2	4	1	3	5	1	4	0	2	4	1	3	0	2	5	1	4	0	3	5	2	4
2	4	0	3	5	2	4	0	2	5	1	3	5	2	4	1	3	0	2	5	1	4	0	3	5
3	5	1	4	0	3	5	1	3	0	2	4	0	3	5	2	4	1	3	0	2	5	1	4	0
4	0	2	5	1	4	0	2	4	1	3	5	1	4	0	3	5	2	4	1	3	0	2	5	1
5	1	3	0	2	5	1	3	5	2	4	0	2	5	1	4	0	3	5	2	4	1	3	0	2

FIG. 7.1. *Using modules for* 3*-interleaving.* (a) *The* 6 *modules,* (b) *tiling the modules.*

Clearly, since $25 = 5 \times 5 + 8 \times 0 + 12 \times 0$, another choice for tiling the 11×25 torus is $\begin{bmatrix} A & A & A & A & A \\ A' & A' & A' & A' & A' \end{bmatrix}$.

Construction 7.1 constructs a 3-interleaved $l_1 \times l_2$ torus by tiling copies of the six module-tori shown in Figure 7.1(a). It can be readily verified that when those six tori are tiled following the rule in Construction 7.1, the resulting interleaving on the $l_1 \times l_2$ torus is indeed a 3-interleaving. There are only a limited number of cases to analyze for the verification, so we skip the details. We comment that Construction 7.1 does not work for the case $l_2 = 19$, because 19 cannot be written as a linear combination of 5, 8, and 12 with nonnegative coefficients, and therefore an $l_1 \times 19$ torus cannot be obtained by tiling the module-tori. We present the construction for the case $l_2 = 19$ below.

Construction 7.2. Optimally 3-interleave an $l_1 \times 19$ torus, where $l_1 \geq 20$.

F

0	2	4	1	3	5	1	3	0	2	4	0	2	5	1	3	5	1	4
1	3	0	2	4	0	2	5	1	3	5	1	4	0	2	4	0	3	5
2	5	1	3	5	1	4	0	2	4	0	3	5	1	3	5	2	4	0
4	0	2	4	0	3	5	1	3	5	2	4	0	2	4	1	3	5	1
5	1	3	5	2	4	0	2	4	1	3	5	1	3	0	2	4	0	3

F'

0	2	4	⓪	3	5	1	3	⑤	2	4	0	2	④	1	3	5	1	4
1	3	⑤	1	4	0	2	④	0	3	5	1	③	5	2	4	0	②	5
2	④	0	2	5	1	③	5	1	4	0	②	4	0	3	5	①	3	0
③	5	1	3	0	②	4	0	2	5	①	3	5	1	4	⓪	2	4	1
4	0	2	4	①	3	5	1	3	⓪	2	4	0	2	⑤	1	3	5	②
5	1	3	5	2	4	0	2	4	1	3	5	1	3	0	2	4	0	3

FIG. 7.2. *Two modules used for* 3*-Interleaving an* $l_1 \times 19$ *torus, where* $l_1 \geq 20$.

Construction: Find nonnegative integers x_1 and x_2 such that $l_1 = 5x_1 + 6x_2$. There are two tori shown in Figure 7.2: a 5×19 torus F and a 6×19 torus F'. Construct an $l_1 \times 19$ torus by tiling x_1 copies of F and x_2 copies of F' vertically (whose order can be arbitrary). The resulting interleaving on the $l_1 \times 19$ torus is a 3-interleaving.

The correctness of Construction 4.5 can be easily verified, so we skip the details. Based on the previous two constructions, we readily get the following conclusion for 3-interleaving.

THEOREM 7.1. *When* $l_1 \geq 20$ *and* $l_2 \geq 15$*, or when* $l_1 \geq 15$ *and* $l_2 \geq 20$*, an* $l_1 \times l_2$ *torus' 3-interleaving number is either* $|S_3|$ *or* $|S_3| + 1$.

We comment that the result obtained here is comparatively better than the result derived in section 4. For example, if Theorem 4.7 is applied for the case $t = 3$, then the bound for l_2 would be 19, but here our bound for l_2 is 15. However, we should notice that the t-interleaving method used here is the same as the method used for $t > 3$ per se. We can see that the module-tori A, B, C in Figure 7.1(a) and F in Figure 7.2 are obtained by removing a zigzag row from A', B', C', and F'. The zigzag rows are shown in circles in those two figures. Both the interleaving method here and the method in section 4 are based on torus tiling. The improvement attained here is made by better tuning of construction parameters and more careful analysis of the bounds. The construction used for $t = 3$ does not follow all the requirements used in section 4. For example, the zigzag row in Figure 7.2 does not follow Rule 3. In section 4, while endeavoring to optimally tune all the parameters, we also need to ensure that the construction will work for all the cases of $t > 3$. If the interleaving problem is analyzed case by case (specifically, for each value of t, l_1, and l_2), the interleaving construction has room for further optimization.

8. Appendix II. In this appendix, we show how to optimally t-interleave large tori when t is even. The process is similar to the case where t is odd, differing only in details. For this reason, we just present a succinct description of the process and results. This appendix's content is parallel to that of the first three subsections of section 4, so comparative reading should help the understanding greatly.

We assume that t is even throughout the remainder of this appendix. The definitions of a *zigzag row* and *removing a zigzag row* are the same as in Definitions 4.1 and 4.2.

Let B be an $l_0 \times l_2$ torus which is t-interleaved by Construction 3.1 utilizing the offset sequence $S =$ "$s_0, s_1, \ldots, s_{l_2-1}$." Let H be an $l_1 \times l_2$ torus obtained by tiling several copies of B vertically. Let $m = \frac{t}{2}$. There are four rules to follow for devising a zigzag row (denoted by $\{(a_0, 0), (a_1, 1), \ldots, (a_{l_2-1}, l_2 - 1)\}$) in H:

- Rule 1. For any j such that $0 \le j \le l_2 - 1$, if the integers s_j, $s_{(j+1) \bmod l_2}$, $\ldots$, $s_{(j+m-1) \bmod l_2}$ do not all equal $t-1$, then $a_j \ge a_{(j+m) \bmod l_2} + m - 1$.
- Rule 2. For any j such that $0 \le j \le l_2 - 1$, if exactly one of the integers $s_j, s_{(j+1) \bmod l_2}, \ldots, s_{(j+m) \bmod l_2}$ equals t, then $a_j \le a_{(j+m+1) \bmod l_2} - (m-2)$.
- Rule 3. For any j such that $0 \le j \le l_2 - 1$, if $s_j = t - 1$, then $a_j \le a_{(j+1) \bmod l_2} - (2m-2)$.
- Rule 4. For any j such that $0 \le j \le l_2 - 1$, $2m - 2 \le a_j \le l_1 - 1 - (2m-2)$.

LEMMA 8.1. *Let B be a torus t-interleaved by Construction* 3.1. *Let H be a torus obtained by tiling copies of B vertically, and let T be a torus obtained by removing a zigzag row in H, where the zigzag row in H follows the four rules listed above. Let G be a torus obtained by tiling copies of B and T vertically. Then, both T and G are t-interleaved.*

Now we present two constructions for finding a zigzag row, which are the counterparts of Construction 4.1 and 4.2. Let B be an $l_0 \times l_2$ torus which is t-interleaved by Construction 3.1 utilizing the offset sequence $S =$ "$s_0, s_1, \ldots, s_{l_2-1}$." Let H be an $l_1 \times l_2$ torus obtained by tiling z copies of B vertically. We say the offset sequence S consists of p P's and q Q's, where $p > 0$ and $q > 0$. We require that in S the P's and Q's are interleaved very evenly, and that S starts with a P and ends with a Q. Let $m = \frac{t}{2}$. Let $L = (2m-2) + (m-1)\lceil \frac{p}{q} \rceil$ if $p \ge q$, and let $L = (2m-2) + (m-2)\lceil \frac{q}{p} \rceil + 1$ if $p < q$. We require that $l_1 \ge (\lceil \frac{p}{q} \rceil + 1)m^2 + (3 - \lceil \frac{p}{q} \rceil)m - 3$ if $p \ge q$ and that $l_1 \ge (\lceil \frac{q}{p} \rceil + 1)m^2 + (3 - \lceil \frac{q}{p} \rceil)m - (2\lceil \frac{q}{p} \rceil + 1)$ if $p < q$. Below we present two constructions for constructing a zigzag row, which is denoted by $\{(a_0, 0), (a_1, 1), \ldots, (a_{l_2-1}, l_2 - 1)\}$, in H, applicable respectively when $p \ge q$ and $p < q$.

Construction 8.1. Constructing a zigzag row in H, when t is even, $t > 2$, and $p \ge q > 0$.

1. Let $s_{x_1}, s_{x_2}, \ldots, s_{x_{p+q}}$ be the integers such that $0 = x_1 < x_2 < \cdots < x_{p+q} = l_2 - m - 1$ and each s_{x_i} $(1 \le i \le p+q)$ is the first element of a P or Q in the offset sequence S.

Let $a_{x_1} = L$. For $i = 2$ to $p+q$, if $s_{x_{i-1}}$ is the first element of a Q, let $a_{x_i} = L$.

For $i = 2$ to $p+q$, if $s_{x_{i-1}}$ is the first element of a P, then let $a_{x_i} = a_{x_{i-1}} - (m-1)$.

2. For $i = 2$ to m and for $j = 1$ to $p+q$, let $a_{x_j+i-1} = a_{x_j+i-2} + L - m + 1$.

3. Let $s_{y_1}, s_{y_2}, \ldots, s_{y_q}$ be the integers such that $y_1 < y_2 < \cdots < y_q = l_2 - 1$ and each s_{y_i} $(1 \le i \le q)$ is the last element of a Q in the offset sequence S.

For $i = 1$ to q, $a_{y_i} = L + (m-1)(L - m + 1) + (m-1)$.

Now we have fully determined the zigzag row, $\{(a_0, 0), (a_1, 1), \ldots, (a_{l_2-1}, l_2 - 1)\}$, in the torus H.

Construction 8.2. Constructing a zigzag row in H, when t is even, $t > 2$, and $0 < p < q$.

1. Let $s_{x_1}, s_{x_2}, \ldots, s_{x_{p+q}}$ be the integers such that $0 = x_1 < x_2 < \cdots < x_{p+q} = l_2 - m - 1$ and each s_{x_i} $(1 \le i \le p+q)$ is the first element of a P or Q in the offset sequence S.

Let $a_{x_1} = L$. For $i = 2$ to $p + q$, if s_{x_i} is the first element of a P, then let $a_{x_i} = L$; if $s_{x_{i-1}}$ is the first element of a P, then let $a_{x_i} = L - \lceil \frac{q}{p} \rceil (m-2) - 1$; otherwise, let $a_{x_i} = a_{x_{i-1}} + (m-2)$.

2. For $i = 2$ to m and for $j = 1$ to $p + q$, let $a_{x_j+i-1} = a_{x_j+i-2} + L - m + 1$.

3. Let $s_{y_1}, s_{y_2}, \ldots, s_{y_q}$ be the integers such that $y_1 < y_2 < \cdots < y_q = l_2 - 1$ and each s_{y_i} is the last element of a Q in the offset sequence S.

For $i = 1$ to q, $a_{y_i} = a_{y_i-1} + L - m + 1$.

Now we have fully determined the zigzag row, $\{(a_0, 0), (a_1, 1), \ldots, (a_{l_2-1}, l_2 - 1)\}$, in the torus H.

THEOREM 8.2. *The zigzag rows constructed by Constructions* 8.1 *and* 8.2 *follow all four rules: Rules* 1, 2, 3, *and* 4.

LEMMA 8.3. *In equation set* (3.2) (*which is in Construction* 3.1)*, let the values of* t, m, *and* l_2 *be fixed. Let* $p = p_0, q = q_0$ *be a solution that satisfies* (3.2)*. Then, another solution* $p = p_1, q = q_1$ *also satisfies* (3.2) *if and only if there exists an integer* c *such that* $p_1 = p_0 + c(m+1)(2m^2+1) \geq 0$ *and* $q_1 = q_0 - cm(2m^2+1) \geq 0$.

LEMMA 8.4. *In equation set* (3.2) (*which is in Construction* 3.1)*, let the values of* t, m, *and* l_2 *be fixed. Let* $\Delta_P = (m+1)(2m^2+1)$ *and* $\Delta_Q = m(2m^2+1)$*. If there exists a solution of* p *and* q *that satisfies* (3.2)*, then there exists a solution* $p = p^*, q = q^*$ *that satisfies not only* (3.2) *but also one of the following two inequalities:*

$$\frac{l_2}{2m+1} - \frac{\Delta_Q}{2} < q^* \leq p^* < \frac{l_2}{2m+1} + \frac{\Delta_P}{2}, \tag{8.1}$$

$$\frac{l_2}{2m+1} - \frac{\Delta_P}{2} \leq p^* < q^* \leq \frac{l_2}{2m+1} + \frac{\Delta_Q}{2}. \tag{8.2}$$

The above results lead to the main conclusion, Theorem 4.8.

We skip the specific construction of optimally t-interleaving large tori here, because of its similarity to Construction 4.3. But we present its sketch: If the torus can be perfectly t-interleaved, then it can be optimally t-interleaved using Construction 2.2. If the torus cannot be perfectly t-interleaved and $t \geq 4$, then it can be optimally t-interleaved using the tori-tiling method. The only remaining case is if the torus cannot be perfectly t-interleaved and $t = 2$. In that case, we can optimally t-interleave the torus (say it is an $l_1 \times l_2$ torus) using $|S_t| + 1 = 3$ distinct colors in the following way: First, interleave a ring of l_1 vertices and a ring of l_2 vertices using three colors (0, 1, and 2) such that no two adjacent vertices in those two rings are assigned the same color. Second, for $i = 1, 2, \ldots, l_1$ (respectively, for $i = 1, 2, \ldots, l_2$), use $I(i)$ (respectively, use $J(i)$) to denote the color assigned to the ith vertex in the ring of l_1 (respectively, l_2) vertices. Third, for $i = 0, 1, \ldots, l_1 - 1$ and $j = 0, 1, \ldots, l_2 - 1$, color the vertex (i, j) in the $l_1 \times l_2$ torus with color $(I(i+1) + J(j+1)) \bmod 3$. This yields an optimal 2-interleaving of the torus.

Acknowledgments. The authors thank the anonymous reviewers for their very careful and thoughtful comments.

REFERENCES

[1] K. A. S. ABDEL-GHAFFAR, *Achieving the Reiger bound for burst errors using two-dimensional interleaving schemes*, in Proceedings of the IEEE International Symposium on Information Theory, Germany, 1997, IEEE Press, Piscataway, NJ, p. 425.

[2] B. F. AlBdaiwi and B. Bose, *Quasi-perfect Lee distance codes*, IEEE Trans. Inform. Theory, 49 (2003), pp. 1535–1539.
[3] C. Almeida and R. Palazzo, *Two-dimensional interleaving using the set partition technique*, in Proceedings of the IEEE International Symposium on Information Theory, Trondheim, Norway, 1994, IEEE Press, Piscataway, NJ, p. 505.
[4] J. Astola, *An Elias-type bound for Lee codes over large alphabets and its applications to perfect codes*, IEEE Trans. Inform. Theory, 28 (1982), pp. 111–113.
[5] E. R. Berlekamp, *Algebraic Coding Theory*, Aegean Park Press, Walnut Creek, CA, 1984.
[6] M. Blaum and J. Bruck, *Correcting two-dimensional clusters by interleaving of symbols*, in Proceedings of the IEEE International Symposium on Information Theory, Trondheim, Norway, 1994, IEEE Press, Piscataway, NJ, p. 504.
[7] M. Blaum, J. Bruck, and P. G. Farrell, *Two-dimensional Interleaving Schemes with Repetitions*, Electrical Technical Report 016, Distributed Information Systems Group, California Institute of Technology, 1997; available online at http://www.paradise.caltech.edu/papers/etr016.pdf.
[8] M. Blaum, J. Bruck, and A. Vardy, *Interleaving schemes for multidimensional cluster errors*, IEEE Trans. Inform. Theory, 44 (1998), pp. 730–743.
[9] S. Borkar, R. Cohn, G. Cox, S. Gleason, T. Gross, H. T. Kung, M. Lam, B. Moore, C. Perterson, J. Pieper, L. Rankin, P. S. Tseng, J. Sutton, J. Urbanski, and J. Webb, *iWarp: An integrated solution to high-speed parallel computing*, in Proceedings of IEEE Supercomputing'88, Orlando, FL, 1988, IEEE Press, Piscataway, NJ, 1988, pp. 330–339.
[10] T. Etzion and A. Vardy, *Two-dimensional interleaving schemes with repetitions: Constructions and bounds*, IEEE Trans. Inform. Theory, 48 (2002), pp. 428–457.
[11] 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.
[12] A. Jiang and J. Bruck, *Diversity coloring for distributed data storage in networks*, IEEE Trans. Inform. Theory, 2003, submitted.
[13] A. Jiang and J. Bruck, *Multicluster interleaving on paths and cycles*, IEEE Trans. Inform. Theory, 51 (2005), pp. 597–611.
[14] A. Jiang and J. Bruck, *Network file storage with graceful performance degradation*, ACM Trans. Storage, 1 (2005), pp. 171–189.
[15] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-correcting Codes*, Elsevier Science, New York, 1977.
[16] R. J. McEliece, *The Theory of Information and Coding*, Cambridge University Press, Cambridge, UK, 2002.
[17] Y. Merksamer and T. Etzion, *On the optimality of coloring with a lattice*, in Proceedings of IEEE International Symposium on Information Theory, Chicago, 2004, IEEE Press, Piscataway, NJ, 2004, p. 21.
[18] W. Oed, *Massively Parallel Processor System CRAY T3D*, technical report, Cray Research Inc., Seattle, WA, 1993.
[19] M. Schwartz and T. Etzion, *Optimal 2-dimensional 3-dispersion lattices*, in Lecture Notes in Comput. Sci. 2643, Springer, NY, 2003, pp. 216–225.
[20] C. L. Seitz, W. C. Athas, K. M. Chandy, A. J. Martin, M. Rem, and S. Taylor, *Submicron Systems Architecture Project Semi-annual Technical Report*, Caltech-CS-TR-88-18, California Institute of Technology, Pasadena, CA, 1988.
[21] A. Slivkins and J. Bruck, *Interleaving schemes on circulant graphs with two offsets*, IEEE Trans. Inform. Theory, to appear; available online at http://www.paradise.caltech.edu/papers/etr054.pdf.
[22] Tera Computer Systems, *Overview of the Tera Parallel Computer*, technical report, 1993.
[23] D. B. West, *Introduction to Graph Theory*, Prentice–Hall, Englewood Cliffs, NJ, 1996.
[24] W. Xu and S. W. Golomb, *Optimal interleaving schemes for correcting 2-d cluster errors*, in Proceedings of IEEE International Symposium on Information Theory, Chicago, IEEE Press, Piscataway, NJ, 2004, p. 23.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 880–892

ON BUDGETED OPTIMIZATION PROBLEMS*

ALPÁR JÜTTNER†

Abstract. In this paper we give a method for solving certain budgeted optimization problems in *strongly polynomial* time. The method can be applied to several known budgeted problems, and in addition we show two new applications. The first one extends Frederickson's and Solis-Oba's result [G. N. Frederickson and R. Solis-Oba, *Combinatorica*, 18 (1998), pp. 503–518] to (poly)matroid intersections from single matroids. The second one is the budgeted version of the minimum cost circulation problem.

Key words. budgeted optimization, inverse problems, matroid intersections, submodular flows

AMS subject classifications. 90C31, 90C35

DOI. 10.1137/S0895480104445071

1. Introduction.

A typical optimization problem in combinatorial optimization consists of minimizing (or maximizing) a linear objective function over some combinatorial objects. Classical problems, like shortest paths, maximum flows, and minimum cost circulations can be interpreted this way.

There are, however, results dealing with other type of optimization problems over the same combinatorial objects. For example, it is well known that if we are given a directed graph $G = (V, E)$, a capacity function $w : E \longrightarrow \mathbb{R}$, and two nodes $s, t \in V$, then the minimum s-t cut can be found algorithmically in strongly polynomial time. Fulkerson [12] introduced and solved the so-called *budgeted version* of this minimum cut problem. In this case we are also given a cost function $c : E \longrightarrow \mathbb{R}$ and a budget constraint $B > 0$, and the task is to increase the amount of the minimum s-t cut as much as possible by increasing the components of the capacity function w individually. If we increase the capacity of an edge e by δ, it costs us $\delta c(e)$ and the total cost of increasing the capacities of the edges is bounded by B. Later Ahuja and Orlin [1] gave a more efficient polynomial algorithm for this problem.

Other papers have also been devoted to similar problems. Fulkerson and Harding [13] and Harding [16] solved the same budgeted version of the shortest s-t path problem. Later Frederickson and Solis-Oba solved the budgeted version of the minimum spanning tree problem [8]. In contrast with the budgeted versions of the minimum cut and the shortest s-t path problem—which were both transformed essentially to the minimum cost flow problem—the solution of the budgeted version of the minimum spanning tree problem needed deeper considerations. Later they extended their method to arbitrary matroids in [9, 10].

It would be natural to extend of this result to the problem of decreasing the maximum weight common base of two matroids. The algorithm presented in [10] does not seem to extend to this case. As a main contribution of this paper, we propose a different approach that can solve this latter problem in strongly polynomial time. In contrast with [10], this solution does not depend on any substantial property

*Received by the editors July 8, 2004; accepted for publication (in revised form) February 28, 2006; published electronically December 5, 2006. This research was supported by the Hungarian National Foundation for Scientific Research grant OTKA T 037547.
http://www.siam.org/journals/sidma/20-4/44507.html
†Department of Operations Research, Eötvös University, Pázmány Péter sétány 1/C, Budapest, Hungary H-1117, and Ericsson Traffic Laboratory, Irinyi J. u. 4-20, Budapest, Hungary H-1117 (alpar@cs.elte.hu).

of matroids or matroid intersections; thus the same scheme can be also used to solve other budgeted optimization problems including all the problems mentioned above.

Therefore it is worth considering the problem above in the following general form. (In order to be consistent with [9, 10], we use an equivalent form, where the problem is to decrease the weight of the maximum weight element.) We are given an underlying set E and a combinatorial optimization problem called the *basic problem* (i.e., the common bases of two given matroids in the latter example). We assume that the convex hull $\mathcal{P}$ of the feasible solution is a polyhedron. (This holds in the case of the problems mentioned above.) Then, for a given weight function $w : E \longrightarrow \mathbb{R}$, cost function $c : E \longrightarrow \mathbb{R}$, and budget constraint $B > 0$, the corresponding *budgeted optimization problem* is to compute the value

$$\alpha := \min\{\max\{(w - y)x : x \in \mathcal{P}\} : y \in \mathbb{R}^E, y \geq 0, cy \leq B\} \tag{1.1}$$

along with the minimizing vector y^*.

It will be pointed out that this problem essentially leads to the *bounded version* of the basic problem, that is, the problem of finding a maximum weight element in the polyhedron

$$\mathcal{P}^u := \mathcal{P} \cap \{x \in \mathbb{R}^E : x \leq u\}, \tag{1.2}$$

where $u \in \mathbb{R}^E$ is an arbitrary constraint vector. More exactly, the following will be proven.

THEOREM 1.1. *If there exists an algorithm which is able to find a maximum weight element in $\mathcal{P}^u$ along with a dual optimal solution (with respect to a linear programming description of $\mathcal{P}^u$) in time T for an arbitrary vector u and this algorithms satisfies the so-called* linearity condition *(see section* 2.1 *for the precise definition), then there exists an algorithm for solving problem* (1.1) *in time $O(T^2)$.*

Note bene: in this paper the word "*linear*" does not refer to the running time.

In section 3, problem (1.1) will be reduced to a parametric problem using Lagrangian relaxation; that is, we will show that problem (1.1) can be transformed to the maximization of the function

$$L(\lambda) := \max\{wx : x \in \mathcal{P}^{\lambda c}\} - \lambda B \tag{1.3}$$

with only one variable.

To do this maximization, one may use, e.g., the binary search technique. This gives the value of the optimal solution, but more effort is still necessary to find the optimal solution itself. In addition, the focus is on strongly polynomial time algorithms in this paper, and binary search does not give such an algorithm.

Thus, in section 4 we will give another way to construct an algorithm $\mathcal{A}'$ to maximize $L(\lambda)$ and to find an optimal solution to (1.1) in strongly polynomial time. The algorithm is based on Megiddo's parametric search method. This technique uses a separation subroutine, and not only does it consider this subroutine, but also it uses its *inner structure* to construct the algorithm $\mathcal{A}'$. Advantages of this technique are the straightforward bound on the running time and that it can be used in very general context (see, e.g., [20, 4]), but the separation subroutine must satisfy the linearity condition. This assumption is not a strong restriction in the sense that most combinatorial optimization problems that can be solved in strongly polynomial time can also be solved with linear algorithms. In section 2 we give a sketch of

Megiddo's technique in general, give the definition of linear algorithms, and show linear algorithms for some combinatorial optimization problems.

In section 5 a slightly more general problem will be discussed, when the modification of some components of the weight function can be prohibited.

In section 6 the presented method will be applied to some already solved problems, and we will show two new applications. The first new problem is the budgeted optimization problem of the minimum cost circulation problem, and the second one extends the Frederickson and Solis-Oba result [8, 9, 10] to matroid intersections from single matroids. Moreover, this method can be used for budgeted submodular optimization problems as well.

Finally, let us mention a similar class of budgeted optimization problem examined by Burkard, Klinz, and Zhang [2]. In this case the set $\mathcal{B} \subseteq \{0,1\}^E$; i.e., $\mathcal{B}$ is a family of subsets of E. A bottleneck-type objective function is used instead of a linear one, and the overall cost of an increment is defined in a rather general way. This cost model is able to handle linear and nonlinear cost functions such as, for example, componentwise increasing separable cost, maximum-like (time limit), or step-like cost function. (See [2] for more details.) This problem is also transformed to a parametric problem, which is solved using Megiddo's scheme.

2. Megiddo's principle.

DEFINITION 2.1. *A real number λ^* is said to be given by a* separation algorithm *$\mathcal{A}(\lambda)$ if $\mathcal{A}(\lambda)$ decides whether $\lambda < \lambda^*$, $\lambda = \lambda^*$, or $\lambda > \lambda^*$ by answering -1, 0, or $+1$.*

Solving combinatorial problems, one often comes across the problem of computing the explicit value of a number $\lambda^* \in \mathbb{R}$ given by a strongly polynomial time separation algorithm. Without some restrictions, only approximation-like algorithms can be given for this problem. For example, it is easy to see that for $\lambda^* := \sqrt{2}$ there exists a separation algorithm that uses only comparisons, additions, and multiplications of rational numbers and the input number, but λ^* cannot be obtained through these operations.

Megiddo [18] showed that, roughly, if one can avoid multiplications in the separation algorithm, then λ^* can also be computed in strongly polynomial time. Namely, he proved the following theorem.

THEOREM 2.2. *Suppose that we are given a $\lambda^* \in \mathbb{R}$ through a separation algorithm $\mathcal{A}(\lambda)$. If $\mathcal{A}$ is* linear in λ *and works in time T, then there exists an algorithm that runs in time $O(T^2)$ and computes the explicit value of λ^*.*

The idea of this method is to simulate the *steps* of the execution of $\mathcal{A}$ on the input λ^* by using only the necessary partial information about the input in each step of the algorithm. At the end of this procedure from this partial information we will be able to determine the right value of λ^*.

For this, however, it is necessary to require the linearity of the algorithm. The precise definition of this assumption comes in the next section; then the theorem will be proven in section 2.2.

2.1. Linear algorithms. To define the notion of a linear algorithm we use a RAM machine which has an additional storage called *limited access memory* (LAM). It may store real numbers, but an algorithm which runs on this machine has only a limited access to this storage. Namely, it can reach the contents of the LAM only through the following operations:

- It can write an element of the RAM or LAM into an element of the LAM,
- it can multiply an element of the LAM with an element of the RAM and store the result in the LAM,

- it can add an element of the LAM to another element of the LAM and store the result in the LAM,
- it can compare two elements of the LAM.

However, for example, it cannot multiply two elements of the LAM, and it cannot read them (that is it cannot copy an element of LAM into the RAM).

DEFINITION 2.3. *Let $\mathcal{A}(x, y)$ be an algorithm, where x and y are its input vectors. We say that $\mathcal{A}$ is linear in $\boldsymbol{x}$ if it gets x in the LAM and also puts the output in the LAM. The algorithm has full access to the other part of the input; in other words, it gets it in the RAM.*

It can be seen that most of the basic operations of data structures can be implemented on a LAM machine; for example, we can choose the minimal element of a set of numbers stored in the LAM, and we can also sort its elements. However, we cannot compute the determinant of a matrix with a LAM machine because we cannot avoid the multiplications of two elements of the matrix. (The determinant itself is a nonlinear polynomial of the elements of the matrix.)

The following claims are given to demonstrate the capability of linear algorithms.

CLAIM 2.1. *There exists an implementation of Dijkstra's algorithm to find a shortest s-t path in a weighted directed graph $G = (V, E, w)$ which is linear in the weight function.*

Proof. During its execution, Dijkstra's algorithm builds a shortest path tree T from the node s. First, T consists of only the node s. Then in each round it finds the node $v \in V \setminus V(T)$ whose path constructed from a path in T and one additional edge has the smallest weight, and it puts v and the last edge of the corresponding path towards T. This can be done, because we can calculate the weight of these paths in the LAM, and we can also choose the path having the smallest weight using only comparisons. □

CLAIM 2.2. *The strongly polynomial time Edmonds–Karp algorithm* [7] *for the MAX-FLOW problem can be implemented in such a way that it is linear in the capacity function.*

Proof. The algorithm starts with the zero flow, and it repeats the following steps until the maximum flow is found. First, it constructs an auxiliary digraph from the original graph by comparing the current edge-values of the flow with the edge-capacities and with zero. Then it looks for a shortest source-sink path in this graph and increases the amount of the flow by modifying its value on the edges of this paths, which can be done using a minimum calculation and some additions. □

CLAIM 2.3. *Goldberg and Tarjan's strongly polynomial time cycle canceling algorithm* [14] *for the minimum cost circulation problem can be implemented in such a way that it is linear in the cost function. It also can be implemented in such a way that it is linear in the bounding vectors. In addition, these algorithms compute the corresponding linear programming dual optimal solution.*

Proof. This algorithm starts with an arbitrary feasible circulation; then in each iteration an auxiliary digraph is constructed and the circulation is improved on the edges of a minimum mean cycle of the auxiliary graph. All this can be done without multiplying the elements of the cost function to each other and also without multiplying the elements of the bounding vectors to each other. □

It is worth mentioning that there cannot exist a strongly polynomial time algorithm for this problem which is linear *both* in the cost function and in the bounding vectors, because we cannot avoid multiplying the elements of the bounding vectors with their corresponding costs.

A common generalization of maximum cost circulations, maximum cost bases of a matroid, and maximum cost common base of two polymatroids is *submodular flows* [6]. One of its several equivalent definitions is the following. (See, e.g., [11].)

DEFINITION 2.4. *A family $\mathcal{F} \subseteq 2^V$ is called a* crossing family *over the underlying set V if for each crossing $X, Y \in \mathcal{F}$ (i.e., $X, Y \in \mathcal{F}$, $X \cap Y \neq \emptyset$, $X \setminus Y \neq \emptyset$, $Y \setminus X \neq \emptyset$, and $X \cup Y \neq V$) we have $X \cup Y$, $X \cap Y \in \mathcal{F}$. A function $b : \mathcal{F} \longrightarrow \mathbb{R}$ on the crossing family $\mathcal{F}$ is called* crossing-submodular *if for each crossing $X, Y \in \mathcal{F}$ we have the submodularity inequality*

$$b(X) + b(Y) \geq b(X \cup Y) + b(X \cap Y). \tag{2.1}$$

DEFINITION 2.5. *Let $G = (V, E)$ be a directed graph with a vertex set V and an edge set E. Let $f, g : E \longrightarrow \mathbb{R} \cup \{\pm\infty\}$ be a lower and an upper capacity function, and let $w : E \longrightarrow \mathbb{R}$ be a cost function. Let $\mathcal{F} \subseteq 2^V$ be a crossing family with $\emptyset, V \in \mathcal{F}$, and $b : \mathcal{F} \longrightarrow \mathbb{R}$ be a crossing-submodular function with $b(\emptyset) = b(V) = 0$. The* submodular flow problem *is described as follows:*

$$\max \quad wx \tag{2.2a}$$

$$f \leq x \leq g, \tag{2.2b}$$

$$\varrho_x(X) - \delta_x(X) \leq b(X) \quad \text{for all } X \subseteq V, \tag{2.2c}$$

where $\varrho_x(X)$ and $\delta_x(X)$ are the sum of the components of x corresponding to the edges entering X and leaving X, respectively.

The following claim is straightforward to check.

CLAIM 2.4. *The strongly polynomial time algorithm for the minimum cost submodular flow problem described in* [11] *can be implemented in such a way that it will be linear in the bounding functions f and g. This algorithm also gives the corresponding linear programming dual optimal solution.* □

2.2. Parametric search. In this section Theorem 2.2 is proven by giving an algorithm $\mathcal{A}'$ that computes the value of λ^*. The new idea of the parametric search developed by Megiddo [18] is that instead of making an independent "optimizer" algorithm that repeatedly *calls* the separation algorithm $\mathcal{A}$ during its execution, $\mathcal{A}'$ will "simulate" how $\mathcal{A}$ would run on the value λ^*. Therefore we now give a scheme of how to construct the algorithm $\mathcal{A}'$ from the linear separation algorithm $\mathcal{A}$.

First, we replace each element of the LAM with a pair of real numbers called *parametric numbers*. Addition and subtraction of these numbers are defined in the same way as the usual vector addition and vector subtraction, while the multiplications and divisions of these numbers are avoided by the linearity assumption on the algorithm $\mathcal{A}$. When a real number x is put into an element of the LAM, it is converted to the parametric number $(x, 0)$. The multiplication of a real number c and a parametric one (x, y) is defined to be (cx, cy).

The only undefined part of $\mathcal{A}'$ is when it compares two values in the LAM, namely when it inquires about whether $(x_1, y_1) \leq (x_2, y_2)$. In this case

- if $y_1 = y_2$, then we answer *true* if and only if $x_1 \leq x_2$;
- if $y_1 > y_2$, then we call $\mathcal{A}$ with the value $\lambda := \frac{x_2 - x_1}{y_1 - y_2}$; we answer *true* if and only if $\mathcal{A}$ returns that $\lambda^* < \lambda$;
- if $y_1 < y_2$, then we call $\mathcal{A}$ with the value $\lambda := \frac{x_2 - x_1}{y_1 - y_2}$. We answer *true* if and only if $\mathcal{A}$ returns that $\lambda^* > \lambda$.

If $\mathcal{A}$ happens to return that $\lambda = \lambda^*$, we stop, since we found λ^*.

The idea behind this construction is that a parametric number (x, y) means the linear expression $x + y\lambda^*$. The algebraic operations are defined to be consistent with the result of the same operations on the corresponding linear expressions.

A comparison $(x_1, y_1) \leq (x_2, y_2)$ means that $x_1 + y_1\lambda^* \leq x_2 + y_2\lambda^*$, which is equivalent to $\lambda^* \leq \frac{x_2 - x_1}{y_1 - y_2}$, $\lambda^* \geq \frac{x_2 - x_1}{y_1 - y_2}$ or $x_1 \leq x_2$, depending on the value of $\operatorname{sgn}(y_1 - y_2)$. So, we can make decision by a simple comparison or using the original $\mathcal{A}$ with $\lambda := \frac{x_2 - x_1}{y_1 - y_2}$ as input.

Now, let us run $\mathcal{A}'$ with the parametric number $(0, 1)$ as the input.

CLAIM 2.5. *$\mathcal{A}'((0,1))$ calls $\mathcal{A}$ with λ^* during its execution.*

Proof. Let us suppose that it is not so. In this case $\mathcal{A}'$ returns with a value when it terminates. $\mathcal{A}$ stored the return value originally in the LAM, so $\mathcal{A}'$ returns a parametric number, i.e., a linear function of λ.

During the execution each comparison involved in calling $\mathcal{A}$ gives us a half-line as a set of possible places of λ^*. Let Λ be the intersection of these closed half-lines. Λ is a (closed) interval and $\lambda^* \in \Lambda$. On the other hand, it is easy to see that for any $\lambda \in \Lambda$ the steps of $\mathcal{A}'((0,1))$ correspond to the steps of $\mathcal{A}(\lambda)$. So, the value returned by $\mathcal{A}'$ is right for all $\lambda \in \Lambda$. Moreover, it is a linear expression of λ, and for λ^* it is equal to 0. These, together with the fact that the result is ± 1 or 0 for all λ, yield that Λ should be equal to $\{\lambda^*\}$. However, this is only possible if $\mathcal{A}$ was called with λ^* during the execution of $\mathcal{A}'$, contradicting to the assumption. □

Finally, let T denote the running time of $\mathcal{A}$. The running time of the main algorithm is the sum of the time used by $\mathcal{A}'$ itself and the time required by the comparisons. Each comparison can be computed by a simple execution of $\mathcal{A}$, the number of the comparisons is at most T, and the number of the steps taken by $\mathcal{A}'$ is $O(T)$. Thus the total time required by the algorithm is $O(T^2)$.

In special cases the running time often can be significantly improved in several different ways. See [19] or [21] for more detail.

We also mention a theorem of Norton, Plotkin, and Tardos, which extends this result to any higher (but fixed) dimension.

THEOREM 2.6 (see [20]). *Let the closed convex set $\mathcal{P} \in \mathbb{R}^d$ be given through a separation algorithm which is linear in its input and runs in time T. Then there is an algorithm which in $O(T^{d+1})$ time either finds a point $x \in \mathcal{P}$ maximizing cx, or concludes that $\max\{cx : x \in \mathcal{P}\}$ is unbounded.* □

3. The Lagrangian relaxation of the problem. In this section we show how problem (1.1) can be transformed to a parametric problem. Let

$$L(\lambda) = \max_{z \in \mathcal{P}^{\lambda c}} wz - \lambda B \tag{3.1}$$

and

$$L^* := \max_{\lambda \geq 0} L(\lambda). \tag{3.2}$$

THEOREM 3.1. *$L^* = \alpha$, where α is the optimal solution of the budgeted optimization problem* (1.1).

Proof. From now on let $Ax \leq b$ be a linear description of $\mathcal{P}$; that is,

$$\mathcal{P} = \{x \in \mathbb{R}^n : Ax \leq b\}. \tag{3.3}$$

It is worth mentioning that the inequalities defining $\mathcal{P}$ need not be given explicitly, and there is no constraint on the number of them.

By the duality theorem,

$$\max_{x\in\mathcal{P}}(w-y)x = \min\{\pi b : \pi \geq 0, \pi A = w - y\}. \tag{3.4}$$

So, the problem (1.1) is equivalent to the following linear program:

$$\begin{aligned} \min \quad & \pi b \\ \pi, y &\geq 0, \\ \pi A + y &= w, \\ yc &\leq B. \end{aligned} \tag{3.5}$$

From (3.1) and (3.2) it follows that

$$\begin{aligned} L^* &= \max wz - \lambda B, \\ \lambda &\geq 0, \\ z &\leq \lambda c, \\ Az &\leq b, \end{aligned} \tag{3.6}$$

which is the dual problem of (3.5). □

4. Maximization of $L(\lambda)$. In this section we give an algorithm which maximizes the function $L(\lambda)$ and gives an optimal solution to (3.5).

First, using the duality theorem, we get that

$$\begin{aligned} L(\lambda) &= \min \pi b + \lambda(cy - B) \\ \pi, y &\geq 0, \\ \pi A + y &= w. \end{aligned} \tag{4.1}$$

Let L^{opt} denote the set of λ's maximizing $L(\lambda)$. Obviously L^{opt} is a (closed) interval. For the sake of simplicity the following notation is used.

DEFINITION 4.1. *$\lambda \leq L^{opt}$ if and only if $\lambda \leq x$ for all $x \in L^{opt}$. $\lambda \geq L^{opt}$, $\lambda < L^{opt}$, and $\lambda > L^{opt}$ are defined similarly.*

CLAIM 4.1. *Let π^0, y^0 be an optimal solution to* (4.1) *for some $\lambda \geq 0$. Then*

- *if $cy^0 > B$, then $\lambda \leq L^{opt}$;*
- *if $cy^0 < B$, then $\lambda \geq L^{opt}$.*

Proof. Let us suppose that $cy^0 > B$ and $\lambda' < \lambda$. Therefore

$$L(\lambda) = \pi^0 b + \lambda(cy^0 - B) > \pi^0 b + \lambda'(cy^0 - B) \geq L(\lambda'), \tag{4.2}$$

proving that $\lambda' \notin L^{opt}$. The second implication can be proven similarly. □

CLAIM 4.2. *Let π^0, y^0 be an optimal solution to* (4.1) *for some $\lambda \geq 0$, and let $cy^0 = B$. Then it is also an optimal solution to* (3.5); *that is, y^0 is an optimal decrement of the weight vector w in problem* (1.1).

Proof. The feasibility of π^0, y^0 is clear. Let π^*, y^* be an optimal solution to (3.5). Then

$$\alpha = \pi^* b \leq \pi^0 b = \pi^0 b + \lambda(cy^0 - B) \leq \pi^* b + \lambda(cy^* - B) \leq \pi^* b = \alpha \tag{4.3}$$

implies the optimality. □

If $L(\lambda) = -\infty$, that is, the polyhedron $\mathcal{P}^{\lambda c}$ is empty, a *certification of its emptiness* is a vector $(\pi^\neg, y^\neg)$ for which

$$\begin{aligned} \pi^\neg, y^\neg &\geq 0, \\ \pi^\neg A + y^\neg &= 0, \\ \pi^\neg b + y^\neg \lambda c &< 0. \end{aligned} \tag{4.4}$$

CLAIM 4.3. *Suppose that $\mathcal{P}^{\lambda c}$ is empty for some $\lambda \geq 0$, and let $\pi^\neg, y^\neg$ be a certification of its emptiness (i.e., a solution to* (4.4)*). Then*

- *if $cy^\neg \geq 0$, then $\lambda < L^{opt}$;*
- *if $cy^\neg \leq 0$, then $\lambda > L^{opt}$.*

Proof. $\pi^\neg, y^\neg$ is also certification of emptiness of $\mathcal{P}^{\lambda' c}$ for all $\lambda' \leq \lambda$ or for all $\lambda' \geq \lambda$, depending on whether $cy^\neg \geq 0$. ∎

Now, we are ready to prove our original theorem.

THEOREM 1.1 (Restated). *Let $\mathcal{A}(\lambda)$ be an algorithm linear in λ and with running time T, which computes $L(\lambda)$ along with a dual optimal solution* (4.1) *or a certification of the emptiness of $\mathcal{P}^{\lambda c}$ if $L(\lambda) = -\infty$. Then there exists an algorithm for solving problem* (1.1) *in time $O(T^2)$.*

Proof. Claim 4.2 shows that to solve problem (1.1) it is enough to find a multiplier λ^* maximizing the function $L(\lambda)$ and an optimal solution π^*, y^* to (4.1) with λ^* in place of λ with the property that $cy^* = B$.

We apply Megiddo's parametric search for the maximization of the function $L(\lambda)$, but a small technical difficulty arises because we are not able to check directly whether or not $\lambda \in I\!\!E^{pt}$ for an arbitrary λ.

Apart from the comparisons, we construct $\mathcal{A}'$ from $\mathcal{A}$ in the same way as in section 2.

The only difference occurs when the algorithm makes a comparison which is involved in calling $\mathcal{A}$ with the input $\lambda := \frac{x_2 - x_1}{y_1 - y_2}$. In this case we do the following:

- If we get $L(\lambda) = -\infty$, then we use Claim 4.3 to decide whether $\lambda < \lambda^*$ or $\lambda > \lambda^*$.
- If $\mathcal{A}$ happens to consider λ to be optimal, that is, we get a solution π^λ, y^λ of (4.1) such that $cy^\lambda = B$, then by Claim 4.2, (π^λ, y^λ) is also an optimal solution to (3.5), and so the execution can be finished.
- If $cy^\lambda > B$, then by Claim 4.1 it means that $\lambda < I\!\!E^{pt}$ unless $\lambda \in I\!\!E^{pt}$, so we accept if the question was $\lambda \leq \lambda^*$ and reject if the question was $\lambda \geq \lambda^*$.
- In the case when $cy^\lambda < B$ we give opposite answers.

For some technical reasons, the algorithm also has to take care to avoid the redundant executions of $\mathcal{A}$; i.e., if the outcome of a comparison can be derived from the outcomes of the previous ones, then we make decision based on the previous comparisons rather then on calling $\mathcal{A}$ once more.

Now let us run $\mathcal{A}'$. During the execution, each comparison which is involved in calling $\mathcal{A}$ gives us a half-line as a set of possible values of λ^*. Let Λ be the intersection of these closed half-lines. Λ is a (closed) interval and $I\!\!E^{pt} \subseteq \Lambda$. Since we never made redundant comparisons, $\mathrm{int}\Lambda \neq \emptyset$. ($\mathrm{int}\Lambda$ is the set of the interior points of Λ.)

On the other hand, it is easy to see that for any $\lambda \in \mathrm{int}\Lambda$ the steps of $\mathcal{A}'$ correspond to the steps of $\mathcal{A}(\lambda)$. So, at the end of its execution $\mathcal{A}'$ gives us the optimum value, which parametric number. Thus, it corresponds to a linear function of λ and is equal to $L(\lambda)$ for all $\lambda \in \mathrm{int}\Lambda$. Moreover, because of the continuity of $L(\lambda)$, this holds for all $\lambda \in \Lambda$. $\mathcal{A}'$ also gives us an optimal solution $(\pi(\lambda), y(\lambda))$ to (4.1) as two vectors

of linear functions of λ. These are right optimal solutions to (4.1) for all $\lambda \in \text{int}\Lambda$. Again, because of continuity, it follows that they are right for all $\lambda \in \Lambda$.

From the above considerations it follows that one of the extrema of Λ maximizes the function $L(\lambda)$. Denote it by λ^o and let (π^o, y^o) be the solution $\mathcal{A}$ returned when it was called with λ^o. Because Λ is a subset of the half-line defined by y^o and because λ^o maximizes $L(\lambda)$ over the set Λ, it follows that $cy^o - B$ and $cy(\lambda^o) - B$ have opposite signs. So, there exists a suitable coefficient $0 \leq \mu \leq 1$ such that $cy^* = B$, where $\pi^* := \mu\pi^o + (1-\mu)\pi(\lambda^o)$ and $y^* = \mu y^o + (1-\mu)y(\lambda^o)$. Since both (π^o, y^o) and $(\pi(\lambda^o), y(\lambda^o))$ are optimal solutions to (4.1) with λ^o in place of λ, (π^*, y^*) is also an optimal solution. Finally, Claim 4.2 ensures that (π^*, y^*) is an optimal solution to (3.5) as well. To sum up, y^* is an optimal solution to problem (1.1).

The bound on the running time can be obtained in the same way used in section 2.2. □

Remark 1. Using Theorem 2.6, Norton, Plotkin, and Tardos also proved the following.

THEOREM 4.2 (see [20]). *Suppose that for a certain vector a and matrix A there exists an algorithm to solve the linear program* $\max\{ax : Ax \leq b\}$ *for arbitrary b, that runs in time t, and is linear in b. Then for any fixed d, there is an algorithm which runs in $O(t^{d+2})$ time and solves the linear program* $\max\{ax + cz : Ax + Cz \leq b\}$ *for any vector c and matrix C with d columns.*

Although this theorem could be applied to (3.6) to compute the value L^*, a self-contained proof was given for two reasons. First, obtaining the dual solution is very essential to us, but in [20] the authors do not deal with this problem. Second, it reduces the running time from $O(T^3)$ to $O(T^2)$.

On the other hand, a straightforward extension of the above algorithm gives an alternative proof to Theorem 4.2 when $d = 1$ in time t^2. Moreover, it extends to arbitrary d. The resulting algorithm runs in time $O(t^{d+1})$, which improves the running time $O(t^{d+2})$ presented in [20]. See [17] for more details.

5. When some components of the cost are fixed. In this section we extend the problem (1.1), so that the modification of some components of the weight function can be prohibited. The main benefit of this extension is that it enables us to use auxiliary variables to define the polyhedron $\mathcal{P}$ for the basic problem.

Let $\mathcal{P} := \{(x_1, x_2) \in \mathbb{R}^{n+m} : Ax_1 + Cx_2 \leq b\}$. We are looking for the optimal solution to the problem

$$\alpha := \min\left\{\max\left\{(w_1 - y)x_1 + w_2x_2 : (x_1, x_2) \in \mathcal{P}\right\} : y \geq 0, cy \leq B\right\}. \tag{5.1}$$

As in section 3 this can be transformed to the following linear program:

$$\begin{aligned} \min \quad & \pi b \\ \pi, y &\geq 0, \\ \pi A + y &= w_1, \\ \pi B &= w_2, \\ yc &\leq B. \end{aligned} \tag{5.2}$$

The corresponding Lagrangian relaxation of this problem is

$$
\begin{aligned}
L(\lambda) = \min \quad & \pi b + \lambda(cy - B) \\
& \pi, y \geq 0, \\
& \pi A + y = w_1, \\
& \pi B = w_2,
\end{aligned}
\tag{5.3}
$$

and by the duality theorem,

$$
L(\lambda) = \max\{w_1 z_1 + w_2 z_2 : (z_1, z_2) \in \mathcal{P}^\lambda\} - \lambda B, \tag{5.4}
$$

where $\mathcal{P}^\lambda := \{(x_1, x_2) \in \mathcal{P} : x_1 \leq \lambda c\}$. The following theorem can be proven similarly to Theorem 3.1.

THEOREM 5.1. *$L^* = \alpha$, where*

$$
L^* := \max_{\lambda \geq 0} L(\lambda). \qquad \square \tag{5.5}
$$

Finally, the maximization of $L(\lambda)$ can be done in the same way as in section 4.

6. Applications. The algorithm presented in the previous sections can be used for a wide range of problems. It is enough to check whether we are able to optimize on the corresponding bounded polyhedron using an algorithm which is linear in the bounding vector (or more generally, using an algorithm that uses λc as the bounding vector and which is linear in λ).

First, as an example, we show how Theorem 1.1 can be applied to Fulkerson's and Harding's problem. Then, sections 6.2 and 6.3 present two new applications, the budgeted minimum cost circulation and the budgeted polymatroid intersection problem. The later one extends the problem examined in [10].

Let us mention that the direct use of Theorem 1.1 gives worse running time than those obtained in [1, 13] and in [10] for the budgeted maximum flow, minimum source-sink path, and matroid optimization problems. However, these running times can be improved in several ways in special cases. See [19] and [21] for these techniques.

6.1. Maximizing the minimum source-sink path. Fulkerson and Harding [13] and Harding [16] solved the case of budgeted optimization problems when we have a non-negative length and a cost function on the edges of a directed graph and we want to increase the length of the minimum length path between two predefined nodes as much as possible by increasing the lengths of the edges keeping the budget constraint.

The minimum length s-t path problem can be formulated as an uncapacitated minimum cost flow problem. So, the computation of $L(\lambda)$ is a capacitated minimum cost flow problem, for which there exists a strongly polynomial time algorithm which is linear in its constraint vector. It also gives back the dual solution which is in the suitable form we need in (4.1).

6.2. Increasing the cost of the minimum cost circulation. We are given a directed graph $G = (V, E)$, a cost function $c : E \longrightarrow \mathbb{R}$, and a lower and an upper bound $l, u : E \longrightarrow \mathbb{R}$ on its edges. Moreover, we are given a cost function $c^m : E \longrightarrow \mathbb{R}$ of the modification and a budget constraint $B > 0$. The task is to find an increment of the cost function c within the budget constraint which increases the cost of the c-minimal cost circulation as much as possible.

This problem can also be handled with the method because the corresponding parametric problem is a minimum cost circulation problem on the graph G with the cost function c, lower bound l, and upper bound $u^\lambda(e) := \min(u(e), \lambda c^m(e))$, which can be computed by a strongly polynomial time algorithm which is linear in λ. The dual optimal solution returned by this algorithm can be easily transformed to a solution of (4.1) in the same way as is discussed in the following section.

6.3. Polymatroid intersection and submodular flows. A natural extension of the problem examined in [10] is the *budgeted matroid intersection problem*. In this case we are given two matroids $M_1 = (E, \mathcal{I}_1)$ and $M_2 = (E, \mathcal{I}_2)$ with a common ground set, a weight and a cost function $w, c : E \longrightarrow \mathbb{R}$, and a budget constraint $B > 0$, and we want to decrease the weight of the maximum weight common base of M_1 and M_2 as much as possible by decreasing independently the weight of the elements of the ground set with the side constraint that the total cost of the decreasing must be at most B.

The common generalization of this problem and the problem presented in section 6.2 is the case of a budgeted optimization problem when we are given a submodular flow problem and we are looking for a modification of its cost function which increases the cost of the minimum cost feasible flow as much as possible.

Namely, using the notation of Definition 2.5, we are given an additional cost function $c : E \longrightarrow \mathbb{R}$ and a budget constraint B, and the problem is to find

$$\min\{\max\{(w-y)x : x \in \mathcal{P}\} : y \geq 0, cy \leq B\}, \tag{6.1}$$

where

$$\mathcal{P} := \{x \in \mathbb{R}^E : f \leq x \leq g \ , \ \varrho_x(X) - \delta_x(X) \leq b(X) \quad \text{for all } X \subseteq V\}. \tag{6.2}$$

The corresponding bounded problem is

$$\max \quad wx \tag{6.3a}$$

$$\varrho_x(X) - \delta_x(X) \leq b(X) \quad \text{for all } X \subseteq V, \tag{6.3b}$$

$$x \leq -f, \tag{6.3c}$$

$$x \leq g, \tag{6.3d}$$

$$x \leq u, \tag{6.3e}$$

where f and g are the lower and the upper capacities and u is the bounding vector.

This is also a submodular flow problem with $g^u(e) := \min(g(e), u(e))$ in place of g. Using Claim 2.4, we get a strongly polynomial time algorithm that is linear in u and computes an optimal flow x and an optimal dual solution (π, z_1, z_2), that is, a solution to

$$\min \sum_{X \subseteq V} \pi_X b(X) - z_1 f + z_2 g^u, \tag{6.4a}$$

$$\pi \in \mathbb{R}^{2^V}, \quad z_1, z_2 \in \mathbb{R}^E, \tag{6.4b}$$

$$\pi, z_1, z_2 \geq 0, \tag{6.4c}$$

$$\sum_{X : e \in \delta(X)} \pi_X - \sum_{X : e \in \varrho(X)} \pi_X - z_1(e) + z_2(e) = w(e) \quad \text{for all } e \subseteq E, \tag{6.4d}$$

where π_X denotes the component of π corresponding to the subset $X \subseteq V$.

Although π is an exponential-size vector, the optimal solution computed by the algorithm consists of at most $|E|$ nonzero elements and is given by the set of nonzero coordinates and their corresponding values.

Finally, (π, z_1, z_2) can be transformed into a dual optimal solution (π, ν_1, ν_2, y) to (6.3), where

$$\nu_1 := z_1, \tag{6.5a}$$

$$\nu_2(e) := \begin{cases} z_2(e) & \text{if } x(e) = g(e), \\ 0 & \text{otherwise,} \end{cases} \tag{6.5b}$$

$$y(e) := \begin{cases} z_2(e) & \text{if } x(e) < g(e), \\ 0 & \text{otherwise.} \end{cases} \tag{6.5c}$$

To sum up, we get a strongly polynomial time algorithm for the bounded version of the submodular flow problem, so Theorem 1.1 provides a strongly polynomial time algorithm for the budgeted submodular flow problem (problem (6.1)).

7. An open problem. Finally, we pose a natural problem to which no strongly polynomial time algorithm is known.

The *budgeted maximum matching problem* is the following. We are given a graph $G = (V, E)$, a cost and a weight function $c, w : E \longrightarrow \mathbb{R}$ on its edges, and a budget constraint B, and we are looking for a vector $z \geq 0$, $cz \leq B$ which minimizes the weight of the $(w - z)$-maximal weight matching.

If G is a bipartite graph, then the problem can be reduced either to the budgeted minimum cost circulation problem or to the budgeted matroid intersection problem, so it can be solved in strongly polynomial time, but the general case is still open.

According to Theorem 1.1 it would be enough to give a linear strongly polynomial time algorithm to the following *bounded maximum weight matching problem.* We are given a graph $G = (V, E)$, a weight, and a capacity function $w, u : E \longrightarrow \mathbb{R}$ on its edges, and the problem is to find

$$\max\{wx : x \in \mathcal{P}, x \leq u\}, \tag{7.1}$$

where $\mathcal{P}$ is a matching polyhedron, that is, the convex hull of the incidence vectors of the matchings of G.

Acknowledgment. The author would like to thank András Frank and Zsuzsa Weiner for their several very useful comments.

REFERENCES

[1] R. K. Ahuja and J. B. Orlin, *A capacity scaling algorithm for the constrained maximum flow problem*, Networks, 25 (1995), pp. 89–98.

[2] R. E. Burkard, B. Klinz, and J. Zhang, *Bottleneck capacity expansion problems with general budget constraints*, RAIRO Oper. Res., 35 (2001), pp. 1–20.

[3] E. Cheng and W. H. Cunningham, *A faster algorithm for computing the strength of a network*, Inform. Process. Lett., 49 (1994), pp. 209–212.

[4] E. Cohen and N. Megiddo, *Strongly polynomial-time and NC algorithms for detecting cycles in dynamic graphs*, J. ACM, 40 (1993), pp. 791–832.

[5] W. H. Cunningham, *Optimal attack and reinforcement of a network*, J. ACM, 32 (1985), pp. 549–561.

[6] J. Edmonds and R. Giles, *A min-max relation for submodular functions on graphs*, Ann. Discrete Math., 1 (1977), pp. 185–204.

[7] J. Edmonds and R. M. Karp, *Theoretical improvements in algorithmic efficiency for network flow problems*, J. ACM, 19 (1972), pp. 248–264.

[8] G. N. Frederickson and R. Solis-Oba, *Increasing the weight of minimum spanning trees*, J. Algorithms, 33 (1999), pp. 244–266.
[9] G. N. Frederickson and R. Solis-Oba, *Efficient algorithms for robustness in matroid optimization*, in Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, 1997, SIAM, Philadelphia, 1997, pp. 659–668.
[10] G. N. Frederickson and R. Solis-Oba, *Algorithms for measuring perturbability in matroid optimization*, Combinatorica, 18 (1998), pp. 503–518.
[11] S. Fujishige, *Submodular Functions and Optimization*, Elsevier Science, New York, 1991.
[12] D. R. Fulkerson, *Increasing the capacity of a network: The parametric budget problem*, Management Sci., 5 (1959), pp. 472–483.
[13] D. R. Fulkerson and G. C. Harding, *Maximizing the minimum source-sink path subject to a budget constraint*, Math. Programming, 13 (1975), pp. 116–118.
[14] A. V. Goldberg and R. E. Tarjan, *Finding minimum-cost circulations by canceling negative cycles*, J. ACM, 36 (1989), pp. 873–886.
[15] M. Grötschel, L. Lovász, and A. Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 169–197.
[16] G. C. Harding, *Some Budgeted Optimization Problems and the Edge Disjoint Branchings Problem*, Ph.D. dissertation, Cornell University, Ithaca, NY, 1977.
[17] A. Jüttner, *Optimization with additional variables and constraints*, Oper. Res. Lett., 33 (2005), pp. 305–311.
[18] N. Megiddo, *Combinatorial optimization with rational objective functions*, Math. Oper. Res., 4 (1979), pp. 414–424.
[19] N. Megiddo, *Applying parallel computation algorithms in the design of serial algorithms*, J. ACM, 30 (1983), pp. 852–865.
[20] C. H. Norton, S. A. Plotkin, and É. Tardos, *Using separation algorithms in fixed dimension*, J. Algorithms, 13 (1992), pp. 79–98.
[21] T. Radzik, *Fractional combinatorial optimization*, in Handbook of Combinatorial Optimization, Vol. 1, D. Z. Du and P. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.
[22] É. Tardos, *A strongly polynomial minimum cost circulation algorithm*, Combinatorica, 5 (1985), pp. 247–255.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 893–912

$\{0, \frac{1}{2}\}$-CUTS AND THE LINEAR ORDERING PROBLEM: SURFACES THAT DEFINE FACETS*

SAMUEL FIORINI†

Abstract. We find new facet-defining inequalities for the linear ordering polytope generalizing the well-known Möbius ladder inequalities. Our starting point is to observe that the natural derivation of the Möbius ladder inequalities as $\{0, \frac{1}{2}\}$-cuts produces triangulations of the Möbius band and of the corresponding (closed) surface, the projective plane. In that sense, Möbius ladder inequalities have the same "shape" as the projective plane. Inspired by the classification of surfaces, a classic result in topology, we prove that a surface has facet-defining $\{0, \frac{1}{2}\}$-cuts of the same "shape" if and only if it is nonorientable.

Key words. linear ordering problem, $\{0, \frac{1}{2}\}$-cut, surface, cyclic order, matching theory

AMS subject classifications. 05C70, 05C20, 05C62, 05C10, 90C57

DOI. 10.1137/S0895480104440985

1. Introduction. Let X be a finite set of cardinality $n \geq 3$, and let $D_n = (X, A_n)$ denote a complete digraph with node set X and arc set A_n. Given nonnegative weights w_{ij} for each arc $ij \in A_n$, the *minimum linear ordering problem (MIN-LOP)* is to find a linear order $\preccurlyeq$ on X whose total weight $\sum_{i \prec j} w_{ij}$ is minimum. The *maximum linear ordering problem (MAX-LOP)* is defined similarly. Both problems are strongly NP-hard [14]. Because a linear order $\preccurlyeq$ is an optimum solution of a MIN-LOP instance if and only if its reverse $\succcurlyeq$ is an optimum solution of the MAX-LOP instance with the same weights, both problems are equivalent as regards exact algorithms. Nevertheless, computing approximate solutions seems to be easier for MAX-LOP [22] than for MIN-LOP [25]. Note that MIN-LOP is essentially the *minimum dicycle cover problem* (which is also known as the *minimum feedback arc set problem*), and MAX-LOP is essentially the *maximum acyclic subgraph problem*. Henceforth, we mainly focus on MIN-LOP and prefer to regard the linear ordering problem as a minimization problem. The standard formulation of MIN-LOP as an integer programming problem has one variable x_{ij} per arc $ij \in A_n$, with $x_{ij} = 1$ if $i \prec j$ and $x_{ij} = 0$ otherwise, and reads

$$\text{minimize} \quad \sum_{ij \in A_n} w_{ij} x_{ij}$$

$$\text{subject to} \quad x_{ij} \geq 0 \qquad \forall ij \in A_n, \tag{1.1}$$

$$x_{ij} + x_{jk} + x_{ki} \geq 1 \qquad \forall ij, jk, ki \in A_n, \tag{1.2}$$

$$x_{ij} + x_{ji} = 1 \qquad \forall ij \in A_n, \tag{1.3}$$

$$x_{ij} \in \mathbb{Z} \qquad \forall ij \in A_n. \tag{1.4}$$

The standard formulation of MAX-LOP as an integer program is identical to the one above, except that the goal is to maximize and that constraints (1.1) and (1.2) are

*Received by the editors February 11, 2004; accepted for publication (in revised form) March 30, 2006; published electronically December 5, 2006.
http://www.siam.org/journals/sidma/20-4/44098.html

†Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Current address: Département de Mathématique, Université Libre de Bruxelles, CP 216, Boulevard du Triomphe, B-1050 Brussels, Belgium (sfiorini@ulb.ac.be). This work was supported by a Fellowship of the Belgian American Educational Foundation and the Fonds National de la Recherche Scientifique.

usually written in an equivalent form, as $x_{ij} \leq 1$ and $x_{ij} + x_{jk} + x_{ki} \leq 2$, respectively. The MAX-LOP formulation was introduced by Grötschel, Jünger, and Reinelt [12, 13] and Reinelt [24], and studied more recently by Goemans and Hall [11] and Newman and Vempala [23]. The convex hull of the points satisfying (1.1)–(1.4) is denoted by P^n_{LO}, or sometimes P^X_{LO}, and is known as the *linear ordering polytope* or *binary choice polytope*; see [9, 8] for a survey. This polytope has one vertex per linear ordering on X; hence the name.

A fair number of facet-defining inequalities of the linear ordering polytope have been determined, including *k-fence inequalities* [13, 5], *t-reinforced k-fence inequalities* [26, 18], *α-critical fence inequalities* [15], *Möbius ladder inequalities* [13], and the inequalities obtained from these by symmetries of the polytope [2, 7]. In this list, the only class of inequalities for which a polynomial time separation algorithm has been published are the Möbius ladder inequalities [3]. By making $n + 1$ calls to any such algorithm, one can solve the separation problem for all inequalities obtained from Möbius ladder inequalities by symmetries. For a more direct approach, see [8]. It is very tempting to look for generalizations of the Möbius ladder inequalities. This is the aim of the present article. The following examples illustrate our approach.

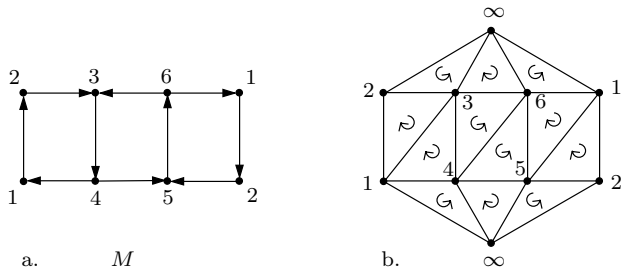


FIG. 1.1. *A Möbius ladder and the corresponding triangulation of the projective plane.*

Example 1. Let $X = \{1, 2, 3, 4, 5, 6\}$ and $M = \{12, 23, 34, 41, 45, 56, 63, 61, 25\}$ (see Figure 1.1(a)). Note that in the figure, some vertices have to be identified. The inequality

$$\sum_{ij \in M} 2x_{ij} \geq 4 \tag{1.5}$$

is a Möbius ladder inequality. (A definition of these inequalities is given below, in subsection 4.2.) It defines a facet of the linear ordering polytope. We now give a cutting plane proof of the fact that the inequality is valid. More precisely, we show that it is a $\{0, \frac{1}{2}\}$-cut for the system (1.1)–(1.3).

If we sum (1.1) for $ij \in \{23, 41, 45, 63, 61, 25\}$ and (1.2) for $ijk \in \{123, 341, 634, 456, 561, 125\}$, and substract (1.3) for $ij \in \{31, 46, 15\}$, the resulting valid inequality reads

$$\sum_{ij \in M} 2x_{ij} \geq 3. \tag{1.6}$$

Because at a vertex of the linear ordering polytope the left-hand side of (1.6) is an even integer, we can add 1 to the right-hand side of (1.6) while preserving its validity. Hence we have proved that (1.5) is valid. In order to visualize the derivation better, we associate with each inequality $x_{ij} \geq 0$ that was used the oriented triangle $ij\infty$, where

$\infty \notin X$, and to each inequality $x_{ij} + x_{jk} + x_{ki} \geq 1$ that was used the oriented triangle ijk. The resulting collection of oriented triangles is represented in Figure 1.1(b). Now the crucial observation is that our cutting plane proof produces a triangulation of a surface, namely, the projective plane (see Figure 1.2(a)).

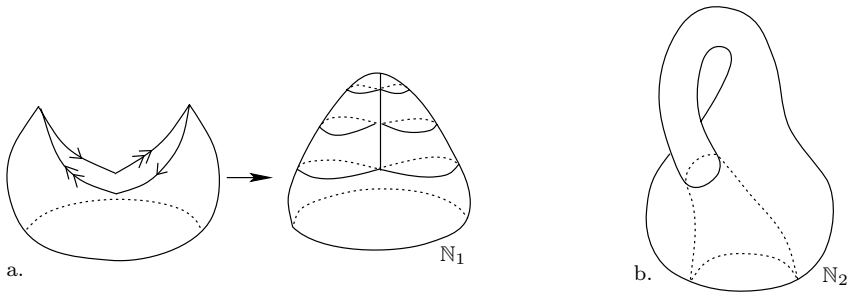


FIG. 1.2. *A representation of the projective plane (left) and the Klein bottle (right).*

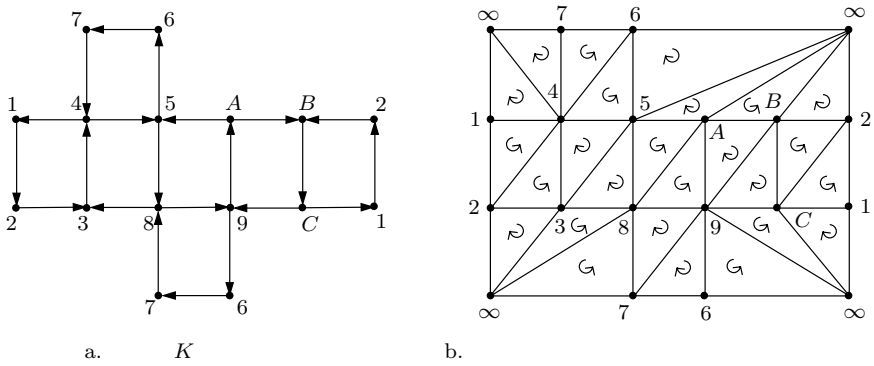


FIG. 1.3. *The support graph of a new facet-defining inequality and the corresponding triangulation of the Klein bottle.*

Example 2. Let $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C\}$ and $K = \{12, 23, 2B, 34, 41, 45, 56, 58, 67, 74, 78, 83, 89, 96, 9A, A5, AB, BC, C1, C9\}$ (see Figure 1.3(a)). By a cutting plane proof similar to that used in Example 1, the inequality

$$\sum_{ij \in K} 2x_{ij} \geq 8$$

can be proved to be valid. This time, we sum (1.1) for $ij \in \{23, 2B, 41, 56, 78, 74, 83, 96, A5, AB, C1, C9\}$ and (1.2) for $ijk \in \{124, 234, 345, 358, 456, 467, 679, 789, 58A, 89A, 9AB, 9BC, 2BC, 12C\}$, and substract (1.3) for $ij \in \{24, 35, 46, 79, 8A, 9B, 2C\}$. If we use the same convention as above to represent the derivation, a triangulation is revealed (see Figure 1.3(b)). This time the corresponding surface is the Klein bottle (see Figure 1.2(b)). It is an interesting exercise to show that the inequality above—which was unknown before—defines a facet of the linear ordering polytope (see the beginning of the proof of Proposition 5.5 for a hint).

In this article, we consider {0, $\frac{1}{2}$}-cuts derived from the system (1.1)–(1.3). Our motivation for studying these cuts is threefold. First, the cuts generalize known facet-defining inequalities, including Möbius ladder inequalities, although they are not

guaranteed to be facet-defining in general. This observation raises the possibility of finding a generalization of the Möbius ladder inequalities whose corresponding separation problem is still tractable. Second, they possess interesting structural properties. For instance, some of them naturally define surfaces. It turns out that the topological properties of these surfaces and the polyhedral properties of the corresponding cuts are related. To our knowledge, this is the first connection of this type observed between topology and polyhedral combinatorics. Third, it is interesting to find new facet-defining inequalities which simultaneously have complex structures and short validity proofs. Since they have short cutting plane proofs, $\{0, \frac{1}{2}\}$-cuts are good candidates.

In section 2, we define $\{0, \frac{1}{2}\}$-cuts and then note some basic results on the $\{0, \frac{1}{2}\}$-cuts obtained from (1.1)–(1.3). In section 3, we give some background on simplicial complexes and surfaces. We begin section 4 by relating $\{0, \frac{1}{2}\}$-cuts for the linear ordering problem to certain pure two-dimensional simplicial complexes. The rest of the section focusses on surface-shaped $\{0, \frac{1}{2}\}$-cuts, i.e., cuts whose corresponding complex is a triangulation of some surface. We establish two necessary conditions for such a $\{0, \frac{1}{2}\}$-cut to define a facet of the linear ordering polytope. We then use these necessary conditions to prove that no $\{0, \frac{1}{2}\}$-cut engendered by an orientable surface is facet-defining. Finally, in section 5, we show how to transform any factor-critical graph into a facet-defining $\{0, \frac{1}{2}\}$-cut which is nearly surface-shaped. As a corollary, we prove that for every nonorientable surface, there is a facet-defining cut with the same "shape."

2. $\{0, \frac{1}{2}\}$-cuts. In this section, we formally define $\{0, \frac{1}{2}\}$-cuts. We then gather some initial results on the $\{0, \frac{1}{2}\}$-cuts for the linear ordering problem arising from its standard linear relaxation (1.1)–(1.3). More specifically, we give a system of linear equations on $\mathbb{F}_2 = GF(2)$ describing all cuts for a certain value of n.

2.1. $\{0, \frac{1}{2}\}$-cuts in general. Consider a system $Ax \geq b$ of linear inequalities with $A \in \mathbb{Z}^{p \times q}$ and $b \in \mathbb{Z}^p$, let P be the polyhedron defined by $Ax \geq b$, and let $P_I = \text{conv}(P \cap \mathbb{Z}^q)$ denote the *integer hull* of P. A $\{0, \frac{1}{2}\}$-*cut* [3] for $Ax \geq b$ is an inequality of the form

$$u^T Ax \geq u^T b + 1, \tag{2.1}$$

where $u \in \{0,1\}^p$, each component of $u^T A$ is even, and $u^T b$ is odd. Every $\{0, \frac{1}{2}\}$-cut is valid for P_I. This definition of $\{0, \frac{1}{2}\}$-cut is slightly nonstandard. In the usual definition, u belongs to $\{0, \frac{1}{2}\}^p$, and the resulting inequality is $\frac{1}{2}$ times (2.1).

Perhaps because they rely on a simple, widely applicable principle, $\{0, \frac{1}{2}\}$-cuts are very common in combinatorial optimization; see, e.g., [3, 4]. For recent progress on $\{0, \frac{1}{2}\}$-cuts and their separation, see [17, 16]. A *multiplier* is any 0/1-vector $u \in \{0,1\}^p$ such that $u^T A \equiv \mathbf{0}^T \pmod 2$ and $u^T b \equiv 1 \pmod 2$, where $\mathbf{0}$ denotes a zero column vector of compatible size. We denote by $M(A, b)$ the set of all multipliers of $Ax \geq b$. This set forms an affine subspace of the affine space $\mathbb{F}_2^p = GF(2)^p = AG(p, 2)$ that we call the *multiplier space* of $Ax \geq b$.

2.2. $\{0, \frac{1}{2}\}$-cuts for the linear ordering problem. Henceforth, $Ax \geq b$ denotes the system formed by (1.1), (1.2) and

$$-x_{ij} - x_{ji} \geq -1 \qquad \forall \{i, j\} \subseteq X. \tag{2.2}$$

We could equally well replace (1.3) by pairs of inequalities, but this would make no essential difference in our discussion. We index the inequalities of $Ax \geq b$ as follows. Let $Y = X \cup \{\infty\}$, where ∞ is any element not in X. The first $(n+1)n(n-1)/3$

inequalities are indexed by the *tricycles* on Y, i.e., the triples of distinct elements of Y taken up to cyclic rotations of their coordinates. In the introduction, we have been using "oriented triangle" to mean "tricycle." The tricycle corresponding to (i, j, k) is denoted by ijk. So ijk, jki, and kij denote the same tricycle. In our indexing scheme, inequality $x_{ij} \geq 0$ corresponds to tricycle ∞ij, and inequality $x_{ij} + x_{jk} + x_{ki} \geq 1$ to tricycle ijk. The last $n(n-1)/2$ inequalities are indexed by the unordered pairs of distinct elements in X. Inequality $-x_{ij} - x_{ji} \geq -1$ corresponds to unordered pair $\{i, j\}$. Thus we write any multiplier as $u = \binom{v}{w}$ for some vector v with $(n+1)n(n-1)/3$ components and some vector w with $n(n-1)/2$ components. Our first result describes the structure of the multiplier space $M(A, b)$. For convenience, we let $M = M(A, b)$ for the rest of the text. Below, $\leq$ denotes any linear order on Y whose largest element is ∞.

PROPOSITION 2.1. *The multiplier space M is defined by the following equations on $\mathbb{F}_2$:*

$$w_{\{i,j\}} = \sum_{\substack{k \in Y \\ k \neq i,j}} v_{ijk} \quad \forall i, j \text{ in } X \text{ with } i < j, \tag{2.3}$$

$$\sum_{\substack{k \in Y \\ k \neq i,j}} v_{ijk} + \sum_{\substack{k \in Y \\ k \neq i,j}} v_{jik} = 0 \quad \forall i, j \text{ in } Y \text{ with } i < j, \tag{2.4}$$

$$\sum_{\substack{i,j,k \in Y \\ i<j<k}} v_{ijk} = 1. \tag{2.5}$$

Proof. Let u be a multiplier, and let i, j be two distinct elements of X. Then we have

$$(u^T A)^T_{ij} = \sum_{\substack{k \in Y \\ k \neq i,j}} v_{ijk} - w_{\{i,j\}} \equiv 0 \pmod 2 \quad \text{and}$$

$$(u^T A)^T_{ji} = \sum_{\substack{k \in Y \\ k \neq i,j}} v_{jik} - w_{\{i,j\}} \equiv 0 \pmod 2.$$

Consequently (2.3) hold, as do (2.4), except perhaps for $j = \infty$. Consider the multigraph with vertex set $Y \setminus \{i\}$ in which two vertices j and k are connected by one edge if either $v_{ijk} = 1$ or $v_{jik} = 1$ but not both, and by two parallel edges if $v_{ijk} = v_{jik} = 1$. The degree of vertex j in this graph is given by the left-hand side of (2.4). So all the vertices of the multigraph except perhaps ∞ have even degree. Because every multigraph has an even number of vertices of odd degree, the degree of ∞ is even, and thus (2.4) hold for all i, j in Y.

Because u is a multiplier, it also has to satisfy the condition $u^T b \equiv 1 \pmod 2$. This condition can be rewritten as follows in $\mathbb{F}_2$:

$$\begin{aligned}
&\sum_{\substack{i,j,k \in X \\ i<j<k}} v_{ijk} + \sum_{\substack{i,j,k \in X \\ i<j<k}} v_{kji} + \sum_{\substack{i,j \in X \\ i<j}} w_{\{i,j\}} = 1 \\
\iff \quad &\sum_{\substack{i,j,k \in X \\ i<j<k}} v_{ijk} + \sum_{\substack{i,j,k \in X \\ i<j<k}} v_{kji} + \sum_{\substack{i,j \in X \\ i<j}} \sum_{\substack{k \in Y \\ k \neq i,j}} v_{ijk} = 1 \\
\iff \quad &\sum_{\substack{i,j,k \in X \\ i<j<k}} v_{ijk} + \sum_{\substack{i,j,k \in X \\ i<j<k}} v_{kji} + \sum_{\substack{i,j,k \in X \\ i<j<k}} v_{kji} + \sum_{\substack{i,j \in X \\ i<j}} v_{ij\infty} = 1 \\
\iff \quad &\sum_{\substack{i,j,k \in Y \\ i<j<k}} v_{ijk} = 1. \qquad \square
\end{aligned}$$

Consider a multiplier $u = \binom{v}{w}$ in M. By Proposition 2.1, u is entirely determined by v. In other words, it suffices to specify the set of tricycles ijk for which $v_{ijk} = 1$ holds in order to determine a multiplier. This set of tricycles has to satisfy the two conditions given by (2.4) and (2.5). In particular, it follows from (2.4) that each unordered pair $\{i, j\} \subseteq Y$ has to be contained in an even number of tricycles of the set. As will be shown later, restricting this number of tricycles to be equal to 0 or 2 already gives rise to a host of interesting inequalities.

The next corollary is a simple application of Proposition 2.1. Although it is not of much use here, we state it because it spawns intriguing questions (see section 6).

COROLLARY 2.2. *The dimension and the cardinality of the multiplier space are respectively given by*

$$\dim M = 2\binom{n+1}{3} - \binom{n}{2} - 1 \quad \textit{and} \quad |M| = 2^{\dim M}.$$

Proof. It suffices to show that the matrix of system (2.4)–(2.5) has rank $\binom{n}{2} + 1$. If we order the variables v_{ijk} in such a way that whenever $i < j < k$, v_{ijk} has position ℓ if and only if v_{kji} has position $\ell + \binom{n+1}{3}$, then the matrix of system (2.4)–(2.5) takes the form

$$N = \begin{pmatrix} B & B \\ \mathbf{1}^T & \mathbf{0}^T \end{pmatrix},$$

where the columns of B are the characteristic vectors of the triangles of the complete graph K_{n+1} on Y. So the columns of B span the cycle space of K_{n+1}, and hence B has rank $\binom{n+1}{2} - (n+1) + 1 = \binom{n}{2}$ [6]. So N has rank $\binom{n}{2} + 1$, as claimed. ∎

3. Simplicial complexes and surfaces. In the preceding section, we proved that $\{0, \frac{1}{2}\}$-cuts for the linear ordering problem correspond to sets of tricycles (or oriented triangles) on $Y = X \cup \{\infty\}$ satisfying certain conditions. This section provides some basic notions and results from topology which will help in recognizing facet-defining cuts on the basis of their global structure.

3.1. Simplicial complexes. An *(abstract) simplicial complex* with *vertex set* V is a collection $\mathcal{K}$ of subsets of V such that (i) $F \in \mathcal{K}$ and $G \subseteq F$ imply $G \in \mathcal{K}$, and (ii) $v \in V$ implies $\{v\} \in \mathcal{K}$. We will always assume that V is finite. A set in $\mathcal{K}$ is called a *face*, and a *k-face* if its cardinality is $k + 1$. The *dimension* of a k-face is k. The *dimension* of $\mathcal{K}$ is the maximum dimension of any of its faces. Note that one-dimensional simplicial complexes correspond to simple graphs. A simplicial complex is said to be *pure* if all its inclusionwise maximal faces have the same dimension. Let v be a vertex of $\mathcal{K}$. The *link* of v is the simplicial complex $\text{link}(v, \mathcal{K}) = \{F - v : v \in F \in \mathcal{K}\}$. Every simplicial complex $\mathcal{K}$ with vertex set V can be canonically realized as a topological space, for instance, as a subspace of $\mathbb{R}^{2d+1}$, where d denotes the dimension of $\mathcal{K}$ [20]. Consider any topological space S. If the canonical realization of $\mathcal{K}$ is homeomorphic to S, then $\mathcal{K}$ is referred to as a *triangulation* of S.

3.2. Surfaces: Definition, invariants, and classification. A *combinatorial surface* is a pure two-dimensional simplicial complex such that the link of every vertex, regarded as a simple graph, is a cycle. In particular, in a combinatorial surface, every 1-face is contained in precisely two 2-faces. A *surface* is a connected compact Hausdorff topological space locally homeomorphic to $\mathbb{R}^2$. Every surface has a triangulation; see, e.g., [21] for a short proof. Moreover, any triangulation of a surface is a combinatorial surface.

Let S be a surface and $\mathcal{K}$ be any triangulation of S. The *Euler characteristic* of triangulation $\mathcal{K}$ is defined by

$$\chi(\mathcal{K}) = f_0 - f_1 + f_2, \tag{3.1}$$

where f_k denotes the number of k-faces of $\mathcal{K}$ for $0 \leq k \leq 2$. If $\mathcal{K}'$ is another triangulation of S, then we have $\chi(\mathcal{K}) = \chi(\mathcal{K}')$ [1]. So we can define the *Euler characteristic* of surface S by letting $\chi(S) = \chi(\mathcal{K})$. The second main invariant of surfaces is orientability. An *oriented* 1*-face* is simply an arc, that is, an ordered pair of distinct elements. Arcs uv and vu are said to be *opposite*. An *oriented* 2*-face* or *oriented triangle* is a tricycle, that is, an ordered triple of distinct elements taken up to cyclic rotations of its coordinates. There are two tricycles on three points, namely, $uvw = vwu = wuv$ and its opposite $wvu = vuw = uwv$. Tricycle uvw determines three arcs: uv, vw, and wu. Two tricycles are said to be *adjacent* if they have exactly two elements in common. Two adjacent tricycles are said to be *compatibly oriented* if the arcs they determine on their common elements are opposite. For instance, uvw and wvu' are adjacent and compatibly oriented, provided that $u \neq u'$. Otherwise they are opposite. An *orientation* of $\mathcal{K}$ is a collection $\vec{\mathcal{K}}$ of tricycles such that for each 2-face $F = \{u, v, w\}$ in $\mathcal{K}$ we have either $uvw \in \vec{\mathcal{K}}$ or $wvu \in \vec{\mathcal{K}}$. (This definition also applies in case $\mathcal{K}$ is any pure two-dimensional simplicial complex.) We say that $\vec{\mathcal{K}}$ is *coherent* if all pairs of adjacent tricycles in $\vec{\mathcal{K}}$ are compatibly oriented. Triangulation $\mathcal{K}$ is said to be *orientable* if it has a coherent orientation. Two cases are possible for S: either all its triangulations are orientable, in which case S is *orientable*, or none of its triangulations is coherently orientable, in which case S is *nonorientable* [1].

Let $\mathbb{S}_h$ denote the surface obtained from the sphere by adding $h \geq 0$ handles, and let $\mathbb{N}_b$ denote the surface obtained from the sphere by removing $b > 0$ discs and replacing them by Möbius bands. All these surfaces are well-defined, up to homeomorphism. The surfaces $\mathbb{S}_1$, $\mathbb{N}_1$, and $\mathbb{N}_2$ are known as the *torus*, *projective plane*, and *Klein bottle*, respectively.

THEOREM 3.1 (the classification of surfaces [1, 21]). *Let S be a surface with Euler characteristic χ. If S is orientable, then it is homeomorphic to $\mathbb{S}_h$ for $h = 1 - \frac{1}{2}\chi$. If S is nonorientable, then it is homeomorphic to $\mathbb{N}_b$ for $b = 2 - \chi$. No two of the surfaces $\mathbb{S}_0$, $\mathbb{S}_1$, $\mathbb{N}_1$, $\mathbb{S}_2$, $\mathbb{N}_2$, ... are homeomorphic.* □

4. Surface-shaped cuts. In this section, we use the terminology introduced in the preceding section to motivate, define, and study surface-shaped cuts. Central in our discussion is the question of characterizing the surface-shaped cuts which are facet-defining. Two main necessary conditions are given. Each of these is proved by reinterpreting surface-shaped cuts from a different standpoint. An important implication of the necessary conditions is that no orientable surface can engender a facet-defining cut.

4.1. Regarding cuts as oriented simplicial complexes. Let $Ax \geq b$ be defined as in subsection 2.2. Consider a multiplier $u = \binom{v}{w}$ in $M(A, b)$. Let $\vec{\mathcal{K}} = \vec{\mathcal{K}}(u)$ denote the set of tricycles ijk on $Y = X \cup \{\infty\}$ such that $v_{ijk} = 1$. As was noted above, u is entirely determined by $\vec{\mathcal{K}}$.

LEMMA 4.1. *If $\vec{\mathcal{K}} = \vec{\mathcal{K}}(u)$ contains a tricycle and its opposite, then the cut defined by the multiplier u is implied by* (1.1)–(1.3).

Proof. Without loss of generality, we assume that $\vec{\mathcal{K}}$ contains both ijk and kji, where i, j, and k are three distinct elements of X. Inequality (2.1) is clearly implied by (1.1)–(1.3) and $\bar{u}^T Ax \geq \bar{u}^T b + 1$, where $\bar{u}$ is the vector obtained from u by replacing

all its coordinates by zeroes except the ones corresponding to ijk and kji. Since the inequality $\bar{u}^T Ax \geq \bar{u}^T b + 1$ reads

$$(x_{ij}+x_{jk}+x_{ki})+(x_{kj}+x_{ji}+x_{ik}) \geq 3 \iff (x_{ij}+x_{ji})+(x_{jk}+x_{kj})+(x_{ki}+x_{ik}) \geq 3,$$

it is implied by (1.3). Hence the $\{0, \frac{1}{2}\}$-cut $u^T Ax \geq u^T b + 1$ is implied by (1.1)–(1.3). The lemma follows. □

If $\vec{\mathcal{K}}$ does not contain a pair of opposite tricycles, then we say that u is *simple*. From now on, we will restrict ourselves to simple multipliers. When u is simple, its corresponding set of tricycles $\vec{\mathcal{K}}$ can be regarded as an orientation of the pure two-dimensional simplicial complex $\mathcal{K} = \mathcal{K}(u)$ whose inclusionwise maximal faces are the sets $\{i, j, k\}$ with $v_{ijk} = 1$ or $v_{kji} = 1$. Because u is a multiplier, $\vec{\mathcal{K}}$ satisfies certain conditions. For instance, (2.4) requires that for each 1-simplex $\{i, j\}$ in $\mathcal{K}$ the number of oriented 2-simplices of the form ijk in $\vec{\mathcal{K}}$ and the number of oriented 2-simplices of the form jik in $\vec{\mathcal{K}}$ have the same parity. In particular, it follows that in $\mathcal{K}$ each 1-simplex is contained in an even number of 2-simplices. If, moreover, $\mathcal{K}$ is a combinatorial surface, then we call multiplier u and the corresponding cut *surface-shaped*.

Conversely, we can start with any combinatorial surface $\mathcal{K}$ whose vertex set is included in Y and define a multiplier u such that $\mathcal{K}(u) = \mathcal{K}$, as follows. Consider any orientation $\vec{\mathcal{K}}$ of $\mathcal{K}$. Let $u = \binom{v}{w}$ denote the 0/1-vector with v determined by $v_{ijk} = 1$ if $ijk \in \vec{\mathcal{K}}$, $v_{ijk} = 0$ otherwise, and w determined by (2.3). Then either u is a multiplier or replacing an odd number of tricycles in $\vec{\mathcal{K}}$ by their opposite yields a 0/1-vector u which is a multiplier. By construction, we have $\vec{\mathcal{K}}(u) = \vec{\mathcal{K}}$ and $\mathcal{K}(u) = \mathcal{K}$. Note that the multipliers obtained in this way are always simple.

4.2. The case of Möbius ladder inequalities. A digraph $D = (N(D), A(D))$ is a *Möbius ladder* if there are a positive integer k and dicycles[1] $C_0, C_1, \ldots, C_{k-1}$ in D such that $A(D) = C_0 \cup C_1 \cup \cdots \cup C_{k-1}$ and the following conditions are satisfied for all i, j:

(M1) $k \geq 3$ and k is odd;
(M2) $C_i \cap C_{i+1}$ contains exactly one arc, denoted by e_i;
(M3) $C_i \cap C_j = \varnothing$ if $j \notin \{i-1, i, i+1\}$;
(M4) $|C_i| \in \{3, 4\}$;
(M5) the total degree of each node in D is greater or equal to 3;
(M6) if C_i and C_j have a node v in common and $i \neq j$, then either C_i, C_{i+1}, …, C_{j-1}, C_j have node v in common, or C_j, C_{j+1}, …, C_{i-1},C_i have node v in common, but not both;
(M7) $D - \{e_{i+1}, e_{i+3}, \ldots, e_{i-2}\}$ contains exactly one dicycle, namely, C_i.

The above definition is due to Reinelt [24]. It is perhaps not very intuitive. Notably, (M1)–(M7) imply that $A(D) - \{e_0, \ldots, e_{k-1}\}$ is a *semicycle*, that is, a set of arcs obtained by reversing certain arcs of a dicycle of length at least three. Whenever $N(D) \subseteq X$, the Möbius ladder $D = (N(D), A(D))$ has a corresponding *Möbius ladder inequality* which reads

$$\sum_{ij \in A(D)} x_{ij} \geq \frac{k+1}{2}. \tag{4.1}$$

Every Möbius ladder inequality defines a facet of the linear ordering polytope [24] and can be derived as a $\{0, \frac{1}{2}\}$-cut from (1.1)–(1.3), as in Example 1. The resulting

[1] Throughout this article, dicycles are regarded as sets of arcs.

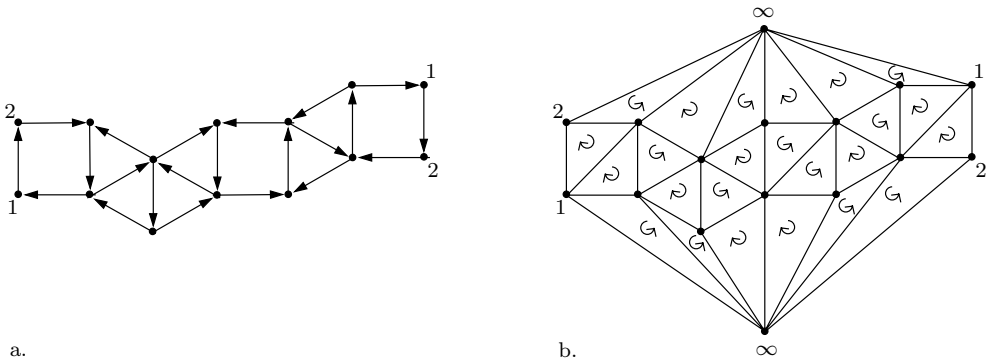


FIG. 4.1. *A Möbius ladder and a corresponding triangulation of* $\mathbb{N}_1$.

collections of tricycles yield triangulations of the projective plane (see Figure 4.1 for a further example). In other words, the following result holds.

PROPOSITION 4.2. *Every Möbius ladder inequality is a surface-shaped* $\{0, \frac{1}{2}\}$*-cut whose underlying surface is the projective plane* $\mathbb{N}_1$. □

4.3. Interpreting the cuts using complete cyclic orders. Let u be a surface-shaped multiplier, and let $\mathcal{K} = \mathcal{K}(u)$ and $\vec{\mathcal{K}} = \vec{\mathcal{K}}(u)$. Consider the graph $G = G(u)$ which has one vertex per 2-face of $\mathcal{K}$ and in which two 2-faces form an edge if the corresponding tricycles in $\vec{\mathcal{K}}$ are adjacent and compatibly oriented. Each connected component of G determines a subcomplex of $\mathcal{K}$, which is referred to as a *zone* of u. The *zone graph* of u has one vertex per zone and one edge per pair of zones containing a common 1-simplex, and is denoted by $Z(u)$. The aim of this subsection is to prove the following lemma. Quite naturally, we call a multiplier *facet-defining* if the corresponding $\{0, \frac{1}{2}\}$-cut defines a facet of the linear ordering polytope.

LEMMA 4.3. *Let* u *be a facet-defining surface-shaped multiplier. Then every zone of* u *is a triangulated cycle.*

The meaning of "triangulated cycle" should be clear. If not, a formal definition is given below. Triangulated cycles are the simplicial complexes which are recursively defined as follows. The simplicial complex $\{\emptyset, \{i_0\}, \{i_1\}, \{i_2\}, \{i_0, i_1\}, \{i_0, i_2\}, \{i_1, i_2\}, \{i_0, i_1, i_2\}\}$ is a *triangulated cycle* with vertex sequence $i_0 i_1 i_2 i_0$. If a simplicial complex $\mathcal{L}$ is a triangulated cycle with vertex sequence $i_0 i_1 \cdots i_{m-1} i_0$, then for each $\alpha \in \{0, \ldots, m-1\}$ and all j not in the vertex set of $\mathcal{L}$, the simplicial complex $\mathcal{L} \cup \{\{j\}, \{i_\alpha, j\}, \{j, i_{\alpha+1}\}, \{i_\alpha, j, i_{\alpha+1}\}\}$ is a triangulated cycle with vertex sequence $i_0 i_1 \cdots i_\alpha j i_{\alpha+1} \cdots i_{m-1} i_0$ (indices are taken modulo m).

Note that Lemma 4.3 in particular implies that every facet-defining surface-shaped multiplier has at least two zones. This is due to the fact that a triangulated cycle is not a combinatorial surface because it has a boundary. The technique we use to prove Lemma 4.3 generalizes that used in the proof of Lemma 4.1. Namely, if the cut defined by a multiplier u is facet-defining, then replacing one or several nonzero coordinates of u by zeroes should cause (2.1) to lose its validity. In order to formalize this idea in the most informative way, we resort to complete cyclic orders.

A set C of tricycles is said to be *asymmetric* if $ijk \in C$ implies $kji \notin C$, *transitive* if $ijk, ik\ell \in C$ and $j \neq \ell$ imply $ij\ell \in C$, a *cyclic order* if it is asymmetric and transitive, and *complete* if $ijk \notin C$ implies $kji \in C$. Complete cyclic orders are combinatorial structures encoding the relative positions of distinct points on a oriented closed curve.

Given a set of distinct points on such a curve, we obtain a complete cyclic order by setting $ijk \in C$ whenever j lies in the open path which goes from i to k in the prescribed orientation. A set of tricycles is said to be *extendable* if it is contained in some complete cyclic order. Determining whether a set of tricycles is extendable or not is an NP-complete problem [10]. We call a set of tricycles *minimally nonextendable* if it is nonextendable and each of its proper subsets is extendable.

The *complete cyclic order polytope*, denoted by P^Y_{CCO}, is the convex hull of the 0/1 characteristic vectors of all complete cyclic order orders on $Y = X \cup \{\infty\}$ in the real vector space which has one coordinate y_{ijk} per tricycle ijk on Y. The polytopes P^X_{LO} and P^Y_{CCO} are affinely equivalent, the equivalence being given by

$$x \mapsto y \quad \text{with} \quad y_{ijk} = \begin{cases} x_{ij} + x_{jk} + x_{ki} - 1 & \text{if } i,j,k \neq \infty, \\ x_{ij} & \text{if } k = \infty. \end{cases} \tag{4.2}$$

A set C of tricycles on Y is nonextendable if and only if its *dual* $C^d = \{kji : ijk \in C\}$ is nonextendable, that is, if and only if the *nonextendable set of tricycles* (*NEST*) *inequality*,

$$\sum_{ijk \in C} y_{ijk} \geq 1, \tag{4.3}$$

is valid for the complete cyclic order polytope. Indeed, the inequality is valid if and only if every vertex of the polytope has $y_{ijk} = 1$ for some $ijk \in C$. Since all vertices of P^Y_{CCO} satisfy $y_{ijk} + y_{kji} = 1$, the latter condition holds if and only if C^d is nonextendable or, equivalently, if and only if C is nonextendable. NEST inequalities were introduced by the author in [8]. Note that (4.3) is valid for P^Y_{CCO} if and only if the inequality

$$\sum_{ij\infty \in C} x_{ij} + \sum_{\substack{ijk \in C \\ i,j,k \neq \infty}} (x_{ij} + x_{jk} + x_{ki}) \geq |\{ijk \in C : i,j,k \neq \infty\}| + 1 \tag{4.4}$$

obtained from it by expressing the y variables in terms of the x variables using (4.2) is valid for P^X_{LO}. We also refer to (4.4) as a *NEST inequality*. Now the key observation is that, modulo (1.3), the cut determined by a multiplier u is exactly the NEST inequality (4.4) with $C = \vec{\mathcal{K}}(u)$. Hence, $\vec{\mathcal{K}}(u)$ has to be minimally nonextendable whenever (2.1) is facet-defining.

Proof of Lemma 4.3. Let $\mathcal{K} = \mathcal{K}(u)$, $\vec{\mathcal{K}} = \vec{\mathcal{K}}(u)$, and $G = G(u)$. Consider any inclusionwise maximal subset U of $V(G)$ such that $G[U]$ is connected and U determines a subcomplex $\mathcal{L}$ of $\mathcal{K}$ which is a triangulated cycle. Let $i_0 i_1 \cdots i_{m-1} i_0$ denote the vertex sequence of $\mathcal{L}$, and let $\vec{\mathcal{L}}$ denote the orientation of $\mathcal{L}$ determined by u. If U is a connected component of G, then there is nothing to prove. Otherwise, there is an index $\alpha \in \{0, \ldots, m-1\}$ and a vertex j of $\mathcal{K}$ such that the 2-face $\{i_\alpha, j, i_{\alpha+1}\}$ belongs to $\mathcal{K}$ but not to U and is adjacent to some element of U in G. By maximality of U, vertex j has to belong to $\mathcal{L}$. It follows that $\vec{\mathcal{L}}$ is nonextendable, and hence $\vec{\mathcal{K}}$ is not minimally nonextendable, a contradiction. □

4.4. Interpreting the cuts in terms of matching theory. As in the preceding subsection, we reconsider surface-shaped cuts from a different angle. Again, this yields a necessary condition for a cut to be facet-defining. An important consequence is that no orientable surface can give rise to a facet-defining cut. We begin with some classic definitions and results from matching theory.

Let $G = (V, E)$ be a graph. A *matching* is a set of pairwise independent edges. When a matching covers every vertex, it is said to be *perfect*. An *edge cover* is a

set of edges covering every vertex. The maximum cardinality of a matching and the minimum cardinality of an edge cover are respectively denoted by $\nu(G)$ and $\rho(G)$. Whenever G has no isolated vertex, we have $\nu(G)+\rho(G) = |V|$. If $G-v$ has a perfect matching for all vertices v, then G is called *factor-critical*. A set of vertices S is said to be *matchable* to $G - S$ if the graph with vertex set $S \cup \mathcal{C}(G - S)$ and edge set $\{\{s, C\} : \exists c \in C \text{ s.t. } sc \in E(G)\}$ contains a matching covering S, where $\mathcal{C}(G - S)$ denotes the collection of all connected components of $G-S$. We will use the following structural result on matchings [6].

THEOREM 4.4. *Every graph G contains a set of vertices S with the following two properties:* (i) *S is matchable to $G - S$, and* (ii) *every component of $G - S$ is factor-critical.* □

The link between surface-shaped $\{0, \frac{1}{2}\}$-cuts and matching theory relies on the concept of a 2-*packing*, i.e., a collection of dicycles on some finite set such that each arc is contained in at most two dicycles of the collection. Whenever $\mathcal{C}$ is a 2-packing with an odd number of dicycles whose ground set is included in X, the 2-*packing inequality*

$$\sum_{ij \in \cup \mathcal{C}} 2x_{ij} \geq |\mathcal{C}| + 1 \tag{4.5}$$

is valid for the linear ordering polytope. By Lemma 4.3, if a surface-shaped multiplier u is facet-defining, then each zone of u determines a dicycle on $Y = X \cup \{\infty\}$. Let $\mathcal{C}$ denote the collection of all those dicycles which do not contain ∞. Then $\mathcal{C}$ is a 2-packing, and it is easy to check that (2.1) and (4.5) coincide. For $i \in Y$, let $Z'_i(u)$ denote the subgraph of the zone graph of u induced by the zones which do not contain i. It emerges from our discussion that $Z'_\infty(u)$ plays a special role. We call it the *restricted zone graph* of u. We are now ready to state and prove our second necessary condition for a surface-shaped cut to define a facet of the linear ordering polytope.

LEMMA 4.5. *Let u be a facet-defining surface-shaped multiplier. Then the following hold:*

(i) *the restricted zone graph $Z'_\infty(u)$ is factor-critical;*
(ii) *the graph $Z'_i(u)$ is factor-critical for all $i \in Y$;*
(iii) *the zone graph $Z(u)$ is factor-critical.*

Proof. We claim that it suffices to prove (i). Indeed, as we can exchange the roles of any element of X and ∞ by a symmetry of the linear ordering polytope [7], (ii) follows from (i). Moreover, we can assume that ∞ is not a vertex of $\mathcal{K}(u)$ by adding one new element to X and then exchanging the roles of this new element and ∞ by a symmetry of the polytope. In virtue of the trivial lifting lemma [24], the resulting surface-shaped multiplier is still facet-defining. Hence (iii) also follows from (i).

We now prove (i). Again, let $\mathcal{C}$ denote the collection of dicycles on Y defined by the zones of u which do not contain ∞. By contradiction, suppose that the restricted zone graph of u is not factor-critical. Then, by Theorem 4.4, there is a partition of $\mathcal{C}$ into nonempty subsets $\mathcal{S}$, $\mathcal{C}_1$, ..., $\mathcal{C}_m$ such that $|\mathcal{C}_\alpha|$ is odd for $1 \leq \alpha \leq m$ and no dicycle of $\mathcal{C}_\alpha$ has an arc in common with any dicycle of $\mathcal{C}_\beta$ if $\alpha \neq \beta$. This is easily seen by considering the graph which has one vertex per dicycle of $\mathcal{C}$, two vertices being adjacent when the corresponding dicycles share an arc. The parity condition on the cardinality of $\mathcal{C}_\alpha$ for $1 \leq \alpha \leq m$ is due to the (trivial) fact that factor-critical graphs have an odd number of vertices. Note that by assertion (i) of Theorem 4.4, we have $m \geq |\mathcal{S}|$. Moreover, note that in (2.1), $u^T b$ exactly counts the number of dicycles in $\mathcal{C}$, so $|\mathcal{C}| = u^T b$ is odd. It follows that we have $m \geq |\mathcal{S}| + 1 \geq 2$. By

summing the 2-packing inequalities corresponding to $\mathcal{C}_1, \ldots, \mathcal{C}_m$ and perhaps some trivial inequalities of the form $x_{ij} \geq 0$, we obtain the inequality

$$\sum_{ij \in \cup \mathcal{C}} 2x_{ij} \geq \sum_{\alpha=1}^{m} |\mathcal{C}_\alpha| + m = |\mathcal{C}| - |\mathcal{S}| + m.$$

Because the right-hand side of the latter inequality is at least $|\mathcal{C}| + 1$, it follows that the $\{0, \frac{1}{2}\}$-cut determined by u, which coincides with inequality (4.5), is implied by the 2-packing inequalities of $\mathcal{C}_1, \ldots, \mathcal{C}_m$ and the trivial inequalities, a contradiction. In conclusion, the restricted zone graph of u has to be factor-critical. □

We can now prove the consequential result which was announced in the beginning of this subsection.

THEOREM 4.6. *Let u be a surface-shaped multiplier. If its associated complex is orientable, then u is not facet-defining.*

Proof. Suppose otherwise. The zone graph of u has to be factor-critical by Lemma 4.5, and bipartite because $\mathcal{K}(u)$ is orientable. Hence the zone graph of u is a one-vertex graph, so u has only one zone. This contradicts Lemma 4.3. □

5. Facet-defining cuts for nonorientable surfaces. In the preceding section, we gave conditions that all facet-defining surface-shaped cuts have to satisfy. In particular, we showed that the underlying surface of any such cut is nonorientable. It is then natural to ask which nonorientable surfaces admit a facet-defining cut. As we show in this section, all of them do. For each nonorientable surface, we will construct a surface-shaped facet-defining cut whose corresponding surface is the given surface. Before diving into the details, we give the intuition behind the construction. The idea is to prove a partial converse to Lemma 4.5(i). We fix a nontrivial factor-critical graph and try to find a facet-defining multiplier whose restricted zone graph is the given graph. We show that this can be done if we first modify the given graph by substituting a path of length 3 for each edge. Despite this restriction, and despite the fact that not all obtained multipliers are surface-shaped, our constructive results allow us to easily build facet-defining surface-shaped cuts of any (nonorientable) "shape."

5.1. Prescribing the restricted zone graph. Let G be any graph. Later on, we will assume that G is a nontrivial factor-critical graph, but for the moment we assume just that G has minimum degree at least 2 and an odd number of vertices. A digraph D without isolated nodes is a *representation* of G if it has a collection $\mathcal{C} = \{C_v : v \in V(G)\}$ of dicycles satisfying the following properties for all vertices v and w of G:

(R1) the length of C_v equals $2\deg(v)$;
(R2) every arc of D is either contained in one dicycle of $\mathcal{C}$ (*simple* arc) or in two dicycles of $\mathcal{C}$ (*double* arc);
(R3) if v and w are nonadjacent, then C_v and C_w are node-disjoint, and if v and w are adjacent, then C_v and C_w have two nodes and one arc in common.

As is easily verified, every graph without pending or isolated vertices has at least one representation. We now state some key properties of representations following from (R1)–(R3). Let D be any representation of G, and let $\mathcal{C} = \{C_v : v \in V(G)\}$ denote the corresponding collection of dicycles. By (R3), each edge $\{v, w\}$ of G uniquely determines a double arc in D, namely, the arc shared by C_v and C_w. Conversely, (R2) and (R3) together imply that every double arc in D uniquely determines an edge in G. Since the dicycle C_v contains one double arc per neighbor of v in G, the respective

positions of these double arcs in C_v determine a complete cyclic order on the neighborhood of each vertex v of G (and also on the edges of G incident to v). In fact, these complete cyclic orders determine the representation up to isomorphism. It follows from (R3) that in each dicycle of $\mathcal{C}$ simple and double arcs alternate. Therefore, every vertex of D has either indegree one and outdegree two or indegree two and outdegree one. Each arc of D contains one vertex of each type, so D is bipartite. Moreover, in every dipath or dicycle of D simple and double arcs alternate.

Condition (R2) obviously implies that the collection $\mathcal{C}$ of dicycles associated with the representation D is a 2-packing. By triangulating arbitrarily each dicycle of $\mathcal{C}$ (without new vertices), we obtain a certain set of tricycles. We then add to this set of tricycles the tricycle ∞ij for each simple arc ij of D. Let $\vec{\mathcal{K}}$ denote the resulting set of tricycles, and let $u = \binom{v}{w}$ denote the 0/1-vector with v determined by $v_{ijk} = 1$ if $ijk \in \vec{\mathcal{K}}$ and $v_{ijk} = 0$ otherwise, and w determined by (2.3).

LEMMA 5.1. *Let G, D, $\mathcal{C}$, and u be defined as above, and let $\mathcal{K} = \mathcal{K}(u)$. Then the following hold:*

(i) *u is a multiplier;*
(ii) *the restricted zone graph of u is precisely G;*
(iii) *the cut determined by u coincides with the 2-packing inequality of $\mathcal{C}$,*
(iv) *the link of every vertex in $\mathcal{K}$ is a cycle, except perhaps that of ∞;*
(v) *the Euler characteristic of $\mathcal{K}$ is $|V(G)| - |E(G)| + 1$.*

Proof.[2] As is easily verified, the zones of u not containing ∞ are in one-to-one correspondence with the dicycles of $\mathcal{C}$. Moreover, two zones have a common 1-face if and only if the corresponding dicycles share a double arc in D. Assertion (ii) follows. Now let $Ax \geq b$ denote the system defined in subsection 2.2. Equations (2.3) hold by definition of u. Since in $\mathcal{K}$ every 1-face is contained in exactly two 2-faces, equations (2.4) hold. Finally, (2.5) holds because $u^T b$ counts the number of zones of u, which is an odd number (recall that we assume that G has an odd number of vertices). Assertion (i) thus follows from Proposition 2.1. We already observed that (iii) holds in subsection 4.4.

We now turn to (iv). Let v be a vertex of $\mathcal{K}$ distinct from ∞. Then v is contained in exactly two zones of u not containing ∞, say $\mathcal{P}$ and $\mathcal{Q}$. These two zones intersect in a common 1-face. Let $i_0 i_1 \cdots i_{p-1} i_0$ and $j_0 j_1 \cdots j_{q-1} j_0$ respectively denote the vertex sequences of $\mathcal{P}$ and $\mathcal{Q}$, with $i_0 = j_0 = v$ and $i_1 = j_1$. Vertex v is contained in exactly two 2-faces of $\mathcal{K}$ through ∞, namely, $\{i_0, i_{p-1}, \infty\} = \{v, i_{p-1}, \infty\}$ and $\{j_0, j_{q-1}, \infty\} = \{v, j_{q-1}, \infty\}$. Now we see that the link of v in $\mathcal{K}$ is some i_1-i_{p-1} path in $\mathcal{P}$ followed by the path with vertex sequence $i_{p-1} \infty j_{q-1}$ followed by some j_{q-1}–j_1 path in $\mathcal{Q}$ (see Figure 5.1). Hence $\text{link}(v, \mathcal{K})$ is a cycle, and (iv) holds.

Finally, in order to prove (v), we compute the number of 0-faces (vertices), 1-faces, and 2-faces of $\mathcal{K}$ as follows:

$$\begin{aligned} f_0 &= 1 + \sum_{v \in V(G)} \deg v &&= 1 + 2|E(G)|, \\ f_1 &= \sum_{v \in V(G)} \left(\frac{9}{2} \deg v - 3 \right) &&= 9|E(G)| - 3|V(G)|, \\ f_2 &= \sum_{v \in V(G)} (3 \deg v - 2) &&= 6|E(G)| - 2|V(G)|. \end{aligned}$$

[2] At several places in this proof we implicitly use the properties of representations stated above. The reader is encouraged to form a mental image of what a representation looks like before reading on.

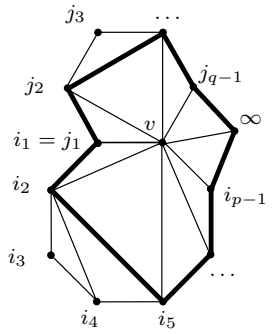


FIG. 5.1. *A view of $\mathcal{K}$ around vertex $v \neq \infty$.*

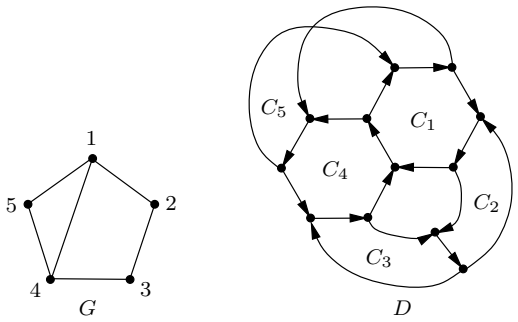


FIG. 5.2. *A graph G and a representation D of graph G.*

Therefore, we have $\chi(\mathcal{K}) = |V(G)| - |E(G)| + 1$, and (v) holds. □

Note that $\text{link}(\infty, \mathcal{K})$ is not always a cycle. For instance, if we start with the representation depicted in Figure 5.2, the link of ∞ in $\mathcal{K}$ is the disjoint union of two cycles.

5.2. Turning factor-critical graphs into facets. In this subsection we show that the multipliers u that we have constructed in the preceding subsection are facet-defining, provided that G is obtained from a nontrivial factor-critical graph G_0 by replacing each edge by a path of length 3, and that the representation we choose for G renders no vertex "extra-bad."

Let G be a nontrivial factor-critical graph, let D be a representation of G, and let $\mathcal{C} = \{C_v : v \in V(G)\}$ denote the corresponding collection of dicycles of D. We begin by noting further useful properties of representations. Consider an s-t dipath P in D. Then P determines a subgraph $H = H(P)$ of G. The edges of H are those which correspond to double arcs in P, and the vertices of H are the endpoints of these edges. We say that a vertex v of H is *primary* if P contains a simple arc of C_v and *secondary* otherwise. If H has at most one primary vertex, then $P \subseteq C_v$ for some v. Otherwise there is a sequence of vertices and edges $v_0e_0v_1e_1 \dots e_{m-1}v_m$ in H such that v_α is primary for all $\alpha \leq m$, v_0 and v_m are respectively the first and last primary vertices of H, $e_\alpha = \{v_\alpha, v_{\alpha+1}\}$ for all $\alpha < m$, and $e_\alpha \neq e_\beta$ for all distinct α and β less than m. Consequently, H always contains a v_0-v_m path on its primary vertices. The above definitions and observations can be readily adapted to the case $s = t$, that is, when P is a dicycle in D.

Now let D be any digraph. A *feedback arc set* (or *dicycle cover*) of D is a set of arcs F such that $D - F$ is acyclic. The minimum cardinality of a feedback arc set of D is denoted by $\tau(D)$. The next lemma is a first step towards the main result of this subsection, namely, Proposition 5.5.

LEMMA 5.2. *Let G_0 be a nontrivial factor-critical graph, let G be the graph obtained from G by replacing each edge by a path of length 3, let D be any representation of G, and let $\mathcal{C} = \{C_v : v \in V(G)\}$ denote the corresponding collection of dicycles of D. Then G is a nontrivial factor-critical graph, and we have*

$$\tau(D) = \rho(G) = (|V(G)| + 1)/2 = (|\mathcal{C}| + 1)/2. \tag{5.1}$$

Therefore, the face of the linear ordering polytope defined by the 2-packing inequality of $\mathcal{C}$ is nonempty.

Proof. It is obvious that G is a nontrivial factor-critical graph. If we show that (5.1) holds, then the face defined by the 2-packing inequality of $\mathcal{C}$, inequality (4.5), is necessarily nonempty. This is due to the fact that the minimum value of the left-hand side of (4.5) for a point of the linear ordering polytope is $2\tau(\cup\mathcal{C}) = 2\tau(D)$. Note that the second equality in (5.1) directly follows from the fact that G is factor-critical, and that the third holds by the definition of a representation.

It remains to prove that we have $\tau(D) = \rho(G)$. Let F be a feedback arc set of D containing only double arcs. Such a feedback arc set exists because if F contains some simple arc, we can replace it with some double arc contained in the same dicycle of $\mathcal{C}$. Feedback arc set F determines a set of edges of G which necessarily covers all vertices of G. So we have $\rho(G) \leq \tau(D)$. In order to prove the converse inequality, consider any minimum edge cover N of G. Then N determines a set of arcs F in D, namely, the set of double arcs corresponding to the edges of N. We claim that F is a feedback arc set. By contradiction, suppose that $D - F$ has a dicycle C. Because N is an edge cover, F hits all dicycles in $\mathcal{C}$. Hence C is not a member of $\mathcal{C}$. It follows that $H(C)$ contains a cycle. By construction of G, this cycle has to contain a vertex v with $\deg_G(v) = 2$. In particular, one of the two edges incident to v has to belong to N, so the corresponding double arc belongs to F, but it also belongs to C, a contradiction. □

As above, let G be a nontrivial factor-critical graph. A vertex v of G is said to be *bad* if we can partition $\delta_G(v) = \{e \in E(G) : v \in e\}$ into two nonempty subsets B and R such that no minimum edge cover of G intersects B and R simultaneously. Now consider some representation D of G. Then a vertex v is called *extra-bad* if it is bad and, moreover, B and R are intervals in the complete cyclic order on $\delta_G(v)$ determined by D (see the paragraph following the definition of representation in subsection 5.1). The following lemma characterizes factor-critical graphs with a bad vertex.

LEMMA 5.3. *Let G be a factor-critical graph, and let v be a vertex of G such that there is a partition of $\delta_G(v)$ into two possibly empty subsets B and R such that in every minimum edge cover of G the edges incident to v are contained either in B or in R. Then $G = G_B \cup G_R$ for some factor-critical graphs G_B and G_R having only vertex v in common and such that $\delta_{G_B}(v) = B$ and $\delta_{G_R}(v) = R$.*

Before proving Lemma 5.3, we state the following theorem on ear decompositions of factor-critical graphs [19]. It plays a central role in the proof of the lemma.

THEOREM 5.4. *Let G be a factor-critical graph. There is a sequence $G_1, \ldots, G_r$ of graphs such that G_1 is the one-vertex graph, G_i is obtained from G_{i-1} by gluing a single path with an odd number of edges having only its endvertices v and w in common with G_{i-1} (we allow the case $v = w$), and $G_r = G$. All graphs $G_1, \ldots, G_r$ are factor-critical.* □

Proof of Lemma 5.3. In the proof, we will refer to edges in B and R as *blue* and *red* edges, respectively. If a subgraph of G through v intersects both B and R, then it will be called *bichromatic*; otherwise it will be called *monochromatic*. We prove the lemma by induction on the number r of ears in an ear decomposition of G; see Theorem 5.4. The result holds trivially if $r = 0$. Now suppose that G can be obtained from some of its factor-critical subgraphs H by the addition of one ear P. If v is not a vertex of H, then the result holds. Assume now that v is a vertex of H. Note that H cannot have a bichromatic minimum edge cover, because otherwise the same would be true for G. By the induction hypothesis, H has two factor-critical subgraphs H_B and H_R such that $H = H_B \cup H_R$, H_B and H_R have only vertex v in common, $\delta_{H_B}(v) = B \cap E(H)$, and $\delta_{H_R}(v) = R \cap E(H)$. Up to symmetry, we have to treat four cases.

Case 1. The endpoints of P are both equal to v. Ear P has to be monochromatic because otherwise there would be a minimum edge cover of G that intersects both B and R. If $\delta_P(v) \subseteq B$, then we let $G_B = H_B \cup P$ and $G_R = H_R$. Else $\delta_P(v) \subseteq R$, and we let $G_B = H_B$ and $G_R = H_R \cup P$.

Case 2. One endpoint of P is v and the other in $H_B - v$. In this case the edge of P incident to v has to be blue because otherwise G would have a bichromatic minimum edge cover. We take $G_B = H_B \cup P$ and $G_R = H_R$.

Case 3. Both endpoints of P are in $H_B - v$. Then we simply let $G_B = H_B \cup P$ and $G_R = H_R$.

Case 4. One endpoint of P is in $H_B - v$ and the other in $H_R - v$. This case is impossible because we can easily construct a bichromatic minimum edge cover of G. □

The next result is the main result of this subsection. It enables us, with the help of Lemma 5.3, to transform any nontrivial factor-critical graph into a facet-defining $\{0, \frac{1}{2}\}$-cut for the linear ordering polytope which is nearly surface-shaped.

PROPOSITION 5.5. *Let G_0 be a nontrivial factor-critical graph, let G be the graph obtained from G_0 by replacing each edge by a path of length* 3*, let D be any representation of G with vertex included in X, and let $\mathcal{C} = \{C_v : v \in V(G)\}$ denote the collection of dicycles associated to D. Then the* 2*-packing inequality of $\mathcal{C}$ is facet-defining for the linear ordering polytope whenever G has no extra-bad vertex with respect to D.*

Proof. By a standard technique for proving that certain inequalities define facets of the linear ordering polytope (see Reinelt [24]), it suffices to show the following claims:

(i) for each dicycle C_v in $\mathcal{C}$ there is a perfect matching of $G - v$ and a corresponding set of arcs in D whose removal kills all dicycles of D except C_v;
(ii) whenever s and t are nodes of D such that neither st nor ts is an arc of D, there is a minimum feedback arc set which intersects every s-t dipath and every t-s dipath.

It is fairly easy to prove claim (i) by adapting the proof of Lemma 5.2. Indeed, let M be a perfect matching of $G - v$, and let F be the corresponding set of double arcs in D. Then $D - F$ cannot contain a dicycle other than C_v because otherwise there would exist a cycle in G and a vertex w on this cycle with $\deg_G(w) = 2$ which is not covered by M and distinct from v, a contradiction.

We now prove claim (ii). Let $e_s = \{v_s, w_s\}$ and $e_t = \{v_t, w_t\}$ be the unique edges of G such that s is incident to the double arc corresponding to e_s and t is incident to the double arc corresponding to e_t. Because neither st nor ts is an arc of D, we have $e_s \neq e_t$. Let d denote the minimum distance in G between an endvertex of e_s and an endvertex of e_t.

Case 1. $d \geq 3$. Let N be a minimum edge cover of G, and let F be the corresponding minimum feedback arc set of D. For every s-t dipath or t-s dipath P in D, the corresponding subgraph $H(P)$ of G contains a path whose length is at least three. Because of the way G was constructed, this path has an internal vertex v of degree 2 in G. One of the two edges incident to v has to be included in N, so F intersects P.

Case 2. $d = 2$. There is a path in G from e_s to e_t that has length 2. Let z be the intermediate vertex of this path. Any length-2 path from e_s to e_t must coincide with the latter path because the girth of G is at least 9. Let N be a minimum edge cover of G containing one of the two edges of the length-2 path from e_s to e_t, and let F denote the corresponding minimum feedback arc set. Now it is not difficult to verify that F intersects every s-t dipath and every t-s dipath.

Case 3. $d = 1$. Without loss of generality we can assume that v_s and v_t are adjacent. Any other path from e_s to e_t has length at least 6 because G has girth at least 9. Let N be any minimum edge cover of G containing $\{v_s, v_t\}$, and let F denote the corresponding minimum feedback arc set of D. Again, it is quite clear that F intersects every s-t dipath and every t-s dipath.

Case 4. $d = 0$. Without loss of generality, we can assume that $v_s = v_t$. For convenience, let us refer to the vertex $v_s = v_t$ as vertex v. Then e_s and e_t determine two intervals in the complete cyclic order at v, namely, the intervals determined by the double arcs on sC_vt and tC_vs, respectively. Because G has no extra-bad vertex, v is not extra-bad, and there is a minimum edge cover N of G containing edges from both intervals. Let F be the minimum feedback arc set of D corresponding to N. Then F intersects every s-t dipath and every t-s dipath in D. □

Avoiding extra-bad vertices in G is always possible. Indeed, if G has no cutvertex, then Lemma 5.3 implies that G has no bad vertices. Whenever a cutvertex v of G is extra-bad, we can "repair" it with the help of Lemma 5.3 by moving one of the blue edges into the middle of the interval of red edges in the complete cyclic order on $\delta_G(v)$ determined by the representation.

Assume now that G_0 is any graph with minimum degree at least 2. Again, let G be the graph obtained from G_0 by substituting a path of length 3 for each edge. Then G admits a representation D. Let $\mathcal{C}$ denote the associated 2-packing. Since it may be that G has an even number of vertices, instead of considering the 2-packing inequality of $\mathcal{C}$ we consider the valid inequality

$$\sum_{ij \in D} 2x_{ij} \geq 2\tau(D) \iff \sum_{ij \in D} x_{ij} \geq \tau(D). \tag{5.2}$$

Using essentially the same arguments as above, we can show that (5.2) is facet-defining only if G (and hence G_0) is factor-critical and has no extra-bad vertices. In this case, (5.2) coincides with the 2-packing inequality of $\mathcal{C}$.

5.3. Constructing a facet for each nonorientable surface. Let G_0, G, D, and $\mathcal{C}$ be as in Proposition 5.5. By Lemma 5.3, representation D can always be chosen in such a way that G has no extra-bad vertices. Then, by Proposition 5.5, the 2-packing inequality of $\mathcal{C}$ is facet-defining. It follows from Lemma 5.1 that this inequality is a $\{0, \frac{1}{2}\}$-cut. Furthermore, the same lemma implies that the corresponding multiplier u is surface-shaped, provided that the link of ∞ in $\mathcal{K} = \mathcal{K}(u)$ is a cycle. A last consequence of Lemma 5.1 is that we have $\chi(\mathcal{K}) = 2 - r$, where $r = |E(G)| - |V(G)| + 1$ denotes the number of ears in any ear decomposition of G. Therefore, proving our final result is just a matter of choosing G_0 and D carefully enough.

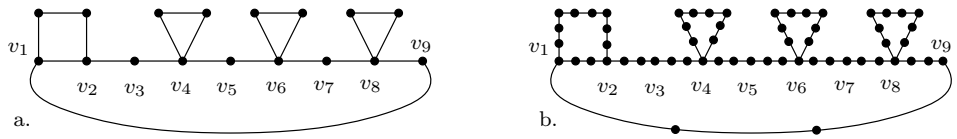


FIG. 5.3. *The graphs G_0 and G used in the proof of Theorem* 5.6.

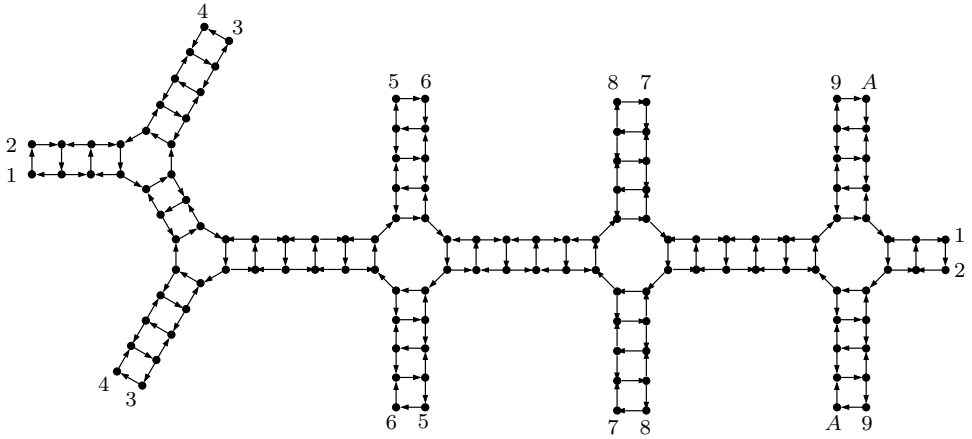


FIG. 5.4. *A representation of the graph G in Figure* 5.3.

THEOREM 5.6. *Each nonorientable surface S has a triangulation $\mathcal{K}$ such that $\mathcal{K} = \mathcal{K}(u)$ for some facet-defining multiplier u.*

Proof. Let $b = 2 - \chi(S)$. If $b = 1$, then S is homeomorphic to $\mathbb{N}_1$ and the theorem follows from Proposition 4.2. Else, consider the graph G_0 obtained from a odd cycle with vertices $v_1, v_2, \ldots, v_{2b-1}$ by attaching $b-1$ ears $P_1, \ldots, P_{b-1}$ of length 3 to the cycle, with endpoints v_1 and v_2 for P_1, and with both endpoints equal to $v_{2\alpha}$ for P_α, $\alpha > 1$. Note that G_0 is factor-critical. An example for $b = 5$ is given in Figure 5.3(a). Let G be the graph obtained from G_0 by replacing each edge by a path of length 3 (see Figure 5.3(b)). Now let D be a representation of G with the following properties. First, the node set of D has to be included in X (this is obviously always possible if we assume that n is large enough). Second, none of the vertices $v_4, v_6, \ldots, v_{2b-2}$ should be extra-bad. There is essentially one way to achieve this (see Figure 5.4). By Lemma 5.3, if none of the latter vertices is extra-bad, then no vertex of G is extra-bad. Let u denote any multiplier obtained from D as in subsection 5.1 and let $\mathcal{K} = \mathcal{K}(u)$. Our third and last requirement on representation D is that the cyclic orderings on the neighborhoods of v_1 and v_2 determined by the representation should be such that the link of ∞ in $\mathcal{K}$ is a cycle. Once again, this can be done (see Figure 5.4). The theorem now follows from Lemma 5.1 and Proposition 5.5. □

6. Conclusion. We have studied $\{0, \frac{1}{2}\}$-Chvátal–Gomory cuts derived from the standard relaxation of the linear ordering polytope. Certain of these cuts correspond to triangulated surfaces. We have shown that a surface has a triangulation yielding a facet of the linear ordering polytope if and only if it is nonorientable. Along the way, we have obtained a host of new facets. Indeed, most facets produced by Proposition 5.5 were not known before. Among the many questions raised by our findings, we note the following three:

(Q1) How can we estimate the number of facet-defining $\{0, \frac{1}{2}\}$-cuts, as a function of n? (Simulation is possible here.)

(Q2) Let S be a nonorientable surface and $\mathcal{K}$ be a triangulation of S. Is there always a facet-defining orientation of $\mathcal{K}$? More generally, what are the facet-defining orientations of $\mathcal{K}$?

(Q3) Is there a polynomial time algorithm solving the separation problem for a superclass of the facet-defining inequalities produced by Proposition 5.5?

Acknowledgments. We thank the two anonymous referees for providing many suggestions which greatly helped the author to improve the article. We also thank Jean-Paul Doignon, Michel Goemans, and Andreas Schulz for their early interest in the results of this article.

REFERENCES

[1] M. Armstrong, *Basic Topology*, Undergrad. Texts Math. 42, Springer-Verlag, New York, 1983.

[2] G. Bolotashvili, M. Kovalev, and E. Girlich, *New facets of the linear ordering polytope*, SIAM J. Discrete Math., 12 (1999), pp. 326–336.

[3] A. Caprara and M. Fischetti, $\{0, \frac{1}{2}\}$-*Chvátal-Gomory cuts*, Math. Program., 74A (1996), pp. 221–235.

[4] A. Caprara, M. Fischetti, and A. Letchford, *On the separation of maximally violated mod-k-cuts*, Math. Program., 87A (2000), pp. 37–56.

[5] M. Cohen and J.-C. Falmagne, *Random utility representation of binary choice probabilities: A new class of necessary conditions*, J. Math. Psychol., 34 (1990), pp. 88–94.

[6] R. Diestel, *Graph Theory*, 2nd ed., Grad. Texts in Math., 173, Springer-Verlag, New York, 2000.

[7] S. Fiorini, *Determining the automorphism group of the linear ordering polytope*, Discrete Appl. Math., 112 (2001), pp. 121–128.

[8] S. Fiorini, *Polyhedral Combinatorics of Order Polytopes*, Ph.D. thesis, Department of Mathematics, Université Libre de Bruxelles, Brussels, Belgium, 2001.

[9] P. Fishburn, *Induced binary probabilities and the linear ordering polytope: A status report*, Math. Social Sci., 23 (1992), pp. 67–80.

[10] Z. Galil and N. Megiddo, *Cyclic ordering is NP-complete*, Theoret. Comput. Sci., 5 (1977), pp. 179–182.

[11] M. Goemans and L. Hall, *The strongest facets of the acyclic subgraph polytope are unknown*, Integer Programming and Optim., 1084 (1996), pp. 415–429.

[12] M. Grötschel, M. Jünger, and G. Reinelt, *A cutting plane algorithm for the linear ordering problem*, Oper. Res., 32 (1984), pp. 1195–1220.

[13] M. Grötschel, M. Jünger, and G. Reinelt, *Facets of the linear ordering polytope*, Math. Programming, 33 (1985), pp. 43–60.

[14] R. Karp, *Reducibility among combinatorial problems*, in Complexity of Computer Computations, Plenum Press, New York, 1972, pp. 85–103.

[15] M. Koppen, *Random utility representation of binary choice probabilities: Critical graphs yielding critical necessary conditions*, J. Math. Psychol., 39 (1995), pp. 21–39.

[16] A. Letchford, *Binary clutter inequalities for integer programs*, Math. Programming, 98 (2003), pp. 201–221.

[17] A. Letchford and A. Lodi, *Polynomial-time separation of simple comb inequalities*, in Integer Programming and Combinatorial Optimization 9, W. Cook and A. Schulz, eds., Lecture Notes in Comput. Sci., 2337, Springer, NY, 2002, pp. 93–108.

[18] J. Leung and J. Lee, *More facets from fences for linear ordering and acyclic subgraph polytopes*, Discrete Appl. Math., 50 (1994), pp. 185–200.

[19] L. Lovász and M. Plummer, *Matching Theory*, Ann. Discrete Math., 29, North–Holland, Amsterdam, 1986.

[20] J. Matousek, *Using the Borsuk–Ulam Theorem*, Lect. Topol. Methods Combin. Geom., Springer, New York, 2003.

[21] B. Mohar and C. Thomassen, *Graphs on Surfaces*, John Hopkins University Press, Baltimore, MD, 2001.

[22] A. Newman, *Approximating the Maximum Acyclic Subgraph*, Master's thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology,

Cambridge, MA, 2000.
[23] A. Newman and S. Vempala, *Fences are futile: On relaxations for the linear ordering problem*, in Integer Programming and Combinatorial Optimization, K. Aardal and B. Gerards, eds., Lecture Notes in Comput. Sci., 2081, 2001, pp. 333–347.
[24] G. Reinelt, *The Linear Ordering Problem: Algorithms and Applications*, Res. Exp. Math., 8, Heldermann-Verlag, Berlin, 1985.
[25] P. Seymour, *Packing directed circuits fractionally*, Combinatorica, 15 (1995), pp. 281–288.
[26] R. Suck, *Geometric and combinatorial properties of the polytope of binary choice probabilities*, Math. Social Sci., 23 (1992), pp. 81–102.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 913–919

MOD-2 CUTS GENERATION YIELDS THE CONVEX HULL OF BOUNDED INTEGER FEASIBLE SETS*

C. GENTILE†, P. VENTURA†, AND R. WEISMANTEL‡

Abstract. This paper focuses on the outer description of the convex hull of all integer solutions to a given system of linear inequalities. It is shown that if the given system contains lower and upper bounds for the variables, then the convex hull can be produced by iteratively generating so-called mod-2 cuts only. This fact is surprising and might even be counterintuitive, since many integer rounding cuts exist that are not mod-2, i.e., representable as the $\{0, \frac{1}{2}\}$ combination of the given constraint system. The key, however, is that in general many more rounds of mod-2 cut generation are necessary to produce the final description than in the traditional integer rounding procedure.

Key words. integer programming, mod-2 cuts, convex hull

AMS subject classifications. 90C10, 90C57, 52B05

DOI. 10.1137/04061831X

1. Introduction.

One of the fundamental results in the theory of linear integer programming states that the convex hull of all integer points in the intersection of finitely many rational half-spaces is a polyhedron. This polyhedron, which we denote by $\mathcal{P}_I$ in the following, can be described by linear inequalities that one obtains in finitely many steps by integer rounding [8].

Let $\mathcal{P}^0 = \mathcal{P} = \{x \in \mathbb{R}^n | Ax \leq b\}$ be a relaxation of $\mathcal{P}_I$; a single step of the integer rounding procedure consists of taking all inequalities $u^T Ax \leq \lfloor u^T b \rfloor$ with $u \in \mathbb{R}^n_+$ and $u^T A \in \mathbb{Z}^n$ and adding them to $\mathcal{P}^0$, obtaining the next relaxation $\mathcal{P}^1$, to which we refer as the first closure of $\mathcal{P}$.

It has been recently shown in [6] that optimizing over the first closure of a polyhedron is $\mathcal{NP}$-hard. This explains that one cannot expect to turn this nice concept of integer rounding into an effective and stand-alone algorithmic tool. The question emerges whether instead of considering the first closure of a polyhedron, one can resort to a weaker relaxation that is algorithmically more tractable. One relaxation that appears particularly appealing for many combinatorial optimization problems is defined as the closure of a polyhedron associated with a special family of rounding cuts. These cuts have been introduced in [3] and are referred to as mod-2 cuts.

More precisely, if $\mathcal{P} = \{x \in \mathbb{R}^n | Ax \leq b\}$ with $A \in \mathbb{Z}^{m \times n}$, then a mod-2 cut is an inequality of the form $\frac{1}{2}u^T Ax \leq \lfloor \frac{1}{2}u^T b \rfloor$, where $u_i \in \{0, 1\}$ for all $i = 1, \ldots, m$ and $\frac{1}{2}u^T A \in \mathbb{Z}^n$; i.e., $u^T A \equiv 0 \bmod 2$.

Among the many important examples of mod-2 cuts we mention the blossom inequalities for the matching problem, the comb inequalities for the traveling salesman problem, the odd-cycle inequalities for the stable set problem or for the set covering problem, and the odd-cycle inequalities in quadratic 0-1 optimization [1].

*Received by the editors May 13, 2005; accepted for publication (in revised form) June 5, 2006; published electronically December 5, 2006. This work has been partially supported by the UE Marie Curie Research Training Network 504438 ADONET.

http://www.siam.org/journals/sidma/20-4/61831.html

†Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti" del CNR, Viale Manzoni 30, 00185 Rome, Italy (gentile@iasi.cnr.it, ventura@iasi.cnr.it).

‡Department for Mathematics/IMO, Otto-von-Guericke-Universität Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany (weismant@imo.math.uni-magdeburg.de).

Mod-2 cuts are a particular subclass of the more general mod-k cuts, which are defined as the inequalities of the form $\frac{1}{k}u^TAx \leq \lfloor \frac{1}{k}u^Tb \rfloor$, with $u_i \in \{0, \dots, k-1\}$ for all $i = 1, \dots, m$ and $\frac{1}{k}u^TA \in \mathbb{Z}^n$; i.e., $u^TA \equiv 0 \bmod k$.

Although the problem of separating mod-2 cuts is $\mathcal{NP}$-hard in general, it can be solved in polynomial time if the constraint matrix meets certain properties (see [3] and [9]). Interestingly, [4] showed that there is a polynomial time algorithm for separating a subclass of mod-k cuts for any prime number k. Computational studies about the effectiveness of mod-k cuts and mod-2 cuts in particular are shown in [7], [12], and [11].

These results suggest that mod-2 cuts are an interesting object to study in further depth. Our paper contributes to this topic by showing that under mild assumptions a description of the integer polyhedron can be obtained by iteratively generating mod-2 cuts only.

In the remainder of this paper we will focus on bounded integer programming problems in inequality form. We will, in addition, assume that lower and upper bounds for the variables are available. More precisely, for $A \in \mathbb{Z}^{m \times n}$, $b \in \mathbb{Z}^m$, and $v \in \mathbb{Z}^n$, the feasible set of integer points is described as

$$\mathcal{P} = \{x \in \mathbb{Z}^n : Ax \leq b,\ -Ix \leq 0,\ Ix \leq v\}.$$

We define $\mathcal{P}_I = conv(\mathcal{P} \cap \mathbb{Z}^n)$.

DEFINITION 1.1. *Let $A \in \mathbb{Z}^{m \times n}$, $b \in \mathbb{Z}^m$, $v \in \mathbb{Z}^n$,*

$$\tilde{A} = \begin{pmatrix} A \\ -I \\ I \end{pmatrix} \quad \text{and} \quad \tilde{b} = \begin{pmatrix} b \\ 0 \\ v \end{pmatrix}.$$

We denote an initial system with

$$S = S^{(0)} = (\tilde{A}, \tilde{b}).$$

The first mod-2 closure of the system S is

$$S^{(1)} = \begin{pmatrix} \tilde{A}, & \tilde{b} \\ \frac{1}{2}u^T\tilde{A}, & \lfloor \frac{1}{2}u^T\tilde{b} \rfloor \quad \text{for all } u \in \{0,1\}^{m+2n} \text{ s.t. } \frac{1}{2}u^T\tilde{A} \in \mathbb{Z}^n \end{pmatrix}.$$

For $t \in \mathbb{Z}_+$, $t \geq 2$, we define recursively $S^{(t)} = (S^{(t-1)})^{(1)}$ to be t-th mod-2 closure of S.

Given any system $S = (A, b)$, let $\mathcal{P}(S)$ denote the corresponding polyhedron $\{x \in \mathbb{R}^n : Ax \leq b\}$.

Remark 1.1. Without loss of generality, we can assume that each column of matrix A contains at least one positive entry. In fact, if this is not the case, i.e., there exists a column $a^i \leq 0$, we can apply the variable substitution $x'_i = v_i - x_i$.

The main result of this paper is a proof of the fact that, by generating mod-2 cuts iteratively, we can produce the convex hull of the integer feasible solutions.

THEOREM 1.1. *There exists $t \in \mathbb{Z}_+$ such that $\mathcal{P}(S^{(t)}) = \mathcal{P}_I$.*

Our proof requires that we make use of properties of the mod-2 closure that we summarize in section 2. Section 3 is devoted to the proof of the main theorem.

2. Properties of the mod-2 closure. This section develops structural properties of mod-2 closures of polyhedra. In particular, we first show that the iterative applications of mod-2 cuts provide a respective dominating inequality for any inequality of the starting system $S^{(0)}$ that is not a lower bound.

LEMMA 2.1. *Let $S^{(0)}$ be a system as introduced in Definition* 1.1*, and let $x_i \le v_i$ be an upper bound inequality contained in $S^{(0)}$. There exists a finite integer t such that $S^{(t)}$ contains both the inequalities $x_i \le v_i$ and $x_i \le v_i'$, with $v_i' \in \mathbb{Z}$ and $v_i' \le v_i$.*

Proof. By Remark 1.1, there exists an inequality

$$a^T x \le a_0 \tag{1}$$

of the system $Ax \le b$ such that $a_i > 0$. Then, $S^{(1)}$ contains the inequality ${a'}^T x \le a_0'$, obtained as a mod-2 cut from the sum of (1), lower bound inequalities for variables x_j with $j \ne i$ such that $a_j > 0$ and a_j odd, and upper bounds for variables x_j such that $a_j < 0$ and a_j odd. Iterating this procedure, after a finite number of steps t' we get an upper bound inequality $x_i \le v_i + \delta$. If $\delta \le 0$, the proof is complete; otherwise, in subsequent rounds we generate mod-2 cuts with multipliers $\frac{1}{2}$ from

$$\begin{aligned} x_i &\le v_i, \\ x_i &\le v_i + \delta. \end{aligned}$$

This gives

$$x_i \le v_i + \left\lfloor \frac{1}{2}\delta \right\rfloor.$$

The argument applies iteratively and shows that after $\lceil \log_2(\delta) \rceil$ steps a second copy of $x_i \le v_i$ is included in some system $S^{(t)}$. ∎

LEMMA 2.2. *Let $S^{(0)}$ be a system as introduced in Definition* 1.1*, and let $a^T x \le a_0$ be an inequality of the system $Ax \le b$ that is not an upper or a lower bound. There exists $t \in \mathbb{Z}_+$ such that $S^{(t)}$ contains both the inequalities $a^T x \le a_0$ and $a^T x \le a_0'$, with $a_0' \in \mathbb{Z}$ and $a_0' \le a_0$.*

Proof. The system $S^{(1)}$ contains the inequality

$$\sum_{\substack{i=1 \\ a_i \text{ even}}}^{n} \frac{a_i}{2} x_i + \sum_{\substack{i=1 \\ a_i \text{ odd}}}^{n} \frac{a_i - 1}{2} x_i \le \left\lfloor \frac{a_0}{2} \right\rfloor. \tag{2}$$

Then the system $S^{(2)}$ contains two copies of inequality (2). If we consider the original inequality $a^T x \le a_0$, the two copies of inequality (2), and the upper bounds constraints $x_i \le v_i$ for all i such that a_i is odd, and sum them up with multipliers $\frac{1}{2}$, we derive that an inequality of the form $a^T x \le a_0 + \delta$, where $\delta \in \mathbb{Z}$, is contained in $S^{(3)}$. If $\delta \le 0$, we are done; otherwise, we apply the same procedure described in the proof of Lemma 2.1, obtaining a second copy of $a^T x \le a_0$ included in the system $S^{(t)}$, for some $t \in \mathbb{Z}_+$. ∎

Our next example illustrates that upper bounds on the variables are needed for Lemma 2.2 to be true.

Example 2.1. Consider the feasible set described as

$$\{(x_1, x_2) \in \mathbb{Z}_+^2 \mid -3x_1 + 5x_2 \le 8\}. \tag{3}$$

One may observe that, using only lower bounds and the initial inequality, it is not possible to derive a copy of $-3x_1 + 5x_2 \leq 8$. The reason is that both numbers -3 and 5 are odd. Therefore, all the inequalities belonging to any mod-2 closure attain a ratio of the two coefficients that is strictly less than $-3/5$. In order to prove this, we use induction on the number t. Suppose that the generic system $S^{(t)}$ does not contain any inequality $a^T x \leq a_0$ with $a_1 \leq 0$, $a_2 \geq 0$, and $\frac{a_1}{a_2} \geq -\frac{3}{5}$, except the original inequality $-3x_1 + 5x_2 \leq 8$. We prove that this property also applies to $S^{(t+1)}$. To this end let $\bar{a}^T x \leq \bar{a}_0$ be any inequality in the system $S^{(t+1)}$. In order to achieve a highest possible ratio $\bar{a}_1/\bar{a}_2$, we can assume that $\bar{a}^T x \leq \bar{a}_0$ is a mod-2 cut from the initial constraint $-3x_1 + 5x_2 \leq 8$ and some other inequalities of $S^{(t)}$. Let $b^T x \leq b_0$ denote the sum of these other inequalities in $S^{(t)}$ that we need to derive the mod-2 cut $\bar{a}^T x \leq \bar{a}_0$. From the hypothesis of the induction we conclude that $\frac{b_1}{b_2} < -\frac{3}{5}$ with b_1 and b_2 odd. It then follows from elementary algebraic manipulations that $\frac{\bar{a}_1}{\bar{a}_2} = \frac{(-3+b_1)/2}{(5+b_2)/2} < -\frac{3}{5}$. This shows that upper bounds on the variables are needed for Lemma 2.2 to be true.

By Lemma 2.1 and Lemma 2.2, the two observations below easily follow.

Observation 2.1. Let $a^T x \leq a_0$ be derived by the inequalities of system $S^{(0)}$ summed up with multipliers $u \in \{0, 1, \ldots, k-1\}^{m+2n}$, with $k \in \mathbb{Z}_+$. By Lemma 2.1 and Lemma 2.2, there exists $t \in \mathbb{Z}_+$ such that $S^{(t)}$ contains, for each inequality of $S^{(0)}$ that is not a lower bound, $2u_i$ copies of a dominating constraint. Therefore, $S^{(t+1)}$ contains an inequality dominating $a^T x \leq a_0$.

Observation 2.2. Consider a mod-2 cut $a^T x \leq a_0$ obtained with multipliers $u \in \{0, \frac{1}{2}\}^{m+2n}$ from the system $S^{(0)}$ of inequalities. If we substitute one of the inequalities of the system with an inequality that dominates it, we obtain a mod-2 cut dominating $a^T x \leq a_0$. This is possible by adding or removing some lower bound inequalities on the variables with odd coefficients in the left-hand side.

To finally prove the key lemma, we first need the following result.

LEMMA 2.3. *If x is prime and y mod $x \neq 0$, there exist $i, j, \gamma \in \mathbb{Z}_+$ such that $ix + jy = 2^\gamma$; i.e., $ix + jy$ is a power of 2.*

Proof. Fermat's little theorem (FLT) states that if p is prime and $a \in \mathbb{Z}$, with a mod $p \neq 0$, then there exists $\alpha \in \mathbb{Z}_+$ such that $a^{p-1} = 1 + \alpha p$. Therefore, if we first apply FLT with $p = x$ and $a = y$, then there exists $\alpha \in \mathbb{Z}$ such that

$$y^{x-1} = 1 + \alpha x. \tag{4}$$

Then by applying FLT with $a = 2^q$, for some number q and $p = x$, there exists $\beta \in \mathbb{Z}$ such that

$$(2^q)^{x-1} = 1 + \beta x, \tag{5}$$

and, for a sufficiently large value of q, $\beta > \alpha$. Let us now subtract (4) from (5),

$$(\beta - \alpha)x + y^{x-1} = (2^q)^{x-1} = 2^{q(x-1)},$$

and then fix $i = \beta - \alpha$, $j = y^{x-2}$, and $\gamma = q(x-1)$. This proves the lemma. □

Resorting to Lemmas 2.1, 2.2, 2.3, and to Observations 2.1 and 2.2, we are now ready to prove that every mod-k cut can be obtained by generating mod-2 cuts iteratively.

LEMMA 2.4. *Let $S^{(0)}$ be a system as introduced in Definition* 1.1. *Let $a^T x \leq a_0$ be a mod-k cut for $\mathcal{P}(S)$; i.e., $a^T = \frac{1}{k} u^T \tilde{A} \in \mathbb{Z}^n$ and $a_0 = \lfloor \frac{1}{k} u^T \tilde{b} \rfloor$ with $u \in$*

$\{0,1,\dots,k-1\}^{m+2n}$. *There exists a number* $t \in \mathbb{Z}_+$ *such that an inequality dominating* $a^T x \le a_0$ *is part of the system* $S^{(t)}$.*

Proof. The inequality $u^T \tilde{A} x \le u^T \tilde{b}$ can be represented as

$$ka^T x \le ka_0 + r, \tag{6}$$

where $r \in \{0,1,\dots,k-1\}$.

Without loss of generality we may assume that k is prime. In fact, if k is not prime, by using induction on the prime factorization of k, let $k_1 > 1$ and $k_2 > 1$ be two integers such that $k_1 k_2 = k$. By Observation 2.1, there exists $t' \in \mathbb{Z}_+$ such that $S^{(t')}$ contains an inequality $q^T x \le q_0$ that dominates (6). By Observation 2.2, the inequality obtained from $q^T x \le q_0$ by applying two successive integer roundings with k_1 and k_2 dominates $a^T x \le a_0$; indeed, $a^T x \le a_0 + \lfloor \frac{\lfloor r/k_1 \rfloor}{k_2} \rfloor$ and $0 \le \lfloor \frac{\lfloor r/k_1 \rfloor}{k_2} \rfloor \le \lfloor \frac{r}{k} \rfloor = 0$.

In the following we will sometimes refer to Observation 2.2 in order to claim that a certain inequality $a^T x \le a_0$ is included in a certain system $S^{(t)}$. In order to be precise, we mean that $S^{(t)}$ either contains the inequality $a^T x \le a_0$ or contains an inequality dominating it. We do not distinguish between these cases, in order to keep our notation simple.

By Observation 2.1, there exists a finite integer t' such that $S^{(t')}$ contains $ka^T x \le ka_0 + r$. The mod-2 cut obtained from the latter inequality, lower bounds for variables x_i with $a_i > 0$ and a_i odd, and upper bounds for variables x_j with $a_j < 0$ and a_j odd, dominates an inequality of the form $\lfloor \frac{k}{2} \rfloor a^T x \le c_0$, for some $c_0 \in \mathbb{Z}$. Therefore, we assume in the following that $\lfloor \frac{k}{2} \rfloor a^T x \le c_0$ is contained in the system $S^{(t'+1)}$.

Let π_0 be the integer odd (and positive) number such that $k = 2^{\alpha_0}\pi_0 + 1$, for some $\alpha_0 \in \mathbb{Z}_+$. Then the inequality

$$\pi_0 a^T x \le \pi_0 a_0 + \delta_0 \tag{7}$$

is contained in $S^{(t'+\alpha_0)}$. If $\delta_0 \le 0$, then the proof is finished. So we assume $\delta_0 > 0$.

In the next iterations, by considering mod-2 cuts obtained from inequalities (6) and (7) with multipliers $\frac{1}{2}$, we will produce a mod-2 inequality of the form $\pi_1 a^T x \le \pi_1 a_0 + \delta_1$, where $\pi_1 2^{\alpha_1} = (k + \pi_0)$ with $\alpha_1 \in \mathbb{Z}_+$, $\pi_1 \in \mathbb{Z}_+$, π_1 odd, and, since δ_1 is obtained by α_1 consecutive mod-2 roundings, $0 \le \delta_1 \le \lfloor \frac{r+\delta_0}{2^{\alpha_1}} \rfloor$.

The crucial observation here is that, if $\frac{\delta_0}{\pi_0} \ge 1$, then $\frac{\delta_1}{\pi_1} < \frac{\delta_0}{\pi_0} - \frac{1}{2k}$; i.e., the ratio $\frac{\delta_0}{\pi_0}$ is decreased by, at least, the fixed amount $\frac{1}{2k}$. In fact, since $r \le k-1$ and $k > \pi_0$, we obtain this relation by applying the following manipulations:

$$\begin{aligned}
\frac{\delta_0}{\pi_0} - \frac{\delta_1}{\pi_1} &\ge \frac{\delta_0}{\pi_0} - \left\lfloor \frac{r+\delta_0}{2^{\alpha_1}} \right\rfloor \frac{1}{\pi_1} = \frac{\delta_0}{\pi_0} - \left\lfloor \frac{(r+\delta_0)\pi_1}{k+\pi_0} \right\rfloor \frac{1}{\pi_1} \ge \frac{\delta_0}{\pi_0} - \frac{r+\delta_0}{k+\pi_0} \\
&\ge \frac{\delta_0}{\pi_0} - \frac{k-1+\delta_0}{k+\pi_0} = \frac{\delta_0 k - k\pi_0 + \pi_0}{\pi_0(k+\pi_0)} \\
&\ge \frac{\pi_0 k - k\pi_0 + \pi_0}{\pi_0(k+\pi_0)} = \frac{1}{k+\pi_0} > \frac{1}{2k}.
\end{aligned}$$

*Note added in proof: Recently Sanjeeb Dash communicated to us that a similar result was shown in [2], also for mod-q cuts with $q > 2$ (note that this generalization is easily derivable also from our proof). However, the proof of Lemma 2.4 is stronger as (a) we use only $\{0, \frac{1}{2}\}$ multipliers, while in [2] all values $k/2$, for any $k \in \mathbb{Z}_+$, are considered; (b) in [2], the upper bounds $-x_i \ge -1$ are used to determine the necessary inequality $\alpha \ge s_0$, while we use the upper bounds on the variables only to produce copies of the original inequalities. In particular, if one assumes that copies of the initial inequalities can be generated someway (e.g., by considering the hypothesis of $\{0, \frac{1}{2}\}$ multipliers), then the result also applies to the unbounded case.

Therefore, repeating the above procedure, after a finite number β of iterations we will produce a system $S^{(t'')}$ that contains the inequalities

$$\begin{array}{ll} ka^T x \le ka_0 + r & \text{and} \\ \pi_\beta a^T x \le \pi_\beta a_0 + \delta_\beta & \text{with } \delta_\beta/\pi_\beta < 1. \end{array}$$

Then, since k is prime and $\pi_\beta < k$, by Lemma 2.2 and Lemma 2.3, there exists a number $t''' \ge t''$ such that $S^{(t''')}$ contains i copies of $ka^T x \le ka_0 + r$ and j copies of $\pi_\beta a^T x \le \pi_\beta a_0 + \delta_\beta$, where $ik + j\pi_\beta = 2^\gamma$.

Then $S^{(t'''+\gamma)}$ contains the inequality

$$\frac{ik + j\pi_\beta}{2^\gamma} a^T x \le \frac{ik + j\pi_\beta}{2^\gamma} a_0 + \delta,$$

with $0 \le \delta \le \lfloor \frac{ir + j\delta_\beta}{2^\gamma} \rfloor$. Since $r \le k-1$ and $\delta_\beta \le \pi_\beta - 1$, $\lfloor \frac{ir + j\delta_\beta}{2^\gamma} \rfloor = 0$. This completes the proof. □

Example 2.2. Consider the feasible set described as

$$\{(x_1, x_2) \in \mathbb{Z}^2_+ | 7x_1 + 14x_2 \le 20\}.$$

The inequality $x_1 + 2x_2 \le 2$ can be derived from multiplying $7x_1 + 14x_2 \le 20$ by $1/7$ and rounding the right-hand-side. Following Lemma 2.4, with one mod-2 operation, we obtain the first inequality of type (7):

$$3x_1 + 6x_2 \le 10.$$

We then produce the next inequality by using the previous two,

$$5x_1 + 10x_2 \le 15.$$

Iterating the procedure, we generate

$$3x_1 + 6x_2 \le 8,$$

which is another inequality of type (7), where $\pi_2 = 3$, $\delta_2 = 2$, and $\beta = 2$; that is, $\delta_2/\pi_2 < 1$. Finally, we consider one copy of $7x_1 + 14x_2 \le 20$ and three copies of $3x_1 + 6x_2 \le 8$, we divide by 16 (corresponding to 4 consecutive mod-2 operations), and obtain $x_1 + 2x_2 \le 2$.

3. Proof of the main theorem.

THEOREM 3.1. *Let $S^{(0)}$ be a system as introduced in Definition* 1.1. *There exists $t \in \mathbb{Z}_+$ such that $\mathcal{P}(S^{(t)}) = \mathcal{P}_I$.*

Proof. It suffices to show that there exists $t_1 \in \mathbb{Z}_+$ such that the inequalities describing the first Chvátal–Gomory closure $\mathcal{P}^1$ are part of the system $S^{(t_1)}$. The polyhedron $\mathcal{P}^1$ is described by the Gomory cuts

$$\mathcal{P}^1 = \{x \in \mathbb{R}^n | u^T Ax \le \lfloor u^T b \rfloor \text{ for all } u \ge 0, u^T A \in \mathbb{Z}^n\}.$$

Every such inequality $u^T Ax \le \lfloor u^T b \rfloor$ with $u = (p_1/q_1, \ldots, p_m/q_m)$ and $p_i \in \mathbb{Z}_+$, $q_i \in \mathbb{Z}_+ \setminus \{0\}$ is a mod-k cut with $k = \prod_{i=1}^m q_i$. In fact, there is a finite representation for $\mathcal{P}^1$ (see [10]) as $\mathcal{P}^1 = \{x \in \mathbb{R}^n |\ u^T Ax \le \lfloor u^T b \rfloor, \text{for all } u \in H(\mathcal{C})\}$, where $H(\mathcal{C})$ is the Hilbert basis of the cone $\mathcal{C} = \{u^T A | u \in \mathbb{R}^m_+\}$.

By Lemma 2.4 every inequality $u^T Ax \leq \lfloor u^T b \rfloor$ with $u \in H(\mathcal{C})$ is contained in $S^{(t')}$ for some $t' \in \mathbb{Z}_+$. Therefore, there exists $t_1 \in \mathbb{Z}_+$ such that $S^{(t_1)}$ contains all the inequalities $u^T Ax \leq \lfloor u^T b \rfloor$ for all $u \in H(\mathcal{C})$, i.e., $\mathcal{P}(S^{(t_1)}) \subseteq \mathcal{P}^1$.

By a theorem of Chvátal [5], $\mathcal{P}_I = \mathcal{P}^\tau$ for some integer $\tau \in \mathbb{Z}_+$. Therefore, we can repeat the same argument for $\mathcal{P}^2, \ldots, \mathcal{P}^\tau$ by finding systems $S^{(t_2)}, \ldots, S^{(t_\tau)}$ such that $\mathcal{P}(S^{(t_i)}) \subseteq \mathcal{P}^i$ for all $i = 2, \ldots, \tau$. This gives the result. □

Our proof of Theorem 3.1 strongly relies on Lemma 2.2. As Example 2.1 illustrates, Lemma 2.2 is not true if upper bounds on the variables are not present. As a consequence, the proof of Theorem 3.1 does not apply to systems without upper bounds. It is, however, straightforward to extend the proof to the case in which we allow multipliers $\{0, \frac{1}{2}, 1\}$ for generating cuts as opposed to having $\{0, \frac{1}{2}\}$ multipliers only.

However, in the case of Example 2.1, for instance, the facet defining inequality $-2x_1 + 3x_2 \leq 4$, derivable as a mod-5 cut from the starting system, can still be obtained as a mod-2 cut in the system $S^{(7)}$. This fact might indicate that even an extension of Theorem 3.1 to the unbounded integer programming case could be true.

Acknowledgments. We thank Giovanni Rinaldi for his helpful suggestions. We also thank the anonymous referees for their useful remarks.

REFERENCES

[1] E. Boros, Y. Crama, and P.L. Hammer, *Chvátal cuts and odd cycle inequalities in quadratic 0-1 optimization*, SIAM J. Discrete Math., 5 (1992), pp. 163–177.

[2] S. R. Buss and P. Clote, *Cutting planes, connectivity, and threshold logic*, Arch. Math. Logic, 35 (1996), pp. 33-62.

[3] A. Caprara and M. Fischetti, *$\{0, \frac{1}{2}\}$-Chvátal–Gomory cuts*, Math. Program., 74 (1996), pp. 221–235.

[4] A. Caprara, M. Fischetti, and A. N. Letchford, *On the separation of maximally violated mod-k cuts*, Math. Program., 87 (2000), pp. 37–56.

[5] V. Chvátal, *Edmonds polytopes and a hierarchy of combinatorial problems*, Discrete Math., 4 (1973), pp. 305–337.

[6] F. Eisenbrand, *On the membership problem for the elementary closure of a polyhedron*, Combinatorica, 19 (1999), pp. 297–300.

[7] M. Fischetti and A. Lodi, *Optimizing over the first Chvátal closure*, in Integer Programming and Combinatorial Optimization, Proceedings of the 11th International IPCO Conference, Berlin, Germany, 2005, Lecture Notes in Comput. Sci., M. Jünger and V. Kaibel, eds., Springer, New York, 2005, pp. 12–22.

[8] R. E. Gomory, *Outline of an algorithm for integer solutions to linear programs*, Bull. Amer. Math. Soc., 64 (1958), pp. 275–278.

[9] A. N. Letchford, *Binary clutter inequalities for integer programs*, Math. Program., 98 (2003), pp. 201–221.

[10] A. Schrijver, *Theory of Linear and Integer Programming*, Wiley, New York, 1986.

[11] A. M. Verweij, *Selected Applications of Integer Programming: A Computational Study*, Ph.D. thesis, Utrecht University, Utrecht, The Netherlands, 2000.

[12] K. Wenger, *Generic Cut Generation Methods for Routing Problems*, Ph.D. thesis, University of Heidelberg, Heidelberg, Germany, 2003.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 920–931

FACTORING FINITE ABELIAN GROUPS BY SUBSETS WITH MAXIMAL SPAN*

SÁNDOR SZABÓ†

Abstract. We say that a finite abelian group has the Rédei property if it does not admit factorization into two normalized subsets that both span the whole group. It will be shown that subgroups inherit the Rédei property from the group. Then four constructions are described to exhibit groups without the Rédei property. Using these we further narrow the list of p-groups that might have the Rédei property.

Key words. factorization of finite abelian groups, Rédei property, full-rank tilings

AMS subject classifications. Primary, 20K01; Secondary, 52C22

DOI. 10.1137/05063828X

1. Introduction. Let G be a finite abelian group written multiplicatively with identity element e. Let $A_1, \dots, A_n$ be subsets of G. If each $g \in G$ is uniquely representable in the form

$$g = a_1 \cdots a_n, \quad a_1 \in A_1, \dots, a_n \in A_n,$$

then we say that the equation $G = A_1 \cdots A_n$ is a factorization of G. A subset A of G is called normalized if $e \in A$. A factorization is called normalized if each factor is a normalized subset. In 1965, Rédei [7] proved that if $G = A_1 \cdots A_n$ is a normalized factorization, where G is a finite abelian group and each $|A_i|$ is a prime, then at least one of the factors is a subgroup of G.

If G is a direct product of cyclic groups of orders $t_1, \dots, t_n$, then we say that G is of type $(t_1, \dots, t_n)$. In order to avoid trivial direct factors we assume that $t_i \geq 2$ for each i, $1 \leq i \leq n$. When each of $t_1, \dots, t_n$ is a power of a prime p, then G is a p-group and n is called the rank of G. A group of type $(p, \dots, p)$ is called an elementary p-group. In 1970, Rédei [8, 9] asked if G is of type (p, p, p) and $G = AB$ is a normalized factorization, then does it follow that either $\langle A \rangle \neq G$ or $\langle B \rangle \neq G$. Here $\langle A \rangle$ stands for the span of A in G, that is, for the smallest subgroup of G that contains A. (See Problem 5 in [8].) In general we say that a finite abelian group has the Rédei property if it does not admit factorization into two normalized subsets that both span the whole group. If $G = \{e\}$, then from a normalized factorization $G = AB$ it follows that $A = B = \{e\}$ and $\langle A \rangle = \langle B \rangle = G$. Therefore by our definition G does not have the Rédei property. It seems reasonable to modify the definition of the Rédei property such that in the $|G| = 1$ singular case G has the Rédei property by definition.

In 1972 Swenson [15] (independently of Rédei) raised the question whether each finite cyclic group has the Rédei property. In 1979 Sands [10] proved that groups of type (p^α, q^β), where p, q are distinct primes have the Rédei property. Fraser and Gordon [5] have shown that elementary p-groups of rank $(p + 1)$ do not have the Rédei property provided $p \geq 5$. In 1985 Szabó [11] exhibited cyclic groups without

*Received by the editors August 17, 2005; accepted for publication (in revised form) July 14, 2006; published electronically December 5, 2006. This work was supported by TET Foundation grant OMFB 00746/06.

http://www.siam.org/journals/sidma/20-4/63828.html

†Institute of Mathematics and Informatics, University of Pécs, Ifjúság u. 6, 7624 Pécs, Hungary (sszabo7@hotmail.com).

the Rédei property. The question of which elementary 2-groups possess the Rédei property originated in coding theory around 1996. (See [1] and [3].) This happened without knowledge of the above-mentioned developments. Recently Östergård and Vardy [6] settled this problem. They proved that an elementary 2-group has the Rédei property if and only if its rank is at most 9. The Rédei property of finite cyclic groups has appeared in a work of Tijdeman [16] in connection with Fourier analysis and in a work of De Felice [4] related to variable length codes. Further details on the history can be found in [2].

In [12] it was established that only a small fraction of the finite abelian groups can have the Rédei property. In this paper we will show that subgroups inherit the Rédei property of the group. Then we present constructions to further narrow the family of finite abelian groups that can have the Rédei property.

2. Extending a factorization. In this section we will show that the family of finite abelian groups with the Rédei property is closed under the operations of forming subgroups and factor groups.

THEOREM 1. *Let G be a finite abelian group and let H be a subgroup of G. If H does not have the Rédei property, then neither does G.*

Proof. We divide the proof into smaller steps.

(1) Let G be a finite abelian group and let H be a subgroup of G with prime order. Let A, B be normalized subsets of G such that $G/H = (AH)/H \cdot (BH)/H$ is a factorization of G/H. Here

$$\begin{aligned}(AH)/H &= \{aH : a \in A\},\\ (BH)/H &= \{bH : b \in B\}.\end{aligned}$$

There are various choices for A and B. However, we require that A and B have only one element from each coset of G/H. From a given coset we can choose any element as a representative freely. Assume that

$$\langle (AH)/H \rangle = \langle (BH)/H \rangle = G/H$$

and there is an element $a \in A \setminus \{e\}$ such that $[(A \setminus \{a\})H]/H$ spans G/H. In other words we assume that $(AH)/H$ spans G/H even if we remove the coset aH. Choose an $h \in H \setminus \{e\}$ and set $A_1 = \big(A \setminus \{a\}\big) \cup \{ah\}$, $B_1 = BH$. We claim the following.

(i) $G = A_1B_1$ is a normalized factorization.

(ii) $\langle B_1 \rangle = G$.

(iii) There is a choice of A and h for which $\langle A_1 \rangle = G$.

In order to prove (i) note that from the factorization $G/H = (AH)/H \cdot (BH)/H$ it follows that for each $g \in G$ the coset gH can be represented in the form

$$gH = (aH)(bH), \quad a \in A, \quad b \in B,$$

and so the elements of the product AB form a complete set of representatives in G modulo H. This means $G = (AB)H$. The computation

$$\begin{aligned}A_1B_1 &= A_1(BH)\\ &= (A_1H)B\\ &= \big[(A \setminus \{a\}) \cup \{ah\}\big]HB\\ &= \big[(AH \setminus aH) \cup ahH\big]B\\ &= \big[(AH \setminus aH) \cup aH\big]B\\ &= (AH)B\\ &= (AB)H\\ &= G\end{aligned}$$

gives that $G = A_1B_1$. On the other hand, the equations

$$\begin{aligned} |G| &= |A||B||H|, \\ |A_1| &= |A|, \\ |B_1| &= |B||H| \end{aligned}$$

clearly hold because we required that A and B have only one element from each coset of G/H. Therefore, $G = A_1B_1$ is a factorization of G.

To prove (ii) note that from $\langle (BH)/H\rangle = G/H$ it follows that for each $g \in G$ there are elements $b_1, \dots, b_s \in B$ and integers $\beta(1), \dots, \beta(s)$ such that gH can be represented in the form $gH = (b_1^{\beta(1)}H)\cdots(b_s^{\beta(s)}H)$. Thus $g = b_1^{\beta(1)}\cdots b_s^{\beta(s)}h_1$ with some $h_1 \in H$. From $b_1^{\beta(1)}\cdots b_s^{\beta(s)} \in \langle B\rangle$ and $h_1 \in H$ we can see that $g \in \langle BH\rangle = \langle B_1\rangle$. This gives that $\langle B_1\rangle = G$.

To verify (iii) set $A^* = A\setminus\{a\}$. Since $\langle (A^*H)/H\rangle = G/H$, for each $g \in G$ there are $a_1, \dots, a_s \in A^*$ and integers $\alpha(1), \dots, \alpha(s)$ such that $gH = (a_1^{\alpha(1)}H)\cdots(a_s^{\alpha(s)}H)$ and consequently $g = a_1^{\alpha(1)}\cdots a_s^{\alpha(s)}h_1$ for some $h_1 \in H$. This means that each coset gH contains an element from $\langle A^*\rangle$. Therefore $G = \langle A^*\rangle H$ and consequently $G = \langle A_1\rangle H$. If there is a coset gH that contains two distinct elements c_1, c_2 from $\langle A^*\rangle$, then there is an $h_1 \in H$ such that $c_1c_2^{-1} = h_1$. From $c_1, c_2 \in \langle A^*\rangle$ it follows that $h_1 \in \langle A^*\rangle \subset \langle A_1\rangle$. Here $h_1 \neq e$ since $c_1 \neq c_2$. In this case $H \subset \langle A_1\rangle$ and we get $G = \langle A_1\rangle$. We may assume that each coset gH contains exactly one element from $\langle A^*\rangle$. In particular the coset aH contains exactly one element c from $\langle A^*\rangle$. Suppose first that $ah \neq c$ for some $h \in H \setminus \{e\}$. Then there is an $h_1 \in H$ for which $(ah)c^{-1} = h_1$. From $ah \in \langle A_1\rangle$ and $c \in \langle A^*\rangle \subset \langle A_1\rangle$ it follows that $h_1 \in \langle A_1\rangle$. But $h_1 \neq e$ as $ah \neq c$. Therefore $H \subset \langle A_1\rangle$ and so $\langle A_1\rangle = G$. Let us turn to the case when $ah = c$ for each $h \in H \setminus \{e\}$. It can happen only when $|H| = 2$. In this case let us choose A such that c is the representative in A chosen from the coset aH. Now $a = c$ and consequently $ah \neq c$ for $h \neq e$. We can conclude as before that $\langle A_1\rangle = G$.

(2) For a finite abelian group G a factor group of G is always isomorphic to a subgroup of G. Thus the result in step (1) can be reformulated in the following way. Let H be a subgroup of G with prime index. If $H = AB$ is a normalized factorization, $\langle A\rangle = \langle B\rangle = H$, there is an element $a \in A\setminus\{e\}$ such that $\langle A\setminus\{a\}\rangle = H$, or there is an element $b \in B \setminus \{e\}$ such that $\langle B \setminus \{b\}\rangle = H$, then there is a normalized factorization $G = A_1B_1$ with $\langle A_1\rangle = \langle B_1\rangle = G$.

For each subgroup H of G there is a chain of subgroups

$$H = H_0 \subset H_1 \subset \cdots \subset H_n = G$$

such that the index $|H_{i+1} : H_i|$ is a prime for each i, $1 \leq i \leq n-1$. The construction above can be used several times unless each element of $A_i \setminus \{e\}$ is needed to span H_i and each element of $B_i \setminus \{e\}$ is needed to span H_i.

(3) Let G be a finite abelian group and let H be a subgroup of G. We assume that H does not have the Rédei property and we want to show that G does not have the Rédei property either. If $|H| = 1$, then by definition H has the Rédei property and so for the remaining part of the proof we may assume that $|H| \geq 2$. We claim that if $H = AB$ is a normalized factorization, where $|A| \geq 2$, $|B| \geq 2$, and each element of $A \setminus \{e\}$ is needed to span H and each element of $B \setminus \{e\}$ is needed to span H, then H has the Rédei property. Phrasing it differently, if the construction described in step (2) is obstructed, then H has the Rédei property under which circumstances one does not wish to carry out the construction. This claim has already been proved by Dinitz in the proof of Theorem 9 of [2].

This completes the proof. □

3. Four constructions. In this section we describe four factorization constructions we will use later.

THEOREM 2. *Let G be a group whose type is one of the following:*

$$\begin{array}{lll}
(ab, cd, 2, 2, 2), & a \geq 2, & b \geq 3, \quad c \geq 2, \quad d \geq 2, \\
(a, bc, 2, 2, 2, 2), & a \geq 3, & b \geq 2, \quad c \geq 2, \quad a \ \textit{is odd}, \\
(ab, 2, 2, 2, 2, 2, 2), & a \geq 2, & b \geq 4, \\
(a, 2, 2, 2, 2, 2, 2, 2), & a \geq 5, & a \ \textit{is odd}.
\end{array}$$

Then G does not have the Rédei property.

Proof. The proof is elementary and constructive.

(1) Let G be a group of type $(ab, cd, 2, 2, 2)$, where $a \geq 2$, $b \geq 3$, $c, d \geq 2$ with basis elements x, y, z_1, z_2, z_3 such that $|x| = ab$, $|y| = cd$, $|z_1| = |z_2| = |z_3| = 2$. Set

$$\begin{array}{ll}
A_1 = \{e, x, x^2, \ldots, x^{a-1}\}, & H_1 = \langle x^a \rangle, \\
A_2 = \{e, y, y^2, \ldots, y^{c-1}\}, & H_2 = \langle y^c \rangle, \\
A_3 = \{e, z_2, z_3, z_1 z_2 z_3\}, & H_3 = \langle z_1 \rangle, \\
A \ = A_1 A_2 A_3, & H \ = H_1 H_2 H_3.
\end{array}$$

Note that

$$\begin{aligned}
A_1 H_1 &= \langle x \rangle, \\
A_2 H_2 &= \langle y \rangle, \\
A_3 H_3 &= \langle z_1, z_2, z_3 \rangle.
\end{aligned}$$

We will use these observations several times later. The computation

$$\begin{aligned}
AH &= (A_1 A_2 A_3)(H_1 H_2 H_3) \\
&= (A_1 H_1)(A_2 H_2)(A_3 H_3) \\
&= \langle x \rangle \langle y \rangle \langle z_1, z_2, z_3 \rangle \\
&= G
\end{aligned}$$

shows that $G = AH$ is a factorization of G.

We modify the subgroup H to get a subset B. We remove the subsets

$$\begin{array}{ll}
x^{a(b-1)} H_2, & \\
y^{c(d-1)} H_3, & x^a y^{c(d-1)} H_3, \\
z_1 H_1 &
\end{array}$$

from H and add the subsets

$$\begin{array}{ll}
x^{a(b-1)} H_2 y, & \\
y^{c(d-1)} H_3 z_2, & x^a y^{c(d-1)} H_3 z_3, \\
z_1 H_1 x &
\end{array}$$

to H to get the subset B from H. We claim that

$$\begin{array}{lcl}
x^{a(b-1)} H_2 A & = & x^{a(b-1)} y H_2 A, \\
y^{c(d-1)} H_3 A & = & y^{c(d-1)} z_2 H_3 A, \\
x^a y^{c(d-1)} H_3 A & = & x^a y^{c(d-1)} z_3 H_3 A, \\
z_1 H_1 A & = & z_1 x H_1 A.
\end{array}$$

The next routine computation verifies the first equation. The remaining three can be checked in a similar way and we leave them for the reader.

$$\begin{aligned} x^{a(b-1)}yH_2A &= x^{a(b-1)}yH_2(A_1A_2A_3)\\ &= x^{a(b-1)}y(H_2A_2)A_1A_3\\ &= x^{a(b-1)}y\langle y\rangle A_1A_3\\ &= x^{a(b-1)}\langle y\rangle A_1A_3\\ &= x^{a(b-1)}(H_2A_2)A_1A_3\\ &= x^{a(b-1)}H_2(A_1A_2A_3)\\ &= x^{a(b-1)}H_2A. \end{aligned}$$

We claim that the subsets

$$\begin{array}{ll} x^{a(b-1)}H_2A, & \\ y^{c(d-1)}H_3A, & x^ay^{c(d-1)}H_3A,\\ z_1H_1A & \end{array}$$

are pairwise disjoint. Since the arguments are similar we will show that the first two subsets are disjoint and leave the remaining ones for the reader. First note that

$$\begin{aligned} x^{a(b-1)}H_2A &= x^{a(b-1)}H_2(A_1A_2A_3)\\ &= x^{a(b-1)}(H_2A_2)A_1A_3\\ &= x^{a(b-1)}\langle y\rangle A_1A_3,\\ y^{c(d-1)}H_3A &= y^{c(d-1)}H_3(A_1A_2A_3)\\ &= y^{c(d-1)}(H_3A_3)A_1A_2\\ &= y^{c(d-1)}\langle z_1,z_2,z_3\rangle A_1A_2. \end{aligned}$$

Assume the contrary—that g is a common element of these subsets. Considering the x-component of g leads to a contradiction.

It follows that $G = AB$ is a normalized factorization of G. It is plain that

$$x, y, z_2, z_3, z_1z_2z_3 \in A,$$

and so $\langle A\rangle = G$. Then

$$\begin{array}{llllll} y^{c(d-1)}z_2 & \in B, & y^{c(d-1)}z_1z_2 & \in \langle B\rangle & \text{imply} & z_1 \in \langle B\rangle,\\ xz_1 & \in B, & z_1 & \in \langle B\rangle & \text{imply} & x \in \langle B\rangle,\\ x^{a(b-1)}y & \in B, & x & \in \langle B\rangle & \text{imply} & y \in \langle B\rangle,\\ y^{c(d-1)}z_2 & \in B, & y & \in \langle B\rangle & \text{imply} & z_2 \in \langle B\rangle,\\ x^ay^{c(d-1)}z_3 & \in B, & x, y & \in \langle B\rangle & \text{imply} & z_3 \in \langle B\rangle. \end{array}$$

Therefore $\langle B\rangle = G$. This completes the construction.

We illustrate the construction in the $a = 2$, $b = 4$, $c = d = 2$ special case. The type of G is $(8, 4, 2, 2, 2)$. The element $x^\alpha y^\beta z_1^{\gamma(1)}z_2^{\gamma(2)}z_3^{\gamma(3)}$ of G is recorded simply by the exponents $(\alpha, \beta, \gamma(1), \gamma(2), \gamma(3))$. The elements of the sets A, H, B are listed in Tables 1(a)–1(c). The modified subsets in H and B are highlighted using ( ), [ ], { }, ⌊ ⌋, respectively.

(2) Let a, b, c be integers such that $a \geq 3$, $b, c \geq 2$, and a is odd. Let G be a group of type $(a, bc, 2, 2, 2, 2)$ with basis elements x, y, z_1, z_2, z_3, u, where $|x| = a$, $|y| = bc$, $|z_1| = |z_2| = |z_3| = |u| = 2$. Set

$$\begin{array}{ll} A_1 = \{e, xu\}, & H_1 = \langle x\rangle,\\ A_2 = \{e, y, y^2, \dots, y^{b-1}\}, & H_2 = \langle y^b\rangle,\\ A_3 = \{e, z_2, z_3, z_1z_2z_3\}, & H_3 = \langle z_1\rangle,\\ A = A_1A_2A_3, & H = H_1H_2H_3. \end{array}$$

TABLE 1(a)

A			
00000	00010	00001	00111
10000	10010	10001	10111
01000	01010	01001	01111
11000	11010	11001	11111

TABLE 1(b)

H			
00000	[00100]	{02000}	{02100}
20000	[20100]	⌊22000⌋	⌊22100⌋
40000	[40100]	42000	42100
(60000)	[60100]	(62000)	62100

TABLE 1(c)

B			
00000	[10100]	{02010}	{02110}
20000	[30100]	⌊22001⌋	⌊22101⌋
40000	[50100]	42000	42100
(61000)	[70100]	(63000)	62100

Note that

$$\begin{aligned} A_1H_1 &= \langle x, u\rangle, \\ A_2H_2 &= \langle y\rangle, \\ A_3H_3 &= \langle z_1, z_2, z_3\rangle. \end{aligned}$$

We can see that $G = AH$ is a factorization of G.

We modify the subgroup H to get a subset B. We remove the subsets

$$\begin{array}{ll} x^{a-1}H_2, & \\ y^{b(c-1)}H_3, & xy^{b(c-1)}H_3, \\ z_1H_1 & \end{array}$$

from H and add the subsets

$$\begin{array}{ll} x^{a-1}H_2y, & \\ y^{b(c-1)}H_3z_2, & xy^{b(c-1)}H_3z_3, \\ z_1H_1u & \end{array}$$

to H to get the subset B from H. It turns out that

$$\begin{aligned} x^{a-1}H_2A &= x^{a-1}yH_2A, \\ y^{b(c-1)}H_3A &= y^{b(c-1)}z_2H_3A, \\ xy^{b(c-1)}H_3A &= xy^{b(c-1)}z_3H_3A, \\ z_1H_1A &= z_1uH_1A \end{aligned}$$

and that the subsets

$$\begin{array}{ll} x^{a-1}H_2A, & \\ y^{b(c-1)}H_3A, & xy^{b(c-1)}H_3A, \\ z_1H_1A & \end{array}$$

are pairwise disjoint. From this it follows that $G = AB$ is a normalized factorization of G. It is clear that $\langle B\rangle = G$. Note that $xu \in A_1$. Since a is odd it follows that

Table 2(a)

A			
000000	000100	000010	001110
100001	100101	100011	101111
010000	010100	010010	011110
110001	110101	110011	111111

Table 2(b)

H			
000000	[001000]	{020000}	{021000}
100000	[101000]	⌊120000⌋	⌊121000⌋
(200000)	[201000]	(220000)	221000

Table 2(c)

B			
000000	[001001]	{020100}	{021100}
100000	[101001]	⌊120010⌋	⌊121010⌋
(210000)	[201001]	(230000)	221000

$(xu)^a = u$. This gives that $x, u \in \langle A\rangle$. Then one can verify that $\langle A\rangle = G$, completing the construction.

We illustrate the construction in the $a = 3$, $b = c = 2$ numerical case. The type of G is $(3, 4, 2, 2, 2, 2)$. The elements of the sets A, H, B are listed in Tables 2(a)–2(c).

(3) Let G be a group of type $(ab, 2, 2, 2, 2, 2)$, where $a \geq 2$, $b \geq 4$ with basis elements x, y_1, y_2, y_3, z_1, z_2, z_3 such that $|x| = ab$, $|y_1| = |y_2| = |y_3| = 2$, $|z_1| = |z_2| = |z_3| = 2$. Set

$$\begin{aligned} A_1 &= \{e, x, x^2, \dots, x^{a-1}\}, & H_1 &= \langle x^a\rangle,\\ A_2 &= \{e, y_2, y_3, y_1y_2y_3\}, & H_2 &= \langle y_1\rangle,\\ A_3 &= \{e, z_2, z_3, z_1z_2z_3\}, & H_3 &= \langle z_1\rangle,\\ A &= A_1A_2A_3, & H &= H_1H_2H_3. \end{aligned}$$

Let us observe that

$$\begin{aligned} A_1H_1 &= \langle x\rangle,\\ A_2H_2 &= \langle y_1, y_2, y_3\rangle,\\ A_3H_3 &= \langle z_1, z_2, z_3\rangle. \end{aligned}$$

Note that $G = AH$ is a factorization of G.

We modify the subgroup H to get a subset B by removing the subsets

$$\begin{aligned} &x^{a(b-1)}H_2, & &x^{a(b-2)}H_2,\\ &y_1H_3, & &x^ay_1H_3,\\ &z_1H_1 \end{aligned}$$

from H and adding the subsets

$$\begin{aligned} &x^{a(b-1)}H_2y_2, & &x^{a(b-2)}H_2y_3,\\ &y_1H_3z_2, & &x^ay_1H_3z_3,\\ &z_1H_1x \end{aligned}$$

TABLE 3(a)

A			
0000000	0000010	0000001	0000111
1000000	1000010	1000001	1000111
0010000	0010010	0010001	0010111
1010000	1010010	1010001	1010111
0001000	0001010	0001001	0001111
1001000	1001010	1001001	1001111
0111000	0111010	0111001	0111111
1111000	1111010	1111001	1111111

TABLE 3(b)

H			
0000000	[0000100]	{0100000}	{0100100}
2000000	[2000100]	⌊2100000⌋	⌊2100100⌋
⌈4000000⌉	[4000100]	⌈4100000⌉	4100100
(6000000)	[6000100]	(6100000)	6100100

TABLE 3(c)

B			
0000000	[1000100]	{0100001}	{0100101}
2000000	[3000100]	⌊2100010⌋	⌊2100110⌋
⌈4001000⌉	[5000100]	⌈4101000⌉	4100100
(6010000)	[7000100]	(6110000)	6100100

to H to get the subset B from H. One can verify that

$$\begin{aligned} x^{a(b-1)}H_2A &= x^{a(b-1)}H_2y_2A, \\ x^{a(b-2)}H_2A &= x^{a(b-2)}H_2y_3A, \\ y_1H_3A &= y_1H_3z_2A, \\ x^a y_1H_3A &= x^a y_1H_3z_3A, \\ z_1H_1A &= z_1xH_1A \end{aligned}$$

and that the subsets

$$\begin{aligned} &x^{a(b-1)}H_2A, \quad x^{a(b-2)}H_2A, \\ &y_1H_3A, \qquad\quad x^a y_1H_3A, \\ &z_1H_1A \end{aligned}$$

are pairwise disjoint. From this we can deduce that $G = AB$ is a normalized factorization of G. The reader can show that $\langle A\rangle = G$ and $\langle B\rangle = G$ which completes the construction.

We work out the $a = 2, b = 4$ case in detail. In this case G is of type $(8, 2, 2, 2, 2, 2, 2)$. The sets A, H, B are depicted in Tables 3(a)–3(c).

(4) Let a be an integer such that $a \geq 5$ and a is odd. Let G be a group of type $(a, 2, 2, 2, 2, 2, 2)$ with basis elements x, y_1, y_2, y_3, z_1, z_2, z_3, u, where $|x| = a$, $|y_1| = |y_2| = |y_3| = 2$, $|z_1| = |z_2| = |z_3| = |u| = 2$. Set

$$\begin{aligned} A_1 &= \{e, xu\}, & H_1 &= \langle x\rangle, \\ A_2 &= \{e, y_2, y_3, y_1y_2y_3\}, & H_2 &= \langle y_1\rangle, \\ A_3 &= \{e, z_2, z_3, z_1z_2z_3\}, & H_3 &= \langle z_1\rangle, \\ A &= A_1A_2A_3, & H &= H_1H_2H_3. \end{aligned}$$

The basic observations we use are the following:

$$\begin{aligned} A_1H_1 &= \langle x, u\rangle, \\ A_2H_2 &= \langle y_1, y_2, y_3\rangle, \\ A_3H_3 &= \langle z_1, z_2, z_3\rangle. \end{aligned}$$

Obviously $G = AH$ is a factorization of G.

We modify the subgroup H to get a subset B by removing the subsets

$$\begin{array}{ll} x^{a-1}H_2, & x^{a-2}H_2, \\ y_1H_3, & xy_1H_3, \\ z_1H_1 & \end{array}$$

from H and adding the subsets

$$\begin{array}{ll} x^{a-1}H_2y_2, & x^{a-2}H_2y_3, \\ y_1H_3z_2, & xy_1H_3z_3, \\ z_1H_1u & \end{array}$$

to H to get the subset B from H. One can verify that

$$\begin{array}{lcl} x^{a-1}H_2A & = & x^{a-1}H_2y_2A, \\ x^{a-2}H_2A & = & x^{a-2}H_2y_3A, \\ y_1H_3A & = & y_1H_3z_2A, \\ xy_1H_3A & = & xy_1H_3z_3A, \\ z_1H_1A & = & z_1uH_1A \end{array}$$

and that the subsets

$$\begin{array}{ll} x^{a-1}H_2A, & x^{a-2}H_2A, \\ y_1H_3A, & xy_1H_3A, \\ z_1H_1A & \end{array}$$

are pairwise disjoint. It is a consequence that $G = AB$ is a normalized factorization of G. The reader can check that $\langle B\rangle = G$. Using $xu \in A_1$ and that a is odd it follows that $(xu)^a = u \in \langle A\rangle$. Then $x \in \langle A\rangle$. One can see that $\langle A\rangle = G$ which completes the construction. □

The $a = 5$ case serves as an illustration. In this case G is of type $(5, 2, 2, 2, 2, 2, 2, 2)$. Tables 4(a)–4(c) list the elements of the sets A, H, B.

Table 4(a)

A			
00000000	00000100	00000010	00001110
10000001	10000101	10000011	10001111
00100000	00100100	00100010	00101110
10100001	10100101	10100011	10101111
00010000	00010100	00010010	00011110
10010001	10010101	10010011	10011111
01110000	01110100	01110010	01111110
11110001	11110101	11110011	11111111

TABLE 4(b)

H			
00000000	[00001000]	{01000000}	{01001000}
10000000	[10001000]	⌊11000000⌋	⌊11001000⌋
20000000	[20001000]	21000000	21001000
(30000000)	[30001000]	(31000000)	31001000
⌈40000000⌉	[40001000]	⌈41000000⌉	41001000

TABLE 4(c)

B			
00000000	[00001001]	{01000100}	{01001100}
10000000	[10001001]	⌊11000010⌋	⌊11001010⌋
20000000	[20001001]	21000000	21001000
(30100000)	[30001001]	(31100000)	31001000
⌈40010000⌉	[40001001]	⌈41010000⌉	41001000

4. *p*-groups. Let p be a prime. Let F_p be a family of p-groups whose types are on the following list or a subgroup of such a group.

$$\begin{array}{lll}
p=2, & (2^\alpha, 2^\beta, 2, 2), & \alpha \geq 3, \quad \beta \geq 2, \\
& (2^\alpha, 2, 2, 2, 2, 2), & \alpha \geq 3, \\
& (2^2, 2^2, 2, 2, 2, 2, 2, 2, 2), & \\
p=3, & (3^\alpha, 3^\beta, 3), & \alpha \geq 2, \quad \beta \geq 2, \\
& (3^\alpha, 3, 3, 3), & \alpha \geq 2, \\
& (3, 3, 3, 3, 3), & \\
p \geq 5, & (p^\alpha, p^\beta, p), & \alpha \geq 1, \quad \beta \geq 1.
\end{array}$$

THEOREM 3. *Let p be a prime and let G be a finite abelian p-group. If G has the Rédei property, then G is a member of the F_p family.*

Proof. Let G be a finite abelian p-group with the Rédei property. Let H be a subgroup of G. By Theorem 1, if H does not posses the Rédei property, then neither does G. So we may assume that H has the Rédei property. In the remaining part of the proof we deal with the $p = 2$, $p = 3$, $p \geq 5$ cases separately.

(1) In the $p = 2$ case H does not have the Rédei property if its type is one of the following:

$$\begin{array}{ll}
(2^2, 2^2, 2^2), & (2^3, 2^2, 2, 2, 2), \\
(2^3, 2, 2, 2, 2, 2, 2), & (2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2).
\end{array}$$

Theorem 4 of [12] and Theorem 1 of [6] settle the first and the last cases, respectively. The remaining cases follow from Theorem 2. Suppose that the type of G is

$$\left(2^{\alpha(1)}, \ldots, 2^{\alpha(r)}, 2^{\beta(1)}, \ldots, 2^{\beta(s)}, 2^{\gamma(1)}, \ldots, 2^{\gamma(t)}\right),$$

where

$$\alpha(1), \ldots, \alpha(r) \geq 3, \quad \beta(1) = \cdots = \beta(s) = 2, \quad \gamma(1) = \cdots = \gamma(t) = 1.$$

If $r + s \geq 3$, then the type of H can be chosen to be $(2^2, 2^2, 2^2)$, and so by Theorem 1, G does not have the Rédei property. This contradiction shows that $r + s \leq 2$. There are three possible choices for $r + s$ and six choices for r and s.

If $r = s = 0$ and $t \geq 10$, then H can be taken to be an elementary 2-group of rank 10 and so G does not have the Rédei property. Therefore in the $r = s = 0$ case

$t \le 9$. If $r = 1$, $s = 0$, $t \ge 6$, then the type of H can be chosen to be $(2^3, 2, 2, 2, 2, 2, 2)$, and so G does not have the Rédei property. Thus in the $r = 1$, $s = 0$ cases $t \le 5$. Continuing in this way we can verify the claim of the theorem in the $p = 2$ particular case.

(2) In the $p = 3$ case by Theorem 2 of [12], H does not have the Rédei property if its type is one of the following:

$$(3^2, 3^2, 3^2), \quad (3^2, 3^2, 3, 3),$$
$$(3^2, 3, 3, 3, 3), \quad (3, 3, 3, 3, 3, 3).$$

Suppose that the type of G is

$$\big(3^{\alpha(1)}, \dots, 3^{\alpha(r)}, 3^{\beta(1)}, \dots, 3^{\beta(s)}\big),$$

where

$$\alpha(1), \dots, \alpha(r) \ge 2, \quad \beta(1) = \cdots = \beta(s) = 1.$$

If $r \ge 3$, then the type of H can be set to be $(3^2, 3^2, 3^2)$, and so G does not have the Rédei property. Thus $r \le 2$. This leaves three choices for r. If $r = 0$ and $s \ge 6$, then H can be chosen to be an elementary 3-group of rank 6 and consequently G does not have the Rédei property. This contradiction implies that in the $r = 0$ case $s \le 5$ must hold. The remaining two cases can be treated in a similar way.

(3) In the $p \ge 5$ case we proceed as in the $p = 3$ case. However in [2] it was established that a group of type (p, p, p, p) does not have the Rédei property. This sorts out further groups.

The proof is complete. □

Let p be a prime and let G be an elementary p-group of rank n. When $p = 2$, the picture is complete. Namely, by the main result of [6], G has the Rédei property for $n \le 9$ and G does not have the Rédei property for $n \ge 10$. When $p = 3$ by Theorem 2.3 of [14], G has the Rédei property for $n \le 4$ and by Theorem 3, G does not have the Rédei property for $n \ge 6$. The $n = 5$ case is undecided. When $p \ge 5$ by Rédei's theorem, G has the Rédei property for $n \le 2$ and by a construction of [2], G does not have the Rédei property for $n \ge 4$. The $n = 3$ case is undecided. It is a conjecture of Rédei from 1970 that G does have the Rédei property for $n = 3$. This conjecture is verified for $p \le 11$ in [13].

PROBLEM 1. *Characterize all elementary p-groups with the Rédei property.*

PROBLEM 2. *Find all p-groups with the Rédei property.*

REFERENCES

[1] G. D. COHEN, S. LITSYN, A. VARDY, AND G. ZÉMOR, *Tiling of binary spaces*, SIAM J. Discrete Math., 9 (1996), pp. 393–412.

[2] M. DINITZ, *Full rank tilings of finite abelian groups*, SIAM J. Discrete Math., 20 (2006), pp. 160–170.

[3] T. ETZION AND A. VARDY, *On perfect codes and tilings: Problems and solutions*, SIAM J. Discrete Math., 11 (1998), pp. 205–223.

[4] C. DE FELICE, *An application of Hajós factorization to variable length codes*, Theoret. Comput. Sci., 164 (1996), pp. 223–252.

[5] O. FRASER AND B. GORDON, *Solution to a problem of A. D. Sands*, Glas. Math. J., 20 (1979), pp. 115–117.

[6] P. R. J. ÖSTERGÅRD AND A. VARDY, *Resolving the existence of full-rank tilings of binary Hamming spaces*, SIAM J. Discrete Math., 18 (2004), pp. 382–387.

[7] L. Rédei, *Die neue Theorie der endlichen abelschen Gruppen und Verallgemeinerung des Hauptsatzes von Hajós*, Acta Math. Acad. Sci. Hungar., 16 (1965), pp. 329–373.
[8] L. Rédei, *Lückenhafte Polynome über Endlichen Körpern*, Birkhäuser Verlag, Basel, 1970.
[9] L. Rédei, *Lacunary Polynomials over Finite Fields*, North–Holland, Amsterdam, 1973.
[10] A. D. Sands, *On Keller's conjecture for certain cyclic groups*, Proc. Edinburgh Math. Soc., 22 (1979), pp. 17–21.
[11] S. Szabó, *A type of factorization of finite abelian groups*, Discrete Math., 54 (1985), pp. 121–125.
[12] S. Szabó, *Constructions related to the Rédei property of groups*, J. London Math. Soc., 73 (2006), pp. 701–715.
[13] S. Szabó and C. Ward, *Factoring elementary groups of prime cube order into subsets*, Math. Comp., 67 (1998), pp. 1199–1206.
[14] S. Szabó and C. Ward, *Factoring groups having periodic maximal subgroups*, Bol. Soc. Mat. Mexicana 3, 5 (1999), pp. 327–333.
[15] C. B. Swenson, *Direct Sum Subset Decompositions of Abelian Groups*, Ph.D. thesis, Washington State University, Pullman, WA, 1972.
[16] R. Tijdeman, *Decomposition of the integers as a direct sum of two subsets*, in Number Theory (Paris 1992–1993), S. David, ed., London Math. Soc. Lecture Note Ser. 215, Cambridge University Press, Cambridge, UK, 1995, pp. 261–276.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 932–946

COMPUTING THE TUTTE POLYNOMIAL ON GRAPHS OF BOUNDED CLIQUE-WIDTH*

OMER GIMÉNEZ†, PETR HLINĚNÝ‡, AND MARC NOY†

Abstract. The Tutte polynomial is a notoriously hard graph invariant, and efficient algorithms for it are known only for a few special graph classes, like for those of bounded tree-width. The notion of clique-width extends the definition of cographs (graphs without induced P_4), and it is a more general notion than that of tree-width. We show a subexponential algorithm (running in time $\exp O(n^{1-\varepsilon})$) for computing the Tutte polynomial on graphs of bounded clique-width. In fact, our algorithm computes the more general U-polynomial.

Key words. Tutte polynomial, cographs, clique-width, subexponential algorithm, U polynomial

AMS subject classifications. 05C85, 68R10

DOI. 10.1137/050645208

1. Introduction. The Tutte polynomial $T(G;x,y)$ of a graph G is a powerful invariant with many applications, not only in graph theory but also in other fields such as knot theory and statistical physics. One important feature of the Tutte polynomial is that by evaluating $T(G;x,y)$ at special points in the plane one obtains several parameters of G. For example, $T(G;1,1)$ is the number of spanning trees of G and $T(G;2,1)$ is the number of forests (that is, spanning acyclic subgraphs) of G.

A question that has received much attention is whether the evaluation of $T(G;x,y)$ at a particular point of the (x,y) plane can be done in polynomial time. Jaeger, Vertigan, and Welsh [9] showed that evaluating the Tutte polynomial of a graph is #P-hard at every point except those lying on the hyperbola $(x-1)(y-1)=1$ and eight special points, including at $(1,1)$ which gives the number of spanning trees. In each of the exceptional cases the evaluation can be done in polynomial time. On the other hand, the Tutte polynomial can be computed in polynomial time for graphs of bounded tree-width. This was obtained independently by Andrzejak [2] and Noble [13]. Recently Hliněný [8] has obtained the same result for matroids of bounded branch-width representable over a fixed finite field, which is a substantial generalization of the previous results; see [6] for additional references on this subject.

In this paper we study the problem of computing the Tutte polynomial for cographs and, more generally, for graphs of bounded clique-width. A graph has clique-width $\leq k$ if it can be constructed using k labels and the following four operations: create a new vertex with label i, take the disjoint union of several labeled graphs, add all edges between vertices of label i and label j, and relabel all vertices with label i to have label j. An expression defining a graph G built from the above four operations

*Received by the editors November 15, 2005; accepted for publication (in revised form) May 18, 2006; published electronically December 11, 2006. An extended abstract has been published in [15]. http://www.siam.org/journals/sidma/20-4/64520.html

†Department of Applied Mathematics, Technical University of Catalonia, Jordi Girona 1–3, 08034 Barcelona, Spain (omer.gimenez@upc.edu, marc.noy@upc.edu). The work of the first author was supported by Beca Fundació Crèdit Andorrà and Project MTM2005-08618-C02-01. The work of the second author was supported by Project MTM2005-08618-C02-01.

‡Faculty of Informatics, Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic (hlineny@fi.muni.cz). This author's work was supported by Czech research grant GAČR 201/05/0050 (VŠB–TU Ostrava), and by the Institute of Theoretical Computer Science, project 1M0545.

using k labels is a *k-expression* for G. When we say that a graph G has clique-width $\leq k$, we always assume that a k-expression for G is given.

A *cograph* is a graph of clique-width at most two; equivalently, it is a graph containing no induced path P_4 on four vertices (see section 4).

Although a class of graphs with bounded tree-width has also bounded clique-width, the converse is not true. For instance, complete graphs have clique-width two. It is well known that all problems expressible in monadic second order logic of incidence graphs become polynomial time solvable when restricted to graphs of bounded tree-width. For bounded clique-width less is true: all problems become polynomial time solvable if they are expressible in monadic second-order logic using quantifiers on vertices but not on edges (adjacency graphs) [3].

Our main results are as follows.

THEOREM 1.1. *The Tutte polynomial of a cograph with n vertices can be computed in time* $\exp\left(O(n^{2/3})\right)$.

THEOREM 1.2. *Let G be a graph with n vertices of clique-width k along with a k-expression for G as an input. Then the Tutte polynomial of G can be computed in time* $\exp\left(O(n^{1-1/(k+2)})\right)$.

Theorem 1.2 is not likely to hold for the class of all graphs, since it would imply the existence of a subexponential algorithm for 3-coloring, hence also for 3-SAT; which is considered highly unlikely in the computer science community. Of course, the main open question is whether there exists a *polynomial time* algorithm for computing the Tutte polynomial of graphs of bounded clique-width. We discuss this issue in the last section.

In fact, our algorithms compute not only the Tutte polynomial, but the so-called U polynomial (see [14]), which is a stronger polynomial invariant. Moreover, we may skip the requirement of having a k-expression for G as an input in Theorem 1.2, if we do not care about an asymptotic behavior in the exponent: Just to prove a subexponential upper bound we may use the approximation algorithm for clique-width by Oum [15] and Seymour [16] (see section 4).

Since our algorithms are quite complicated, for an illustration, we first present in section 2 a simplified algorithm computing the number of forests in a cograph, that is, evaluating $T(G; 2, 1)$ for graphs of clique-width ≤ 2. In section 3 we extend the algorithm to the computation of the full Tutte polynomial on cographs. Section 4 then discusses more closely the notion of graph clique-width. Finally, in section 5 we prove our main result, Theorem 1.2.

2. Forests in cographs. The problem of computing the number of spanning forests in an arbitrary graph is $\#P$-hard [9]. In this section we show the existence of a subexponential algorithm for the class of cographs.

2.1. Definition and signatures. The class of *cographs* is defined recursively as follows:

1. A single vertex is a cograph.
2. A disjoint union of two cographs is a cograph.
3. A complete union of two cographs is a cograph.

Here a *complete union* of two graphs $G \boxtimes H$ means the operation of taking a disjoint union $G \dot\cup H$, and adding all edges between $V(G)$ and $V(H)$. (We will avoid using the notation $\oplus$ at all since in the context of clique-width it is used to denote a disjoint union while in some other areas it denotes a complete union.) A cograph G can be represented by a tree, whose internal nodes correspond to operations (2)

and (3) above, and whose leaves correspond to single vertices. We call such a tree an *expression* for G.

For example, all cliques are cographs, and the complement of a cograph is a cograph again. Cographs have a long history of theoretical and algorithmic research. In particular, they are known to be exactly the graphs without induced paths on four vertices (P_4-free).

Let us call a *signature* a multiset of positive integers. The *size* $\|\boldsymbol{\alpha}\|$ of a signature $\boldsymbol{\alpha}$ is the sum of all elements in $\boldsymbol{\alpha}$, respecting repetition in the multiset. A signature $\boldsymbol{\alpha}$ of size n is represented by the *characteristic vector* $\boldsymbol{\alpha} = (a_1, a_2, \ldots, a_n)$, where there are $a_i \geq 0$ elements i in $\boldsymbol{\alpha}$, and $\sum_{i=1}^{n} i \cdot a_i = n$. (On the other hand, the *cardinality* of $\boldsymbol{\alpha}$ is $|\boldsymbol{\alpha}| = \sum_{i=1}^{n} a_i$, as usual.) An important fact we need is the following. Recall that $\Theta(f)$ is a usual shortcut for all functions having the same asymptotic growth rate as f.

LEMMA 2.1. *There are* $2^{\Theta(\sqrt{n})}$ *distinct signatures of size* n.

Proof. Each signature actually corresponds to a partition of n into an unordered sum of positive integers. It is well known [11, Chapter 15] that there are $2^{\Theta(\sqrt{n})}$ of those. □

We call a *double-signature* a multiset of ordered pairs of nonnegative integers, excluding the pair $(0,0)$. The *size* $\|\boldsymbol{\beta}\|$ of a double-signature $\boldsymbol{\beta}$ is the sum of all $(x+y)$ for $(x,y) \in \boldsymbol{\beta}$, respecting repetition in the multiset. We, moreover, need to prove the following.

LEMMA 2.2. *There are* $\exp\left(\Theta(n^{2/3})\right)$ *distinct double-signatures of size* n.

Lemma 2.2 is a particular case of Lemma 5.1, which is proved in section 5.

LEMMA 2.3. *A double-signature* $\boldsymbol{\beta}$ *of size* n *has at most* $\exp\left(O(n^{2/3})\right)$ *different submultisets (i.e., of different characteristic vectors).*

Proof. Just count all double-signatures of size $\leq n$. □

2.2. Forest signature table. Let us now consider a graph G and a forest $U \subset G$. The signature $\boldsymbol{\alpha}$ of U is the multiset of sizes of the connected components of U. (Obviously, $\boldsymbol{\alpha}$ has size $|V(G)|$ if U spans all the vertices.) We call a *(spanning) forest signature table* of the graph G a vector $\boldsymbol{T}$ (realized as an array $\boldsymbol{T}[\ldots]$); such that $\boldsymbol{T}$ records, for each signature $\boldsymbol{\alpha}$ of size $|V(G)|$, the number of spanning forests $U \subset G$ having signature $\boldsymbol{\alpha}$ (as $\boldsymbol{T}[\boldsymbol{\alpha}]$). For simplicity we usually skip the word "spanning" if it is clear from the context. We are going to compute the forest signature table of a cograph G recursively along the way G has been constructed. For that we describe two algorithms.

Let us denote by Σ_G the set of all signatures of size $|V(G)|$. It is important to keep in mind that signatures are considered as multisets, which also concerns set operations. For instance, a *multiset union* $\boldsymbol{\gamma} \uplus \boldsymbol{\delta}$ is obtained as the sum of the characteristic vectors of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, and a *multiset difference* $\boldsymbol{\gamma} \setminus \boldsymbol{\delta}$ is defined by the nonnegative difference of those.

ALGORITHM 2.4. *Combining the spanning forest signature tables of graphs* F *and* G *into the one of the disjoint union* $H = F \dot{\cup} G$.

`Input:` *Graphs* F, G, *and their forest signature tables* $\boldsymbol{T}_F, \boldsymbol{T}_G$.
`Output:` *The forest signature table* $\boldsymbol{T}_H$ *of* $H = F \dot{\cup} G$.

```
create empty table T_H of forest signatures of size |V(H)|;
for all signatures α_F ∈ Σ_F, α_G ∈ Σ_G do
   set α = α_F ⊎ α_G (a multiset union);
   add T_H[α] += T_F[α_F] · T_G[α_G];
done.
```

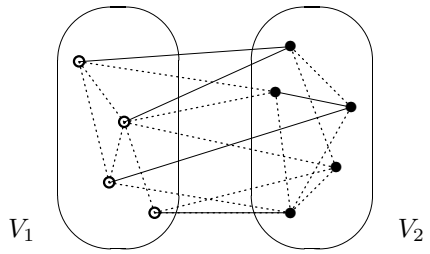


FIG. 2.1. *An illustration of a spanning forest (solid edges) in a graph partitioned into V_1 (white) and V_2 (black); this forest has double-signature $\{(2,1),(1,2),(0,1),(1,1)\}$.*

The running time of this algorithm is proportional to the number of pairs of signatures (α_F, α_G), which is $\exp\big(O(n^{2/3})\big)$, where $n = |V(H)|$; this is due to Lemma 2.2 and the fact that we have the $O(\,)$ expression in the exponent.

The second algorithm is, on the other hand, more complicated. It involves double-signatures with the following meaning: Consider a graph H with vertices partitioned into two parts $V(H) = V_1 \cup V_2$, and a forest $U \subset H$. The double-signature of U (w.r.t. V_1, V_2) is the multiset of pairs $\big(|V(C) \cap V_1|, |V(C) \cap V_2|\big)$ over all connected components C of U; see an illustration in Figure 2.1.

The idea behind the algorithm is to obtain the double-signatures (for $V_1 = V(F)$ and $V_2 = V(G)$) of the spanning forests in $H = F \boxtimes G$ from the signatures of the spanning forests in F and G. For every pair of forests $U_F \subset F$ and $U_G \subset G$, the algorithm iteratively counts the different ways in which each component of U_G can be joined to components of U_F. During the process, double-signatures are needed to distinguish between former vertices of F and of G in already joined components. In fact, the algorithm works with pairs of signatures $\boldsymbol{\alpha}_F$ and $\boldsymbol{\alpha}_G$, that is, with whole classes of forests instead of particular forests. We also remark that a submultiset is considered among all possible selections of repeated elements, as if they were pairwise distinct.

ALGORITHM 2.5. *Combining the spanning forest signature tables of graphs F and G into the one of the complete union $H = F \boxtimes G$.*

`Input:` *Graphs F, G, and their forest signature tables $\boldsymbol{T}_F, \boldsymbol{T}_G$.*
`Output:` *The forest signature table $\boldsymbol{T}_H$ of $H = F \boxtimes G$.*

```
    create empty table T_H of forest signatures of size |V(H)|;
    for all signatures α_F ∈ Σ_F, α_G ∈ Σ_G do
       set z = |V(F)|;
       create empty table X of forest double-signatures of size z;
          // Imagine particular forests U_F ⊂ F, U_G ⊂ G of signature α_F, α_G,
          // and a selected component C ⊂ U_G of size c.
       set X[double-signature {(a,0) : a ∈ α_F}] = 1;
       for each c ∈ α_G (with repetition) do
          create empty table X' of forest double-signatures of size z + c;
          for all double-signatures β of size z s.t. X[β] > 0 do
(†)          for all submultisets γ ⊆ β (with repetition) do
                set d_1 = Σ_{(x,y)∈γ} x, d_2 = Σ_{(x,y)∈γ} y;
                set double-signature β' = (β \ γ) ⊎ {(d_1, d_2 + c)};
(*)             add X'[β'] += X[β] · Π_{(x,y)∈γ} cx;
             done
          done
```

```
      set X = X', z = z + c;  dispose X';
    done
    for all double-signatures β of size |V(H)| do
      set signature α_0 = {x + y : (x, y) ∈ β};
      add T_H[α_0] += X[β] · T_F[α_F] · T_G[α_G];
    done
  done.
```

Proof of Algorithm 2.5. We now explain the algorithm and show its correctness. It is more easily understood if one imagines particular forests (representatives) $U_F \subset F$ and $U_G \subset G$ in the place of the signatures $\boldsymbol{\alpha}_F$ and $\boldsymbol{\alpha}_G$ chosen in the first `for` cycle. Then one may routinely verify that all subsequent computations depend only on the forest signatures $\boldsymbol{\alpha}_F$, $\boldsymbol{\alpha}_G$ (not on the particular forests), and hence it is correct to finally multiply the computed values in $\boldsymbol{X}$ by the numbers $\boldsymbol{T}_F[\boldsymbol{\alpha}_F] \cdot \boldsymbol{T}_G[\boldsymbol{\alpha}_G]$.

In the tables $\boldsymbol{X}, \boldsymbol{X}'$ we iteratively compute the numbers of all spanning forests in H that result by adding some edges between the forests U_F and U_G. For a particular iteration, imagine a new forest component of size c in G which is to be joined with another forest component in $F \boxtimes G$ which has x vertices in $V(F)$. There are $c \cdot x$ edges to consider between the components, and we have to select exactly one connecting edge to maintain acyclicity, so there are cx choices there. These numbers are naturally multiplied (*) when joining more components together; see the steps in Figure 2.2.

So the core of the algorithm in the second cycle "`for` each $c \in \boldsymbol{\alpha}_G \ldots$" reads: We consider an arbitrary order $C_1, C_2, \ldots, C_k$ on the connected components of U_G. For $i = 1, 2, \ldots, k$, we take the component C_i, and count all possible ways how to connect C_i by selected edges to a subset (†) of components of each of the previously constructed forests on $V(F \cup C_1 \cup \cdots \cup C_{i-1})$ which are recorded in the table $\boldsymbol{X}$. The other ends of those selected edges are considered only among vertices in $V(F)$. (Recall that the complete union $H = F \boxtimes G$ has added *all* edges between $V(F)$ and $V(C_i)$.) We then record (*) numbers of all the new forests on $V(F \cup C_1 \cup \cdots \cup C_i)$ in a new table $\boldsymbol{X}'$ that will play the role of $\boldsymbol{X}$ in the next iteration.

More precisely, after finishing iteration $i = 1, 2, \ldots, k$ described in the previous paragraph, each entry $\boldsymbol{X}'[\boldsymbol{\beta}]$ equals the number of all forests U' of signature $\boldsymbol{\beta}$ spanning $V(F \cup C_1 \cup \cdots \cup C_i)$ such that $U' \restriction V(F) = U_F$ and $U' \restriction V(G) = U_G \restriction C_1 \cup \cdots \cup C_i$. This follows easily by an induction from the previous arguments. At the end we count each spanning forest $U \subseteq H$ such that $U \restriction V(F) = U_F$ and $U \restriction V(G) = U_G$ exactly once. Finally, the double-signatures in the table $\boldsymbol{X}$ partition the vertices into $V(F)$

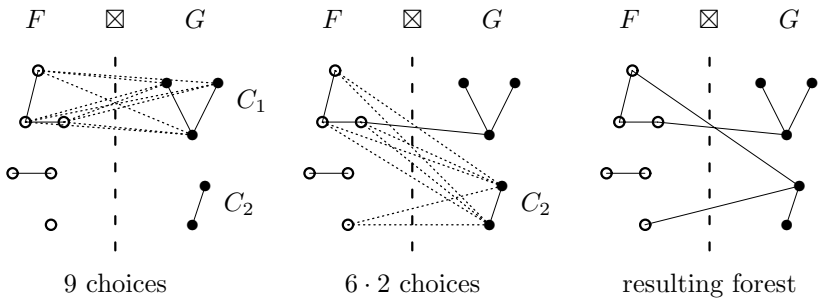


FIG. 2.2. *An illustration of inner iterations of Algorithm* 2.5*: Particular spanning forests U_F, U_G are chosen in F and G (other edges are not shown here), and the components C_1, C_2 of U_G are then joined to some of the four components of U_F. Possible choices of edges for these joins are shown in dotted lines.*

and $V(G)$, but that is no longer needed so we "simplify" them—we record the resulting numbers only by the (single) forest signatures in the resulting table $\boldsymbol{T}_H$. □

2.3. Time analysis. To get a fine time-complexity analysis of Algorithm 2.5, we have to insert a slight modification. (A problem may occur in the original Algorithm 2.5 in the fourth nested cycle "`for` all submultisets $\boldsymbol{\gamma} \subseteq \boldsymbol{\beta}$" if $\boldsymbol{\beta}$ consists, say, of $n/2$ copies of the element 2. Then there are up to $\exp\big(\Theta(n)\big)$ submultisets γ to consider.)

ALGORITHM 2.6. *Same as Algorithm* 2.5, *except the program line (†) now reads*

$$\texttt{for } \textit{all different submultisets } \boldsymbol{\gamma} \subseteq \boldsymbol{\beta} \texttt{ do},$$

and the line (*) reads

$$\texttt{add } \boldsymbol{X}'[\boldsymbol{\beta}'] \mathrel{+}= \boldsymbol{X}[\boldsymbol{\beta}] \cdot \prod_{(x,y)\in\boldsymbol{\gamma}} cx \cdot \prod_{(x,y)\in\langle\boldsymbol{\beta}\rangle} \binom{\mu_{\boldsymbol{\beta}}(x,y)}{\mu_{\boldsymbol{\gamma}}(x,y)},$$

where $\langle\boldsymbol{\alpha}\rangle$ denotes the ordinary set formed by elements of a multiset $\boldsymbol{\alpha}$, and $\mu_{\boldsymbol{\alpha}} z$ is the repetition of an element z in $\boldsymbol{\alpha}$.

Proof of Algorithm 2.6. We prove that this algorithm computes the same results as Algorithm 2.5. Notice that the outcome of the computation between the lines (†) and (*) depends only on the characteristic vector of $\boldsymbol{\gamma}$. Hence instead of all $\boldsymbol{\gamma} \subseteq \boldsymbol{\beta}$, it is enough to consider (much less of) pairwise different submultisets $\boldsymbol{\gamma} \subseteq \boldsymbol{\beta}$, and then multiply the resulting number by all possible choices (combinations) of repeated elements of $\boldsymbol{\gamma}$ from $\boldsymbol{\beta}$, as we do here in Algorithm 2.6. □

Now, since we use $O(\ )$ in the exponent, it is enough to argue that each of the `for` cycles in Algorithm 2.6 (2.5) is iterated at most $\exp\big(O(n^{2/3})\big)$ times. This follows easily from Lemmas 2.1, 2.2, and 2.3.

LEMMA 2.7. *Algorithm* 2.6 *runs in time* $\exp\big(O(n^{2/3})\big)$, *where* $n = |V(H)|$.

We remark that the improvement presented in Algorithm 2.6 has been fully incorporated in the subsequent algorithms, without further notices.

THEOREM 2.8. *The number of spanning forests in an n-vertex cograph can be computed in time* $\exp\big(O(n^{2/3})\big)$.

Proof. Consider a cograph G and a tree expression defining it. The forest signature table of a single vertex is trivial, and by Algorithms 2.4 and 2.6, the forest signature tables of a union or a complete union of two cographs can be computed in time claimed. Finally, knowing the forest signature table $\boldsymbol{T}$ of G, the number of all spanning forests of G is computed by adding up the entries of $\boldsymbol{T}$. □

Here we should note that the expression defining a cograph can be found in linear time [5], and hence we do not require it on the input.

3. The Tutte polynomial of a cograph. The Tutte polynomial can be defined in a number of equivalent ways. For our purposes, given a graph $G = (V, E)$ we define the Tutte polynomial as

$$T(G; x, y) = \sum_{F\subseteq E} (x-1)^{r(E)-r(F)} (y-1)^{|F|-r(F)},$$

where $r(F) = |V| - k(F)$ and $k(F)$ is the number of connected components of the spanning subgraph induced by the edge-subset F. It is clear that knowing $T(G; x, y)$ is the same as knowing, for every i and j, how many spanning subgraphs with the edge set F in G are there with $|F| = i$ and $k(F) = j$.

Consider a spanning subgraph $W \subset G$ determined on $V(W) = V(G)$ by an arbitrary subset $F \subset E(G)$, $F = E(W)$. The sizes of the connected components of W define a signature of size $|V(G)|$. In the *(spanning) subgraph signature table* $\boldsymbol{S}$ of G, for each signature $\boldsymbol{\alpha}$ of size $|V(G)|$ and each number of edges $f \in \{0, 1, 2, \ldots, |E(G)|\}$, we record the number $\boldsymbol{S}[\boldsymbol{\alpha}, f]$ of all spanning subgraphs of G having f edges and having component sizes according to the signature $\boldsymbol{\alpha}$. We abbreviate by $\boldsymbol{\gamma} \restriction_i$ the multiset formed by all the ith coordinates (repetitions accounted for) of the elements of a double-signature $\boldsymbol{\gamma}$.

In order to prove Theorem 1.1 we need analogues of Algorithms 2.4 and 2.5 for computing subgraph signature tables. The algorithm for disjoint unions is again straightforward and we omit it; the one for complete unions comes next.

Besides adding an edge number as the second index to the signature tables, the only other major difference of this algorithm from Algorithm 2.5 is that the single line (*) is replaced with another `for` cycle calling a procedure `CellSel` of further Algorithm 3.2.

ALGORITHM 3.1. *A modification of Algorithm* 2.6 *(*2.5*) for computing the (spanning) subgraph signature table of the complete union* $H = F \boxtimes G$.

`Input:` *Graphs* F, G, *and their subgraph signature tables* $\boldsymbol{S}_F, \boldsymbol{S}_G$.

`Output:` *The subgraph signature table* $\boldsymbol{S}_H$ *of* $H = F \boxtimes G$.

`create` *empty table* $\boldsymbol{S}_H$ *of subgraph signatures of size* $|V(H)|$`;`
`for` *all* $\boldsymbol{\alpha}_F \in \Sigma_F$, *and* $e_F = 0, 1, \ldots, |E(F)|$ *s.t.* $\boldsymbol{S}_F[\boldsymbol{\alpha}_F, e_F] > 0$ `do`
  `for` *all* $\boldsymbol{\alpha}_G \in \Sigma_G$, *and* $e_G = 0, \ldots, |E(G)|$ *s.t.* $\boldsymbol{S}_G[\boldsymbol{\alpha}_G, e_G] > 0$ `do`
    `set` $z = |V(F)|$`;`
    `create` *empty table* $\boldsymbol{Y}$ *of subgraph double-signature of size* z`;`
    `set` $\boldsymbol{Y}\big[$*double-signature* $\{(a, 0) : a \in \boldsymbol{\alpha}_F\},\ e_F\big] = 1$`;`
    `for` *each* $c \in \boldsymbol{\alpha}_G$ *(with repetition)* `do`
      `create` *empty table* $\boldsymbol{Y}'$ *of subgraph double-signature of size* $z + c$`;`
      `for` *all* $\boldsymbol{\beta}$ *of size* z, *and* e *s.t.* $\boldsymbol{Y}[\boldsymbol{\beta}, e] > 0$ `do`
        `for` *all different submultisets* $\boldsymbol{\gamma} \subseteq \boldsymbol{\beta}$ `do`
          `set` $r = \displaystyle\prod_{(x,y)\in\langle\boldsymbol{\beta}\rangle} \binom{\mu_{\boldsymbol{\beta}}(x,y)}{\mu_{\boldsymbol{\gamma}}(x,y)}$`;`
          `set` $d_1 = \|\boldsymbol{\gamma} \restriction_1\| = \sum_{(x,y)\in\boldsymbol{\gamma}} x$, $d_2 = \|\boldsymbol{\gamma} \restriction_2\| = \sum_{(x,y)\in\boldsymbol{\gamma}} y$`;`
          `set` *double-signature* $\boldsymbol{\beta}' = (\boldsymbol{\beta} \setminus \boldsymbol{\gamma}) \uplus \{(d_1, d_2 + c)\}$`;`
          `for` $f = |\boldsymbol{\gamma}|, |\boldsymbol{\gamma}| + 1, \ldots, c \cdot d_1$ `do`
            `set` *multiset* $D = c \cdot (\boldsymbol{\gamma} \restriction_1) = \{cx : (x, y) \in \boldsymbol{\gamma}\}$`;`
            `call` *Algorithm* 3.2*:* $p =$ `CellSel`(D, f)`;`
            `add` $\boldsymbol{Y}'[\boldsymbol{\beta}', e + f] \mathrel{+}= \boldsymbol{Y}[\boldsymbol{\beta}, e] \cdot r \cdot p$`;`
          `done`
        `done`
      `done`
      `set` $\boldsymbol{Y} = \boldsymbol{Y}'$, $z = z + c$`;` `dispose` $\boldsymbol{Y}'$`;`
    `done`
    `for` *all double-signature* $\boldsymbol{\beta}$ *of size* $|V(H)|$, *and* f, *s.t.* $\boldsymbol{Y}[\boldsymbol{\beta}, f] > 0$ `do`
      `set` *signature* $\boldsymbol{\alpha}_0 = \{x + y : (x, y) \in \boldsymbol{\beta}\}$`;`
      `add` $\boldsymbol{S}_H[\boldsymbol{\alpha}_0, f + e_G] \mathrel{+}= \boldsymbol{Y}[\boldsymbol{\beta}, f] \cdot \boldsymbol{S}_F[\boldsymbol{\alpha}_F, e_F] \cdot \boldsymbol{S}_G[\boldsymbol{\alpha}_G, e_G]$`;`
    `done`
  `done`
`done.`

Proof of Algorithm 3.1. This algorithm is similar to the improved version of Algorithm 2.6, and so we only sketch the proof here. The main new difficulty lies in counting the different ways in which a connected component of c vertices in $\boldsymbol{\alpha}_G$ can be connected with f edges to the selected components of signatures $(x, y) \in \boldsymbol{\gamma}$. Recall that when counting forests we had no such difficulty, since we joined the component of $\boldsymbol{\alpha}_G$ to each component of $\boldsymbol{\gamma}$ with exactly one edge; thus we used exactly $f = |\boldsymbol{\gamma}|$ edges chosen in $\prod_{(x,y)\in\boldsymbol{\gamma}} cx$ different ways. The procedure "`CellSel`(D, f)" counts this for spanning subgraphs, and we defer the explanation to Algorithm 3.2; see also Figure 3.1.

Finally, notice that the edge numbers in tables $\boldsymbol{Y}$, $\boldsymbol{Y}'$ do not account for the edges from $E(G)$, since we do not know how many edges each one has of the components of $\boldsymbol{\alpha}_G$. Those edges are summed up at the end, when obtaining the signatures for H from the double-signatures stored in $\boldsymbol{Y}$. □

ALGORITHM 3.2. *Computing the number of cellular selections: We are selecting ℓ elements from the union $C_1 \cup C_2 \cup \cdots \cup C_k$, where C_i for $i = 1, 2, \ldots, k$ are pairwise disjoint cells of sizes $d_i = |C_i|$, and we require that some element is selected from every cell.*

`Input:` *A multiset $D = \{d_1, d_2, \ldots, d_k\}$ of cell sizes, and a number ℓ.*
`Output:` *The number* `CellSel`(D, ℓ) *of all such possible selections.*

```
create table u[1..k][1..ℓ], filled with 0;
for j = 1, 2, ..., d_1 do  set u[1][j] = (d_1 choose j);
set z = d_1;
for i = 2, 3, ..., k do
   add z += d_i;
   for j = i, i+1, ..., min(ℓ, z) do
      for s = 1, 2, ..., min(j − (i − 1), d_i) do
         add u[i][j] += u[i − 1][j − s] · (d_i choose s);
      done
   done
done
return u[k][ℓ].
```

Proof of Algorithm 3.2. Let $u_{i,j} = \boldsymbol{u}[i][j]$ be the number of cellular selections of j elements chosen among the first i cells. These numbers satisfy the recurrence relation

$$u_{i,j} = \sum_{s=1}^{r} u_{i-1,j-s} \cdot \binom{d_i}{s},$$

where r is the maximum number of elements than can be selected from the ith cell to obtain a total of j elements. Since the ith cell has d_i elements available, and the $i-1$ previous cells contributed at least one element each to the resulting j elements, it follows that $r = \min\{j - (i-1), d_i\}$.

Algorithm 3.2 applies the previous recurrence in a correct order, and avoids useless computations like with values of j too small or too large. It runs in $O(k\ell^2)$ steps. □

Proof of Theorem 1.1. As in Theorem 2.8, the subgraph signature table $\boldsymbol{S}$ of a cograph can be computed in time proportional to the number of all possible double-signatures of size n, i.e., in $\exp\big(O(n^{2/3})\big)$. Then, summing the entries of $\boldsymbol{S}$, we compute the numbers of spanning subgraphs with a given number of edges and a number of components. As we have remarked previously, these numbers give (efficiently) the Tutte polynomial. □

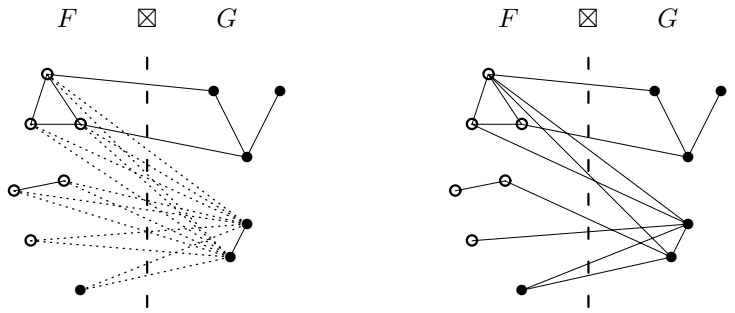


FIG. 3.1. *How cellular selections arise in Algorithm* 3.1 *when adding edges to a spanning subgraph; here we are selecting* $f = 7$ *edges out of cell sizes* $\{6, 4, 2, 2\}$ *(possible edge choices shown in dotted lines).*

The U polynomial of an n-vertex graph G is defined in [14] as

$$U(G; \mathbf{x}, y) = \sum_{F \subseteq E} x_{n_1} \cdots x_{n_k} (y-1)^{|F|-r(F)},$$

where $n_1, \ldots, n_k$ are the vertex sizes of the components of the spanning subgraph (V, F). If we let $x_1 = \cdots = x_n = x - 1$ in the expression above, we recover the Tutte polynomial $T(G; x, y)$ up to a power of $x - 1$. It is clear that the subgraph signature table of a graph is precisely equivalent to the U polynomial, hence in the statement of Theorem 1.1 we can replace "U polynomial" for "Tutte polynomial."

4. Graph clique-width. In this section we give a more precise definition of clique-width and some of its properties.

Graphs from now on are labeled on the vertices; $V_i(G)$ denotes the set of vertices in G that have label i. A graph has clique-width $\leq k$ if it can be constructed using k labels and the following four operations:

1. $v(i)$: creates a new vertex with label i.
2. $\dot\cup$: produces the union of several disjoint graphs, without modifying the labels.
3. $\eta_{i,j}$, $i \neq j$: joins all the vertices labeled i to all the vertices labeled j. This operation does not create multiple edges.
4. $\rho_{i,j}$: all vertices labeled i are relabeled to have label j. It allows one to merge two label classes into one, thus freeing a label for later use.

An expression defining a graph G built from the above four operations using k labels is a *k-expression* for G.

Now we explain why cographs have clique-width at most two. Operations 1 and 2 are analogous to rules 1 and 2 for cographs introduced in section 2.1. In order to perform the complete union $G \boxtimes H$ of two cographs (rule 3), one relabels all vertices of G to have label 1 and all vertices of H to have label 2 (using operations $\rho_{i,j}$), and then applies the operation $\eta_{1,2}$. The interested reader may check why the path P_4 has clique-width greater than two. Furthermore, we show in Figure 4.1 examples of the graphs C_5 and C_7 having clique-width 3 and 4, respectively. (Actually, all cycles have clique-width at most 4.)

A k-expression for G is *irredundant* if, whenever operation $\eta_{i,j}$ is applied, no vertex with label i has been previously joined to any vertex of label j. It can be shown [4] that for every k-expression, one can construct an irredundant k-expression defining the same graph. Hence in the next section we assume that graphs with bounded clique-width are defined by means of irredundant expressions.

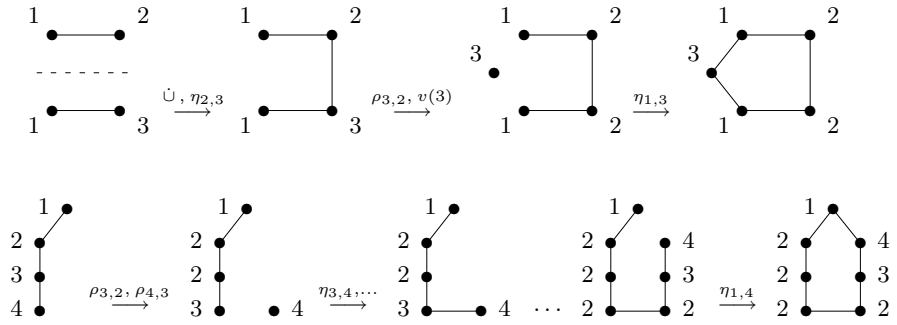


FIG. 4.1. *An example—optimal expressions defining the cycles C_5 and C_7 (starting from trivially constructed subgraphs on the left).*

Another important question concerns computing the clique-width of a graph, and more importantly, finding a defining k-expression. Until recently [16], algorithms running on graphs of bounded clique-width *needed* a corresponding k-expression on the input. The first (and currently only known) efficient way of approximating [16, 15] the expression for a graph of bounded clique-width uses a new notion of rank-width [16]. It is remarkable how close is the computation of rank-width on graphs [15] to the computation of branch-width on binary matroids [7]. It is that rank-width of a bipartite graph which equals branch-width of the matroid formed by the associated binary adjacency matrix minus one, and a simple translation can be used for general graphs. However, this interesting topic is far beyond the scope of our paper, and so we refer interested readers to the cited papers.

5. The Tutte polynomial for bounded clique-width. In this section we prove Theorem 1.2. Analogously to section 2, the most involved part is Algorithm 5.4, which corresponds to adding edges in the operation $\eta_{i,j}$.

Instead of double-signatures, we need k-signatures to mark the different labels of the vertices belonging to the components of a subgraph. A k-signature is a multiset of k-tuples of nonnegative integers, excluding the k-tuple $(0,\ldots,0)$. The *size* $\|\boldsymbol{\beta}\|$ of a k-signature $\boldsymbol{\beta}$ is the sum of all $(x_1+\cdots+x_k)$ for $(x_1,\ldots,x_k)\in\boldsymbol{\beta}$, respecting repetition in the multiset. As in the case of double-signatures we have the following lemma.

LEMMA 5.1. *There are* $\exp\big(\Theta(n^{k/(k+1)})\big)$ *distinct k-signatures of size n, for each fixed k.*

Since the subsequent proof is quite involved, it is worth mentioning that an easy encoding argument gives an upper bound of $\exp\big(O(n^{k/(k+1)}\log n)\big)$, which is almost as good: We limit the number of nonzero entries of the characteristic vector of a k-signature. In the worst case, at most all those $\Theta(t^k)$ entries corresponding to $(c_1,\ldots,c_k)$ are nonzero where $0\le c_i\le t$, and t is such that $\Theta(t^{k+1})=n$ (the size of the signature). Hence the characteristic vector of a k-signature has at most $\Theta\big(n^{k/(k+1)}\big)$ nonzero entries, and we may encode all k-signatures of size n by choosing those nonzero entries in all possible ways, and then trying all values between 1 and n for each;

$$\binom{n^k}{\Theta(n^{k/(k+1)})}\cdot n^{\Theta(n^{k/(k+1)})}=n^{\Theta(k\cdot n^{k/(k+1)})}=\exp\big(O(n^{k/(k+1)}\log n)\big).$$

Proof of Lemma 5.1. The proof is based on generating functions and complex analysis. Let p_n be the number of distinct k-signatures of size n. The associated generating function is equal to

$$P(z) = \sum_{n\geq 0} p_n z^n = \prod_{n\geq 1} \frac{1}{(1-z^n)^{\binom{n+k-1}{k-1}}}.$$

The reason is that the number of nonnegative (ordered) solutions of $x_1 + \cdots + x_k = n$ is equal to $\binom{n+k-1}{n} = \binom{n+k-1}{k-1}$. The infinite product encodes the fact that we are taking multisets of those k-tuples.

According to a result of Meinardus (see [1, Theorem 6.2]), the asymptotic behavior of p_n is determined by the associated Dirichlet series

$$D(s) = \sum_{n\geq 1} \frac{\binom{n+k-1}{k-1}}{n^s}.$$

Provided some analytical conditions on $D(s)$ hold, that in our case are easy to check, we have

$$p_n \sim C n^{\gamma} \exp\left(K n^{\rho/(\rho+1)}\right),$$

where C, γ, K are constants and ρ is the unique (simple) pole of $D(s)$ in a suitable region $\mathrm{Re}(s) > -C_0$, where $0 < C_0 < 1$.

Now it is clear that $D(s)$ can be expressed as a linear combination of $\zeta(s-k+1), \zeta(s-k+2), \ldots, \zeta(s)$, where $\zeta(s) = \sum_{n\geq 1} n^{-s}$ is the Riemann zeta function. Since $\zeta(s)$ has a unique simple pole at $s = 1$ for $\mathrm{Re}(s) > 0$, the pole of $D(s)$ we are looking for is at $\rho = k$, and this proves the result. □

The next two algorithms, which correspond to the operations $\dot\cup$ and $\rho_{i,j}$, need no special analysis.

ALGORITHM 5.2. *Combining the spanning subgraph k-signature tables of k-labeled graphs F and G into the one of the disjoint union $H = F \dot\cup G$.*

`Input:` *Graphs F, G, and their subgraph k-signature tables $\boldsymbol{S}_F, \boldsymbol{S}_G$.*
`Output:` *The subgraph k-signature table $\boldsymbol{S}_H$ of $H = F \dot\cup G$.*
    `create` *empty table $\boldsymbol{S}_H$ of subgraph k-signatures of size $|V(H)|$*;
    `for` *all k-signatures $\boldsymbol{\alpha}_F \in \Sigma_F^k$, and $e_F = 0, 1, \ldots, |E(F)|$* `do`
        `for` *all k-signatures $\boldsymbol{\alpha}_G \in \Sigma_G^k$, and $e_G = 0, 1, \ldots, |E(G)|$* `do`
            `set` $\boldsymbol{\alpha} = \boldsymbol{\alpha}_F \uplus \boldsymbol{\alpha}_G$ *(a multiset union)*;
            `add` $\boldsymbol{S}_H[\boldsymbol{\alpha}, e_F + e_G] \mathrel{+}= \boldsymbol{S}_F[\boldsymbol{\alpha}_F, e_F] \cdot \boldsymbol{S}_G[\boldsymbol{\alpha}_G, e_G]$;
    `done.`

ALGORITHM 5.3. *Modifying the spanning subgraph k-signature table of a k-labeled graph G into the one of $\rho_{i,j}$, the relabeling $1 \to 2$ of G.*

`Input:` *A k-labeled graph G and its subgraph k-signature table $\boldsymbol{S}_G$.*
`Output:` *The subgraph k-signature table $\boldsymbol{S}_H$ of $H = \rho_{1,2}(G)$.*
    `create` *empty table $\boldsymbol{S}_H$ of subgraph k-signatures of size $|V(G)|$*;
    `for` *all k-signatures $\boldsymbol{\alpha}_G \in \Sigma_G^k$, and $e_G = 0, 1, \ldots, |E(G)|$* `do`
        `set` *k-signature* $\boldsymbol{\alpha}'_G = \{(0, a_1 + a_2, \ldots, a_k) : (a_1, a_2, \ldots, a_k) \in \boldsymbol{\alpha}_G\}$;
        `add` $\boldsymbol{S}_H[\boldsymbol{\alpha}'_G, e_G] \mathrel{+}= \boldsymbol{S}_G[\boldsymbol{\alpha}_G, e_G]$;
    `done.`

The next algorithm computes the k-signature table of the graph $H = \eta_{1,2}(G)$ from that of G, assuming that there were no edges in G between vertices with labels

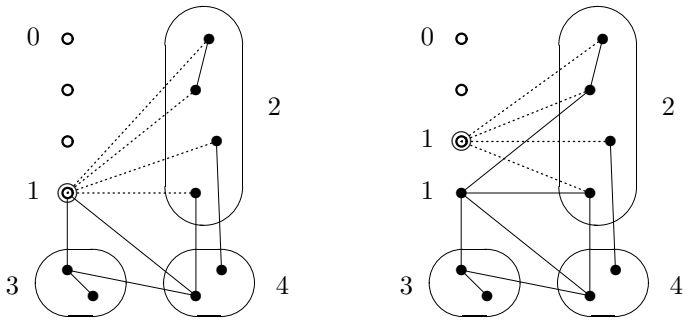


FIG. 5.1. *Two steps illustrating the method of counting all spanning subgraphs of $H = \eta_{1,2}(G)$; the hollow vertices are not processed yet (label 0), and the dotted edges show possible choices of new edges from the marked vertex (v).*

1 and 2 (like in an irredundant expression). To understand the main idea, let us consider the following method for obtaining all the spanning subgraphs in H such that their restriction to G gives a certain subgraph G_α. Start by relabeling every vertex with label 1 in G to 0, meaning that these vertices have not been processed yet. Then choose a vertex v with label 0, relabel to 1, and consider all the different subgraphs of H we can generate from the original one by adding new edges from v to some (selected) vertices with label 2. Obviously, the restriction of any of these subgraphs to G still gives G_α. Iterate the process for every generated subgraph and for every vertex with label 0 until none remains; see an illustration of the method in Figure 5.1.

Algorithm 5.4 essentially follows this process, but it improves the running time by working with the signatures instead of subgraphs, and with connected components instead of single vertices, as has been done in Algorithms 2.5 and 3.1.

The algorithm counts all spanning subgraphs of H such that their restriction to G is of a signature $\boldsymbol{\alpha}$, for every possible $\boldsymbol{\alpha}$. A $(k+1)$-signature table $\boldsymbol{Y}$ is used to store all intermediate results of the computation (together). Instead of processing a vertex v one at a time, we choose a component B with some labels 0 from a signature $\boldsymbol{\beta}$ in Y, and then we choose a submultiset $\boldsymbol{\gamma} \subseteq \boldsymbol{\beta}$ signing the components that B will be joined to, using exactly f of the new edges of H. Procedure `CellSel` (Algorithm 3.2) computes efficiently the number of ways this can be done. We add the resulting signatures and numbers to the table $\boldsymbol{Y}$ and we repeat until no signature with vertices labeled 0 remains in $\boldsymbol{Y}$. We then update the table $\boldsymbol{S}_H$ with the signatures computed in the table $\boldsymbol{Y}$ and we start again from a new signature $\boldsymbol{\alpha}$ of G.

ALGORITHM 5.4. *Updating the subgraph k-signature table of a k-labeled graph G, such that there is no edge between the labels 1 and 2 in G, into the one of the graph H obtained from G by adding all edges between the labels 1 and 2.*

`Input:` *A k-labeled graph G, and its subgraph k-signature table* $\boldsymbol{S}_G$.

`Output:` *The subgraph k-signature table* $\boldsymbol{S}_H$ *of* $H = \eta_{1,2}(G)$.

`create` *empty table* $\boldsymbol{S}_H$ *of subgraph k-signatures of size* $|V(H)|$`;`

`for` *all* $\boldsymbol{\alpha} \in \Sigma_G^k$, *and* $e = 0, 1, \ldots, |E(G)|$ *s.t.* $\boldsymbol{S}_G[\boldsymbol{\alpha}, e] > 0$ `do`

`//` *Imagine a particular spanning subgraph in G of a k-signature* $\boldsymbol{\alpha}$ *with e edges.*

`create` *empty table* $\boldsymbol{Y}$ *of subgraph* $(k+1)$*-signatures of size* $|V(H)|$`;`

`set` $(k+1)$*-signature* $\boldsymbol{\alpha}_0 = \{(a_1, 0, a_2, \ldots, a_k) : (a_1, a_2, \ldots, a_k) \in \boldsymbol{\alpha}\}$`;`

`set` $\boldsymbol{Y}[\boldsymbol{\alpha}_0, e] = 1$`;`

```
while  there exists β ∈ Σ_H^{k+1}, b = (b_0, b_1, ..., b_k) ∈ β, and e',
              such that Y[β, e'] > 0 and b_0 > 0  do
  select such β, e', and b, with maximal ‖β ⌈_0‖;
    // Imagine a particular spanning subgraph in H with e' edges and
    // a (k+1)-signature β, and its component B corresponding to b:
    // b_0 of the vertical of B are going to be "joined to" labels 2 in H.
  set y = Y[β, e'], Y[β, e'] = 0, β' = β \ {b};
  for all different submultisets γ ⊆ β' such
          that c_2 > 0 for each (c_0, c_1, c_2, ..., c_k) ∈ γ  do
    // We connect B to all components in γ.
    set r = ∏_{d∈⟨β'⟩} C(#_{β'} d, #_γ d);
    for i = 0, 1, ..., k do d_i = ‖γ ⌈_i‖ = Σ_{(c_0,...,c_k)∈γ} c_i;
    set d = (d_0, d_1 + b_0 + b_1, d_2 + b_2, ..., d_k + b_k);
    set (k+1)-signature δ = (β' \ γ) ⊎ {d};
    set multiset D = b_0 · (γ ⌈_2) = {b_0 c_2 : (c_0, c_1, c_2, ..., c_k) ∈ γ};
    for f = |γ|, |γ|+1, ..., b_0 · (d_2 + b_2) do
      // We count all the cellular selections from D ⊎ {b_2} here,
      // but we allow to select nothing from {b_2} as well.
      call Algorithm 3.2:
(‡)       p = CellSel(D ⊎ {b_2}, f) + CellSel(D, f);
      add Y[δ, e' + f] += y · r · p;
    done
  done
done
for all β ∈ Σ_H^{k+1}, and f, such that Y[β, f] > 0 do
  set signature α_0 = {(b_1, ..., b_k) : (b_0, b_1, ..., b_k) ∈ β};
  add S_H[α_0, f] += Y[β, f] · S_G[α, e];
done
done.
```

Proof of Algorithm 5.4. The idea is analogous to the proofs of Algorithms 2.5 and 3.1. The only noticeable differences are the following two:

- Since we are now adding edges inside the same graph (instead of composing two previously disjoint graphs), we need an artificial new label 0 for marking those vertices that still have to be processed, among those having original label 1. One little advantage of the extra label is that now we can store all intermediate results of our computation in the same $(k+1)$-signature table $\boldsymbol{Y}$, and to control the computation we just need one large `while` cycle.
- Unlike for cographs, we now have to consider the possibility that our selected component B has vertices of both labels 0 and 2, and hence we may want to add some edges induced on $V(B)$ as well. This is taken care of on the line (‡) of the algorithm.

Again, it is easier to imagine a particular spanning subgraph G_α with e edges in the place of the signature $\boldsymbol{\alpha}$, since subsequent computations do not depend on a particular choice of G_α. Under this assumption table $\boldsymbol{Y}$ counts $(k+1)$-labeled subgraphs of H whose restriction to G is G_α. It is convenient to see table $\boldsymbol{Y}$ as a `set` of subgraphs, stored according to signature for the sake of efficiency.

This set $\boldsymbol{Y}$ contains at the beginning the subgraph G_α, where vertices of label 1 receive label 0 to mark that they are unprocessed. At every iteration of the `while` cycle

we choose some subgraphs of $\boldsymbol{Y}$ and process some of its vertices of label 0, marking them with label 1 and considering all possible ways of joining them to vertices of label 2. The resulting subgraphs are stored in the table $\boldsymbol{Y}$ in place of the former ones. The `while` loop ends when no subgraph in $\boldsymbol{Y}$ has vertices of label 0.

To be more precise, at every iteration of the `while` loop we choose subgraphs of $(k+1)$-signature $\boldsymbol{\beta}$ and e' edges. (Imagine again a particular subgraph G_β.) We select a component B with $b_0 > 0$ vertices of label 0. We process component B by joining its vertices of label 0 to some of the $\|\boldsymbol{\beta} \restriction_2\|$ vertices of label 2 in G_β. Hence the components containing the vertices of label 2 may be joined to B. The `for` loop iterates over all suitable subsets $\boldsymbol{\gamma}$ of such components to account for all possible resulting signatures. Components not in $\boldsymbol{\gamma}$ receive no edge from B, while components in $\boldsymbol{\gamma}$ receive at least one edge, so we call procedure `CellSel` to count efficiently the number of ways $\boldsymbol{\gamma}$ may be joined to B. In fact, `CellSel` is called twice, since the vertices of label 0 and 2 in B itself may be joined by either none or some edges.

Observe that a pair of descendants of, say, G_β differ in at least one edge joining vertices of labels 1 and 2, and further descendants of them will still differ at the same edge, since new edges are only added between vertices of labels 0 and 2. This implies that $\boldsymbol{Y}$ is free of duplicates, because all subgraphs have a common ancestor G_α. So, at the end, in the Table $\boldsymbol{Y}$, we count exactly once each spanning subgraph $W \subseteq H$ such that $W \restriction G = G_\alpha$. After multiplying by $\boldsymbol{S}_G[\boldsymbol{\alpha}, e]$, we record in $\boldsymbol{S}_H$ the number of all spanning subgraphs of H having their restriction to G of signature $\boldsymbol{\alpha}$, for each $\boldsymbol{\alpha}$. □

Proof of Theorem 1.2. The idea is the same as in the proof of Theorem 1.1 at the end of section 3. In the `while` cycle we always choose signature $\boldsymbol{\beta}$ among those with the maximal number of unprocessed vertices ($\|\boldsymbol{\beta} \restriction_0\|$), and then we strictly decrease the number of those. Hence we never process the same pair $(\boldsymbol{\beta}, e)$ twice. So time complexity is again dominated by the lengths of the tables involved. Since we need a table of $(k+1)$-signatures, the complexity $\exp\left(O(n^{(k+1)/(k+1+1)})\right) = \exp\left(O(n^{1-\frac{1}{k+2}})\right)$ follows from Lemma 5.1. □

Finally we remark that, exactly as in the case of cographs, the algorithm computes the full U polynomial on graphs of bounded clique-width.

6. Concluding remarks. We have shown that the Tutte and U polynomials can be computed in subexponential time for cographs, and more generally for graphs with bounded clique-width. Such a result is very unlikely to hold for all graphs. Of course, the important question of whether the Tutte polynomial can be computed in polynomial time, or the problem is #P-hard even for graphs of bounded clique-width, remains open. (The U polynomial is obviously not computable in polynomial time due to its size.)

On the other hand, the *chromatic* polynomial for graphs of bounded clique-width can be computed in polynomial time (although not FPT). This follows by adapting the algorithm in [10] for computing the chromatic number, keeping track also of the number of r-colorings for $r = 1, \ldots, n$, where n is the number of vertices; see in [12]. To our knowledge, that is possibly the only currently known natural example of graph classes other than chordal graphs, where the chromatic polynomial can be computed in polynomial time, but the complexity of computing the Tutte polynomial is undecided.

REFERENCES

[1] G. E. Andrews, *The Theory of Partitions*, Cambridge University Press, Cambridge, UK, 1984.

[2] A. Andrzejak, *An algorithm for the Tutte polynomials of graphs of bounded treewidth*, Discrete Math., 190 (1998), pp. 39–54.
[3] B. Courcelle, J. A. Makowsky, and U. Rotics, *Linear time solvable optimization problems on graphs of bounded clique-width*, Theory Comput. Systems, 33 (2000), pp. 125–150.
[4] B. Courcelle and S. Olariu, *Upper bounds to the clique width of graphs*, Discrete Appl. Math., 101 (2000), pp. 77–114.
[5] D. G. Corneil, Y. Perl, and L. K. Stewart, *A linear recognition algorithm for cographs*, SIAM J. Comput., 14 (1985), pp. 926–934.
[6] O. Giménez and M. Noy, *On the complexity of computing the Tutte polynomial of bicircular matroids*, Combin. Probab. Comput., 15 (2006), pp. 385–395.
[7] P. Hliněný, *A parametrized algorithm for matroid branch-width*, SIAM J. Comput., 35 (2005), pp. 259–277 (electronic).
[8] P. Hliněný, *The Tutte polynomial for matroids of bounded branch-width*, Combin. Probab. Comput., 15 (2006), pp. 397–409.
[9] F. Jaeger, D. L. Vertigan, D. J. A. Welsh, *On the computational complexity of the Jones and Tutte polynomials*, Math. Proc. Cambridge Philos. Soc., 108 (1990), pp. 35–53.
[10] D. Kobler and U. Rotics, *Edge dominating set and colorings on graphs with fixed clique-width*, Discrete Appl. Math., 126 (2003), pp. 197–221.
[11] J. H. van Lint and R. M. Wilson, *A Course in Combinatorics*, Cambridge University Press, Cambridge, UK, 1992.
[12] J. A. Makowsky and U. Rotics, *Computing the Chromatic Polynomial on Graphs of Bounded Clique-Width*, preprint, 2005.
[13] S. D. Noble, *Evaluating the Tutte polynomial for graphs of bounded tree-width*, Combin. Probab. Comput., 7 (1998), pp. 307–321.
[14] S. D. Noble and D. J. A. Welsh, *A weighted graph polynomial from chromatic invariants of knots*, Ann. Inst. Fourier (Grenoble), 49 (1999), pp. 1057–1087.
[15] S.-I. Oum, *Approximating rank-width and clique-width quickly*, in Lecture Notes in Comput. Sci. 3787, Springer, Berlin, 2005, pp. 49–58.
[16] S.-I. Oum and P. D. Seymour, *Approximating clique-width and branch-width*, J. Combin. Theory Ser. B, 96 (2006), pp. 514–528.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 947–959

MULTIVARIABLE CODES OVER FINITE CHAIN RINGS: SERIAL CODES*

E. MARTÍNEZ-MORO† AND I. F. RÚA‡

Abstract. The structure of multivariate serial codes over a finite chain ring R is established using the structure of the residue field $\bar{R}$. Multivariate codes extend in a natural way the univariate cyclic and negacyclic codes and include some nontrivial codes over R. The structure of the dual codes in the serial abelian case is also derived, and some conditions for the existence of self-dual codes over R are studied.

Key words. finite chain ring, multivariate codes, serial codes

AMS subject classifications. 11T71, 13M10, 94B99

DOI. 10.1137/050632208

1. Introduction.

The relevance of finite rings in algebraic coding theory, originally restricted to codes over binary (or finite field) alphabets, has been progressively noticed. For instance, many classical codes can be seen as ideals in certain algebras over a finite field [1, 7, 18]. On the other hand, the theory of error-correcting codes over finite rings has gained certain relevance since the realization that some nonlinear codes over finite fields can be constructed from linear codes over such rings [4, 9, 13, 14, 15]. This paper is a contribution to both lines of research, and its purpose is to describe a class of multivariate codes over a finite chain ring R.

Throughout the paper a multivariate serial code over R is an ideal in a particular type of R-algebras. Generally, this ideal is not a semisimple module over R, but its image code over the residue ring $\bar{R}$ is semisimple. We use the known machinery for semisimple codes to decompose our codes as a direct sum of uniserial (or chain) modules; hence the name serial [20].

Our study is based on Poli's decomposition of the roots of the defining ideal in cyclotomic classes [19, 18]. Thus our codes extend the definition of cyclic and negacyclic codes over finite chain rings to the multivariable case [8]. Gröbner bases over rings could be also used in the study of our codes (see, for example, [3]), but we follow Poli's approach since his technique reveals directly the underlying cyclotomic structure of the codes.

The paper is organized as follows. In section 2 we collect the basic results needed on finite chain rings. Section 3 is devoted to the definition of the codes and their ambient space as well as the description of their structure. In section 4 we study the duals of abelian semisimple codes. Finally in section 5 we characterize those nontrivial abelian semisimple codes that are self-dual.

*Received by the editors May 24, 2005; accepted for publication (in revised form) May 30, 2006; published electronically December 11, 2006.

http://www.siam.org/journals/sidma/20-4/63220.html

†Departamento de Matemática Aplicada, Universidad de Valladolid, 47002 Valladolid, Spain (edgar@maf.uva.es). This author's work was partially supported by MEC MTM2004-00876 and MTM2004-00958 I+D projects.

‡Departamento de Matemáticas, Estadística y Computación Universidad de Cantabria, 39005 Santander, Spain (i.f.rua@unican.es). This author's work was partially supported by MTM2004-08115-C04-01 and FICYT (IB05-186) I+D projects.

2. Preliminaries. In this section we fix our notation and recall some basic facts about finite chain rings (see [2, 12] for a complete account). In this paper all rings will be associative, commutative, and with identity. A ring R is called a *local ring* if it has a unique maximal ideal. A local ring is a *chain ring* if its lattice of ideals is a chain. In this case, since the ideals are linearly ordered by inclusion, the ring is also called *uniserial* [20]. It can be shown [8, Proposition 2.1] that R is a finite commutative chain ring if and only if R is a local ring and its maximal ideal is principal. In such a case, let $a \in R$ be a fixed generator of the maximal ideal, and let t be its nilpotency index (notice that a is a nilpotent element). Then the chain of ideals of R is

$$\langle 0\rangle = \left\langle a^t\right\rangle \subsetneq \left\langle a^{t-1}\right\rangle \subsetneq \cdots \subsetneq \left\langle a^1\right\rangle \subsetneq \left\langle a^0\right\rangle = R. \tag{2.1}$$

In what follows, R will always be a finite commutative chain ring and a will be the generator of its maximal ideal. Also, $\mathbb{F}_q = \bar{R} = R/\langle a\rangle$ will be the residue field of R, where $q = p^l$, for a prime number p. We will denote by $R[X]$ the polynomial ring in the indeterminate X with coefficients in R. We can extend the natural ring homomorphism $r \mapsto \bar{r} = r + \langle a\rangle$ as follows:

$$\begin{array}{ccc} R & \hookrightarrow & R[X] \\ {\scriptstyle -}\downarrow & & {\scriptstyle -}\downarrow \\ \mathbb{F}_q & \hookrightarrow & \mathbb{F}_q[X]. \end{array} \tag{2.2}$$

Two polynomials $f_1, f_2 \in R[X]$ are *coprime* if $\langle f_1, f_2\rangle = R[X]$, where $\langle f_1, f_2\rangle$ is the ideal generated by the polynomials f_i, $i = 1, 2$. A polynomial $f \in R[X]$ is called *basic irreducible* if it is not a zero divisor and $\bar{f} \in \mathbb{F}_q[X]$ is irreducible. We will use the known Hensel lemma in the following form. (See [2, Theorem 3.2.6] for a proof.)

THEOREM 2.1 (Hensel's lemma). *Let $f \in R[X]$ be a monic polynomial such that $\bar{f} = g_1 g_2 \dots g_r$, where the polynomials $g_i \in \bar{R}[X]$ are monic and pairwise coprime. Then there exist pairwise coprime monic polynomials $f_i \in R[X]$ such that $f = f_1 f_2 \dots f_r$ and $\bar{f}_i = g_i$ for all $i = 1, \dots, r$. This decomposition is unique up to a permutation of the factors, which are called* lifting factors *of f.*

Let S be an *extension* of R, i.e., a ring containing R. If $T \subseteq S$, with $T \neq \emptyset$ of finite cardinality, then the extension of R *generated* by T, denoted $R(T)$, is the smallest subring of S containing $R \cup T$. If S is a finite local ring with residue field K, then K is a field extension of $\mathbb{F}_q$. If this field extension is separable, then S is called a *separable extension* of R.

If $f(X) \in R[X]$ is a basic irreducible polynomial, then $\langle a, f(X)\rangle$ is a maximal ideal of $R[X]$ [2, Remark after Lemma 3.2.10]. Therefore, the homomorphism (2.2) induces the following isomorphism:

$$R[X]/\langle a, f(X)\rangle \cong \mathbb{F}_q[X]/\left\langle \overline{f}(X)\right\rangle \cong \mathbb{F}_q(\alpha),$$

where $\overline{f}(\alpha) = 0$. Hence $S = R[X]/\langle f(X)\rangle$ is a local ring with maximal ideal $\langle a, f(X)\rangle + \langle f(X)\rangle = \langle a\rangle + \langle f(X)\rangle$; i.e., it is a finite local chain ring. Moreover, it is a separable extension of R, since the field extension $\mathbb{F}_q(\alpha)|\mathbb{F}_q$ is separable. Since the element $A = X + \langle f(X)\rangle \in S$ is a root of the polynomial $f(X) \in S[X]$ that lifts α (i.e., $\overline{A} = \alpha \in \overline{S}$), we can write $S = R(A)$.

In our paper we consider monic polynomials $t_i(X_i) \in R[X_i]$ $(i = 1, \dots, r)$. Usually, we will require $\bar{t}_i(X_i) \in \mathbb{F}_q[X_i]$ to be *square-free*, i.e., $\bar{t}_i(X_i) = \prod_{j=1}^{r_i} g_{ij}(X_i)$, where $g_{ij}(X_i) \in \mathbb{F}_q[X_i]$ are pairwise coprime. In this case, from Hensel's lemma,

$$t_i(X_i) = \prod_{j=1}^{r_i} f_{ij}(X_i), \tag{2.3}$$

where $f_{ij}(X_i)$ are pairwise coprime monic basic irreducible polynomials such that $\overline{f}_{ij} = g_{ij}$. This decomposition is unique up to a relabeling of the factors. If K is an extension of $\mathbb{F}_q$ and μ is a root of $\bar{t}_i(X_i)$, we shall denote by $\mathrm{Irr}(\mu, K)$ the minimal polynomial of μ over the field K. This polynomial divides $\bar{t}_i(X_i)$ and, more specifically, divides one of the factors $g_{ij}(X_i)$.

3. Multivariable serial codes. In this section we will obtain the structure of a multivariable serial code over the finite chain ring R; i.e., we will describe explicitly the ideals of the quotient ring $\mathcal{R} = R[X_1, \ldots, X_r]/\langle t_1(X_1), \ldots, t_r(X_r)\rangle$. This structure has been studied in [18, 19], in the case where the ring R is the finite field $\mathbb{F}_q$. In that case, the square-free condition on the polynomials $t_i(X_i)$ is known as the "semisimple condition," because of the semisimple structure of the ring $\mathcal{R}$. (A ring is called *semisimple* if it can be decomposed as a direct sum of simple ideals.) In the general case, the square-free condition on the polynomials $t_i(X_i)$ leads to a similar decomposition of the ring $\mathcal{R}$. It is a direct sum of finite chain rings, and so it is a *serial* ring [20]. This decomposition is based on the corresponding decomposition of the semisimple ring $\mathbb{F}_q[X_1, \ldots, X_r]/\langle \bar{t}_1(X_1), \ldots, \bar{t}_r(X_r)\rangle$ (obtained in [18, 19]).

3.1. Decomposition of $\mathcal{R}$. From now on

$$I = \langle t_1(X_1), \ldots, t_r(X_r)\rangle \lhd R[X_1, \ldots, X_r]$$

will be the ideal generated by the polynomials $t_i(X_i)$, $i = 1, \ldots, r$, where $\bar{t}_i(X_i)$ is square-free. Let H_i be the set of roots of $\bar{t}_i(X_i)$ in a suitable extension field K_i of $\mathbb{F}_q$ (notice that $\bar{t}_i(X_i)$ has no multiple roots).

DEFINITION 3.1. *Let $\mu = (\mu_1, \ldots, \mu_r) \in H_1 \times \cdots \times H_r$; then we define the* class *of μ as*

$$C(\mu) = \left\{ (\mu_1^{q^s}, \ldots, \mu_r^{q^s}) \mid s \in \mathbb{N} \right\}. \tag{3.1}$$

PROPOSITION 3.2 (see [19]). *If $\mu = (\mu_1, \ldots, \mu_r) \in H_1 \times \cdots \times H_r$ and d_i is the degree of $\mathrm{Irr}(\mu_i, \mathbb{F}_q)$ for $i = 1, \ldots, r$, then*

$$|C(\mu)| = \text{l.c.m.}(d_1, d_2, \ldots, d_r) = [\mathbb{F}_q(\mu_1, \ldots, \mu_r) : \mathbb{F}_q],$$

where l.c.m. stands for least common multiple. Moreover, the set of classes $C(\mu)$ is a partition of $H_1 \times \cdots \times H_r$, and for any ideal $J \lhd \mathbb{F}_q[X_1, \ldots, X_r]/\langle \bar{t}_1(X_1), \ldots, \bar{t}_r(X_r)\rangle$ the affine variety $V(J)$ of common zeros of the elements in J is a union of $C(\mu)$ classes.

Proof. See [19, Chapter 5, Propositions 1, 2, and 3]. □

DEFINITION 3.3. *If $\mu = (\mu_1, \ldots, \mu_r) \in H_1 \times \cdots \times H_r$, then for all $i = 1, \ldots, r$ let $p_{\mu,i}(X_i)$ denote the polynomial $\mathrm{Irr}(\mu_i, \mathbb{F}_q)$. Additionally, for all $i = 2, \ldots, r$, consider the polynomials $b_{\mu,i}(X_i) = \mathrm{Irr}(\mu_i, \mathbb{F}_q(\mu_1, \ldots, \mu_{i-1})) \in \mathbb{F}_q(\mu_1, \ldots, \mu_{i-1})[X_i] = \mathbb{F}_q[\mu_1, \ldots, \mu_{i-1}, X_i]$ and $\widetilde{b_{\mu,i}}(X_i) = \frac{p_{\mu,i}(X_i)}{b_{\mu,i}(X_i)}$. Then, define the polynomials*

$$w_{\mu,i}(X_1, \ldots, X_i), \pi_{\mu,i}(X_1, \ldots, X_i) \in \mathbb{F}_q[X_1, \ldots, X_i]$$

obtained from $b_{\mu,i}(X_i)$ and $\widetilde{b_{\mu,i}}(X_i)$, substituting μ_i by X_i.

Remark 1. Notice that, if $\mu' \in C(\mu)$, then $p_{\mu,i} = p_{\mu',i}$, $b_{\mu,i} = b_{\mu',i}$, $\widetilde{b_{\mu,i}} = \widetilde{b_{\mu',i}}$, $w_{\mu,i} = w_{\mu',i}$, and $\pi_{\mu,i} = \pi_{\mu',i}$. So, we can write $p_{C,i} = p_{\mu,i}$, $b_{C,i} = b_{\mu,i}$, $\widetilde{b_{C,i}} = \widetilde{b_{\mu,i}}$, $w_{C,i} = w_{\mu,i}$, and $\pi_{C,i} = \pi_{\mu,i}$, where $C = C(\mu)$ is the class of μ.

Remark 2. For all $i = 1, \dots, r$, since $p_{C,i} \mid \bar{t}_i$, we have that $\bar{t}_i = p_{C,i}\widehat{p_{C,i}}$, where the polynomials $p_{C,i}, \widehat{p_{C,i}}$ are coprime. Also, for all $i = 2, \dots, r$, we have that $b_{C,i}, \widetilde{b_{C,i}}$ are coprime.

Remark 3. The following ring isomorphism holds:

$$\mathbb{F}_q[X_1, \dots, X_r]/\langle p_{C,1}, w_{C,2}, \dots, w_{C,r}\rangle \cong \mathbb{F}_q(\mu_1, \dots, \mu_r). \tag{3.2}$$

DEFINITION 3.4. *If C is the class of $\mu \in H_1 \times \cdots \times H_r$, then for all $i = 1, \dots, r$ we define $q_{C,i}$ as the Hensel's lifting of the polynomial $p_{C,i}$ to $R[X_i]$ with respect to the factorization of Remark 2, i.e., $t_i = q_{C,i}\widehat{q_{C,i}}$. Also, for all $i = 2, \dots, r$, consider the lifting factors $v_{C,i}, \widetilde{v_{C,i}}$ of the factorization $p_{C,i} = b_{C,i}\widetilde{b_{C,i}} \in \mathbb{F}_q(\mu_1, \dots, \mu_{i-1})[X_i]$ to $R_{i-1}[X_i]$, where R_{i-1} is the local ring $R(\theta_1, \dots, \theta_{i-1})$ (θ_i is a root of $q_{C,i}$ lifting μ_i). Since $v_{C,i}, \widetilde{v_{C,i}} \in R[\theta_1, \dots, \theta_{i-1}, X_i]$, we can substitute θ_i by X_i to get polynomials*

$$z_{C,i}(X_1, \dots, X_i), \sigma_{C,i}(X_1, \dots, X_r) \in R[X_1, \dots, X_i].$$

Remark 4. The ring $T = R[X_1, \dots, X_r]/\langle q_{C,1}, z_{C,2}, \dots, z_{C,r}\rangle$ is local with maximal ideal $M = \langle a, q_{C,1}, z_{C,2}, \dots, z_{C,r}\rangle + \langle q_{C,1}, z_{C,2}, \dots, z_{C,r}\rangle$ and quotient ring

$$T/M \cong \mathbb{F}_q(\mu_1, \dots, \mu_r). \tag{3.3}$$

From now on, we shall denote the ideal $\langle q_{C,1}, z_{C,2}, \dots, z_{C,r}\rangle$ by I_C.

LEMMA 3.5. *If C is the class of $\mu \in H_1 \times \cdots \times H_r$, then the quotient ring $R[X_1, \dots, X_r]/I_C$ is a finite commutative chain ring with maximal ideal $\langle a + I_C\rangle$, residue field $\mathbb{F}_q(\mu_1, \dots, \mu_r)$, and precisely the following ideals:*

$$\langle 0\rangle = \left\langle a^t + I_C\right\rangle \subsetneq \left\langle a^{t-1} + I_C\right\rangle \subsetneq \cdots \subsetneq \left\langle a^1 + I_C\right\rangle = M \subsetneq \left\langle a^0 + I_C\right\rangle. \tag{3.4}$$

Proof. It is a straightforward conclusion of the discussion in Remark 4. ∎

DEFINITION 3.6. *If C is the class of $\mu \in H_1 \times \cdots \times H_r$, then we define the following polynomial in $R[X_1, \dots, X_r]$:*

$$h_C(X_1, \dots, X_r) = \prod_{i=1}^{r} \frac{t_i(X_i)}{q_{C,i}(X_i)} \prod_{i=2}^{r} \sigma_{C,i}(X_2, \dots, X_r). \tag{3.5}$$

PROPOSITION 3.7. *If C is the class of $\mu \in H_1 \times \cdots \times H_r$, then the annihilator of $\langle h_C + I\rangle$ in $R[X_1, \cdots, X_r]/I$ is*

$$\operatorname{Ann}(\langle h_C + I\rangle) = I_C + I. \tag{3.6}$$

Proof. Clearly $I_C + I \subseteq \operatorname{Ann}(\langle h_C + I\rangle)$. On the other hand, if $g + I \in \operatorname{Ann}(\langle h_C + I\rangle)$, then $\overline{gh}_\mu \in \bar{I} = \langle \bar{t}_1(X_1), \dots, \bar{t}_r(X_r)\rangle$, and so $\bar{g} + \bar{I} \in \operatorname{Ann}(\langle \bar{h}_C + \bar{I}\rangle) = \langle \bar{q}_{C,1}, \bar{z}_{C,2}, \dots, \bar{z}_{C,r}\rangle$ [19, Chapter 5, Proposition 6]. Hence $g + I \in \langle I_C + \langle a\rangle\rangle + I$ and thus $\operatorname{Ann}(\langle h_C + I\rangle) = \langle I_C + \langle a^s\rangle\rangle + I$ for some $s \in \{0, \dots, t\}$. Now, if θ_i is a root of $q_{C,i}$ lifting μ_i and we denote $\Theta = (\theta_1, \dots, \theta_r)$, then $h_C(\Theta) \notin \langle a\rangle$ (since $\bar{h}_C(\mu) \neq 0$ [19, Chapter 5, Proposition 7]), and therefore we can conclude $\operatorname{Ann}(\langle h_C + I\rangle) = I_C + I$ as desired (otherwise $s < t$, and so $a^{t-1} = a^s a^{t-1-s} \in \operatorname{Ann}(\langle h_C + I\rangle)$ implies $a^{t-1}h_C \in I$ and $0 = a^{t-1}h_C(\Theta)$, i.e., $h_C(\Theta) \in \langle a\rangle$, a contradiction). ∎

LEMMA 3.8. *Let $\mathcal{C}$ be the set of classes $C(\mu)$, where $\mu \in H_1 \times \cdots \times H_r$, and let $C, C' \in \mathcal{C}$. Then, the set of zeros of $\bar{h}_C$ is $H_1 \times \cdots \times H_r \setminus C$, and the set of zeros of $\bar{I}_C$ is C. Moreover, we have the following:*

1. $\langle t_1(X_1), \ldots, t_r(X_r)\rangle = \bigcap_{C\in\mathcal{C}} I_C$.
2. I_C, $I_{C'}$ *are comaximal if* $C \neq C'$, *i.e.,* $\langle I_C, I_{C'}\rangle = R[X_1, \ldots, X_r]$.

Proof. The proof is a direct translation of [19, Chapter 5, Proposition 7]. Note that the ideal $\bar{I} = \langle \bar{t}_1(X_1), \ldots, \bar{t}_r(X_r)\rangle$ is a radical ideal in $\mathbb{F}_q[X_1, \ldots, X_r]$, and that the following equality of affine varieties holds:

$$V(\langle \bar{t}_1(X_1), \ldots, \bar{t}_r(X_r)\rangle) = \bigsqcup_{C\in\mathcal{C}} C = V(\bar{I}_C). \tag{3.7}$$

Thus, $\langle \bar{t}_1(X_1), \ldots, \bar{t}_r(X_r)\rangle = \bigcap_{C\in\mathcal{C}} \bar{I}_C$.

1. Clearly $\langle t_1(X_1), \ldots, t_r(X_r)\rangle \subseteq \bigcap_{C\in\mathcal{C}} I_C$. On the other hand, if $f \in \bigcap_{C\in\mathcal{C}} I_C$, then by Proposition 3.7 we have that $f + I \in \operatorname{Ann}(\langle h_C + I\rangle)$, i.e., $fh_C \in I$, for all $C \in \mathcal{C}$. The result now follows.
2. This claim arises from the fact that the union in (3.7) is disjoint. ∎

THEOREM 3.9. *Let* $\mathcal{C}$ *be the set of classes* $C(\mu)$, *where* $\mu \in H_1 \times \cdots \times H_r$, *and let* $I = \langle t_1(X_1), \ldots, t_r(X_r)\rangle$. *For all* $C \in \mathcal{C}$, *let* $I_C = \langle q_{C,1}, z_{C,2}, \ldots, z_{C,r}\rangle$ *and* h_C *be as in* (3.5). *Then, the following isomorphism holds:*

$$R[X_1, \ldots, X_r]/I \cong \bigoplus_{C\in\mathcal{C}} \langle h_C + I\rangle, \tag{3.8}$$

where $\langle h_C + I\rangle \cong R[X_1, \ldots, X_r]/I_C$ *is a finite commutative chain ring with maximal ideal* $\langle a + I_C\rangle$.

Proof. The result follows from the Chinese remainder theorem:

$$R[X_1, \ldots, X_r]/I = R[X_1, \ldots, X_r]/\bigcap_{C\in\mathcal{C}} I_C \cong \bigoplus_{C\in\mathcal{C}} R[X_1, \ldots, X_r]/I_C. \qquad ∎$$

Remark 5. This theorem extends the corresponding result for cyclic codes ($R = \mathbb{F}_q, r = 1, t(X) = x^{i_1} - 1$). In that case the decomposition of $\mathbb{F}_q[X]/\langle X^{i_1} - 1\rangle$ can also be obtained from the discrete Fourier decomposition of the algebra. Indeed, the decomposition of the ring $R[X_1, \ldots, X_r]/I$ is equivalent to the existence of primitive orthogonal idempotents elements $e_C \in \mathcal{R} = R[X_1, \ldots, X_r]/I$ (one for each class $C \in \mathcal{C}$) such that $1 = \sum_{C\in\mathcal{C}} e_C$ and $e_C\mathcal{R} \cong \langle h_C + I\rangle$ [2, Proposition 3.1.3]. We consider polynomials g_C, $C \in \mathcal{C}$, such that the idempotent e_C is the element $g_C h_C + I$, so that $g_C h_C + I_C = 1 + I_C$.

3.2. Description of the codes. As mentioned in the Introduction, codes over a finite field are the main object of study of classical coding theory (the textbook [11] is a good introduction to the topic). Natural modifications lead us to codes over finite rings [2].

For the finite commutative chain ring R, let R^n be the R-module of n-uples. We say that a subset $\mathcal{K}$ of R^n is a *linear code* if $\mathcal{K}$ is an R-submodule of R^n. Given an ideal $J \lhd R[X_1, \ldots, X_r]$ such that the algebra $R[X_1, \ldots, X_r]/J$ has finite rank n as an R-module, and given an ordering on the set of terms, each element of $R[X_1, \ldots, X_r]/J$ can be identified with an n-uple in R^n.

The scalar product of the elements $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n) \in R^n$ is $\mathbf{x}\cdot\mathbf{y} = x_1y_1 + \cdots + x_ny_n \in R$. We say that $\mathbf{x}, \mathbf{y}$ are *orthogonal* if $\mathbf{x}\cdot\mathbf{y} = 0$, and for a linear code $\mathcal{K}$ we define the *dual code* as $\mathcal{K}^\perp = \{\mathbf{x} \in R^n \mid \mathbf{x}\cdot\mathbf{c} = 0 \quad \forall \mathbf{c} \in \mathcal{K}\}$. The code $\mathcal{K}$ is called *self-dual* if $\mathcal{K} = \mathcal{K}^\perp$.

DEFINITION 3.10 (multivariable serial code). *Let* $t_i(X_i) \in R[X_i]$ *be monic polynomials over the finite chain ring* R, $i = 1, \ldots, r$. *A* multivariable code *is an ideal* $\mathcal{K}$

of the ring $R[X_1,\dots,X_r]/\langle t_1(X_1),\dots,t_r(X_r)\rangle$. *If the polynomials* $\bar{t}_i(X_i)$ *are square-free, then we shall say that the code is* serial.

We shall see later (Corollary 3.11) that any multivariable serial code is a sum of finite commutative chain rings, i.e., of uniserial rings. This explains the name of the codes.

Multivariable codes include, among others, cyclic and negacyclic semisimple codes [8], as well as multivariable semisimple codes over finite fields [19]. Next we present an example of nontrivial codes that belong to this class too.

Example 1 (see [16]). Let $R = GR(q^2, 2^2)$ $(q = 2^l)$ be the Galois ring of cardinality q^2 and characteristic 2^2 [12], and let $S = GR(q^{2m}, 2^2)$ be its Galois extension of odd degree $m \geq 3$. Both R and S are finite commutative chain rings with maximal ideals $2R$ and $2S$ and residue fields $\overline{R} = GF(q)$ and $\overline{S} = GF(q^m)$, respectively. The set $\Gamma(S) = \{\beta^{q^m} = \beta \mid \beta \in S\}$ is known as Teichmüller coordinate set (TCS). Any element $\beta \in S$ can be decomposed as $\beta = \gamma_0(\beta) + 2\gamma_1(\beta)$, where $\gamma_i(\beta) \in \Gamma(S)$. Moreover, if $\oplus : \Gamma(S) \times \Gamma(S) \to \Gamma(S)$ is defined as $\beta \oplus \varepsilon = \gamma_0(\beta + \varepsilon)$, then $(\Gamma(S), \oplus, \cdot)$ is the finite field $GF(q^m)$. If $\lambda \in \Gamma(S)$ is a generator of the cyclic group $\Gamma(S)^*$, then its multiplicative order is $\tau = q^m - 1$. The TCS of R, $\Gamma(R) = \{\beta^q = \beta \mid \beta \in R\} = \{w_0 = 0, w_1, \dots, w_{q-1}\}$, is the unique subfield $GF(q)$ of $\Gamma(S)$. Let $\mathrm{Tr} : S \to R$ be the trace function from S onto R. We define the (shortened) *R-base linear code*

$$\mathcal{L} = \{(\mathrm{Tr}(\xi) + \beta, \mathrm{Tr}(\xi\lambda) + \beta, \dots, \mathrm{Tr}(\xi\lambda^{\tau-1}) + \beta) \mid \xi \in S, \beta \in R\}.$$

It is an R-linear code of length τ and cardinality $q^{2(m+1)}$. The (shortened) *generalized Kerdock code* is the projection of $\mathcal{L}$ in $\Gamma(R)^{\tau q}$ with the help of τ copies of the RS-map:

$$\gamma_*(\beta) = (\gamma_1(\beta), \gamma_1(\beta) \oplus w_1\gamma_0(\beta), \dots, \gamma_1(\beta) \oplus w_{q-1}\gamma_0(\beta)), \quad \beta \in R.$$

It is a $GF(q)$-nonlinear code of length τq, cardinality $q^{2(m+1)}$, and Hamming distance $\frac{q-1}{q}(n - \sqrt{n}) - q$.

This generalized Kerdock code can be presented in a polycyclic form. To do this we consider a multivariable code over the finite chain ring R. The multiplicative group $U = 1 + 2R = \{u_0 = 1, u_1, \dots, u_{q-1}\}$ is a direct product $\langle\eta_1\rangle \times \cdots \times \langle\eta_l\rangle$ of l subgroups of order 2. Let $r = l + 1$, and consider the ideal I of $R[X_1,\dots,X_r]$ generated by the polynomials

$$t_1(X_1) = X_1^\tau - 1, t_2(X_2) = X_2^2 - 1, \dots, t_r(X_r) = X_r^2 - 1.$$

If we denote $\overrightarrow{U} = (u_0,\dots,u_{q-1})$ and $\overrightarrow{\beta} \otimes \overrightarrow{U} = (\beta_1\overrightarrow{U},\dots,\beta_q\overrightarrow{U}) \in R^{q\tau}$, for any $\overrightarrow{\beta} \in R^\tau$, then the multivariable code $\mathcal{K} \lhd R[X_1,\dots,X_r]/I$ given by

$$\mathcal{K} = \left\{ \sum_{i_1=0}^{\tau-1} \sum_{i_2=0}^{1} \cdots \sum_{i_r=0}^{1} \left((\mathrm{Tr}(\xi\lambda^{i_1}) + \beta)\eta_1^{i_2} \dots \eta_l^{i_r}\right) X_1^{i_1} X_2^{i_2} \dots X_r^{i_r} \mid \xi \in S, \beta \in R \right\}$$

is equivalent to the code $\mathcal{L} \otimes \overrightarrow{U}$. Moreover, the shortened generalized Kerdock code is equivalent to the polycyclic code $\gamma_1^{q\tau}(\mathcal{K})$. Notice that this code is not serial, though.

Next we describe the structure of multivariable serial codes. The following two results are straightforward corollaries of Theorem 3.9.

Corollary 3.11. *If* $I = \langle t_1(X_1),\dots,t_r(X_r)\rangle$ *such that* $\bar{t}_i(X_i)$ *is square-free, then any semisimple serial code* $\mathcal{K} \lhd R[X_1,\dots,X_r]/I$ *is a sum of ideals of the form*

$$\left\langle a^{j_C} h_C + I \right\rangle, \qquad 0 \leq j_C \leq t,\ C \in \mathcal{C}, \tag{3.9}$$

where t is the nilpotency index of the maximal ideal $\langle a\rangle$ of R, and $\mathcal{C}$, h_C are as in Theorem 3.9.

COROLLARY 3.12. *In the conditions of the previous corollary, there are $(t+1)^N$ multivariable serial codes in $R[X_1,\dots,X_r]/I$, where $N=|\mathcal{C}|$.*

We shall now obtain an explicit description of multivariable serial codes in terms of polynomials of the ring $R[X_1,\dots,X_r]$.

THEOREM 3.13. *If $\mathcal{K}$ is a multivariable serial code in $R[X_1,\dots,X_r]/I$, then there exists a family of polynomials $G_0,\dots,G_t\in R[X_1,\dots,X_r]$ such that*

$$I=\bigcap_{i=0}^{t}\operatorname{Ann}\langle G_i+I\rangle \quad \text{and} \quad \mathcal{K}=\left\langle G_1,aG_2,\dots,a^{t-1}G_t\right\rangle+I. \tag{3.10}$$

The ideals $\langle G_i+I\rangle$ are uniquely determined, and the ideals $\operatorname{Ann}\langle G_i+I\rangle$, $\operatorname{Ann}\langle G_j+I\rangle$ are comaximal; i.e., $\langle \operatorname{Ann}(G_i+I),\operatorname{Ann}(G_j+I)\rangle=\mathcal{R}$. Moreover, $\mathcal{K}=\langle G+I\rangle$, where $G=\sum_{i=0}^{t-1}a^iG_{i+1}$.

Proof. By Corollary 3.11, $\mathcal{K}$ is a direct sum of ideals of the form $\left\langle a^{j_C}h_C+I\right\rangle$, where $0\le j_C\le t$ and $C\in\mathcal{C}$. If $N=|\mathcal{C}|$ is the number of classes in $\mathcal{C}$, then, reordering the classes in $\mathcal{C}$ if necessary, we have

$$\begin{aligned}\mathcal{K}=&\left\langle h_{C_{k_1+1}}+I\right\rangle\oplus\cdots\oplus\left\langle h_{C_{k_1+k_2}}+I\right\rangle\\&\oplus\left\langle ah_{C_{k_1+k_2+1}}+I\right\rangle\oplus\cdots\oplus\left\langle ah_{C_{k_1+k_2+k_3}}+I\right\rangle\oplus\cdots\\&\oplus\left\langle a^{t-1}h_{C_{\sum_{i=1}^{t}k_i+1}}+I\right\rangle\oplus\cdots\oplus\left\langle a^{t-1}h_{C_N}+I\right\rangle,\end{aligned}$$

where $k_i\ge 0$, for all $i=1,2,\dots,t$, and $\sum_{i=1}^{t}k_i+1\le N$. Let $k_0=0$ and $k_{t+1}=N-\sum_{i=1}^{t}k_i$, and define

$$G_i=\sum_{j=k_0+\cdots+k_i+1}^{k_0+\cdots+k_{i+1}}g_{C_j}h_{C_j},$$

where $g_{C_j}\in R[X_1,\dots,X_r]$, $j=k_0+\cdots+k_i+1,\dots,k_0+\cdots+k_{i+1}$, $i=0,\dots,t$, are the polynomials defining the primitive orthogonal idempotents of Remark 5. Then

$$\langle G_i+I\rangle=\sum_{j=k_0+\cdots+k_i+1}^{k_0+\cdots+k_{i+1}}\left\langle h_{C_j}+I\right\rangle,$$

and so we have $\mathcal{K}=\left\langle G_1,aG_2,\dots,a^{t-1}G_t\right\rangle+I$ and

$$\bigcap_{i=0}^{t}\operatorname{Ann}\langle G_i+I\rangle=\bigcap_{i=0}^{t}\bigcap_{j=k_0+\cdots+k_i+1}^{k_0+\cdots+k_{i+1}}\operatorname{Ann}\left(\left\langle h_{C_j}+I\right\rangle\right)=\bigcap_{k=0}^{N}I_{C_k}+I=I.$$

Moreover, the ideals $\operatorname{Ann}\langle G_i+I\rangle$, $\operatorname{Ann}\langle G_j+I\rangle$ are comaximal, in view of Lemma 3.8. The uniqueness of the ideals $\langle G_i+I\rangle$ follows from fact that the decomposition in Theorem 3.9 is unique and from Corollary 3.11. Finally, the equality $\mathcal{K}=\langle G+I\rangle$ is satisfied, since each element G_i is a sum of primitive idempotent orthogonals of the ring. Let us notice that adding the elements a^iG_{i-1} to get one single generator is similar to the technique used in [5, Corollary after Theorem 6]. The fact that the ideal is principal was also proved in [6] for certain cases. □

With this description in hand we can obtain the cardinality of any multivariable serial code.

COROLLARY 3.14. *In the conditions of Theorem* 3.13*, $R[X_1,\dots,X_r]/I$ is a principal ideal ring and, for any multivariable serial code $\mathcal{K}$, we have*

$$|\mathcal{K}| = |\bar{R}|^{\sum_{i=0}^{t-1}(t-i)N_i},$$

where N_i is the number of zeros $\mu \in H_1 \times \cdots \times H_r$ of $\bar{G}_i, i = 0,\dots,t-1$.

Proof. For $i = 0,\dots,t-1$, we have

$$\left\langle a^i G_{i+1} + I \right\rangle = \left(\frac{|R|}{|\langle a^i \rangle|} \right)^{\mathrm{rank}_R(\langle \mathrm{G_i + I} \rangle)} = |\bar{R}|^{(t-i)\mathrm{rank}_R(\langle \mathrm{G_i + I} \rangle)}.$$

Since $\mathrm{rank}_R(\langle G_i + I\rangle) = \dim_{\bar{R}} \left\langle \bar{G}_i + \bar{I} \right\rangle$, the result follows from [19]. □

3.3. Hamming distance of the codes. For $\mathbf{c} \in R^n$ we denote by $\mathrm{wt}(c)$ the *Hamming weight* of $\mathbf{c}$, that is, the cardinality of $\mathrm{supp}(\mathbf{c}) = \{i \mid c_i \neq 0\}$, the *support* of $\mathbf{c}$. The *minimum distance* of a code $\mathcal{K} \in R^n$, i.e., the minimum Hamming weight of the nonzero elements in $\mathcal{K}$, will be denoted by $d(\mathcal{K})$.

DEFINITION 3.15. *The* socle $\mathfrak{S}(\mathcal{K})$ *of an R-linear code $\mathcal{K}$ is defined as the sum of all its irreducible R-submodules.*

According to [10], the equality

$$\mathfrak{S}(\mathcal{K}) = \{\mathbf{c} \in \mathcal{K} \mid M\mathbf{c} = 0\}$$

holds for any R-linear code $\mathcal{K}$. So we may consider $\mathfrak{S}(\mathcal{K})$ as a linear space over the field $\mathbb{F}_q$, where $\bar{r} \cdot \mathbf{c} = r\mathbf{c}$, for all $\bar{r} \in \mathbb{F}_q$, $\mathbf{c} \in \mathfrak{S}(\mathcal{K})$.

LEMMA 3.16 (see [10]). *If $\mathcal{K}$ is an R-linear code of length n, then the socle $\mathfrak{S}(\mathcal{K})$ is a linear code of length n over the field $\mathbb{F}_q$ and $d(\mathcal{K}) = d(\mathfrak{S}(\mathcal{K}))$.*

Proof. For the proof, see [10, Proposition 5]. □

PROPOSITION 3.17. *In the conditions of Theorem* 3.13*, $d(\mathcal{K}) = d(\mathcal{L})$, where $\mathcal{L}$ is the code $\left\langle \overline{G_1},\dots,\overline{G_t} \right\rangle + \overline{I}$ in $\mathbb{F}_q[X_1,\dots,X_r]/\left\langle \bar{t}_1(X_1),\dots,\bar{t}_r(X_r) \right\rangle$.*

Proof. The socle of the code $\mathcal{K}$ is $\mathfrak{S}(\mathcal{K}) = \left\langle a^{t-1}G_1, a^{t-1}G_2,\dots,a^{t-1}G_t \right\rangle + I$, which can be seen as a linear code over $\mathbb{F}_q$. Consider the $\mathbb{F}_q$-vector space isomorphism $\phi : a^{t-1}R[X_1,\dots,X_r]/I \to \mathbb{F}_q[X_1,\dots,X_r]/\bar{I}$, given by $a^{t-1}g + I \to \overline{g} + \overline{I}$, to conclude the result. □

In general, we can not state that the minimum distance of a semisimple code $\mathcal{K}$ is equal to the minimum distance of the code $\overline{\mathcal{K}}$. The most we can say is that, if $\overline{\mathcal{K}} \neq 0$, then $d(\mathcal{K}) \leq d(\overline{\mathcal{K}})$. However, there is one class of multivariable serial codes for which the equality holds.

DEFINITION 3.18. *In the conditions of Theorem* 3.13*, the code $\mathcal{K}$ is called the* Hensel lift of a multivariable semisimple code *if $\langle G_1 + I\rangle \neq I$ and $\langle G_i + I\rangle = 0$, for all $i = 2,\dots,t$.*

Notice that, although a Hensel lift code lifts a semisimple code over a field, it is *not* semisimple. It is a serial ring, as we have seen at the beginning of this section.

This notion extends the definition of a *Hensel lift of a cyclic code* introduced in [17]. For this class of codes we have the following result, which naturally generalizes [17, Corollary 4.3].

COROLLARY 3.19. *If $\mathcal{K} \neq 0$ is a Hensel lift of a multivariable semisimple code, then $d(\mathcal{K}) = d(\overline{\mathcal{K}})$.*

Proof. As noticed above, the inequality $d(\mathcal{K}) \leq d(\overline{\mathcal{K}})$ holds. On the other hand, since $\mathcal{K}$ is a Hensel lift of a multivariable semisimple code, we have that $\mathcal{L} = \overline{\mathcal{K}}$, and the result follows from Proposition 3.17. □

All classical bounds on distances for multivariable semisimple codes over fields (Bose–Ray-Chaudhuri–Hocquenghem, Hartmann–Tzeng, Roos, etc.) also apply to their Hensel lifts. Note that these bounds can be stated in the multivariable abelian case due to Proposition 8 in [19, Chapter 6], which we recall in Proposition 3.21 below.

DEFINITION 3.20. *A multivariable code $\mathcal{K} \lhd R[X_1, \ldots, X_r]/I$ is called* abelian *if $I = \langle X_1^{i_1} - 1, \ldots, X_r^{i_r} - 1 \rangle$, where $i_1, \ldots, i_r \in \mathbb{N}$.*

Let $\bigsqcup_{i=1}^{l} \bigsqcup_{j=1}^{s_i} C^{(i,j)}$ be the set of defining roots of a multivariable semisimple abelian code in $\mathbb{F}_q[X_1, \ldots, X_r]/\bar{I}$ [19], where $C^{(i,j)} = C(\mu^{(i,j)}) \in \mathcal{C}$ such that $p_{C^{(i,j)},1} = p_{C^{(k,l)},1}$ if and only if $i = k$. Consider, for all $C^{(i,j)}$, the polynomial

$$\frac{\bar{t}_1(X_1)}{p_{C^{(i,j)},1}(X_1)} \left(\prod_{k=2}^{r} \frac{\bar{t}_k(X_k)}{p_{C^{(i,j)},k}(X_k)} \prod_{k=2}^{r} \pi_{C^{(i,j)},k}(X_2, \ldots, X_r) \right) = \frac{\bar{t}_1(X_1)}{p_{C(i,j),1}(X_1)} \left(F_{ij}(X_2, \ldots, X_r) \right).$$

Here $p_{C^{(i,j)},k}$ and $\pi_{C^{(i,j)},k}$ are as in the previous section. The polynomial $F_{ij} \in \mathbb{F}_q[X_2, \ldots, X_r]$ is uniquely determined by the class $C^{(i,j)}$. Let us consider the finite field

$$\mathbb{F}^{(i)} = \mathbb{F}_q(X_1)/p_{C^{(i,1)},1}(X_1)$$

and the code $\mathcal{J}_i$ generated by $\sum_{j=1}^{s_i} F_{ij}$ in the algebra $\mathbb{F}^{(i)}[X_2, \ldots, X_r]/\langle \bar{t}_2, \ldots, \bar{t}_r \rangle$, $i = 1, \ldots, l$. We have the following result.

PROPOSITION 3.21 (see [19]). *The minimum weight of a multivariable semisimple code over a field $\mathbb{F}_q$ and of its Hensel lift over the ring R is at least $\min_{1 \leq i \leq l}\{d_i \cdot \delta_i\}$, where d_i is the minimum weight of the code in $\mathbb{F}_q[X_1]/\bar{t}(X_1)$ generated by*

$$\frac{\bar{t}_1(X_1)}{p_{C^{(i,1)},1}(X_1) \cdot \cdots \cdot p_{C^{(l,1)},1}(X_1)}$$

and δ_i is the minimum weight of the code $\mathcal{J}_i$.

Proof. The proof is a straightforward generalization of Lemma 3 and Proposition 8 in [19, Chapter 6]. □

Remark 6. Notice that, in view of this result, the computation of the minimum distance of a multivariable serial abelian code in r variables is reduced to computations of minimum distances of multivariable semisimple abelian codes over a finite field in fewer variables.

4. Dual codes of multivariable serial abelian codes. In this section we describe the dual codes of multivariable serial abelian codes. Notice that any defining ideal I of a multivariable serial abelian code, as in Definition 3.20, must satisfy the following property: $(i_j, p) = 1$ for all $j = 1, \ldots, r$, since the code is serial. On the other hand, any multivariable abelian code (not necessarily serial) can be seen also as a *group code*, i.e., as an ideal of a certain group ring, namely the group ring $R(\mathsf{C}_{i_1} \times \cdots \times \mathsf{C}_{i_r})$, where C_s is the cyclic group of order s.

DEFINITION 4.1. *Let $R[X_1, \ldots, X_r]/I$ be a multivariable serial abelian code with $I = \langle x_1^{i_1} - 1, \ldots, X_r^{i_r} - 1 \rangle$. Then we define the ring automorphism τ of $R[X_1, \ldots, X_r]/I$ given by $\tau(f(X_1, \ldots, X_r)) = f(X_1^{-1}, \ldots, X_r^{-1}) = f(X_1^{i_1 - 1}, \ldots, X_r^{i_r - 1})$.*

It is clear that this automorphism preserves the Hamming weights.

THEOREM 4.2. *If* $\mathcal{K} = \langle G_1, aG_2, \ldots, a^{t-1}G_t \rangle + I$ *is a multivariable serial abelian code in the conditions of Theorem* 3.13*, then its dual code is*

$$\mathcal{K}^{\perp} = \langle \tau(G_0), a\tau(G_t), \ldots, a^{t-1}\tau(G_2) \rangle + I,$$

where the polynomials $\tau(G_i)$*,* $i = 0, 2, 3, \ldots, t$*, are also in the conditions of Theorem* 3.13*.*

Proof. Let us first prove that $\mathcal{K}^{\perp} = \tau(\mathrm{Ann}(\mathcal{K}))$. For all $F + I \in R[X_1, \ldots, X_r]/I$ we have that $F + I \in \tau(\mathrm{Ann}(\mathcal{K}))$ if and only if, for all $Q + I \in \mathcal{K}$,

$$\begin{aligned} I &= Q\tau(F) + I \\ &= \sum_{l_1,\ldots,l_r} q_{l_1,\ldots,l_r} X_1^{l_1} \ldots X_r^{l_r} \sum_{j_1,\ldots,j_r} f_{j_1,\ldots,j_r} X_1^{i_1-j_1} \ldots X_r^{i_r-j_r} + I \\ &= \sum_{k_1,\ldots,k_r} \left(\sum_{l_1,\ldots,l_r} q_{l_1,\ldots,l_r} f_{l_1-k_1 \pmod{i_1},\ldots,l_r-k_r \pmod{i_r}} \right) X_1^{k_1} \ldots X_r^{k_r} + I \\ &= \sum_{k_1,\ldots,k_r} (\mathbf{q} \cdot \mathbf{z}_{\mathbf{k_1},\ldots,\mathbf{k_r}}) X_1^{k_1} \ldots X_r^{k_r} + I, \end{aligned}$$

where $\mathbf{q}$ and $\mathbf{z}_{\mathbf{k_1},\ldots,\mathbf{k_r}}$ denote, respectively, the vectors of coefficients of Q and $X_1^{k_1} \ldots X_r^{k_r} F$, in a fixed ordering of the terms in $R[X_1, \ldots, X_r]/I$. Hence, $F + I \in \tau(\mathrm{Ann}(\mathcal{K}))$ if and only if, for all $Q + I \in \mathcal{K}$ and for all $0 \leq k_1 < i_1, \ldots, 1 \leq k_r < i_r$, $\mathbf{q} \cdot \mathbf{z}_{\mathbf{k_1},\ldots,\mathbf{k_r}} = 0$, i.e., $\mathbf{y}_{\mathbf{k_1},\ldots,\mathbf{k_r}} \cdot \mathbf{f} = 0$, where $\mathbf{y}_{\mathbf{k_1},\ldots,\mathbf{k_r}}$ denotes the vector of coefficients of $X_1^{-k_1} \ldots X_r^{-k_r} Q$; that is if and only if $F + I \in \mathcal{K}^{\perp}$.

Notice that the polynomials $\tau(G_i)$, $i = 0, \ldots, t$, are in the conditions of Theorem 3.13, and so it is enough to see that $a^i G_{t+1-i} + I \in \mathrm{Ann}(\mathcal{K})$, $i = 0, \ldots, t-1$, to conclude the result (here we denote $G_{t+1} = G_0$). Let $i, j = 0, \ldots, t-1$, if $i + j \geq t$; then $(a^i G_{t+1-i} + I)(a^j G_{j+1} + I) = a^{i+j}(G_{t+1-i} G_{j+1}) + I = I$ and, if $i + j < t$, then $\langle G_{t+1-i} + I \rangle \neq \langle G_{j+1} + I \rangle$. Thus $(a^i G_{t+1-i} + I)(a^j G_{j+1}) = I$, from the decomposition of $\mathcal{K}$ in Theorem 3.13. ∎

COROLLARY 4.3. *In the conditions of the previous theorem,*

$$|\mathcal{K}^{\perp}| = |\bar{R}|^{\sum_{i=0}^{t-1} iN_i},$$

where N_i *is the number of zeros* $\mu \in H_1 \times \ldots H_r$ *of* $\bar{G}_i, i = 0, \ldots, t-1$*, and* $\mathcal{K}^{\perp} = \langle \tau(G_0) + a\tau(G_t) + \cdots + a^{t-1}\tau(G_2) + I \rangle$*.*

Proof. The result follows from [8, Proposition 2.11] and the fact that the polynomials $\tau(G_i)$ are in the conditions of Theorem 3.13. ∎

Remark 7. In view of Theorem 4.2, all the remarks concerning the distance of a multivariable serial abelian code observed in the previous section can be applied also to its dual. Of course, the results about the minimum distance of a code and the minimum distance of its dual involving the MacWilliams identity for codes over quasi-Frobenius modules [10] also apply in our case. For the sake of brevity we will not get into details, though.

5. Self-dual abelian serial codes. In the previous section we have described explicitly the dual code of a multivariable serial abelian code $\mathcal{K}$. We now want to study conditions on $\mathcal{K}$ to be self-dual. Notice first that if the nilpotency index t of a is even, then there always exists a *trivial self-dual code* $\langle a^{\frac{t}{2}} \rangle$. On the other hand, remember

that any abelian code is also a group code, and so the problem of existence of self-dual multivariable serial abelian codes can be reduced to the existence of self-dual group codes. This problem has been solved for some classes of rings R. In this direction an interesting work is [21], where the existence of self-dual codes is characterized when R is a Galois ring. The proof of this characterization is based on group representation theory, and it can be also applied when R is a finite commutative chain ring. Namely, the following result holds.

THEOREM 5.1. *Let R be a commutative finite chain ring with maximal ideal $\langle a\rangle$ of nilpotency index t, quotient ring $\overline{R} = \mathbb{F}_{p^l}$, and let G be a finite group. Then $R\mathsf{G}$ contains a self-dual group code (that is, an ideal $\mathcal{K} \lhd R\mathsf{G}$ such that $\mathbf{x}\cdot\mathbf{y} = 0$ for all $x, y \in \mathcal{K}$) if and only if either p is odd and t even, or p and $t|\mathsf{G}|$ are even.*

Proof. The proof is exactly the same as in the case of R being a Galois ring [21]. This is due to the following two facts: any finite commutative chain ring R is a Frobenius ring [22], and for any finite group G we have a filtration

$$0 \subsetneq a^{t-1}R\mathsf{G} \subsetneq \cdots \subsetneq a^1 R\mathsf{G} \subsetneq R\mathsf{G}. \qquad \square$$

In view of this result we can only expect to find nontrivial self-dual codes in the serial abelian case if and only if either p and $|\mathsf{G}|$ are even, or t is even. The first case is clearly impossible, since $|\mathsf{G}| = \prod_{j=1}^{r} i_j$ even implies that there exists an exponent i_j even, and the code is not serial (notice that $p = 2$). So we have only to study the case when t is an even number. As a consequence of Theorem 4.2 we have the following result.

COROLLARY 5.2. *Let $\mathcal{K} = \langle G_1, aG_2, \ldots, a^{t-1}G_t\rangle + I$ be a multivariable serial abelian code in the conditions of Theorem* 3.13. *Then $\mathcal{K}$ is self-dual if and only if $\langle G_i + I\rangle = \langle \tau(G_j) + I\rangle$ when $i + j \equiv 1 \pmod{t+1}$.*

Proof. By Theorem 4.2 we have $\mathcal{K}^{\perp} = \langle \tau(G_0), a\tau(G_t), \ldots, a^{t-1}\tau(G_2)\rangle + I$. Therefore, if $\langle G_i + I\rangle = \langle \tau(G_j) + I\rangle$ with $i + j \equiv 1 \pmod{t+1}$, then $\mathcal{K} = \mathcal{K}^{\perp}$, and the code is self-dual. Conversely, if $\mathcal{K} = \mathcal{K}^{\perp}$, then $\langle G_1, aG_2, \ldots, a^{t-1}G_t\rangle + I = \langle \tau(G_0), a\tau(G_t), \ldots, a^{t-1}\tau(G_2)\rangle + I$, and the result follows from the uniqueness of the ideals in Theorem 3.13. $\square$

THEOREM 5.3. *If t is an even number, then there exist nontrivial self-dual multivariable serial abelian codes if and only if there exists $\mu \in H_1 \times \cdots \times H_r$ such that $C(\mu) \neq C(\mu^{-1})$, where $\mu^{-1} = (\mu_1^{-1}, \ldots, \mu_r^{-1})$.*

Proof. Let us first assume that there exists $\mu \in H_1 \times \cdots \times H_r$ such that $C(\mu) \neq C(\mu^{-1})$. Let $G + I$ be a generator of the multivariable serial abelian code $\bigoplus_{C\neq C(\mu), C(\mu^{-1})} \langle h_C + I\rangle$, and consider

$$\mathcal{K} = \left\langle a^{\frac{t}{2}-1}h_{C(\mu)}, a^{\frac{t}{2}}G, a^{\frac{t}{2}+1}h_{C(\mu^{-1})}\right\rangle + I.$$

Since $\langle \tau(h_{C(\mu^{-1})}) + I\rangle = \langle h_{C(\mu)} + I\rangle$ and $\langle \tau(G) + I\rangle = \langle G + I\rangle$ we have, from the previous corollary, that $\mathcal{K}$ is a nontrivial self-dual multivariable serial abelian code.

Conversely, if $\mathcal{K} = \langle G_1, aG_2, \ldots, a^{t-1}G_t\rangle + I$ is a self-dual multivariable serial abelian code, then for all i, j such that $i + j \equiv 1 \pmod{t+1}$ we have that $\langle G_i + I\rangle = \langle \tau(G_j) + I\rangle$. Assume now that $C(\mu) = C(\mu^{-1})$, for any $\mu \in H_1 \times \cdots \times H_r$. Then $\langle h_{C(\mu)} + I\rangle = \langle h_{C(\mu^{-1})} + I\rangle = \langle \tau(h_{C(\mu)}) + I\rangle$, and so $\langle G_j + I\rangle = \langle \tau(G_j) + I\rangle = \langle G_i + I\rangle$ for all i, j such that $i + j \equiv 1 \pmod{t+1}$. From the decomposition of Theorem 3.13 we obtain that $\mathcal{K} = \langle a^{\frac{t}{2}} + I\rangle$ is the trivial self-dual code. $\square$

The existence of nontrivial self-dual multivariable serial abelian codes can be eventually reduced to an arithmetic problem, as the following result shows.

COROLLARY 5.4. *If t is an even number, then there exist nontrivial self-dual multivariable serial abelian codes if and only if $q^j \not\equiv -1 \pmod{\text{l.c.m}(i_1, \ldots, i_r)}$ for all $j \in \mathbb{N}$.*

Proof. From the previous theorem we have that nontrivial self-dual multivariable serial abelian codes do not exist if and only if $C(\mu) = C(\mu^{-1})$ for all $\mu \in H_1 \times \cdots \times H_r$. If ξ_j is an i_jth primitive root of unity, then this condition is equivalent to the following one: for all $0 \leq k_j < i_j$, $j = 1, \ldots, r$, there exists a natural number l such that $\xi_j^{-k_j} = \xi_j^{q^l k_j}$; i.e., $q^l k_j \equiv -k_j \pmod{i_j}$. Therefore nontrivial self-dual multivariable serial abelian codes do not exist if and only if there exists a natural number l such that $q^l \equiv -1 \pmod{i_j}$ for all $j = 1, \ldots, r$; that is, $q^j \equiv -1 \pmod{\text{l.c.m.}(i_1, \ldots, i_r)}$. □

This result extends Theorem 4.4 in [8] for the case of self-dual cyclic codes. In this work it is also included a discussion about pairs of natural numbers (q, n) for which $q^j \not\equiv -1 \pmod{n}$, for all natural numbers j, when q is a prime number. The search of conditions for a pair of numbers to satisfy this property when q is a power of a prime number is an open problem.

REFERENCES

[1] S. D. BERMAN, *On the theory of group codes*, Cybernetics, 3 (1969), pp. 25–31.

[2] G. BINI AND F. FLAMINI, *Finite Commutative Rings and Their Applications*, Kluwer Int. Ser. Engrg. Comput. Sci. 680, Kluwer Academic Publishers, Boston, MA, 2002.

[3] E. BYRNE AND P. FITZPATRICK, *Gröbner bases over Galois rings with an application to decoding alternant codes*, J. Symbolic Comput., 31 (2001), pp. 565–584.

[4] A. R. CALDERBANK, A. R. HAMMONS, JR., P. V. KUMAR, N. J. A. SLOANE, AND P. SOLÉ, *A linear construction for certain Kerdock and Preparata codes*, Bull. Amer. Math. Soc. (N.S.), 29 (1993), pp. 218–222.

[5] A. R. CALDERBANK AND N. J. A. SLOANE, *Modular and p-adic codes*, Des., Codes Cryptogr., 6 (1995), pp. 21–35.

[6] J. CAZARAN AND A. V. KELAREV, *Generators and weights of polynomial codes*, Arch. Math., 69 (1997), pp. 479–486.

[7] P. CHARPIN, *Une généralisation de la construction de Berman des codes de Reed et Muller p-aires*, Comm. Algebra, 16 (1988), pp. 2231–2246.

[8] H. Q. DINH AND S. R. LÓPEZ-PERMOUTH, *Cyclic and negacyclic codes over finite chain rings*, IEEE Trans. Inform. Theory, 50 (2004), pp. 1728–1744.

[9] A. R. HAMMONS, JR., P. V. KUMAR, A. R. CALDERBANK, N. J. A. SLOANE, AND P. SOLÉ, *The $\mathbf{Z}_4$-linearity of Kerdock, Preparata, Goethals, and related codes*, IEEE Trans. Inform. Theory, 40 (1994), pp. 301–319.

[10] V. L. KURAKIN, A. S. KUZMIN, V. T. MARKOV, A. V. MIKHALEV, AND A. A. NECHAEV, *Linear codes and polylinear recurrences over finite rings and modules (a survey)*, in Applied Algebra, Algebraic Algorithms and Error-correcting Codes (Honolulu, HI, 1999), Lecture Notes in Comput. Sci. 1719, Springer, Berlin, 1999, pp. 365–391.

[11] F. J. MACWILLIAMS AND N.J.A. SLOANE, *The Theory of Error-correcting Codes*, North–Holland, Amsterdam, 1977.

[12] B. R. MCDONALD, *Finite Rings with Identity*, Marcel Dekker, New York, 1974.

[13] A. A. NECHAEV, *Trace function in Galois ring and noise stable codes*, in Proceedings of the V. All-Union Symposium on Theory of Rings, Algebras and Modules (Novosibirsk., 1982), p. 97 (in Russian).

[14] A. A. NECHAEV, *Kerdock's code in cyclic form*, Diskret. Mat., 1 (1989), pp. 123–139.

[15] A. A. NECHAEV AND A. S. KUZMIN, *Linearly presentable codes*, in Proceedings of the 1996 IEEE International Symposium or Information Theory and Applications, Victoria, BC, Canada, 1996, IEEE Press, Piscataway, NJ, pp. 31–34.

[16] A. A. NECHAEV AND A. S. KUZMIN, *Formal duality of linearly presentable codes over a Galois field*, in Applied Algebra, Algebraic Algorithms and Error-correcting Codes (Toulouse, 1997), Lecture Notes in Comput. Sci. 1255, Springer, Berlin, 1997, pp. 263–276.

[17] G. H. Norton and A. Sălăgean, *On the Hamming distance of linear codes over a finite chain ring*, IEEE Trans. Inform. Theory, 46 (2000), pp. 1060–1067.
[18] A. Poli, *Important algebraic calculations for n-variables polynomial codes*, Discrete Math., 56 (1985), pp. 255–263.
[19] A. Poli and L. Huguet, *Codes Correcteurs: Théorie et Applications*, Masson, Paris, 1988.
[20] G. Puninski, *Serial Rings*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001.
[21] W. Willems, *A note on self-dual group codes*, IEEE Trans. Inform. Theory, 48 (2002), pp. 3107–3109.
[22] J. A. Wood, *Duality for modules over finite rings and applications to coding theory*, Amer. J. Math., 121 (1999), pp. 555–575.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 960–976

ON IDENTITIES CONCERNING THE NUMBERS OF CROSSINGS AND NESTINGS OF TWO EDGES IN MATCHINGS*

MARTIN KLAZAR†

Abstract. Let M, N be two matchings on $[2n]$ (possibly $M = N$). For an integer $l \geq 0$, let $\mathcal{T}(M, l)$ be the set of those matchings on $[2n + 2l]$ which can be obtained from M by successively adding l times the first edge, and similarly for $\mathcal{T}(N, l)$. Let $s, t \in \{cr, ne\}$, where cr is the statistic of the number of crossings in a matching and ne is the statistic of the number of nestings (possibly $s = t$). We prove that if the statistics s and t coincide, respectively, on the sets of matchings $\mathcal{T}(M, l)$ and $\mathcal{T}(N, l)$ for $l = 0, 1$, then they coincide on these sets for every $l \geq 0$; similar identities hold for the joint statistic of cr and ne. These results are instances of a general identity for crossings and nestings weighted by elements from an abelian group.

Key words. matching, crossing, nesting

AMS subject classifications. 05A15, 05A18

DOI. 10.1137/050625357

1. Introduction and formulation of the main result. In this article we investigate distributions of the numbers of crossings and nestings of two edges in matchings. For example, it is known that for each k and n there are as many matchings M on $\{1, 2, \dots, 2n\}$ with k crossings as there are with k nestings. All matchings form an infinite tree $\mathcal{T}$ rooted in the empty matching $\emptyset$, in which the children of M are the matchings obtained from M by adding to M in all ways the new first edge. The problem we address is this: Given two (not necessarily distinct) matchings M and N on $\{1, 2, \dots, 2n\}$, when is it the case that the numbers of crossings (or nestings, or crossings versus nestings) have identical distributions on the levels of the two subtrees of $\mathcal{T}$ rooted in M and N? Our main result is Theorem 1.1, which determines when this happens, in fact in a more general setting. Before formulating it we give definitions and fix notation.

We denote the set $\{1, 2, 3, \dots\}$ by $\mathbb{N}$, the set $\mathbb{N} \cup \{0\}$ by $\mathbb{N}_0$, and (for $n \in \mathbb{N}$) the set $\{1, 2, \dots, n\}$ by $[n]$. The cardinality of a set A is denoted $|A|$. By a *multiset* we understand a "set" in which repetitions of elements are allowed. This can be modeled by a pair $H = (X, m)$, where X is a set, the *groundset* of the multiset H, and the mapping $m : X \to \mathbb{N}$ determines multiplicities of the elements in H. However, we will not need this formalism and will record multiplicities by repetitions. A *matching* M on $[2n]$ is a set partition of $[2n]$ into n two-element *blocks* which we also call *edges*. The set of all matchings on $[2n]$ is denoted $\mathcal{M}(n)$; we define $\mathcal{M}(0) = \{\emptyset\}$. Two distinct blocks A and B of M form a *crossing* (they *cross*) if $\min A < \min B < \max A < \max B$ or $\min B < \min A < \max B < \max A$. Similarly, they form a *nesting* (they are *nested*) if $\min A < \min B < \max B < \max A$ or $\min B < \min A < \max A < \max B$. We draw a diagram of M in which we put the elements $1, 2, \dots, 2n$ as points on a line, from left to right, and connect by a semicircular arc lying above the line the two points

*Received by the editors February 25, 2005; accepted for publication (in revised form) June 5, 2006; published electronically December 11, 2006.

http://www.siam.org/journals/sidma/20-4/62535.html

†Institute for Theoretical Computer Science (ITI) and Department of Applied Mathematics (KAM), Faculty of Mathematics and Physics of Charles University, Malostranské náměstí 25, 118 00 Praha, Czech Republic (klazar@kam.mff.cuni.cz). ITI is supported as project 1M0021620808 by the Ministry of Education of the Czech Republic.

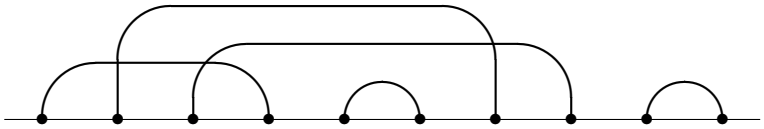


FIG. 1. *Matching with three crossings and two nestings.*

of each block. For two crossing blocks the corresponding arcs intersect, and for two nested blocks one of the arcs covers the other; see Figure 1.

By $cr(M)$, respectively $ne(M)$, we denote the number of crossings, respectively nestings, in M. The n edges of $M \in \mathcal{M}(n)$ are naturally ordered by their first elements. The first edge of M is $\{1, x\}$, and the last edge is one whose first vertex is the last one among the n first vertices.

We investigate distribution of the numbers $cr(M)$ and $ne(M)$ on $\mathcal{M}(n)$ and on the subsets of $\mathcal{M}(n)$ defined by prescribing the matching formed by the last k edges of M. The total number of matchings in $\mathcal{M}(n)$ is

$$|\mathcal{M}(n)| = (2n-1)!! = 1 \cdot 3 \cdot 5 \cdot \cdots \cdot (2n-1).$$

It is known that the number of matchings on $[2n]$ with no crossing equals the number of matchings with no nesting, and that it is the nth Catalan number; see Stanley [15, Problems 6.19o and 6.19ww] (in fact, Problem 6.19ww in [15] encodes nonnesting matchings in $\mathcal{M}(n)$ in the obvious way by standard Young tableaux of shape (n, n)):

$$|\{M \in \mathcal{M}(n) : \ cr(M) = 0\}| = |\{M \in \mathcal{M}(n) : \ ne(M) = 0\}| = \frac{1}{n+1}\binom{2n}{n}.$$

The more general result that for each k and n

$$|\{M \in \mathcal{M}(n) : \ cr(M) = k\}| = |\{M \in \mathcal{M}(n) : \ ne(M) = k\}|$$

was derived by de Sainte-Catherine in [13]. Even more is true because the joint statistic is symmetric:

$$|\{M \in \mathcal{M}(n) : \ cr(M) = k, ne(M) = l\}| = |\{M \in \mathcal{M}(n) : \ cr(M) = l, ne(M) = k\}|$$

for every $k, l \in \mathbb{N}_0$ and $n \in \mathbb{N}$. A simple proof for this symmetry can be given by adapting the Touchard–Riordan method [17], [12], which encodes matchings and their numbers of crossings by weighted Dyck paths (Klazar and Noy [9]), or see Kasraoui and Zeng [6] for more general result. Here we put these results in a more general framework.

By the *tree of matchings* $\mathcal{T} = (\mathcal{M}, E, r)$ we understand the infinite rooted tree with the vertex set

$$\mathcal{M} = \bigcup_{n=0}^{\infty} \mathcal{M}(n),$$

which is rooted in the empty matching $r = \emptyset$ and in which directed edges in $E(\mathcal{T})$ are the pairs (M, N) such that $M \in \mathcal{M}(n)$, $N \in \mathcal{M}(n+1)$, and N arises from M by adding a new first edge; that is, we relabel the vertices of M as $\{2, 3, \ldots, 2n+2\}\backslash\{x\}$ for some $x \in \{2, 3, \ldots, 2n+2\}$ and add to M the block $\{1, x\}$; see Figure 2.

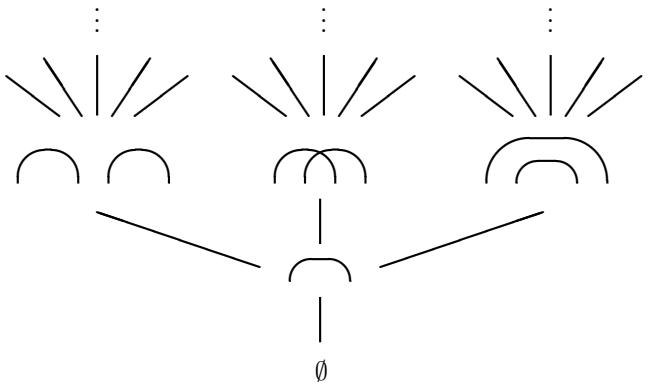


FIG. 2. *Tree of matchings* $\mathcal{T}$.

Each vertex $N \in \mathcal{M}(n)$ has $2n+1$ children and, if $n > 0$, is a child of a unique vertex $M \in \mathcal{M}(n-1)$. A *level* in a rooted tree is the set of vertices with the same distance from the root. In $\mathcal{T}$ the levels are the sets $\mathcal{M}(n)$. The *subtree* $\mathcal{T}(M)$ of $\mathcal{T}$ rooted in $M \in \mathcal{M}(n)$ is the rooted subtree on the vertex set $\mathcal{N} \subset \mathcal{M}$ consisting of M and all its descendants; that is, $\mathcal{N}$ contains M and all matchings obtained from M by successively adding each new first edge. In other words, $\mathcal{T}(M)$ consists of all $N \in \mathcal{M}$ in which the last n edges form a matching (order-isomorphic to) M. Clearly, $\mathcal{T}(\emptyset) = \mathcal{T}$. We denote the lth level of $\mathcal{T}(M)$ by $\mathcal{T}(M,l)$. For $M \in \mathcal{M}(n)$ we have $\mathcal{T}(M,0) = \{M\}$, and $\mathcal{T}(M,1)$ is the set of children of M in $\mathcal{T}$. Also, $|\mathcal{T}(M,l)| = (2n+1)(2n+3)\dots(2n+2l-1)$.

Besides the statistics $cr(M) \in \mathbb{N}_0$ and $ne(M) \in \mathbb{N}_0$ on $\mathcal{M}$ we consider the joint statistics $cn(M) = (cr(M), ne(M)) \in \mathbb{N}_0^2$ and $nc(M) = (ne(M), cr(M)) \in \mathbb{N}_0^2$. Two statistics s, u on two subsets $\mathcal{N}_1, \mathcal{N}_2 \subset \mathcal{M}$ *coincide (have the same distribution)* if $s(\mathcal{N}_1) = u(\mathcal{N}_2)$ as multisets, that is, if for every element e we have

$$|\{M \in \mathcal{N}_1 :\ s(M) = e\}| = |\{M \in \mathcal{N}_2 :\ u(M) = e\}|.$$

Notational convention. If $f :\ X \to Y$ is a mapping and $Z \subset X$, the symbol $f(Z)$ usually denotes the image $\mathrm{Im}(f|Z) = \{f(z) :\ z \in Z\}$. In this article we use $f(Z)$ to denote the multiset whose ground set is $\mathrm{Im}(f|Z)$ and in which each element $y = f(z)$, $z \in Z$, appears with the multiplicity $|f^{-1}(y) \cap Z|$. So in our $f(Z)$ each element y has the proper multiplicity in which it is attained as a value of f on Z.

Let $A = (A, +)$ be an abelian group and $\alpha, \beta \in A$ be its two elements. The most general statistic on matchings that we consider is $s_{\alpha,\beta} :\ \mathcal{M} \to A$ given by

$$s_{\alpha,\beta}(M) = cr(M)\alpha + ne(M)\beta.$$

Our main result is the next theorem.

THEOREM 1.1. *Let $M, N \in \mathcal{M}(n)$ be two (not necessarily distinct) matchings and, for $\alpha, \beta \in A$, $s_{\alpha,\beta}$ be the statistic $s_{\alpha,\beta}(M) = cr(M)\alpha + ne(M)\beta$.*

1. *If $s_{\alpha,\beta}(\mathcal{T}(M,l)) = s_{\alpha,\beta}(\mathcal{T}(N,l))$ for $l = 0$ and 1, then this identity holds for all $l \geq 0$.*
2. *If $s_{\alpha,\beta}(\mathcal{T}(M,l)) = s_{\beta,\alpha}(\mathcal{T}(N,l))$ for $l = 0$ and 1, then this identity holds for all $l \geq 0$.*

In words, for the statistic $s_{\alpha,\beta}$ to coincide level by level on the subtrees $\mathcal{T}(M)$ and $\mathcal{T}(N)$ it suffices if it coincides on the first two levels, and similarly for the pair of statistics $s_{\alpha,\beta}, s_{\beta,\alpha}$.

Specializing, we obtain identities for the statistics cr, ne, cn, and nc.

THEOREM 1.2. *Let $M, N \in \mathcal{M}(n)$ be two (not necessarily distinct) matchings and $s, t \in \{cr, ne\}$, $u, v \in \{cn, nc\}$ be statistics on matchings (we allow $s = t$ and $u = v$).*

1. *If $s(\mathcal{T}(M,l)) = t(\mathcal{T}(N,l))$ for $l = 0, 1$, then $s(\mathcal{T}(M,l)) = t(\mathcal{T}(N,l))$ for all $l \geq 0$.*
2. *If $u(\mathcal{T}(M,l)) = v(\mathcal{T}(N,l))$ for $l = 0, 1$, then $u(\mathcal{T}(M,l)) = v(\mathcal{T}(N,l))$ for all $l \geq 0$.*

Proof. 1. Let $A = (\mathbb{Z}, +)$. Setting $\alpha = 1, \beta = 0$ and $\alpha = 0, \beta = 1$ and using points 1 and 2 of Theorem 1.1, we obtain the identities for cr and ne.

2. Let $A = (\mathbb{Z}^2, +)$. Setting $\alpha = (1,0), \beta = (0,1)$ and $\alpha = (0,1), \beta = (1,0)$ and using 1 and 2 of Theorem 1.1, we obtain the identities for cn and nc. □

We illustrate the last theorem by four examples. We mentioned the first two already, the result of de Sainte-Catherine and the symmetry $cn = nc$.

COROLLARY 1.3. *For every $k \in \mathbb{N}_0$ and $n \in \mathbb{N}$ there are as many matchings on $[2n]$ with k crossings as those with k nestings.*

Proof. Set $M = N = \emptyset$ and $s = cr, t = ne$. The assumption of the theorem is satisfied because $cr(\emptyset) = ne(\emptyset) = 0$ and $cr(\mathcal{M}(1)) = ne(\mathcal{M}(1)) = \{0\}$. □

COROLLARY 1.4. *For every $k, l \in \mathbb{N}_0$ and $n \in \mathbb{N}$ there are as many matchings on $[2n]$ with k crossings and l nestings as there are with l crossings and k nestings; the joint statistic is symmetric.*

Proof. Set $M = N = \emptyset$ and $s = cn, t = nc$. The assumption of the theorem is satisfied because $cn(\emptyset) = nc(\emptyset) = (0,0)$ and $cn(\mathcal{M}(1)) = nc(\mathcal{M}(1)) = \{(0,0)\}$. □

COROLLARY 1.5. *For every $k \in \mathbb{N}_0$ and $n \in \mathbb{N}$ there are as many matchings on $[2n]$ which have k crossings and have the last two edges nested as there are which have k nestings and have the last two edges separated (neither crossing nor nested).*

Proof. Set $M = \{\{1,4\},\{2,3\}\}$, $N = \{\{1,2\},\{3,4\}\}$, $s = cr$, and $t = ne$. The assumption of the theorem is satisfied because $cr(M) = ne(N) = 0$ and the values of cr on the five children of M are $0, 0, 1, 1, 2$, which coincide with the values of ne on the five children of N. □

COROLLARY 1.6. *Let $M = \{\{1,2\},\{3,5\},\{4,6\}\}$ and $N = \{\{1,3\},\{2,4\},\{5,6\}\}$. For every $k, n \in \mathbb{N}$ there are as many matchings on $[2n]$ with k crossings in which the last three edges form a matching order-isomorphic to M as there are in which the last three edges form a matching order-isomorphic to N.*

Proof. Set the matchings M, N as given and $s = t = cr$. Then $cr(M) = cr(N) = 1$ and $cr(\mathcal{T}(M,1)) = cr(\mathcal{T}(N,1)) = \{1,1,1,2,2,2,3\}$. □

We call two matchings $M, N \in \mathcal{M}(n)$ *crossing-similar* and write $M \sim_{cr} N$ if $cr(\mathcal{T}(M,l)) = cr(\mathcal{T}(N,l))$ for all $l \geq 0$. Similarly we define the *nesting-similarity* $\sim_{ne}$. These two relations are equivalences and partition $\mathcal{M}(n)$ into equivalence classes. We use Theorem 1.2 to characterize these classes and to count them. In Theorems 3.3 and 3.5 we prove that the numbers of classes in $\mathcal{M}(n)/\sim_{cr}$ and $\mathcal{M}(n)/\sim_{ne}$ are, respectively,

$$2^{n-2}\left(\binom{n}{2}+2\right) \quad \text{and} \quad 2\cdot 4^{n-1} - \frac{3n-1}{2n+2}\binom{2n}{n}.$$

These two numbers differ; the latter is roughly a square of the former. On the first level of description of the enumerative complexity of crossings and nestings, that of

the numbers $cr(M)$ and $ne(M)$, symmetry reigns as shown in Corollaries 1.3 and 1.4. On the next level of description, that of the similarity classes, symmetry is broken because $|\mathcal{M}(n)/\sim_{ne}|$ is much bigger than $|\mathcal{M}(n)/\sim_{cr}|$; in Proposition 3.7 we show that $\sim_{ne}$ is a refinement of $\sim_{cr}$. From this point of view nestings are definitely more complicated than crossings; see also Theorem 4.4.

We prove Theorem 1.1 in section 2. The method we employ is induction on the number of edges. In section 3 we prove Theorems 3.3 and 3.5, enumerating the crossing-similarity and nesting-similarity classes. In section 4 we give further applications of the main theorem in Proposition 4.1, which characterizes the matchings M, N such that $cr(\mathcal{T}(M,l)) = ne(\mathcal{T}(N,l))$ for every $l \geq 0$; in Corollary 4.3, which deals with the statistic of pairs of separated edges; and in Theorem 4.4, which enumerates the classes of mod 2 crossing-similarity and mod 2 nesting-similarity. In section 5 we give some concluding comments.

2. The proof of Theorem 1.1. For a set X let $\mathcal{S}(X)$ be the set of all *finite multisets* with elements in X. By the sum

$$X_1 + X_2 + \cdots + X_r = \sum_1^r X_i$$

of the multisets $X_1, X_2, \ldots, X_r \in \mathcal{S}(X)$ we mean the union of their groundsets with multiplicities of the elements added. Any function $f\colon X \to \mathcal{S}(Y)$ naturally extends to

$$f:\ \mathcal{S}(X) \to \mathcal{S}(Y) \quad \text{by} \quad f(U) = \sum_{x\in U} f(x),$$

where the summand $f(x)$ appears with the multiplicity of x in U. Now if $Z \subset X$, we can understand the symbol $f(Z)$ in two ways—as the image of $f|Z$ or as the value of the extended f on Z. Due to our convention, both ways give the same result.

In this section, A shall denote an abelian group $(A,+)$, and A^* will be the set of finite sequences over A. We shall work with functions from A^* to $\mathcal{S}(A)$ or to $\mathcal{S}(A^*)$, which we will extend in the mentioned way, often without explicit notice, to functions defined on $\mathcal{S}(A^*)$. If $u = x_1x_2\ldots x_t \in A^*$ and $y \in A$, by $x_1x_2\ldots x_t + y$, we define the sequence $(x_1+y)(x_2+y)\ldots(x_t+y)$ obtained by adding y to each term of u.

Definition 2.1. *For $\alpha, \beta \in A$ and $i \in \mathbb{N}$ we define the mapping $R_{\alpha,\beta,i} : \bigcup_{l\geq i} A^l \to \bigcup_{l\geq i+2} A^l$ by*

$$R_{\alpha,\beta,i}(x_1x_2\ldots x_l) = x_i(x_1x_2\ldots x_i + x_i - x_1 + \alpha)(x_ix_{i+1}\ldots x_l + x_i - x_1 + \beta)$$

and the mapping $R_{\alpha,\beta}:\ A^* \to \mathcal{S}(A^*)$ by

$$R_{\alpha,\beta}(x_1x_2\ldots x_l) = \{R_{\alpha,\beta,i}(x_1x_2\ldots x_l):\ 1 \leq i \leq l\}.$$

Thus $R_{\alpha,\beta}(x_1x_2\ldots x_l)$ is an l-element multiset of sequences with length $l+2$.

Let $M \in \mathcal{M}(n)$ be a matching. The *gaps* of M are the first gap before 1, the $2n-1$ gaps between two consecutive elements of $[2n]$, and the last $(2n+1)$th gap after $2n$; M has $2n+1$ gaps. For $\alpha, \beta \in A$ we assign to every matching $N \in \mathcal{M}(n)$, $n \in \mathbb{N}_0$, a sequence $seq_{\alpha,\beta}(N) \in A^*$ with length $2n+1$. If $n = 0$, we set $seq_{\alpha,\beta}(\emptyset) = 0 = 0_A$. Let $n \geq 1$ and $(M,N) \in E(\mathcal{T})$, $M \in \mathcal{M}(n-1)$, which means that N is obtained from M by adding a new first edge $e = \{1,x\}$, where x is inserted in the ith gap of M for some $i \in [2n-1]$. We set

$$seq_{\alpha,\beta}(N) = R_{\alpha,\beta,i}(seq_{\alpha,\beta}(M)).$$

For example, if $M = \{\{1,3\},\{2,4\}\}$, then $seq_{\alpha,\beta}(M) = \alpha, 2\alpha, 3\alpha, 2\alpha+\beta, \alpha+2\beta$.

For $u \in A^*$ we denote by $R^l_{\alpha,\beta}(u) = R_{\alpha,\beta}(R_{\alpha,\beta}(\dots(R_{\alpha,\beta}(u))\dots))$ the lth iteration of the mapping $R_{\alpha,\beta}$ (which we extend to $\mathcal{S}(A^*)$). The next lemma is immediate from the definitions.

LEMMA 2.2. *For every $\alpha, \beta \in A$, $M \in \mathcal{M}$, and $l \in \mathbb{N}_0$ we have*

$$R^l_{\alpha,\beta}(seq_{\alpha,\beta}(M)) = seq_{\alpha,\beta}(\mathcal{T}(M,l)).$$

The next lemma relates the sequences $seq_{\alpha,\beta}(M)$ and the statistic $s_{\alpha,\beta}$ on $\mathcal{M}$.

LEMMA 2.3. *For every $\alpha, \beta \in A$ and $N \in \mathcal{M}(n)$ the first term of the sequence $seq_{\alpha,\beta}(N)$ equals $s_{\alpha,\beta}(N) = cr(N)\alpha + ne(N)\beta$.*

Proof. For $n = 0$ this holds. For $n \geq 1$ we proceed by induction on n. Suppose that $(M,N) \in E(\mathcal{T})$ and that N arises by adding new first edge $\{1,x\}$ to M, where x is inserted in the ith gap. Let $seq_{\alpha,\beta}(M) = a_1 a_2 \dots a_{2n-1}$.

We claim that in

$$a_j - a_1 = u_j\alpha + v_j\beta$$

the number u_j counts the edges in M covering the jth gap, and v_j counts the edges in M lying to the left of the jth gap.

Suppose that this claim holds. Then $cr(N) = cr(M)+u_i$ and $ne(N) = ne(M)+v_i$. Since $cr(M)\alpha + ne(M)\beta = a_1$ (by induction), the first term of $seq_{\alpha,\beta}(N)$ is $a_i = a_i - a_1 + a_1 = u_i\alpha + v_i\beta + cr(M)\alpha + ne(M)\beta = cr(N)\alpha + ne(N)\beta$, as we wanted to show.

It suffices to prove the claim by induction on n. For $n = 0$ it holds trivially. We assume that it holds for $seq_{\alpha,\beta}(M)$ and deduce it for $seq_{\alpha,\beta}(N)$; M, N, and i are as before. Let $seq_{\alpha,\beta}(N) = b_1 b_2 \dots b_{2n+1}$. We first describe the changes in gaps caused by the addition of $\{1,x\}$ to M. A new first gap appears; it is of course covered by no edge and has no edge to its left. For $1 \leq j \leq i$ the jth gap turns into the $(j+1)$th gap; these gaps get covered by one more edge and have the same numbers of edges to their left as before. The ith gap is split in two, which creates a new gap, the $(i+2)$th; it is covered by as many edges as the ith gap in M, but it has one more edge to its left. For $i+1 \leq j \leq 2n-1$ the jth gap turns into the $(j+2)$th one; these gaps are covered by as may edges as before, but they have one more edge to their left.

By the definition of $R_{\alpha,\beta,i}$, $b_1 = a_i$, $b_j = a_{j-1} + a_i - a_1 + \alpha$ for $2 \leq j \leq i+1$, and $b_j = a_{j-2} + a_i - a_1 + \beta$ for $i+2 \leq j \leq 2n+1$. Thus $b_1 - b_1 = 0$, $b_j - b_1 = a_{j-1} - a_1 + \alpha = (u_{j-1}+1)\alpha + v_{j-1}\beta$ for $2 \leq j \leq i+1$, and $b_j - b_1 = a_{j-2} - a_1 + \beta = u_{j-2}\alpha + (v_{j-2}+1)\beta$ for $i+2 \leq j \leq 2n+1$. This agrees with the described changes in gaps, and so the claim holds for $seq_{\alpha,\beta}(N)$. ∎

Let us denote by $f^0_0 : A^* \to A$ the function taking the first term of a sequence, and by $f^1_0 : A^* \to \mathcal{S}(A)$ the function creating the multiset of all terms of a sequence. By the definitions and Lemmas 2.2 and 2.3, if $seq_{\alpha,\beta}(M) = a_1 a_2 \dots a_{2n+1}$, then

$$s_{\alpha,\beta}(\mathcal{T}(M,1)) = f^0_0(R_{\alpha,\beta}(seq_{\alpha,\beta}(M))) = \{a_1, a_2, \dots, a_{2n+1}\} = f^1_0(seq_{\alpha,\beta}(M)).$$

For the induction argument we will need more complicated functions besides f^0_0 and f^1_0. For an integer $r \geq 0$ and $\gamma \in A$ we define the function $f^r_\gamma \colon A^* \to \mathcal{S}(A)$ by

$$f^r_\gamma(x_1 x_2 \dots x_l) = \{x_{a_1} + x_{a_2} + \dots + x_{a_r} - (r-1)x_1 + \gamma : \ 1 \leq a_1 \leq a_2 \leq \dots \leq a_r \leq l\}.$$

So $f^0_0(x_1 x_2 \dots x_l) = \{x_1\}$, and $f^1_\gamma(x_1 x_2 \dots x_l)$ is the multiset $\{x_1+\gamma, x_2+\gamma, \dots, x_l+\gamma\}$.

LEMMA 2.4. *Let $X, Y \in \mathcal{S}(A^*)$ (possibly $X = Y$) be two multisets such that $f^r_\gamma(X) = f^r_\gamma(Y)$ for every $r \geq 0$ and $\gamma \in A$. Then for every mapping $R_{\alpha,\beta}$ of Definition 2.1 we have*

1. $f^r_\gamma(R_{\alpha,\beta}(X)) = f^r_\gamma(R_{\alpha,\beta}(Y))$,
2. $f^r_\gamma(R_{\alpha,\beta}(X)) = f^r_\gamma(R_{\beta,\alpha}(Y))$

for every $r \geq 0$ and $\gamma \in A$.

Proof. We prove only the second identity with $R_{\alpha,\beta}$ and $R_{\beta,\alpha}$; the proof of the first identity is similar and easier. We proceed by induction on r. The case $r = 0$ is clear since $f^0_\gamma(R_{\alpha,\beta}(X)) = f^1_\gamma(X)$ for every $X \in \mathcal{S}(A^*)$ and $\gamma \in A$. We assume that $r \geq 1$ and that for every s, $0 \leq s < r$, and $\gamma \in A$ we have $f^s_\gamma(R_{\alpha,\beta}(X)) = f^s_\gamma(R_{\beta,\alpha}(Y))$. We consider only the function f^r_0; the proof for general γ is similar.

We split the multisets $U = f^r_0(R_{\alpha,\beta}(X))$ and $V = f^r_0(R_{\beta,\alpha}(Y))$, which arise by summation, into several contributions and show that, after rearranging, the corresponding contributions to U and V are equal. U is the multiset of elements $y_{a_1} + y_{a_2} + \cdots + y_{a_r} - (r-1)y_1$, where the sequence $y_1 y_2 \ldots y_l$ runs through $R_{\alpha,\beta}(X)$ and the indices a_i run through the r-tuples $1 \leq a_1 \leq a_2 \leq \cdots \leq a_r \leq l$, and similarly for V. The first contribution C is defined by the condition $a_1 = 1$. C contributes to U the elements

$$y_1 + y_{a_2} + \cdots + y_{a_r} - (r-1)y_1 = y_{a_2} + \cdots + y_{a_r} - (r-2)y_1,$$

where $y_1 y_2 \ldots y_l$ runs through $R_{\alpha,\beta}(X)$ and the indices a_i run through the $(r-1)$-tuples $1 \leq a_2 \leq a_3 \leq \cdots \leq a_r \leq l$. Thus C contributes $f^{r-1}_0(R_{\alpha,\beta}(X))$. To V it contributes $f^{r-1}_0(R_{\beta,\alpha}(Y))$. Hence C contributes equally to U and V because $f^{r-1}_0(R_{\alpha,\beta}(X)) = f^{r-1}_0(R_{\beta,\alpha}(Y))$ by the inductive assumption.

Each $v = y_1 y_2 \ldots y_l \in R_{\alpha,\beta}(X)$ is in $R_{\alpha,\beta}(u)$ for some $u = x_1 x_2 \ldots x_{l-2} \in X$ and (by the definition of $R_{\alpha,\beta}$) consists of three segments: It starts with a term x_i of u, then it comes $x_1 \ldots x_i$ termwise incremented by $x_i - x_1 + \alpha$, and the third segment of v is $x_i \ldots x_{l-2}$ termwise incremented by $x_i - x_1 + \beta$; similarly for $v \in R_{\beta,\alpha}(Y)$. We split the rest of U and V (in which $a_1 > 1$, i.e., every y_{a_i} lies in the second or in the third segment) into $r+1$ disjoint contributions C_t according to the number t, $0 \leq t \leq r$, of the y_{a_i}'s lying in the second segment. By the definition of $R_{\alpha,\beta}$, C_t contributes to U the elements

$$\begin{aligned} &x_{b_1} + \cdots + x_{b_r} + t(x_i - x_1 + \alpha) + (r-t)(x_i - x_1 + \beta) - (r-1)x_i \\ &= x_{b_1} + \cdots + x_{b_r} + x_i - rx_1 + t\alpha + (r-t)\beta, \end{aligned}$$

where $u = x_1 x_2 \ldots x_{l-2}$ runs through X, the indices b_j run through the r-tuples satisfying $1 \leq b_1 \leq \cdots \leq b_t \leq i \leq b_{t+1} \leq \cdots \leq b_r \leq l-2$, and i runs through $1 \leq i \leq l-2$. (The length $l-2$ depends on u.) Effectively the indices b_j and i run through all weakly increasing $(r+1)$-tuples of numbers from $[l-2]$. Thus C_t contributes to U the elements $f^{r+1}_\gamma(X)$, where $\gamma = t\alpha + (r-t)\beta$. By the definition of $R_{\beta,\alpha}$, C_t contributes to V the elements $f^{r+1}_{\gamma'}(Y)$, where $\gamma' = t\beta + (r-t)\alpha$. So C_t contributes to U and V in general differently, but (by the assumption on X and Y) the contributions of C_t to U and C_{r-t} to V are equal. By symmetry, $\sum_0^r C_i$ contributes the same amount to U and V. Since U and V are covered by equal and disjoint contributions C and $\sum_0^r C_i$, we conclude that $U = V$, i.e., $f^r_0(R_{\alpha,\beta}(X)) = f^r_0(R_{\beta,\alpha}(Y))$.

The proof of point 1 is similar and easier, because now C_t contributes equally to $U = f^r_0(R_{\alpha,\beta}(X))$ and $V = f^r_0(R_{\alpha,\beta}(Y))$. □

Next we show that for the equality of all functions f^r_γ on two one-element sets it in fact suffices that f^0_0 and f^1_0 are equal. We prove it in two lemmas. Let $g^r \colon A^* \to \mathcal{S}(A)$

be defined by

$$g^r(x_1x_2\dots x_l)=\{x_{a_1}+x_{a_2}+\cdots+x_{a_r}:\ 1\le a_1\le a_2\le\cdots\le a_r\le l\}.$$

LEMMA 2.5. *If $u,v\in A^*$ are such that $g^1(u)=g^1(v)$, then $g^r(u)=g^r(v)$ for all $r\ge 1$.*

Proof. Let $g^1(u)=g^1(v)$ and $r\in\mathbb{N}$. For $\bar a=(a_1,\dots,a_s)\in A^s$ we denote by $(\bar a)$ the multiset $\{a_1,\dots,a_s\}$, and if $\bar n=(n_1,\dots,n_s)\in\mathbb{N}^s$, then $\bar n\cdot\bar a=n_1a_1+\cdots+n_sa_s\in A$. For $s\in\mathbb{N}$, $X\in\mathcal{S}(A)$, and $u=x_1x_2\dots x_l\in A^*$ we define

$$S(s,X,u)=\{\bar x=(x_{a_1},\dots,x_{a_s}):\ 1\le a_1<a_2<\cdots<a_s\le l,(\bar x)=X\}.$$

For $r,s\in\mathbb{N}$ we define

$$N(r,s)=\{(n_1,\dots,n_s)\in\mathbb{N}^s:\ n_1+\cdots+n_s=r\}.$$

Now we can rewrite $g^r(u)$ and $g^r(v)$ as

$$g^r(u)=\{\bar n\cdot\bar a:\ s\in[r],X\in\mathcal{S}(A),\bar n\in N(r,s),\bar a\in S(s,X,u)\},$$
$$g^r(v)=\{\bar n\cdot\bar a:\ s\in[r],X\in\mathcal{S}(A),\bar n\in N(r,s),\bar a\in S(s,X,v)\}.$$

We claim that (i) for every fixed $s\in[r]$ and $X\in\mathcal{S}(A)$ the multiset

$$m(\bar a)=\{\bar n\cdot\bar a:\ \bar n\in N(r,s)\}$$

is the same for all $\bar a\in A^s$ with $(\bar a)=X$, and that (ii) for every fixed $s\in[r]$ and $X\in\mathcal{S}(A)$ we have $|S(s,X,u)|=|S(s,X,v)|$. This will prove that $g^r(u)=g^r(v)$.

To show (i), we take $\bar a,\bar b\in A^s$ with $(\bar a)=(\bar b)=X$. Then $\bar a$ can be obtained from $\bar b$ by permuting coordinates: $\bar a=\pi(\bar b)$ for some $\pi\in\mathcal{S}_s$, and $\bar n\cdot\bar b=\pi(\bar n)\cdot\bar a$. If $\bar n$ runs through $N(r,s)$, so does $\pi(\bar n)$. Hence $m(\bar a)=m(\bar b)$. To show (ii), we suppose that X consists of the distinct elements $x_1,\dots,x_t$ with multiplicities $n_1,\dots,n_t$, where $n_1+\cdots+n_t=s$ (else $|S(s,X,u)|=|S(s,X,v)|=0$), and denote by $m_a(u)$ and $m_a(v)$ the numbers of occurrences of $a\in A$ in u and v. Because $m_a(u)=m_a(v)$ for every $a\in A$, we have indeed

$$|S(s,X,u)|=\prod_{i=1}^{t}\binom{m_{x_i}(u)}{n_i}=\prod_{i=1}^{t}\binom{m_{x_i}(v)}{n_i}=|S(s,X,v)|.\qquad\square$$

LEMMA 2.6. *If $X,Y\in\mathcal{S}(A^*)$ are one-element sets such that $f_0^0(X)=f_0^0(Y)$ and $f_0^1(X)=f_0^1(Y)$, then $f_\gamma^r(X)=f_\gamma^r(Y)$ for every $r\ge 0$ and $\gamma\in A$.*

Proof. We need to prove that if $u,v\in A^*$ are two sequences beginning with the same term and having equal numbers of occurrences of each $a\in A$, then $f_\gamma^r(u)=f_\gamma^r(v)$ for every $r\ge 0$ and $\gamma\in A$. It suffices to consider functions f_0^r; the proof with general γ is similar. Since u and v start with the same term, by the definition of f_0^r it suffices to prove that $g^r(u)=g^r(v)$ for every $r\ge 1$. This is true by Lemma 2.5. $\square$

Proof. Proof of Theorem 1.1. We prove only claim 2; the proof of 1 is very similar and easier. Let $s_{\alpha,\beta}(\mathcal{T}(M,l))=s_{\beta,\alpha}(\mathcal{T}(N,l))$ for $l=0,1$. By Lemma 2.3 and the following remark, this means that $f_0^0(seq_{\alpha,\beta}(M))=f_0^0(seq_{\beta,\alpha}(N))$ and $f_0^1(seq_{\alpha,\beta}(M))=f_0^1(seq_{\beta,\alpha}(N))$. By Lemma 2.6, $f_\gamma^r(seq_{\alpha,\beta}(M))=f_\gamma^r(seq_{\beta,\alpha}(N))$ for every $r\in\mathbb{N}_0$ and $\gamma\in A$. By repeated application of Lemma 2.4.2 we get

$$f_\gamma^r(R_{\alpha,\beta}^l(seq_{\alpha,\beta}(M)))=f_\gamma^r(R_{\beta,\alpha}^l(seq_{\beta,\alpha}(N)))$$

for every $l, r \in \mathbb{N}_0$ and $\gamma \in A$. In particular,

$$f_0^0(R^l_{\alpha,\beta}(seq_{\alpha,\beta}(M))) = f_0^0(R^l_{\beta,\alpha}(seq_{\beta,\alpha}(N))).$$

But by Lemma 2.2 we have

$$R^l_{\alpha,\beta}(seq_{\alpha,\beta}(M)) = seq_{\alpha,\beta}(\mathcal{T}(M,l)) \text{ and } R^l_{\beta,\alpha}(seq_{\beta,\alpha}(N)) = seq_{\beta,\alpha}(\mathcal{T}(N,l)).$$

Thus, by Lemma 2.3,

$$s_{\alpha,\beta}(\mathcal{T}(M,l)) = s_{\beta,\alpha}(\mathcal{T}(N,l))$$

for every $l \geq 0$, which we wanted to prove. □

We give a formulation of Theorem 1.1 in terms of the sequences $seq_{\alpha,\beta}(M)$.

THEOREM 2.7. *Let $M, N \in \mathcal{M}(n)$ be two (not necessarily distinct) matchings and $\alpha, \beta \in A$ be two elements of the abelian group.*

1. *We have $s_{\alpha,\beta}(\mathcal{T}(M,l)) = s_{\alpha,\beta}(\mathcal{T}(N,l))$ for all $l \geq 0$ iff $s_{\alpha,\beta}(M) = s_{\alpha,\beta}(N)$ and the sequences $seq_{\alpha,\beta}(M)$ and $seq_{\alpha,\beta}(N)$ are equal as multisets (when order is neglected).*
2. *We have $s_{\alpha,\beta}(\mathcal{T}(M,l)) = s_{\beta,\alpha}(\mathcal{T}(N,l))$ for all $l \geq 0$ iff $s_{\alpha,\beta}(M) = s_{\beta,\alpha}(N)$ and the sequences $seq_{\alpha,\beta}(M)$ and $seq_{\beta,\alpha}(N)$ are equal as multisets.*

3. The numbers of similarity classes. In this section, we determine the cardinalities $|\mathcal{M}(n)/\sim_{cr}|$ and $|\mathcal{M}(n)/\sim_{ne}|$. Let $A = (\mathbb{Z}, +)$. For $M \in \mathcal{M}$ we define its *crossing sequence* $crs(M)$ by $crs(M) = seq_{1,0}(M) - a_1$, where a_1 is the first term of $seq_{1,0}(M)$, and its *nesting sequence* $nes(M)$ by $nes(M) = seq_{0,1}(M) - b_1$, where b_1 is the first term of $seq_{0,1}(M)$. Recall that (by the proof of Lemma 2.3) the ith term of $crs(M)$ is the number of edges in M covering the ith gap, and the ith term of $nes(M)$ is the number of edges lying to the left of the ith gap. For example, $M = \{\{1,4\},\{2,5\},\{3,6\}\}$ has $crs(M) = (0,1,2,3,2,1,0)$ and $nes(M) = (0,0,0,0,1,2,3)$. By Theorems 1.2 and 2.7, $M \sim_{cr} N \iff cr(M) = cr(N)$ and $f_0^1(crs(M)) = f_0^1(crs(N))$; that is, M and N are crossing-similar iff they have the same numbers of crossings and their crossing sequences are equal as multisets; an analogous result holds for the nesting-similarity.

Let $e = \{a,d\}, f = \{b,c\} \in M$, $1 \leq a < b < c < d \leq 2n$, be a nesting in $M \in \mathcal{M}(n)$. We define its *width* as $\min(b-a, d-c)$. We define the width of a crossing in the same way, only $\{a,d\}$ is replaced with $\{a,c\}$ and $\{b,c\}$ with $\{b,d\}$. Suppose that the nesting e, f has the minimum width among all nestings in M and that its width is realized by $b-a$. Switching the first vertices of the edges e and f, we obtain another matching N. If the width of e, f is realized by $d-c$, we switch the second vertices of e and f. This transformation $M \rightsquigarrow N$ is called the *n-c transformation*. In the same way, by switching the first or the second vertices of the edges in a crossing with minimum width, we define the *c-n transformation*.

LEMMA 3.1. *Let $M, N \in \mathcal{M}(n)$, where N is obtained from M by the n-c (c-n) transformation. Then N has the same sets of first and second vertices of the edges as M and $ne(N) = ne(M) - 1, cr(N) = cr(M) + 1$ ($ne(N) = ne(M) + 1, cr(N) = cr(M) - 1$).*

Proof. The first claim about N is obvious. Let $e = \{a,c\}, f = \{b,d\} \in M$, $1 \leq a < b < c < d \leq 2n$, be a crossing in M with the minimum width which is equal to $b-a$ (if it is equal to $d-c$, the argument is similar). The c-n transformation replaces e by $e' = \{b,c\}$ and f by $f' = \{a,d\}$. Because of the minimality of the width, every edge of M that has one endpoint between a and b must have the other endpoint

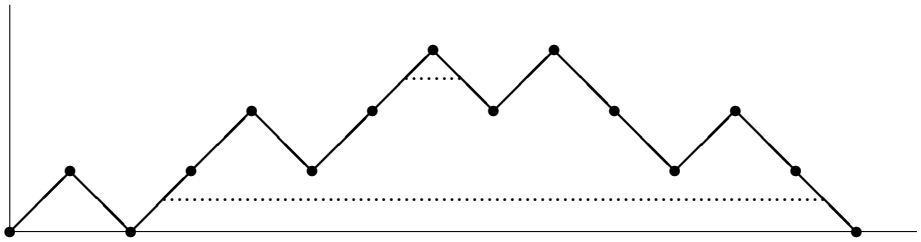


FIG. 3. *Dyck path with semilength seven and two marked tunnels.*

between a and b as well. It follows that e' crosses the same edges distinct from f as e does, and similarly for f' and f. The edge e' is covered by the same edges different from f' as e, and similarly for f' and f. The edge e' does not cover the edges lying between a and b which were covered by e, but these are now covered by f' and were not covered by f. If we do not consider the pairs e, f and e', f', M and N have the same numbers of crossings and the same numbers of nestings. Since e, f is a crossing and e', f' is a nesting, in total N has one less crossing and one more nesting than M. The argument for the n-c transformation is similar and is left to the reader. □

We use *Dyck paths* to encode $crs(M)$ and $nes(M)$. Recall that a Dyck path D with semilength $n \in \mathbb{N}$ is a lattice path $D = (d_0, d_1, \ldots, d_{2n})$, where $d_i \in \mathbb{Z}^2$, from $d_0 = (0,0)$ to $d_{2n} = (2n, 0)$ that makes n up-steps $d_i - d_{i-1} = (1,1)$, n down-steps $d_i - d_{i-1} = (1,-1)$, and never gets below the x axis (so, in fact, $d_i \in \mathbb{N}_0^2$). We denote the set of Dyck paths with semilength n by $\mathcal{D}(n)$; $|\mathcal{D}(0)| = 1$. We think of $D \in \mathcal{D}(n)$ also as a broken line in the plane that connects $(0,0)$ with $(2n,0)$ and consists of $2n$ straight segments $s_i = d_i d_{i+1}$; see Figure 3. A *tunnel* in D is a horizontal segment t that has altitude $n + \frac{1}{2}$ for some $n \in \mathbb{N}_0$, lies below D, and intersects D only in its endpoints. Each $D \in \mathcal{D}(n)$ has exactly n tunnels. Note that projections of two tunnels on the x axis either are disjoint or are in inclusion (as in the example in Figure 3). If the latter happens, we say that the tunnel with larger projection *covers* the other tunnel.

Deleting from $D \in \mathcal{D}(n)$, $n \geq 1$, the first up-step and the first down-step at which D visits the x axis again, we obtain, shifting appropriately the resulting two parts of D, a unique decomposition of D into a pair of Dyck paths E, F, where $E \in \mathcal{D}(m)$ for $0 \leq m < n$ and $F \in \mathcal{D}(n-1-m)$. This *decomposition of Dyck paths* can be used for inductive proofs of their properties.

We associate with every Dyck path $D = (d_0, d_1, \ldots, d_{2n})$ its *sequence of altitudes* $als(D) = (d_0^y, d_1^y, \ldots, d_{2n}^y) \in \mathbb{N}_0^{2n+1}$, where $d_i = (d_i^x, d_i^y)$, and its *profile* $pr(D) = (a_1, a_2, \ldots, a_m) \in \mathbb{N}^m$, where m is the maximum term of $als(D)$ and a_i is half of the number of segments s_i of D that lie in the horizontal strip $i - 1 \leq y \leq i$. It follows that $a_1 + a_2 + \cdots + a_m = n$ and $pr(D)$ is a composition of n. It follows easily by induction on m that for every composition $a = (a_1, a_2, \ldots, a_m)$ of n there is a $D \in \mathcal{D}(n)$ with $pr(D) = a$. For example, the Dyck path in Figure 3 has $als(D) = (0,1,0,1,2,1,2,3,2,3,2,1,2,1,0)$ and $pr(D) = (2,3,2)$.

There is a natural surjective mapping $F : \mathcal{M}(n) \to \mathcal{D}(n)$ defined as follows. We take the diagram of $M \in \mathcal{M}(n)$ and travel the baseline l from $-\infty$ to ∞. Simultaneously we construct, step by step, a lattice path D. We start D at $(0,0)$, and when we encounter on l the first (second) vertex of an edge, we make in D an up-step (down-step). In the end we get a Dyck path $D \in \mathcal{D}(n)$ and set $F(M) = D$. Using

the decomposition of Dyck paths and induction, it is easy to prove that F is surjective. Clearly, the preimages $F^{-1}(D)$ consist exactly of the matchings sharing the same sets of first and second vertices. Another important property of F is that for every $D \in \mathcal{D}(n)$ there is exactly one *noncrossing* (i.e., with $cr(M) = 0$) $M \in F^{-1}(D)$, namely the M whose edges correspond in the obvious way to the tunnels in D. This follows by the decomposition of Dyck paths.

LEMMA 3.2. *Let $n \in \mathbb{N}$ and $F : \mathcal{M}(n) \to \mathcal{D}(n)$ be the aforementioned mapping.*

1. *For every $M \in \mathcal{M}(n)$ we have $crs(M) = als(F(M))$.*

2. *For every $M, N \in \mathcal{M}(n)$ we have $f_0^1(crs(M)) = f_0^1(crs(N))$ iff $pr(F(M)) = pr(F(N))$.*

3. *For every composition $a = (a_1, a_2, \ldots, a_m)$ of n and every $i \in \mathbb{N}_0$, $0 \leq i \leq \sum_{i=1}^m (i-1)a_i$, there is an $M \in \mathcal{M}(n)$ such that $pr(F(M)) = a$ and $cr(M) = i$. There exist no a and no M such that $pr(F(M)) = a$ and $cr(M) > \sum_{i=1}^m (i-1)a_i$.*

Proof. 1. This is clear from the definitions of $crs(M)$ and $als(D)$.

2. Using 1, we look at $f_0^1(als(D))$, where $D = F(M)$. Let $pr(D) = (a_1, a_2, \ldots, a_m)$ and r_i be the multiplicity of $i \in \mathbb{N}_0$ in $als(D)$. It is clear that $r_0 = a_1 + 1$ and $r_m = a_m$. We claim that for $0 < i < m$ we have $r_i = a_i + a_{i+1}$. In the strip $i - 1 \leq y \leq i$ we have $v = 2a_i$ segments $s_1, s_2, \ldots, s_v$ of D, and in the strip $i \leq y \leq i + 1$ we have $w = 2a_{i+1}$ segments $t_1, t_2, \ldots, t_w$. The occurrences of i in $als(D)$ are due to the upper endpoints of the s_j's and due to the lower endpoints of the t_j's. But for each s_j its upper endpoint coincides with the upper endpoint of s_{j-1} or with that of s_{j+1} or with the lower endpoint of some t_k, and similarly for the lower endpoints of the t_j's. So i appears $(v + w)/2 = a_i + a_{i+1}$ times. On the other hand, $a_i = r_{i-1} - r_{i-2} + \cdots + (-1)^i r_1 + (-1)^{i+1}(r_0 - 1)$ for every $1 \leq i \leq m$. Therefore the r_i's are completely determined by the composition $pr(D)$ and vice versa.

3. Let a composition $a = (a_1, a_2, \ldots, a_m)$ of n be given. We take an arbitrary $D \in \mathcal{D}(n)$ with $pr(D) = a$. It follows by the decomposition of Dyck paths and induction that the sum

$$S(a) = \sum_{i=1}^{m} (i-1)a_i$$

counts the ordered pairs t_1, t_2 of distinct tunnels in D where t_1 covers t_2. For the unique noncrossing $M \in F^{-1}(D)$ we have $ne(M) = S(a)$ because nestings in M are in 1-1 correspondence with the pairs of tunnels, one of them covering the other. So $cr(M) = 0$, $ne(M) = S(a)$, $F(M) = D$, $pr(F(M)) = a$. For any given $i \in \{0, 1, \ldots, S(a)\}$, using repeatedly the n-c transformation of Lemma 3.1, we transform M into N such that $cr(N) = i, ne(N) = S(a) - i$, and $F(N) = F(M) = D$. Now suppose that there is an $M \in F^{-1}(D)$ with $cr(M) = c > S(a)$. Using the c-n transformation of Lemma 3.1, we transform it into $N \in F^{-1}(D)$ with $cr(N) = 0$ and $ne(N) = ne(M) + c > S(a)$. This contradicts the unicity of the noncrossing matching in $F^{-1}(D)$. □

THEOREM 3.3. *For $n \in \mathbb{N}$ the set $\mathcal{M}(n)/\sim_{cr}$ of crossing-similarity classes has*

$$2^{n-2}\left(\binom{n}{2} + 2\right)$$

elements.

Proof. By the previous lemma, $|\mathcal{M}(n)/\sim_{cr}|$ equals

$$\sum_a (1 + a_2 + 2a_3 + \cdots + (m-1)a_m) = 2^{n-1} + \sum_a (a_2 + 2a_3 + \cdots + (m-1)a_m),$$

where we sum over all compositions $a_1+a_2+\cdots+a_m=n$, which are 2^{n-1} in number. The last sum is the coefficient of x^n in the expansion of

$$\left(\frac{d}{dy}\sum_{m\geq 0}\frac{x}{1-x}\cdot\frac{xy}{1-xy}\cdot\frac{xy^2}{1-xy^2}\cdot\cdots\cdot\frac{xy^m}{1-xy^m}\right)\Bigg|_{y=1}.$$

Differentiating the product in the summand by the Leibniz rule and using that

$$\left(\frac{d}{dy}\frac{xy^i}{1-xy^i}\right)\Bigg|_{y=1}=\frac{ix}{(1-x)^2},$$

we obtain that the expansion equals

$$\frac{1}{1-x}\sum_{m\geq 0}\binom{m+1}{2}\left(\frac{x}{1-x}\right)^{m+1}.$$

Using the binomial expansion $(1-z)^{-r}=\sum_{n\geq 0}\binom{r+n-1}{n}z^n$, we simplify this to

$$\frac{x^2}{(1-2x)^3}=\sum_{n\geq 0}\binom{n+2}{2}2^n x^{n+2},$$

and the result follows. □

The values of $|\mathcal{M}(n)/\!\sim_{cr}|$ form the sequence $(1,3,10,32,96,276,\dots)$. Subtracting 2^{n-1}, we get the sequence $(0,1,6,24,80,240,\dots)$, which counts crossing-similarity classes in $\mathcal{M}(n)$ for matchings with at least one crossing. This sequence is entry A001788 of [14] and counts, for example, also 4-cycles in the $(n+1)$-dimensional hypercube.

The situation for nestings is simpler, and the number of similarity classes is bigger because nesting sequences are nondecreasing and therefore $f_0^1(nes(M))=f_0^1(nes(N))$ iff $nes(M)=nes(N)$. By Theorems 1.2 and 2.7, $M\sim_{ne}N$ iff M and N have the same numbers of nestings and the same nesting sequences. For $D\in\mathcal{D}(n)$ we define $ne(D)$ to be the number of ordered pairs t_1,t_2 of distinct tunnels in D such that t_1 covers t_2. The *down sequence* $dos(D)$ of $D=(d_0,d_1,\dots,d_{2n})$ is $(v_0,v_1,\dots,v_{2n})$, where v_i is the number of down-steps $d_j-d_{j-1}=(1,-1)$ for $1\leq j\leq i$. For example, for the Dyck path in Figure 3 we have $ne(D)=7$ and $dos(D)=(0,0,1,1,1,2,2,2,3,3,4,5,5,6,7)$.

LEMMA 3.4. *Let $n\in\mathbb{N}$ and $F:\ \mathcal{M}(n)\to\mathcal{D}(n)$ be the mapping defined above.*

1. *For every $M\in\mathcal{M}(n)$ we have $nes(M)=dos(F(M))$. There is a bijection between the sets $\{nes(M):\ M\in\mathcal{M}(n)\}$ and $\mathcal{D}(n)$.*

2. *For every Dyck path $D\in\mathcal{D}(n)$ and every $i\in\mathbb{N}_0$, $0\leq i\leq ne(D)$, there is an $M\in F^{-1}(D)$ such that $ne(M)=i$. There is no $M\in F^{-1}(D)$ with $ne(M)>ne(D)$.*

Proof. 1. The first claim follows at once from the definitions. It is also clear that $dos(D)$ is uniquely determined by D and vice versa.

2. We know from the proof of Lemma 3.2.3 that $ne(D)=ne(M)$ for the unique noncrossing $M\in F^{-1}(D)$. Now we argue as in the proof of Lemma 3.2.3. □

THEOREM 3.5. *For $n\in\mathbb{N}$ the set $\mathcal{M}(n)/\!\sim_{ne}$ of nesting-similarity classes has*

$$2\cdot 4^{n-1}-\frac{3n-1}{2n+2}\binom{2n}{n}$$

elements.

Proof. By the previous lemma,

$$|\mathcal{M}(n)/\sim_{ne}| = \sum_{D\in\mathcal{D}(n)} (1+ne(D)) = |\mathcal{D}(n)| + \sum_{D\in\mathcal{D}(n)} ne(D).$$

We claim that this number is equal to the coefficient of x^n in the expansion of the expression

$$C + x^2(2xC'+C)^2C,$$

where $C = C(x) = \sum_{n\geq 0} |\mathcal{D}(n)|x^n = 1 + x + 2x^2 + 5x^3 + \cdots$. It is well known that $C = (1-\sqrt{1-4x})/2x = \sum_{n\geq 0} \frac{1}{n+1}\binom{2n}{n}x^n$. Using the relations $xC^2 - C + 1 = 0$ and $2xCC' + C^2 = C'$, we simplify the expression to

$$2C(x) + \frac{1/2}{1-4x} - \frac{3/2}{\sqrt{1-4x}}.$$

Using the expansion of $C(x)$, geometric series, and $(1-4x)^{-1/2} = \sum_{n\geq 0}\binom{2n}{n}x^n$, we obtain the formula.

To establish the claim, recall that $\sum_{D\in\mathcal{D}(n)} ne(D)$ counts the triples (D, t_1, t_2), where $D \in \mathcal{D}(n)$ and t_1, t_2 are two distinct tunnels in D such that t_1 covers t_2. Let the segments of D supporting t_i be r_i (up-step) and s_i (down-step). Let the lower endpoints of the segments r_i (s_i) be a_i (b_i), and their upper endpoints be a_i' (b_i'), $i = 1, 2$. The deletion of the interiors of the segments r_1, s_1, r_2, and s_2 splits D into five lattice paths $L_1, \ldots, L_5$, where L_1 starts at $(0,0)$ and ends in a_1, L_2 starts at a_1' and ends at a_2, L_3 starts at a_2' and ends at b_2', L_4 starts at b_2 and ends at b_1', and L_5 starts at b_1 and ends at $(2n, 0)$. Each L_i is nonempty but may be just a single lattice point. The concatenation L_1L_5, where L_5 is appropriately shifted so that a_1 and b_1 are identified in one distinguished point, is a Dyck path, and similarly for L_2L_4 with a_2 and b_2 identified and distinguished. L_3 is a Dyck path by itself (after an appropriate shift). We see that the triples (D, t_1, t_2) in question are in a 1-1 correspondence with the triples (E_1, E_2, E_3), where $E_i \in \mathcal{D}(n_i)$, $n_i \in \mathbb{N}_0$, are such that $n_1 + n_2 + n_3 = n - 2$, and moreover E_1 and E_2 have one distinguished lattice point (out of $2n_1 + 1$, respectively $2n_2 + 1$, points). It follows that the number of the triples (E_1, E_2, E_3) is the coefficient of x^{n-2} in $(2xC' + C)^2C$. □

The values of $|\mathcal{M}(n)/\sim_{ne}|$ form the sequence $(1, 3, 12, 51, 218, 926, \ldots)$. Subtracting the Catalan numbers $C_n = |\mathcal{D}(n)|$, we get the sequence $(0, 1, 7, 37, 176, 794, \ldots)$, which counts nesting-similarity classes in $\mathcal{M}(n)$ for matchings with at least one nesting. This sequence is entry A006419 of [14] and appears in Welsh and Lehman [18, Table VIb] in enumeration of planar maps. We summarize this identity in the next proposition.

PROPOSITION 3.6. *For $n = 1, 2, \ldots$ the formula*

$$2\cdot 4^{n-1} - \frac{3n+1}{2n+2}\binom{2n}{n}$$

counts the following objects:

1. *the triples (D, t_1, t_2), where D is a Dyck path with semilength n and t_1, t_2 are two distinct tunnels in D such that t_1 covers t_2,*
2. *the nesting-similarity classes in $\{M \in \mathcal{M}(n) : ne(M) > 0\}/\sim_{ne}$,*
3. *the vertex-rooted planar maps with two vertices and n faces, which are edge 2-connected and may have loops and multiple edges. See Figure 4 for the case $n = 3$.*

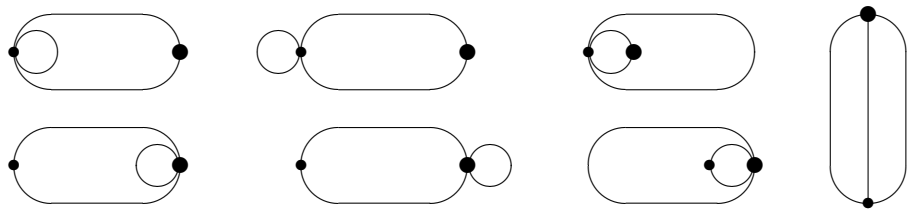


FIG. 4. *Rooted and edge* 2*-connected planar maps with two vertices and three faces.*

Proof. Claims 1 and 2 follow from the proof of Theorem 3.5, and 3 follows by checking the formulas in [18]. Alternatively, it is not too hard to establish bijection between the triples in 1 and the maps in 3. □

The present author proved in [7, Theorem 3.1] that the number of the triples (T, v_1, v_2), where T is a rooted plane tree with n vertices and v_1, v_2 are two (not necessarily distinct) vertices of T such that v_1 lies on the path joining the root of T and v_2, equals

$$\frac{4^{n-1} + \binom{2n-2}{n-1}}{2}.$$

It is straightforward to relate Dyck paths and rooted plane trees and to derive the formula of Theorem 3.5 from this.

Not only are there more equivalence classes in $\mathcal{M}(n)/\sim_{ne}$ than in $\mathcal{M}(n)/\sim_{cr}$, but also $\sim_{ne}$ is in fact a refinement of $\sim_{cr}$. To show this we consider another statistic $ca(M)$, where $M \in \mathcal{M}(n)$, counting *camels*, which are (unordered) pairs of edges in M order-isomorphic to $\{\{1,2\},\{3,4\}\}$. Note that $cr(M) + ne(M) + ca(M) = \binom{n}{2}$ and that $ca(M)$ is uniquely determined by the Dyck path $F(M) = D = (d_0, d_1, \ldots, d_{2n})$ because it is the number of pairs (d_i, d_j), where $i < j$, d_i is a down-step and d_j is an up-step.

PROPOSITION 3.7. *For every* n *the partition* $\mathcal{M}(n)/\sim_{ne}$ *refines the partition* $\mathcal{M}(n)/\sim_{cr}$.

Proof. Suppose that $M, N \in \mathcal{M}(n)$ satisfy $M \sim_{ne} N$. Then $nes(M) = nes(N)$ and $ne(M) = ne(N)$. By Lemma 3.4 we know that $F(M) = F(N) = D$ for a common Dyck path D. By Lemma 3.2, $crs(M) = crs(N)$; hence $f_0^1(crs(M)) = f_0^1(crs(N))$. Also, by the above remark on statistic $ca(\cdot)$, $cr(M) = \binom{n}{2} - ne(M) - ca(M) = \binom{n}{2} - ne(N) - ca(N) = cr(N)$. Thus $M \sim_{cr} N$. □

4. Further applications. Corollary 1.5 presents two matchings M and N such that the distribution of cr on the levels of $\mathcal{T}(M)$ equals the distribution of ne on the levels of $\mathcal{T}(N)$. We show that there are no other substantially different examples.

PROPOSITION 4.1. *Let* $M, N \in \mathcal{M}(n)$ *be two matchings. We have* $cr(\mathcal{T}(M,l)) = ne(\mathcal{T}(N,l))$ *for every* $l \geq 0$ *iff* $M = M_n = \{\{1,2n\},\{2,2n-1\},\ldots,\{n,n+1\}\}$ *and* $N = N_n = \{\{1,2\},\{3,4\},\ldots,\{2n-1,2n\}\}$.

Proof. The *if* part is clear by Theorem 1.2: $cr(M_n) = ne(N_n) = 0$ and

$$cr(\mathcal{T}(M_n,1)) = ne(\mathcal{T}(N_n,1)) = \{0,0,1,1,2,2,\ldots,n-1,n-1,n\}$$

because

$$crs(M_n) = (0, 1, 2, \ldots, n-1, n, n-1, \ldots, 2, 1, 0),$$
$$nes(N_n) = (0, 0, 1, 1, 2, 2, \ldots, n-1, n-1, n).$$

To show the *only if* part, we prove that the only matchings $M, N \in \mathcal{M}(n)$ satisfying $cr(M) = ne(N)$ and $f_0^1(crs(M)) = f_0^1(nes(N))$ are M_n and N_n. Since for every $N \in \mathcal{M}(n)$ the sequence $nes(N)$ ends with n, we must have n in $crs(M)$, which means that the middle gap of M must be covered by all edges. Thus all first vertices of the edges in M must precede all second vertices, and $crs(M) = (0, 1, 2, \ldots, n-1, n, n-1, \ldots, 2, 1, 0)$.

Thus $f_0^1(nes(N)) = \{0, 0, 1, 1, 2, 2, \ldots, n-1, n-1, n\}$, which forces $N = N_n$. Thus $cr(M) = ne(N) = ne(N_n) = 0$, which forces $M = M_n$. □

Therefore we have no other examples of equidistribution of cr and ne on the levels of $\mathcal{T}(M)$ than $M = \emptyset$ and $M = \{\{1, 2\}\}$, because $M_n = N_n$ only for $n = 0, 1$. We call the matchings $M \in \mathcal{M}(n)$ encountered in the proof in which all edges cover the middle gap, equivalently which have $f_0^1(crs(M)) = \{0, 0, 1, 1, 2, 2, \ldots, n-1, n-1, n\}$, *permutational matchings;* they are in 1-1 correspondence with the permutations of $[n]$ and are $n!$ in number.

Because $|\mathcal{M}(n)| = (2n-1)!! = n^n(2/\mathrm{e} + o(1))^n$ and the numbers of crossing-similarity and nesting-similarity classes are only exponential, we have very many examples as in Corollary 1.6 when cr (or ne) has equal distributions on the levels of $\mathcal{T}(M)$ and $\mathcal{T}(N)$ for $M \neq N$. The next corollary follows from the asymptotics of the numbers of similarity classes given in Theorems 3.3 and 3.5.

COROLLARY 4.2. *Every set of matchings $X \subset \mathcal{M}(n)$ contains a subset of $|X|/(2+o(1))^n$ mutually crossing-similar matchings and a subset of $|X|/(4+o(1))^n$ mutually nesting-similar matchings.*

An explicit example of a big similarity class is provided by permutational matchings in $\mathcal{M}(n)$. They all share the same crossing sequence $(0, 1, 2, \ldots, n-1, n, n-1, \ldots, 2, 1, 0)$ and the same nesting sequence $(0, 0, \ldots, 0, 1, 2, \ldots, n-1, n)$. Hence at least

$$\frac{n!}{\binom{n}{2}+1} = n^n(1/\mathrm{e} + o(1))^n$$

of them are mutually crossing-similar, and at least so many of them are mutually nesting-similar.

Recall that $ca(M)$ is the number of pairs of separated edges in M. This statistic behaves on the levels of the subtrees of $\mathcal{T}$ in the same way as cr and ne do.

COROLLARY 4.3. *Let $M, N \in \mathcal{M}(n)$ be two matchings such that ca has the same distribution on the first two levels of the subtrees $\mathcal{T}(M)$ and $\mathcal{T}(N)$. Then ca has the same distribution on all levels.*

Proof. For $M \in \mathcal{M}(n)$ we have $ca(M) = \binom{n}{2} - (cr(M) + ne(M))$. Thus this result follows by Theorem 1.1.1 if we set $A = (\mathbb{Z}, +)$ and $\alpha = \beta = 1$. □

Note that while the number of $M \in \mathcal{M}(n)$ with $cr(M) = 0$ (or with $ne(M) = 0$) is the Catalan number $\frac{1}{n+1}\binom{2n}{n}$, the number of $M \in \mathcal{M}(n)$ with $ca(M) = 0$ is much bigger, namely $n!$ (these are exactly permutational matchings).

It is possible to investigate the general similarity relation $\sim_{A,\alpha,\beta}$ on $\mathcal{M}(n)$ defined, for an abelian group $A = (A, +)$ and two its elements $\alpha, \beta \in A$, by $M \sim_{A,\alpha,\beta} N$ iff $s_{\alpha,\beta}(\mathcal{T}(M, l)) = s_{\alpha,\beta}(\mathcal{T}(N, l))$ for every $l \geq 0$. We consider here only the case $A = (\mathbb{Z}_2, +)$ and define the statistics $cr_2(M), ne_2(M) \in \{0, 1\}$ as parity of the numbers

$cr(M), ne(M)$. We define the sequences $crs_2(M)$ and $nes_2(M)$ of M by reducing $crs(M)$ and $nes(M)$ modulo 2. For two matchings $M, N \in \mathcal{M}(n)$ we define $M \sim_{cr,2} N$ iff $cr_2(\mathcal{T}(M,l)) = cr_2(\mathcal{T}(N,l))$ for every $l \geq 0$, and similarly for $M \sim_{ne,2} N$. By Theorems 1.1 and 2.7, $M \sim_{cr,2} N$ iff $cr_2(M) = cr_2(N)$ and $crs_2(M)$ and $crs_2(N)$ are equal as multisets after forgetting the order of terms, and similarly for $\sim_{ne,2}$. (Now $nes_2(M)$ is not nondecreasing, and we may have $f_0^1(nes_2(M)) = f_0^1(nes_2(N))$ for $nes_2(M) \neq nes_2(N)$.) We determine the numbers of equivalence classes for $\sim_{cr,2}$ and $\sim_{ne,2}$.

THEOREM 4.4. *We have* $|\mathcal{M}(1)/\sim_{cr,2}| = 1$ *and* $|\mathcal{M}(n)/\sim_{cr,2}| = 2$ *for* $n \geq 2$. *The two classes of mod* 2 *crossing-similarity have* $((2n-1)!!+1)/2$ *and* $((2n-1)!!-1)/2$ *elements. We have* $|\mathcal{M}(1)/\sim_{ne,2}| = 1$, $|\mathcal{M}(2)/\sim_{ne,2}| = 3$, *and* $|\mathcal{M}(n)/\sim_{ne,2}| = 2n$ *for* $n \geq 3$.

Proof. By the definition of $crs(M)$, $crs_2(M) = (0,1,0,1,0,\ldots,1,0)$ for every matching M. Thus the classes of mod 2 crossing-similarity are determined only by $cr_2(M)$ and, for $n \geq 2$, we have two of them. The fact that

$$|\{M \in \mathcal{M}(n) : \ cr_2(M) = 0\}| - |\{M \in \mathcal{M}(n) : \ cr_2(M) = 1\}| = 1$$

for every $n \geq 1$ was proved by Riordan [12] by generating functions; a simple proof by involution was given by Klazar [8].

To handle nestings modulo 2, recall that $nes(M) = dos(D)$, where $D = F(M)$ and that nesting sequences of the matchings $M \in \mathcal{M}(n)$ are in bijection with the Dyck paths $D \in \mathcal{D}(n)$ (Lemma 3.4). We claim that the n Dyck paths

$$D_1 = udu^{n-1}d^{n-1}, D_2 = u^2du^{n-2}d^{n-1}, \ldots, D_{n-1} = u^{n-1}dud^{n-1}, D_n = u^nd^n$$

(u is the up-step and d is the down-step) realize all possible numbers of 1's and 0's in the sequences $\{dos_2(D) : \ D \in \mathcal{D}(n)\}$ and hence in the sequences $\{nes_2(M) : M \in \mathcal{M}(n)\}$. The number of 1's (0's) in $dos_2(D_i)$, $i = 1,2,\ldots,n$, is $n + \lceil n/2 \rceil - i$ $(1 + i + \lfloor n/2 \rfloor)$. It suffices to show that no $dos_2(D)$ has fewer than $\lceil n/2 \rceil$ 1's and fewer than $2 + \lfloor n/2 \rfloor$ 0's. In every D each of the n down-steps contributes to $dos_2(D)$ exactly one 1 (by one of its endpoints), and each of these 1's may belong to at most two down- steps. Thus we must have at least $\lceil n/2 \rceil$ 1's. The argument for 0's is similar, but now the 0 contributed by the first down-step is never shared (with the next down-steps) and there is one more 0 contributed by the first up-step. So we have at least $1 + 1 + \lfloor n/2 \rfloor$ 0's. Thus, for every $n \geq 1$, $|\{f_0^1(nes_2(M)) : \ M \in \mathcal{M}(n)\}| = n$. If $n \geq 3$, for each D_i there are $M, M' \in F^{-1}(D_i)$ with $ne(M') = ne(M) - 1$. (We take for M the noncrossing matching in $F^{-1}(D_i)$, which has at least one nesting, and apply the n-c transformation.) Thus, for $n \geq 3$, there are $2n$ classes of mod 2 nesting-similarity. The cases $n = 1, 2$ are easy to treat separately. □

5. Concluding remarks. An interesting result for crossings and nestings of higher order was obtained by Chen et al. in [2], where it is proved that for every $k, l, n \in \mathbb{N}$ the number of matchings in $\mathcal{M}(n)$ with no k-crossing and no l-nesting is the same as the number of matchings with no k-nesting and no l-crossing (the same result is obtained in [2] for set partitions); here k-crossing is a k-tuple of pairwise crossing edges and similarly for k-nesting. Another generalization of crossings and nestings was investigated by Jelínek [4] who is interested in numbers of matchings $M \in \mathcal{M}(n)$ such that M does not contain a fixed permutational matching $N \in \mathcal{M}(3)$ as an ordered submatching. For further recent results on crossings and nestings, see Bousquet-Mélou and Xin [1], Corteel [3], Jonsson [5], Kasraoui and Zeng [6], Krattenthaler [10], de Mier [11], and Stanley [16].

It may be interesting to try to extend the results and methods of the present article to crossings and nestings of higher order. Another research direction may be to apply our method to other structures besides matchings. Finally, one may try to go to higher levels of the description of the enumerative complexity of crossings and nestings—denoting by $G : \mathcal{M} \to \mathcal{M}/\sim_{cr}$ the mapping sending M to its equivalence class, when is it the case that $G(\mathcal{T}(M,l)) = G(\mathcal{T}(N,l))$ for every $l \geq 0$; and similarly for $\sim_{ne}$.

Acknowledgments. I am grateful to Marc Noy for his hospitality during my two visits in UPC Barcelona in 2004 and for stimulating discussions. I thank two anonymous referees for careful reading of my manuscript; one of them suggested Proposition 3.7.

REFERENCES

[1] M. Bousquet-Mélou and G. Xin, *On partitions avoiding* 3*-crossings*, Sém. Lothar. Combin., 54 (2005/06), paper B54e.

[2] W. Y. C. Chen, E. Y. P. Deng, R. R. X. Du, R. P. Stanley, and C. H. Yan, *Crossings and nestings of matchings and partitions*, Trans. Amer. Math. Soc., to appear.

[3] S. Corteel, *Crossings and alignments of permutations*, Adv. Appl. Math., to appear.

[4] V. Jelínek, *Dyck paths and pattern-avoiding matchings*, European J. Combin., 28 (2007), pp. 202–213.

[5] J. Jonsson, *Generalized triangulations and diagonal-free subsets of stack polyominoes*, J. Combin. Theory, Ser. A, 112 (2005), pp. 117–142.

[6] A. Kasraoui and J. Zeng, *Distribution of crossings, nestings, and alignments of two edges in matchings and partitions*, Electron. J. Combin., 13 (2006), paper R33.

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

[8] M. Klazar, *Counting even and odd partitions*, Amer. Math. Monthly, 110 (2003), pp. 527–532.

[9] M. Klazar and M. Noy, *On the Symmetry of Joint Distribution of Crossings and Nestings in Matchings*, manuscript.

[10] Ch. Krattenthaler, *Growth diagrams, and increasing, and decreasing chains in fillings of Ferrers shapes*, Adv. Appl. Math., 37 (2006), pp. 404–431.

[11] A. de Mier, *k-noncrossing and k-nonnesting graphs and fillings of Ferrers diagrams*, online preprint http://arxiv.org/abs/math.CO/0602195.

[12] J. Riordan, *The distribution of crossings of chords joining pairs of* 2*n points on a circle*, Math. Comput., 29 (1975), pp. 215–222.

[13] M. de Sainte-Catherine, *Couplages et Pfaffiens en Combinatoire, Physique et Informatique*, Ph.D. thesis, University of Bordeaux I, Talence, France 1983.

[14] N. J. A. Sloane, *The On-Line Encyclopedia of Integer Sequences*, published online at http://www.research.att.com/~njas/sequences, AT&T Labs, 2005.

[15] R. P. Stanley, *Enumerative Combinatorics*, Vol. 2, Cambridge University Press, Cambridge, UK, 1999.

[16] R. P. Stanley, *Increasing and decreasing subsequences of permutations and their variants*, in Proceedings of the ICM'06, to appear.

[17] J. Touchard, *Sur un problème de configurations et sur les fractions continues*, Canadian J. Math., 4 (1952), pp. 2–25.

[18] T. R. S. Welsh and A. B. Lehman, *Counting rooted maps by genus.* III: *Nonseparable maps*, J. Combin. Theory, Ser. B, 18 (1975), pp. 222–259.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 977–998

ON COST MATRICES WITH TWO AND THREE DISTINCT VALUES OF HAMILTONIAN PATHS AND CYCLES*

SANTOSH N. KABADI† AND ABRAHAM P. PUNNEN‡

Abstract. A polynomial time testable characterization of cost matrices associated with a complete digraph on n nodes such that all the Hamiltonian cycles (tours) have the same cost is well known. Tarasov [*U.S.S.R. Comput. Maths. Math. Phys.*, 21 (1981), pp. 167–174.] obtained a characterization of cost matrices where tour costs take two distinct values. We provide a simple alternative characterization of such cost matrices, which can be tested in $O(n^2)$ time. We also provide analogous results where tours are replaced by Hamiltonian paths. When the cost matrix is skew-symmetric, we provide polynomial time testable characterizations such that the tour costs take three distinct values. Corresponding results for the case of Hamiltonian paths are also given. Using these results, special instances of the asymmetric traveling salesman problem (ATSP) are identified that are solvable in polynomial time and that have improved constant factor approximation schemes. In particular, we observe that the 3/2 performance guarantee of the Christofides algorithm extends to all metric Hamiltonian symmetric matrices. Further, we identify special classes of ATSP for which polynomial ϵ-approximation algorithms are available for $\epsilon \in \{3/2, 4/3, 4\tau, \frac{3\tau^2}{2}, \frac{4+\delta}{3}\}$, where $\tau > 1/2$ and $\delta \geq 0$ are constants.

Key words. combinatorial optimization, graphs, Hamiltonian cycles, Hamiltonian paths, approximation algorithms

AMS subject classifications. 90C27, 90C57, 90C35

DOI. 10.1137/S0895480104445332

1. Introduction.

Let G be a directed graph with node set $V(G) = \{1, 2, \ldots, n\}$ and arc set $E(G)$. For each arc $(i, j) \in E(G)$ a cost c_{ij} is prescribed. Let C be the cost matrix associated with G such that the (i, j)th element of C is c_{ij} if $(i, j) \in E(G)$ and ∞ if $(i, j) \notin E(G)$. If G is an undirected graph, then the matrix C is symmetric. For any Hamiltonian cycle (tour) H of G the cost of H corresponding to C is given by $C(H) = \sum_{(i,j)\in H} c_{ij}$. Cost matrix C is said to be a *k distinct cost tour matrix (DTC(k) matrix)* if and only if the number of distinct values of costs of tours in G is exactly k. In particular, C is a DTC(1) matrix if and only if every tour in G has the same cost.

It is easy to see that, for any digraph G and any arbitrary mappings $a, b : V(G) \to \Re$, if a cost matrix C associated with G is defined as

$$c_{ij} = a_i + b_j \quad \text{for all } (i, j) \in E(G), \tag{1.1}$$

then all tours in G have the same value. Interestingly, Gabovich [6] proved that this condition totally characterizes the class of DTC(1) matrices when G is a complete digraph. He attributes the result for the undirected case to [20] and [24]. For future reference we summarize the characterization of DTC(1) matrices for a complete digraph below.

*Received by the editors July 24, 2004; accepted for publication (in revised form) June 5, 2006; published electronically December 11, 2006. This work was supported by NSERC research grants.
http://www.siam.org/journals/sidma/20-4/44533.html

†Faculty of Business Administration, University of New Brunswick, Fredericton, NB, Canada E3B 5A3 (kabadi@unb.ca).

‡Department of Mathematics, Simon Fraser University, 14th Floor Central City Tower, 13450 102nd Ave., Surrey, BC, Canada V3T 5X3 (punnen@unbsj.ca).

THEOREM 1.1 (after [6]). *If G is a complete digraph with associated cost matrix C, then the following statements are equivalent:*

(1) *All tours in G have the same cost β with respect to C.*
(2) *There exist $\{a_i, b_i : i = 1, 2, \ldots, n\}$ such that $c_{ij} = a_i + b_j$ for all $(i, j) \in E(G)$ and $\sum_{i=1}^{n}(a_i + b_i) = \beta$.*

(Note that by a complete digraph we mean a digraph in which each two vertices are joined by two oppositely oriented arcs.) It is easy to see that if C is symmetric, we can choose $a_i = b_i$ for all i, and if C is skew-symmetric we can choose $a_i = -b_i$. Independent proofs of Theorem 1.1 are reported in [5, 7, 12]. Independent proofs for the undirected case are also reported in [16, 23]. Condition (2) in Theorem 1.1 can be tested in $O(n^2)$ time.

A cost matrix C associated with G is said to be a *k distinct cost Hamiltonian path matrix (DPC(k) matrix)* if and only if the number of distinct values of costs of Hamiltonian paths in G is exactly k. Thus if C is a DPC(1) matrix, then every Hamiltonian path in G has the same cost. The structure of a DPC(1) matrix associated with a complete digraph is much simpler than that of a DTC(1) matrix.

LEMMA 1.2 (see [12]). *A cost matrix C associated with a complete digraph G is a DPC(1) matrix if and only if it is a constant matrix (i.e., all the nondiagonal elements of C are identical).*

In view of Theorem 1.1 and Lemma 1.2, a natural question is to identify the structure of DTC(k) and DPC(k) matrices associated with complete digraphs for $k \geq 2$. Tarasov [26] provided an elegant characterization of DTC(2) matrices. His proof is inductive in nature with a base case for $n = 5$, the validity of which is established by complete enumeration using a computer. In this paper we provide an alternative characterization of DTC(2) matrices associated with complete digraphs. Our characterization is simple and can be tested in $O(n^2)$ time. Further, our proof does not use complete enumeration. We also provide a complete characterization of DPC(2) matrices associated with complete digraphs and establish a relationship between DTC(2) and DPC(2) matrices.

Tarasov [26] also obtained a characterization of cost matrices associated with the assignment problem (minimum weight bipartite matching problem) [18] with three distinct objective function values. However, no corresponding results are known for Hamiltonian cycles. We give a complete characterization of DTC(3) and DPC(3) skew-symmetric cost matrices associated with complete digraphs. It is also shown that there are no skew-symmetric DPC(2) matrices of size greater than 3.

Given a digraph G and an associated cost matrix C, the well known *traveling salesman problem (TSP)* is to find a tour H in G such that its cost $C(H)$ is as small as possible. If the graph G is undirected or, equivalently, the matrix C is symmetric, the resulting TSP is called a symmetric traveling salesman problem (STSP). Thus STSP is a special case of TSP. To emphasize the fact we are considering a directed graph, we some times refer to TSP as an asymmetric traveling salesman problem (ATSP).

The general TSP is well known to be NP-hard [19]. It is NP-hard even to find an ϵ-approximate solution for this problem for any constant $\epsilon > 0$ [19]. However, there are several special cases of TSP that are solvable in polynomial time [11], and several special cases that are solvable using polynomial time approximation schemes (PTAS) or ϵ-approximation algorithms for constant $\epsilon > 0$ in polynomial time. The books edited by Lawler et al. [19] and by Gutin and Punnen [8] provide the state of the art on the topic. Polynomial solvability of an instance of TSP and existence of a polynomial ϵ-approximation algorithm for it depend on the properties of the

associated cost matrix. Note that some polynomially solvable classes of TSP are characterized in terms of the structure of the underlying (di)graph. However, any such characterization can be rephrased in terms of properties of the cost matrix. For a comprehensive study on polynomially solvable cases of TSP, we refer to [11].

The simplest of all the polynomially solvable cases of TSP is the constant TSP, where the cost matrix is a DTC(1) matrix corresponding to a complete digraph, since in this case every tour is optimal. We show that the characterizations of DTC(2) and DTC(3) cost matrices discussed above identify new polynomially solvable cases of the TSP. In addition, we show that these also help us in solving some instances of ATSP as STSP. It is well known that any ATSP on n nodes can be formulated as an STSP on $2n$ nodes. However, we show that for special ATSP, our results make it possible to find equivalent STSP of the same size.

For any tour $H = (u_1, u_2, \ldots, u_n, u_1)$, we denote its *reversal* $(u_n, u_{n-1}, \ldots, u_1, u_n)$ by H^*. We say that a digraph G is symmetrical if and only if for any arc (i, j) in $E(G)$ the arc (j, i) is also in $E(G)$; and we say that G is Hamiltonian symmetrical if and only if for any tour H in G its reversal H^* is also in G. Thus, every symmetrical digraph is Hamiltonian symmetrical. A cost matrix C associated with a Hamiltonian symmetrical digraph G is said to be *Hamiltonian symmetrical* [9] if and only if $C(H) = C(H^*)$ for every tour H in G. Recently, Halskau [9] showed that the cost matrix C associated with a complete digraph G is Hamiltonian symmetrical if and only if there exist mappings $a, b : V(G) \to \Re$ such that

$$c_{ij} = a_i + b_j + d_{ij}, \tag{1.2}$$

where $D = (d_{ij})$ is a symmetric matrix of the same size as C. He also showed that this condition can be tested easily in $O(n^2)$ time. From (1.2) it can be seen that

$$C(H) = D(H) + \alpha \quad \text{for all tours } H \in G. \tag{1.3}$$

Thus, as observed by Halskau [9], solving ATSP with cost matrix C is equivalent to solving STSP with cost matrix D. If $\alpha \neq 0$, the transformation given by (1.2) (and hence (1.3)) does not preserve ϵ-optimality. We construct a simple transformation from ATSP to STSP that characterizes Hamiltonian symmetrical matrices associated with symmetrical digraphs. This transformation preserves ϵ-optimality as well as τ-triangular inequality [1] and range inequality [17]. Thus known performance guarantees of various approximation algorithms for the STSP extend to the more general class of Hamiltonian symmetric TSPs. In particular, for metric ATSP with a Hamiltonian symmetric cost matrix, we observe that a 3/2-approximate solution can be obtained using the Christofides algorithm [19]. It is an open question to find a polynomial ϵ-approximation algorithm for the metric ATSP for any constant ϵ [14]. The best known performance ratio for a polynomial approximation algorithm for such an instance of ATSP is $4/3 \log_3 n \approx 0.842 \log_2 n$ [13]. When C is Hamiltonian symmetrical and satisfies the weak τ-triangle inequality (see section 4), we observe that the performance ratio becomes $\min\{4\tau, \frac{3}{2}\tau^2\}$ for constant $\tau \geq 1$, and when C satisfies the weak range inequality (see section 4) the performance ratio becomes $\frac{4+\delta}{3}$ for constant $\delta \geq 0$.

Kabadi and Punnen [12] introduced a special class of graphs called SC-Hamiltonian graphs, that includes complete (di)graphs, complete bipartite (di)graphs, etc., for which Theorem 1.1 continues to hold. We observe that the transformation from ATSP to STSP discussed above extends easily to all symmetrical SC-Hamiltonian

digraphs. We also consider relationships between ATSP and STSP when G is a complete digraph and $|C(H) - C(H^*)| = 0$ or α for some positive number α. Using the notion of DTC(2) and skew-symmetric DTC(3) matrices, we identify special classes of ATSP on n nodes for which an optimal solution can be obtained by solving $n/2$ symmetric TSPs on n nodes.

The major contributions of the paper are summarized below:

- A simple alternative characterization of DTC(2) matrices associated with complete digraphs is given along with a simple proof. This further enhances the knowledge of structural properties of this class of matrices.
- Complete polynomially testable characterizations of DPC(2) matrices, skew-symmetric DPC(3) matrices, and skew-symmetric DTC(3) matrices associated with complete digraphs are given.
- New special cases of ATSP are identified that can be solved in polynomial time. Further, using known constant factor approximation algorithms for STSP, special classes of ATSP are identified for which polynomial ϵ-approximation algorithms are available for $\epsilon \in \{3/2, 4/3, 4\tau, 3/2\tau^2, \frac{4+\delta}{3}\}$, where $\tau > 1/2$ and $\delta \geq 0$ are constants.

It would be interesting to find simple proofs for our characterization of skew-symmetric DTC(3) and DPC(3) matrices associated with complete digraphs. Further, it is a challenging problem to identify polynomially testable characterizations of DTC(k) and DPC(k) matrices (if they exist) for $k \geq 3$. (It may be noted that for $k = 3$, we give such characterizations only for skew-symmetric matrices.)

The paper is organized as follows. In section 2 we discuss our characterization of DTC(2) and DPC(2) matrices associated with complete digraphs. Section 3 deals with our characterization of DTC(3) and DPC(3) skew-symmetric matrices associated with complete digraphs. Special ATSPs are considered in section 4, and concluding remarks are given in section 5.

We conclude this section by introducing some notation and tour construction schemes. For any (di)graph G, its vertex set is denoted by V(G), and its edge set is denoted by $E(G)$. Unless otherwise specified, throughout the rest of this paper we assume that G is a complete digraph, and thus all the nondiagonal elements of the associated cost matrix C are finite. For any cost matrix C associated with G and any subgraph H of G, we denote its cost $\sum_{ij \in H} c_{ij}$ by $C(H)$. By elements of a cost matrix we mean its nondiagonal elements with finite values. Let $H = (u_1, u_2, \ldots, u_n, u_1)$ be a tour in G. We describe below four schemes for constructing new tours from H. These constructions are used extensively in the subsequent sections. (See Figure 1 for illustrations of these schemes.)

Scheme 1 (*ordered* 3-*exchange*): Let (i, j) be a given arc not in H. Without loss of generality, we assume that $i = u_1$. Let $j = u_r$ for some $2 < r < n$. Choose some integer ℓ such that $r \leq \ell \leq n$. Then the new tour obtained by this construction is given by $H' = (u_1, u_r, u_{r+1}, \ldots, u_\ell, u_2, u_3, \ldots, u_{r-1}, u_{\ell+1}, u_{\ell+2}, u_n, u_1)$.

Scheme 2 (*arc reversal*): Let (i, j) be an arc in H. Without loss of generality, we assume that $i = u_1$ and $j = u_2$. Then the new tour obtained by this construction is given by $\bar{H} = (u_2, u_1, u_3, u_4, \ldots, u_{n-1}, u_n, u_2)$.

Scheme 3 (*inverse arc reversal*): Let (i, j) be an arc in H. Without loss of generality, we assume that $i = u_1$ and $j = u_2$. Then the new tour obtained by this construction is given by $\hat{H} = (u_1, u_2, u_n, u_{n-1}, \ldots, u_4, u_3, u_1)$.

Scheme 4 (*path reversal*): Consider a path $(u_r, u_{r+1}, \ldots, u_s)$ in H for some $1 \leq r, s, \leq n$. (Here, indices of u are taken modulo n.) Then the new tour obtained by this construction is given by $\tilde{H} = (u_s, u_{s-1}, \ldots, u_r, u_{s+1}, \ldots u_{r-1}, u_s)$. Note that in

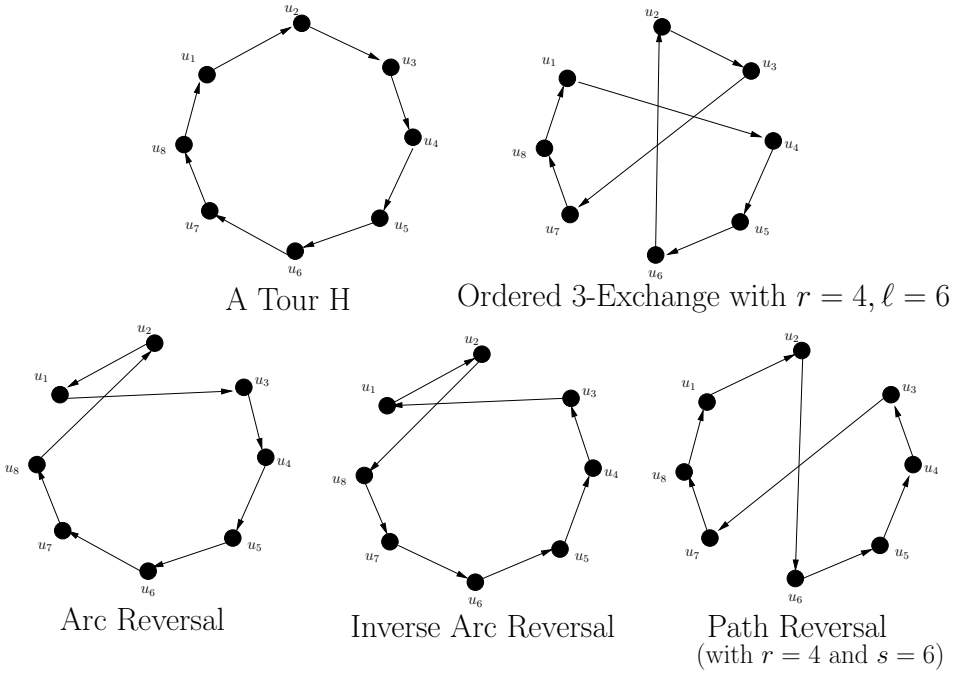


FIG. 1. *Examples of Schemes* 1 *to* 4.

Scheme 4, if we choose $u_r = i$ and $s = r + 1$, we get Scheme 2, and if we set $u_r = j$ and $s = r - 1$, we get Scheme 3.

2. DTC(2) and DPC(2) matrices for complete digraphs. In this section we discuss our characterizations of DTC(2) and DPC(2) matrices associated with complete digraphs. Thus, *throughout this section, all nondiagonal elements of cost matrices considered are finite.*

For any cost matrix C and any $r \in \{1, 2, \ldots, n\}$, define $a_r = 0$, $b_r = 0$, $a_i = c_{ir}$ for $i \neq r$, and $b_i = c_{ri}$ for $i \neq r$. Define the matrix $\hat{C} = (\hat{c}_{ij})_{n\times n}$, as $\hat{c}_{ij} = c_{ij} - a_i - b_j$. We call $\hat{C}$ the *r-reduced matrix of C*. For the r-reduced matrix $\hat{C}$, it can be seen that $\hat{c}_{rj} = \hat{c}_{jr} = 0$ for $j \in \{1, 2, \ldots, n\}$, $j \neq r$. We call the $(n-1) \times (n-1)$ submatrix C^0 of $\hat{C}$ obtained by deleting its rth row and rth column the *r-reduced submatrix of C*. For convenience, we refer to the n-reduced matrix and the n-reduced submatrix of C as simply *the reduced matrix* and *the reduced submatrix* of C, respectively.

To motivate the study of DTC(2) and DPC(2) matrices, let us start with an example. Consider the cost matrix C given below and its reduced matrix $\hat{C}$:

$$C = \begin{bmatrix} \infty & 6 & 4 & 5 & 9 & 1 & 5 \\ 5 & \infty & 5 & 6 & 10 & 2 & 6 \\ 3 & 5 & \infty & 4 & 8 & 0 & 3 \\ 6 & 8 & 6 & \infty & 11 & 3 & 7 \\ 11 & 13 & 11 & 12 & \infty & 8 & 12 \\ 8 & 10 & 8 & 9 & 13 & \infty & 9 \\ 5 & 7 & 4 & 6 & 10 & 2 & \infty \end{bmatrix}, \quad \hat{C} = \begin{bmatrix} \infty & -6 & -5 & -6 & -6 & -6 & 0 \\ -6 & \infty & -5 & -6 & -6 & -6 & 0 \\ -5 & -5 & \infty & -5 & -5 & -5 & 0 \\ -6 & -6 & -5 & \infty & -6 & -6 & 0 \\ -6 & -6 & -5 & -6 & \infty & -6 & 0 \\ -6 & -6 & -5 & -6 & -6 & \infty & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \infty \end{bmatrix}.$$

It is difficult to see that the tour costs corresponding to the cost matrix C above have only two distinct values. Now look at the reduced matrix $\hat{C}$. Its reduced submatrix

contains only two distinct elements -5 and -6. This is no accident, and we will show that the reduced submatrix of a DTC(2) matrix always contains exactly two distinct elements. However, this is not enough to characterize DTC(2) matrices. Obviously, if these two distinct elements are distributed arbitrarily, the corresponding tour costs could have a large number distinct values. So the question is, in what patterns are the two distinct elements distributed within the reduced submatrix, and the answer to this yields the required characterization. The following facts are easy to verify.

Fact 1. For any $r \in \{1, 2, \ldots, n\}$, the r-reduced matrix of a skew-symmetric $n \times n$ matrix is a skew-symmetric matrix, and the r-reduced matrix of a symmetric $n \times n$ matrix is a symmetric matrix.

Fact 2. For any positive integer k, if we subtract a constant from all entries in any row or column of a DTC(k) matrix, then the resulting matrix is also a DTC(k) matrix.

Fact 3. For any positive integer k and any $r \in \{1, 2, \ldots, n\}$, an $n \times n$ matrix C is a DTC(k) matrix if and only if its r-reduced matrix $\hat{C}$ is a DTC(k) matrix.

We now give a characterization of DTC(k) matrices in terms of DPC(k) matrices.

THEOREM 2.1. *If an $n \times n$ cost matrix C is a DTC(k) matrix, then its r-reduced submatrix is a DPC(k) matrix for every $r \in \{1, 2, \ldots, n\}$. Conversely, if there exists an $r \in \{1, 2, \ldots, n\}$ such that the r-reduced submatrix of C is a DPC(k) matrix, then C is a DTC(k) matrix.*

Proof. For any $n \times n$ cost matrix C and any $r \in \{1, 2, \ldots, n\}$, let $\hat{C}$ and C^0 be the r-reduced matrix and the r-reduced submatrix, respectively, of C. Let G^0 be the digraph obtained from G by deleting node r. Then C^0 is a cost matrix associated with complete digraph G^0. Since row r and column r of $\hat{C}$ have all zero nondiagonal entries, any Hamiltonian path P in G^0 can be extended to a Hamiltonian cycle H in G such that $C^0(P) = \hat{C}(H)$, and conversely, for any Hamiltonian cycle H in G, the Hamiltonian path P in G^0 obtained by deleting the node r has $\hat{C}(H) = C^0(P)$. Hence, $\hat{C}$ is a DTC(k) matrix if and only if C^0 is a DPC(k) matrix. The result follows in view of Fact 3. □

The following corollary of Theorem 1.1, Lemma 1.2, and Theorem 2.1 can be easily verified and will be useful in this section and the next.

COROLLARY 2.2. *Let C be an $n \times n$ skew-symmetric cost matrix associated with a complete digraph G. Then C is a DPC(1) matrix if and only if all its nondiagonal elements are zero. If every nondiagonal element of C is either 0 or $\pm x$ for some positive value x, then C is a DTC(1) matrix if and only if there exists $S \subseteq N = \{1, 2, \ldots, n\}$ such that for any $i, j \in \{1, 2, \ldots, n\}$, $i \neq j$,*

$$c_{ij} = \begin{cases} -\alpha & \text{if } i \notin S \text{ and } j \in S, \\ \alpha & \text{if } i \in S \text{ and } j \notin S, \\ 0 & \text{otherwise} \end{cases}$$

for $\alpha \in \{-x, x\}$.

Our main results in this section are polynomially testable characterizations of DTC(2) and DPC(2) matrices. Theorem 2.1 provides a characterization of DTC(2) matrices in terms of DPC(2) matrices. We shall give a characterization of DPC(2) matrices that can be verified in polynomial time. For this, we need a characterization of a special class of DTC(2) matrices with only two types of nondiagonal entries α and β and such that the two distinct values of tour costs are $\{n\alpha, (n-1)\alpha + \beta\}$ or $\{n\beta, (n-1)\beta + \alpha\}$. Such a DTC(2) matrix is called an *elementary DTC(2) matrix*.

THEOREM 2.3. *A cost matrix C is an elementary DTC(2) matrix if and only if the following three conditions are satisfied:*

(1) *The nondiagonal elements of C have only two distinct values, α and β.*
(2) *C is not a DTC(1) matrix.*
(3) *Either the set of arcs with cost α or the set of arcs with cost β is one of the following three types:* (i) $\{(i,j),(j,i)\}$ *for some pair i,j of nodes;* (ii) $\{(i,u) : u \in S\}$, *where $i \in N$, $S \subset V(G)\setminus\{i\}$, and $S \neq \emptyset$;* (iii) $\{(u,i) : u \in S\}$, *where $i \in N$, $S \subset V(G) \setminus \{i\}$, and $S \neq \emptyset$.*

Proof. If a cost matrix C satisfies the conditions of the theorem, then it is easy to verify that the set of distinct tour costs is either $\{n\alpha, (n-1)\alpha+\beta\}$ or $\{n\beta, (n-1)\beta+\alpha\}$.

Conversely, suppose that C is an elementary DTC(2) matrix with nondiagonal elements of values α and β, and neither the set of arcs with cost α nor the set of arcs with cost β is of the above three types. Then, since C is not a DTC(1) matrix, there exists a pair of arcs $(i,j),(u,v)$ of cost α and a pair of arcs $(x,y),(s,t)$ of cost β with $i \neq u$, $j \neq v$, $x \neq s$, and $y \neq t$. It can be verified that there exists a tour containing both the arcs (i,j) and (u,v) and a tour containing both the arcs (x,y) and (s,t), contradicting the fact that D is an elementary DTC(2) matrix. □

LEMMA 2.4. *If C is a DPC(2) matrix and H is any tour in G, then arcs of H have at most two distinct costs.*

(It may be noted that in the statement of Lemma 2.4, C is a DPC(2) matrix, but H is a tour and not a Hamiltonian path.)

Proof. If possible, let H contain three arcs, say e, f, and g, with distinct costs, say x, y, and z. By deleting these arcs one at a time, we get three Hamiltonian paths in G with distinct costs $C(H) - x$, $C(H) - y$, and $C(H) - z$. This contradicts the fact that C is a DPC(2) matrix, and hence the result follows. □

Using Lemma 2.4, we now prove a stronger result.

LEMMA 2.5. *If C is a DPC(2) matrix and $n > 3$, then the arcs of G have exactly two distinct costs.*

Proof. Obviously, arcs of G have at least two distinct costs. If possible, let there be more than two distinct costs. Then it is possible to find a tour H containing arcs with at least two distinct costs, say α and β. By Lemma 2.4, arcs of H have exactly two distinct costs, α and β. Let (i,j) be an arc of G not in H with cost γ, where $\gamma \neq \alpha, \beta$. Using Scheme 1 of the previous section, with $u_1 = i$ and $\ell = n$, generate tour H'. Clearly, H' contains at least one and at most three arcs of cost γ. By Lemma 2.4, all other arcs of H' have cost precisely one of α or β. Let us assume it is α. Then, H contains at most three arcs of cost β. Thus the cost of H is $(n-x)\alpha+x\beta$, and the cost of H' is $(n-y)\alpha+y\gamma$ for some x, y such that $1 \leq x, y \leq 3$. Since $n > 3$, it can be verified that the collection of all Hamiltonian paths, obtained from H and H' by deleting an arc, have at least three distinct costs, contradicting the fact that C is a DPC(2) matrix. □

If $n = 3$, it is possible to have a DPC(2) matrix with more than two distinct elements. For example consider the matrix

$$C^* = \begin{bmatrix} \infty & 0 & 1 \\ 1 & \infty & 2 \\ 0 & 1 & \infty \end{bmatrix}.$$

All the Hamiltonian paths in the complete digraph G on three nodes with C^* as the cost matrix have cost either 0 or 2; yet C^* has nondiagonal elements with three distinct values. However, Lemma 2.5 can be shown to hold even for $n = 3$ if C is a symmetric matrix.

THEOREM 2.6. *For any integer $n > 3$, an $n \times n$ matrix C associated with a complete digraph G is a DPC(2) matrix if and only if it satisfies one of the following two conditions:*

(1) *C is a DTC(1) matrix, and nondiagonal elements of C have exactly two distinct values.*
(2) *C is an elementary DTC(2) matrix.*

Proof. Suppose that C satisfies conditions (1) or (2). We must show that C is a DPC(2) matrix. If condition (1) is satisfied, then all the tours in G have the same cost, say δ, and C contains elements of value, say α and β, only. Then the Hamiltonian paths in G have costs $\delta - \alpha$ and $\delta - \beta$, and hence C is a DPC(2) matrix. Suppose that condition (1) is not satisfied but condition (2) is satisfied. Thus arcs of G have exactly two distinct costs, say α and β, and the set of distinct values of costs of tours in G is $\{n\alpha, (n-1)\alpha + \beta\}$ or $\{n\beta, (n-1)\beta + \alpha\}$. Thus the set of distinct values of costs of Hamiltonian paths in G is $\{(n-1)\alpha, (n-2)\alpha + \beta\}$ or $\{(n-1)\beta, (n-2)\beta + \alpha\}$. Thus C is a DPC(2) matrix.

Conversely, assume that C is a DPC(2) matrix. By Lemma 2.5 the elements of C must be either α or β for some α and β. Suppose that C does not satisfy any of the conditions (1) and (2). Then there exist two tours H^1 and H^2 in G with costs $C(H^1) = x\alpha + (n-x)\beta$ and $C(H^2) = y\alpha + (n-y)\beta$ such that $\{0, 1\} \neq \{x, y\} \neq \{n-1, n\}$. Without loss of generality, let $y < x$. If $x = n$ and $y = 0$, then choose an arc (i, j) in G of cost β and using Scheme 1 given in the previous section, generate tour H' from the tour H^1 with $i = u_1$, $j = u_r$, and $\ell = n$. The cost of the tour H' is $z\alpha + (n-z)\beta$ for some $n - 3 \leq z < n$. It is easy to see that by removing from each of H^1, H^2, and H' an arc of each type, we get Hamiltonian paths of at least three distinct costs. We thus have a contradiction.

We thus have to consider the following cases: (i) $y = 0$ and $1 < x < n$, (ii) $1 \leq y < x - 1$, and (iii) $1 \leq y = x - 1 < n - 1$. In each of these cases, by removing from each of H^1 and H^2 an arc of each type, we get Hamiltonian paths of at least three distinct costs. This contradicts the fact that C is a DPC(2) matrix, and hence the result follows. □

In the case of symmetric matrices, Theorem 2.6 can be shown to hold even for the case $n = 3$.

We get the following interesting corollary of Theorem 2.6, Fact 1, Theorem 2.1, and Corollary 2.2.

COROLLARY 2.7. *There is no DPC(2) skew-symmetric matrix of size $n \geq 4$, and no DTC(2) skew-symmetric matrix of size $n \geq 5$.*

Proof. Let C be an $n \times n$ skew-symmetric matrix for some $n \geq 4$. It follows from Corollary 2.2 that C cannot be a DTC(1) matrix with exactly two distinct types of elements, and it follows from Theorem 2.3 that it cannot be an elementary DTC(2) matrix. Hence, by Theorem 2.6, C cannot be a DPC(2) matrix. The second part of this corollary now follows from Fact 1 and Theorem 2.1. □

The following is an example of a 3×3 skew-symmetric DPC(2) matrix:

$$\begin{bmatrix} \infty & 1 & -1 \\ -1 & \infty & 1 \\ 1 & -1 & \infty \end{bmatrix}.$$

Theorem 2.6 gives a characterization of the class of DPC(2) matrices in terms of the classes of DPC(1) and elementary DTC(2) matrices. A complete characterization of the class of DPC(1) matrices is given in Lemma 1.2 and a complete characterization of elementary DTC(2) matrices is given in Theorem 2.3. Thus we have a

complete characterization of DPC(2) matrices, and hence by Theorem 2.1 we have a complete characterization of the class of DTC(2) matrices. We now observe that, using our characterizations, we can recognize DTC(2) and DPC(2) matrices in strongly polynomial time.

Given an $n \times n$ matrix C, we can check whether it is a DTC(2) matrix as follows. Construct the reduced matrix $\hat{C}$ of C, obtain the $(n-1) \times (n-1)$ reduced submatrix C^0 by deleting row n and column n, and verify whether C^0 satisfies the conditions (i) or (ii) of Theorem 2.6 (which can be easily done in $O(n^2)$ time). By Theorem 2.1, C^0 satisfies one of these conditions if and only if C is a DTC(2) matrix. If C^0 satisfies one of the conditions of Theorem 2.6, then it is possible to construct two Hamiltonian paths in G^0 of distinct costs in $O(n^2)$ time. These paths can be extended to tours in G of distinct costs with respect to C. The smaller of these tours is an optimal solution to the corresponding instance of TSP on G. All these computations can be performed in $O(n^2)$ time.

It may be noted that if C is a DTC(2) matrix, then any tour with objective function value greater than or equal to the average value of all tours is an optimal solution to the corresponding instance of TSP. It is known that several well-known heuristics for TSP produce solutions with objective function value no worse than the average cost of all tours [22]. Thus any such heuristic guarantees an optimal solution when the cost matrix of a given instance TSP is a DTC(2) matrix.

3. Skew-symmetric DTC(3) and DPC(3) matrices for complete digraphs. Tarasov [26] gave a complete characterization of distance matrices for the assignment problem (minimum weight bipartite matching problem) [18] with three distinct objective function values. No corresponding results for Hamiltonian cycles are available. In this section we provide complete characterization of such skew-symmetric cost matrices. As we show in section 4, skew-symmetric matrices are useful in converting some special ATSPs into STSPs.

We first prove several lemmas that allow us to present the proof of our main result in a simple way. *Throughout this section, we assume that the digraph G under consideration is a complete digraph and that the associated cost matrix C is skew-symmetric.* Our first aim is to establish that a skew-symmetric DPC(3) matrix will not contain more than three distinct elements.

The following lemma is easy to verify.

LEMMA 3.1. *For a skew-symmetric cost matrix C, if there exists a Hamiltonian path (tour) of cost α, then its reversal has a cost of $-\alpha$. Hence, if C is a DPC(k) (DTC(k)) matrix for some $k \geq 3$, then the corresponding distinct costs of Hamiltonian paths (tours) will be $\{\pm\alpha_1, \pm\alpha_2, \ldots, \pm\alpha_s\}$ if k is even and $\{0, \pm\alpha_1, \pm\alpha_2, \ldots, \pm\alpha_s\}$ if k is odd, for some $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_s$, where $s = \lfloor \frac{k}{2} \rfloor$. In particular, if C is a DPC(3) (DTC(3)) matrix, then the corresponding distinct costs of Hamiltonian paths (tours) will be $\{0, \pm\alpha_1\}$ for some $\alpha_1 > 0$.*

LEMMA 3.2. *If C is a DPC(3) matrix, then no tour in G contains more than three arcs of distinct costs.*

Proof. Let H be a tour in G containing four arcs with distinct costs w, x, y, z. Then G contains Hamiltonian paths of distinct values $C(H) - w$, $C(H) - x$, $C(H) - y$, $C(H) - z$. This contradicts the fact that C is a DPC(3) matrix. □

LEMMA 3.3. *If C is DPC(3) and $n \geq 5$, then no tour in G contains arcs of costs $x, -x$, and y for any distinct nonzero $x, -x, y$.*

Proof. Without loss of generality assume $x > 0$. Suppose that there exists a tour H in G containing arcs of costs $x, -x$, and y. Renumber the nodes in G if necessary

to make $H = (1, 2, \ldots, n, 1)$. From Lemma 3.2 all arcs of H have weight $x, -x$, or y. Let $C(H) = \theta$. Then by removing different arcs from the tour, we get Hamiltonian paths of distinct costs $\theta - x$, $\theta + x$, and $\theta - y$. It follows from Lemma 3.1 that one of these Hamiltonian paths has cost zero, and hence θ equals x, $-x$, or y. We consider three mutually exclusive and exhaustive cases.

Case 1: $-x < y < x$. In this case, by Lemma 3.1, $\theta = y$ and $\theta - x(= y - x) = -(\theta + x)(= -(y + x))$, which implies that $y = 0$, a contradiction.

Case 2: $x < y$. In this case, by Lemma 3.1, $\theta = x$, $y = 3x$, and the three distinct costs of Hamiltonian paths are $\{-2x, 0, 2x\}$. For each $z \in \{-x, x, y\}$, let $n(z) =$ number of arcs in the tour H of cost z. Then $n(-x) = 3n(y) + n(x) - 1$. We now consider three different subcases.

Subcase 1: $n(y) \geq 2$ *and there exists a pair of nonadjacent arcs in* H*, one of cost* x *and the other of cost* y. From such a pair of arcs, choose the one, say arc (i, j), of cost y. Construct a tour $\hat{H}$ from H using the arc (i, j) and Scheme 3. Then the tour $\hat{H}$ contains at least one arc of cost y, at least $3 + n(x)$ arcs of cost x, and at least one and at most $2 + n(x)$ arcs of cost $-x$. Hence, each arc in $\hat{H}$ has cost x, $-x$, or y, and the cost $\hat{\theta}$ of the new tour satisfies $\hat{\theta} \geq 4x$. The cost of the Hamiltonian path, obtained from $\hat{H}$ by deleting an arc of cost x, is at least $3x$ and therefore is not in the set $\{-2x, 0, 2x\}$, contradicting the fact that C is a DPC(3) matrix.

Subcase 2: $n(y) = 2$ *and the unique arc in* H *of cost* x *is adjacent to both the arcs in* H *of cost* y. Let us assume without loss of generality that the arcs in H of cost y are $(1, 2)$ and $(3, 4)$. The arc in H of cost x is then $(2, 3)$. Construct a tour $\bar{H}$ from H using the arc $(1, 2)$ as (i, j) and Scheme 2. The tour $\bar{H}$ contains arcs of costs $-y$, y, and $-x$. Using this tour and case (i) above, we get a contradiction.

Subcase 3: $n(y) = 1$. In this case, $n(-x) = 2 + n(x) \geq 3$. Let us assume without loss of generality that the arc in H of cost y is $(1, 2)$. We consider five possibilities designated as Subcases 3.1–3.5.

Subcase 3.1: *At least one neighbor and at least one nonneighbor in* H *of the arc* $(1, 2)$ *has cost* x. Construct a tour $\hat{H}$ from H using the arc $(1, 2)$ as (i, j) and Scheme 3. Then the tour $\hat{H}$ contains at least one arc of cost y, at least $1 + n(x)$ arcs of cost x, and at least one and at most $1 + n(x)$ arcs of cost $-x$. Hence, each arc in $\hat{H}$ has cost x, $-x$, or y, and the cost $\hat{\theta}$ of the new tour satisfies $\hat{\theta} \geq 3x$. Repeating Case 2 with the tour $\hat{H}$, we get a contradiction.

Subcase 3.2: $n(x) = 2$ *and both the neighbors in* H *of the arc* $(1, 2)$ *have cost* x. In this case, $n = 7$ and arc $(3, 4)$ has cost $-x$. Construct tour $\bar{H}$ from H using the arc $(3, 4)$ as (i, j) and Scheme 2. Then the tour $\bar{H}$ contains at least one arc of cost y, at least two arcs of cost x, and at least two arcs of cost $-x$, and at least one neighbor and at least one nonneighbor in $\bar{H}$ of the arc $(1, 2)$ have cost x. Hence, each arc in $\hat{H}$ has cost x, $-x$, or y. If the cost $\bar{\theta}$ of the new tour is not x, then by repeating Case 2 with this new tour, we arrive at a contradiction. Else, by repeating the Subcase 3.2 with the new tour, we arrive at a contradiction.

Subcase 3.3: $n(x) = 2$ *and none of the neighbors in* H *of the arc* $(1, 2)$ *has cost* x. Suppose there exists an arc $(s, t) = (a, b)$ (other than the arc $(2, 1)$) that is not in H and has some cost $z \notin \{-x, x\}$. If the set $\{(a, a + 1), (b - 1, b)\}$ contains all the arcs in H of cost x or if it contains the arc $(1, 2)$, then instead of the arc (a, b) consider the arc (b, a) and denote it by (s, t) and its cost by z. The set $\{(b, b + 1), a - 1, a)\}$ will then not contain the arc $(1, 2)$, nor will it contain all the arcs in H of cost x. Construct tour H' from H using the Scheme 1 with arc (s, t) as (i, j) and choosing ℓ such that H' contains the arc $(1, 2)$ and at least one arc of cost x. Then the tour

H' contains at least one arc of each of the costs $-x,\ x,\ y,\ z$. Hence, $z = y$ and the tour H' contains at least two arcs of cost y and at least one arc of each of the costs $-x$ and x. Repeating Subcase 1 or Subcase 2 of Case 2 with H', we then arrive at a contradiction. Hence, the only arcs in G of cost other than $-x$ and x are $(1,2)$ of cost y and arc $(2,1)$ of cost $-y$.

If there exists a tour H_2 in G containing $(1,2)$ and all other arcs of cost x, then by deleting an arc of value x from H_2 we get a Hamiltonian path of value $(n+1)x \geq 6x$, contradicting the fact that C is a DPC(3) matrix. If there exists a tour H_3 in G containing $(1,2)$ and all other arcs of cost $-x$, then by deleting arc $(1,2)$ from H_3 we get a Hamiltonian path of value $(1-n)x \leq -4x$, a contradiction. Thus every tour in G containing the arc $(1,2)$ contains at least one arc of each of the costs x and $-x$. If the cost of any such tour is not x, then by repeating Case 2 with this tour, we arrive at a contradiction. Else, let G^0 be the graph obtained by contracting the arc $(1,2)$ in G to a pseudonode 0. Then every Hamiltonian tour in G^0 has cost $-2x$. Let C^0 be the distance matrix associated with G^0 with rows/ columns arranged in order $(0,3,\ldots,n)$. Then all the nondiagonal elements of C^0 are $\pm x$. Let $S^1 = \{i : 3 \leq i < n : c^0_{in} = x\}$ and $S^2 = \{i : 3 \leq i < n : c^0_{in} = -x\}$. Let D be the reduced submatrix matrix of C^0. Then by Theorem 2.1 and Corollary 2.2, all the nondiagonal elements of D must equal 0. This implies that $c^0_{ij} = 2x$ for all $i \in S^1,\ j \in S^2$; $c^0_{ij} = -2x$ for all $i \in S^2,\ j \in S^1$; and $c^0_{ij} = 0$ for all $i \in S^1,\ j \in S^1$ and for all $i \in S^2,\ j \in S^2$. But this contradicts the fact that all the nondiagonal elements of C^0 are $\pm x$.

Subcase 3.4: $n(y) = 1$, $n(x) = 1$, *and the arcs in* H *of costs* x *and* y *are neighbors.* In this case, $n = 5$. Suppose the arc $(2,3)$ has cost x. The tour $H^1 = (1,2,4,3,5,1)$ has at least one arc of each of the costs $-x$, x, and y. Hence, its total cost must be x, which implies that each of the arcs $(2,4)$ and $(3,5)$ has cost $-x$. Now, the tour $H^2 = (1,2,3,5,4,1)$ has at least one arc of each of the costs $-x$, x, and y. Hence, its total cost must be x, which implies that the cost of the arc $(4,1)$ is $-3x = -y$. The tour H^2 thus contains arcs with four distinct costs, which contradicts Lemma 3.2.

Now, suppose the arc $(5,1)$ has cost x. The tour $H^1 = (1,2,3,5,4,1)$ has at least one arc of each of the costs $-x$, x, and y. Hence, its total cost must be x, which implies that each of the arcs $(3,5)$ and $(4,1)$ has cost $-x$. Now, the tour $H^2 = (1,2,4,3,5,1)$ has at least one arc of each of the costs $-x$, x, and y. Hence, its total cost must be x, which implies that the cost of the arc $(2,4)$ is $-3x = -y$. The tour H^2 thus contains arcs with four distinct costs, a contradiction.

Subcase 3.5: $n(y) = 1$, $n(x) = 1$, *and the arcs in* H *of costs* x *and* y *are non-neighbors.* In this case too, $n = 5$. Suppose that the arc $(3,4)$ has cost x. The tour $H^1 = (1,2,5,4,3,1)$ has at least one arc of each of the costs $-x$, x, and y. Hence, its total cost must be x, which implies that each of the arcs $(2,5)$ and $(3,1)$ has cost $-x$. Similarly, the tour $H^2 = (1,2,5,3,4,1)$ has at least one arc of each of the costs $-x$, x, and y. Hence, its total cost must be x, which implies that each of the arcs $(5,3)$ and $(4,1)$ has cost $-x$. Now, the tour $H^3 = (1,2,3,5,4,1)$ has at least one arc of each of the costs $-x$, x, and y, and its total cost is $3x$, a contradiction.

Now, suppose that the arc $(4,5)$ has cost x. The tour $H^1 = (1,2,5,4,3,1)$ has at least one arc of each of the costs $-x$, x, and y. Hence, its total cost must be x, which implies that each of the arcs $(2,5)$ and $(3,1)$ has cost $-x$. Similarly, the tour $H^2 = (1,2,5,4,3,1)$ has at least one arc of each of the costs $-x$, x, and y. Hence, its total cost must be x, which implies that each of the arcs $(2,4)$ and $(5,3)$ has cost $-x$. Now, the tour $H^3 = (1,2,4,3,5,1)$ has at least one arc of each of the costs $-x$, x, and y, and its total cost is $3x$, a contradiction.

Case 3: $y < -x$. In this case, by considering the reversal H^* of H and repeating Case 2, we arrive at a contradiction.

This completes the proof. □

LEMMA 3.4. *If C is DPC(3) and $n \geq 5$, then no tour in G contains arcs of costs x, y, and z for any distinct nonzero real numbers x, y, z.*

Proof. If possible let G contain a tour H with arcs having costs x, y, and z. Without loss of generality, let us assume that $H = (1, 2, \ldots, n, 1)$. From Lemma 3.2, H does not contain any arc of cost other than x, y, and z.

Case 1: *H contains a pair of nonadjacent arcs $(a, a+1)$ and $(b, b+1)$ with the same cost.* Let the cost of each of these arcs be x. If there exists some arc in H of cost y or z that is not in $\{(a-1, a), (a+1, a+2)\}$, then set $(i, j) = (a, a+1)$. Else, since $n \geq 5$, there exists some arc in H of cost y or z that is not in $\{(b-1, b), (b+1, b+2)\}$, and therefore set $(i, j) = (b, b+1)$. Construct a tour $\bar{H}$ from H using Scheme 2 and the chosen arc (i, j). The tour $\bar{H}$ contains arcs of costs $\{-x,\ x,\ y\}$ or $\{-x,\ x, z\}$. However, this contradicts Lemma 3.3.

The condition of Case 1 is always satisfied if H contains three arcs of the same cost or $n \geq 7$.

Case 2: *$n = 6$ and the tour H contains exactly two arcs of each of the costs $x,\ y,\ z$.* By deleting different arcs from H, we get Hamiltonian paths of three different costs $2x + 2y + z,\ 2x + y + 2z$, and $x + 2y + 2z$. Without loss of generality, let us assume that $x < y < z$. Then Lemma 3.1 implies that $2x + y + 2z = 0$ and $x + 2y + 2z = -(2x + 2y + z)$. But this implies that $x = -z$ and $y = 0$; we have a contradiction.

Case 3: *$n = 5$ and the tour H contains exactly two arcs of cost x that are adjacent and exactly two arcs of cost y that are adjacent.* By deleting different arcs from H, we get Hamiltonian paths of three different costs $2x + 2y$, $2x + y + z$, and $x + 2y + z$. Without loss of generality, let us assume that $x < y$.

Subcase 1: $x < y < z$. In this case, Lemma 3.1 implies that $2x + y + z = 0$ and $x + 2y + z = -(2x + 2y)$. These imply that $y > 0$, $x = -3y$, and $z = 5y$, and therefore the distinct costs of Hamiltonian paths are $\{-4y, 0, 4y\}$. Suppose that the arcs of cost x are $(1, 2)$ and $(2, 3)$ and the arcs of cost y are $(3, 4)$ and $(4, 5)$. Construct a tour $\hat{H}$ from H using Scheme 3 with arc $(2, 3)$ as arc (i, j). Then by Lemma 3.2 it follows that $\hat{H}$ contains only arcs of costs x, $-y$, and $-z$. Deleting from $\hat{H}$ an arc of cost $-y$, we get a Hamiltonian path of cost no more than $-8y$, contradicting the fact that C is a DPC(3) matrix.

Subcase 2: $x < z < y$. In this case, Lemma 3.1 implies that $2x + 2y = 0$, which implies that $x = -y$. We thus have a contradiction.

Subcase 3: $z < x < y$. In this case, consider the reversal H^* of H which satisfies the condition of Subcase 1, a contradiction. This completes the proof of the lemma. □

LEMMA 3.5. *If C is DPC(3), then no tour in G contains arcs of cost $0, x$, and y, where x and y are nonzero values with $|x| \neq |y|$.*

Proof. Suppose that G contains a tour H with arcs of costs $0,\ x$, and y. From Lemma 3.2, no arc in H can have cost other than $0, x$, or y. By deleting different arcs from H, we can generate Hamiltonian paths of distinct costs $C(H),\ C(H) - x,\ C(H) - y$. Without loss of generality, let us assume that $x < y$.

If $0 < x < y$, then $C(H) - y < C(H) - x < C(H)$. Then Lemma 3.1 implies that $C(H) - x = 0$ or $C(H) = x$, which is impossible.

If $x < y < 0$, then Lemma 3.1 implies that $C(H) - y = 0$ or $C(H) = y$, which is again impossible.

If $x < 0 < y$, then Lemma 3.1 implies that $C(H) = 0$ and hence $C(H) - y = -y$, which (by Lemma 3.1) equals $-C(H) + x = x$, a contradiction. This proves the result. □

LEMMA 3.6. *If C is a DPC(3) skew-symmetric matrix with more than three distinct values of nondiagonal elements and $n \geq 5$, then no tour in G contains arcs of cost $0, x$, and $-x$ for any positive value x.*

Proof. Suppose G contains a tour H with arcs with costs $0,\ x$, and $-x$ for some positive value x. Lemma 3.2 implies that no arc in H can have cost other than $0, x$, or $-x$. We can generate from H Hamiltonian paths of distinct costs $C(H)-x,\ C(H)$, and $C(H) + x$. Lemma 3.1 therefore implies that $C(H) = 0$, and therefore H contains an equal number (say α) of arcs of values x and $-x$. Since C contains more than three distinct values of nondiagonal elements, there exists an arc (s, t) in G of some value $y \notin \{-x, 0, x\}$. It is easy to see that we can construct from H using Scheme 1, either arc (s, t) or arc (t, s) as arc (i, j), and a proper choice of ℓ, a Hamiltonian path containing either arcs of costs $x, -x, y$ or arcs of costs $0, x, y$. In either case, we arrive at a contradiction using Lemma 3.3 or Lemma 3.5. This proves the lemma. □

THEOREM 3.7. *If C is an $n \times n$ skew-symmetric DPC(3) matrix where $n \geq 5$, then the number of distinct values of nondiagonal elements of C is no more than three.*

Proof. If possible let C be an $n \times n$ skew-symmetric DPC(3) matrix with $n \geq 5$ and more than three distinct values of nondiagonal elements. Then it will contain nondiagonal elements of four distinct values of the type $\{-x, x, -y, y\}$ for some positive distinct values x and y. It is easy to see that G contains a tour H with at least two arcs of distinct costs.

If H contains three arcs of distinct costs, we have a contradiction by one of Lemmas 3.3, 3.5, 3.5, and 3.6.

Suppose that H contains exactly two arcs of distinct costs, say α and β. Since $n \geq 5$, at least three of the arcs will have same cost, say α. If $\beta \neq -\alpha$, we can construct a tour H_1 as in the proof of Lemma 3.4, containing arcs of costs $\alpha, -\alpha, \beta$. If $\beta = -\alpha$, then again we can construct a tour in G containing $\alpha, -\alpha, \gamma$ for some γ such that $|\gamma| \neq |\alpha|$. In either case we obtain a contradiction to Lemma 3.3. This proves the theorem. □

We shall now give polynomially testable characterizations of DPC(3) and DTC(3) matrices. For this, we need a characterization of a special class of skew-symmetric DTC(3) matrices with nondiagonal elements $0,\ x,\ -x$ for some positive x such that (i) the three distinct values of tour costs are $\{-x,\ 0,\ x\}$ and (ii) every tour of cost x contains precisely one arc of cost x and the remaining arcs of cost 0. (By skew-symmetry, it follows that the same is true for $-x$.) We call such a DTC(3) matrix an *elementary DTC(3) matrix*. The following property of tours associated with a DTC(3) matrix is useful in characterizing this class of matrices.

LEMMA 3.8. *If C is an elementary DTC(3) matrix with nondiagonal elements $0, x$, and $-x$ for $x \neq 0$, then no tour in G has two adjacent arcs of cost x or two adjacent arcs of cost $-x$.*

Proof. If possible, let H be a tour in G where at least two arcs of cost x are adjacent in H. Without loss of generality assume that H contains the path $1 - 2 - 3$ and $c_{12} = c_{23} = x$. Let ℓ and t be two nodes in G such that arcs $(\ell, 1)$ and $(3, t)$ are in H. Let $P(t, \ell)$ denote the path from t to ℓ in H, and let $C(P(t, \ell)) = \theta$. Since C is an elementary DTC(3) matrix, $C(H) = 0$. Let H_1 be the tour obtained from H by

reversing the path $1-2-3$ (Scheme 4). Then

$$C(H) = 2x + \theta + c_{\ell 1} + c_{3t} = 0 \tag{3.1}$$

and

$$C(H_1) = -2x + \theta + c_{1t} + c_{\ell 3} = 0. \tag{3.2}$$

From (3.1) and (3.2) we have

$$4x = c_{1t} + c_{\ell 3} - c_{\ell 1} - c_{3t}. \tag{3.3}$$

Since elements of C are $x, -x$, or 0, from (3.3) we have $c_{1t} = c_{\ell 3} = x$ and $c_{\ell 1} = c_{3t} = -x$, and hence from (3.1), $\theta = 0$. Now construct a tour H_2 from H by reversing arc $(1,2)$ (Scheme 2). We have

$$C(H_2) = -2x + c_{13} + c_{\ell 2} = 0. \tag{3.4}$$

From (3.4), $c_{13} = c_{\ell 2} = x$. Now the tour $\ell - 2 - 3 - 1 - t - P(t,\ell)$ has cost $2x$, a contradiction to the fact that C is an elementary DTC(3) matrix.

Because we are dealing with skew-symmetric matrices, if at least two arcs of cost $-x$ are adjacent in H, then using the reversal of H in place of H in the above argument, we get a contradiction. This completes the proof. □

An immediate consequence of Lemma 3.8 is that if the cost matrix C is an elementary DTC(3) matrix, then for any node i of G, each arc coming into i has cost in $\{0, \alpha\}$ for some $\alpha \in \{x, -x\}$. Otherwise, if there exist arcs (i,j), (i,t) with $c_{ij} = x$ and $c_{it} = -x$, then any tour containing the path $t - i - j$ will violate Lemma 3.8. This property is crucial to the proof of our characterization of elementary DTC(3) matrices. We summarize the matrix version of this property in the following corollary.

COROLLARY 3.9. *If C is an elementary DTC(3) matrix, then no row (or column) contains both x and $-x$.*

THEOREM 3.10. *A skew-symmetric cost matrix C with nondiagonal elements $0,\ x,\ -x$ for some $x > 0$ is an elementary DTC(3) matrix if and only if there exist some $r \in N$ and a nonempty proper subset S of $N - \{r\}$ such that for any $i, j \in \{1, 2, \ldots, n\}$, $i \neq j$,*

$$c_{ij} = \begin{cases} \alpha & \text{if } i \in S \text{ and } j = r, \\ -\alpha & \text{if } i = r \text{ and } j \in S, \\ 0 & \text{otherwise} \end{cases}$$

for $\alpha \in \{x, -x\}$.

Proof. Let C be a skew-symmetric with nondiagonal elements $0,\ x,\ -x$ for some $x > 0$. If C satisfies the condition of the theorem, then consider any tour H in G. Let the arcs in H incident to node r be $\{(i,r),(r,j)\}$. If $\{i,j\} \subseteq S$ or $\{i,j\} \cap S = \emptyset$, then $C(H) = 0$. If $|\{i,j\} \cap S| = 1$, then for some $\alpha \in \{x, -x\}$, $C(H) = \alpha$ and the tour contains precisely one arc of cost α and the other arcs of cost 0. Thus, C is an elementary DTC(3) matrix.

Conversely, suppose C is an elementary DTC(3) matrix. It follows from Corollary 3.9 that for each $i \in V(G)$, the cost of each outgoing arc of node i belongs to $\{0, \alpha\}$, and the cost of each incoming arc of node i belongs to $\{0, -\alpha\}$ for some $\alpha \in \{x, -x\}$. Thus we can renumber the nodes such that for the corresponding reordering of the rows and columns of C, the modified matrix (which we shall also

denote by C for convenience) has the form such that

$$C = \begin{pmatrix} O_1 & O_2 & A \\ O_3 & O_4 & O_5 \\ -A & O_6 & O_7 \end{pmatrix},$$

where O_2, O_3, O_5, O_6 are matrices with all entries zero; O_1, O_4, O_7 are matrices with all nondiagonal entries zero; and A is a matrix with all entries $0, x$ or all entries $0, -x$ and with no row or column with all zero entries. For definiteness, we assume without loss of generality that entries of A are $0, x$. Let the columns of O_1 be indexed by set $S_1 = \{1, 2, \ldots, p\}$, columns of O_2 by set $S_2 = \{p+1, p+2, \ldots, p+q\}$, and columns of A by set $S_3 = \{p+q+1, p+q+2, \ldots, n\}$. If $S_2 = \emptyset$, then the matrix A has at least one zero element. For if A has no zero element, then by subtracting x from each row in S_1 and adding x to each column in S_1, we can reduce all the nondiagonal elements of the matrix to zero, and hence C is a DTC(1) matrix, a contradiction. If A has exactly one row or one column, then C is of the required type. Thus A has at least two rows and columns.

Case 1: *A contains at least one zero entry.* Then there exist $i \in S_1$, $j \in S_3$ such that $c_{ij} = 0$, and there exist $z \in S_1$ and $t \in S_3$ such that $z \neq i, t \neq j$, and $c_{it} = c_{zj} = x$. Now construct a tour containing the path $z - j - i - t$. If $C(H) = x$ or $-x$, we have a contradiction to the fact that C is an elementary DTC(3) matrix. Thus $C(H) = 0$. Now construct the tour $\hat{H}$ from H by reversing arc (i, j) (Scheme 2). Since $c_{zi} = c_{jt} = 0$, $C(\hat{H}) = -2x$, a contradiction.

Case 2: *All entries of A are x.* In this case $S_2 \neq \emptyset$. Choose $i \in S_2$ and $t, z \in S_3$; $t \neq z$. Then $c_{1,i} = c_{zi} = 0$ and $c_{1,t} = c_{2,z} = x$. Construct a tour H in G containing the path $2 - z - i - 1 - t$. If $C(H) = x$ or $-x$, we have a contradiction. Thus $C(H) = 0$. Construct a tour $\hat{H}$ from H by reversing the path $z - i - 1$. Then $C(\hat{H}) = -2x$, a contradiction. This completes the proof. □

THEOREM 3.11. *Let C be a skew-symmetric $n \times n$ cost matrix corresponding to the complete digraph G on node set $N = \{1, 2, \ldots, n\}$, where $n \geq 5$. Then C is a DPC(3) matrix if and only if one of the following holds:*

(1) *C is a DTC(1) matrix with nondiagonal elements $\{0, x, -x\}$ for some $x > 0$.*

(2) *C is an elementary DTC(3) matrix.*

Proof. If C satisfies any one of the two conditions of the theorem, then it can be readily verified that the distinct values of costs of Hamiltonian paths in G are $\{x, -x, 0\}$.

Conversely suppose that C is a skew-symmetric DPC(3) matrix with $n \geq 5$. Then C contains at least two distinct elements, and by Theorem 3.7 it does not contain more than three distinct elements. Thus the distinct elements of C are either of the form $x, -x$ or of the form $0, x, -x$ for some $x > 0$, and the cost of each tour and Hamiltonian path in G is of the form px for some integer p.

If C is a DTC(1) matrix, then by Corollary 2.2 it follows that C must contain nondiagonal elements of values $x, -x, 0$. By Corollary 2.7, C cannot be a DTC(2) matrix.

Suppose that C is a DTC(k) matrix for some $k \geq 3$. Choose a tour H in G of cost px with largest possible value of p. The tour H must contain at least one arc of cost x.

Case 1: $p = n$. The tour H contains only arcs of values x. Hence, there exists a Hamiltonian path of cost $(n-1)x$. Choose any arc (i, j) in H and construct a new tour $\bar{H}$ using Scheme 2. The new tour has arcs of cost $x, -x$ and has a total cost of

$\bar{p}x$ for some $(n-2) \geq \bar{p} \geq (n-6) \geq -1$ (since $n \geq 5$). Hence, there exist Hamiltonian paths of costs $(\bar{p}-1)x, (\bar{p}+1)x$. If $0 \leq \bar{p} < n-2$, then we have Hamiltonian paths of four distinct costs $-(n-1)x, -(\bar{p}+1)x, (\bar{p}+1)x, (n-1)x$. If $\bar{p} = n-2$ or -1, then we have Hamiltonian paths of four distinct costs $-(n-1)x, -(\bar{p}-1)x, (\bar{p}-1)x, (n-1)x$. In either case, we have contradiction to the fact that C is a DPC(3) matrix.

Case 2: $2 \leq p < n$. In this case, there exist Hamiltonian paths of four distinct costs $-(p+\alpha)x, -(p-1)x, (p-1)x, (p+\alpha)x$, where $\alpha = 1$ if the tour H contains arcs of cost $-x$ and $\alpha = 0$ if it contains arcs of cost 0. We thus have a contradiction.

Case 3: $p = 1$. By Lemma 3.1, C must be a DTC(3) matrix with distinct values of tour costs $0, \pm x$. If there exist tours of cost x containing arcs of costs $-x$ and 0, then since such a tour contains an arc of cost x, we get Hamiltonian paths of costs 0, $\pm x, \pm 2x$, contradicting the fact that C is a DPC(3) matrix. Suppose a tour of cost x contains an arc of cost $-x$. Then there exist Hamiltonian paths of costs 0, $\pm 2x$. In this case, if any tour of cost 0 contains an arc of cost x or $-x$, or if any tour of cost x contains an arc of cost 0, then we get Hamiltonian paths of costs $\pm x$, contradicting the fact that C is a DPC(3) matrix. Hence, every tour of cost 0 contains all arcs of cost 0, and every tour of cost x contains only arcs of cost $\pm x$. Choose an arc (i, j) of cost 0 and contract it in G to a pseudonode "0" (i.e., delete nodes i and j from G, add the new node "0," and for every node z, $z \neq i, j, 0$ add an arc $(z, 0)$ with cost equal to c_{zi} and an arc $(0, z)$ with cost equal to c_{jz}.) Then in any tour in the resultant digraph if we replace the pseudonode "0" by the arc (i, j), we get a tour in the original digraph of the same cost and containing the arc (i, j) of cost 0. Since every arc in the resultant digraph lies in some tour, all the arcs in the resultant digraph therefore have cost 0. Since C is skew-symmetric, this implies that all the nondiagonal elements of C are zero, a contradiction. Hence, every tour in G of cost x contains only one arc of cost x and all other arcs of cost 0. Thus, C is an elementary DTC(3) matrix. This proves the theorem. ∎

COROLLARY 3.12. *For any integer $n \geq 6$, an $n \times n$ skew-symmetric matrix C is a DTC(3) matrix if and only if there exists an $r \in \{1, 2, \ldots, n\}$ such that the r-reduced submatrix of C is an elementary DTC(3) matrix.*

Proof. If an r-reduced submatrix of C is an elementary DTC(3) matrix, then by Theorem 3.11, this r-reduced submatrix is a DPC(3) matrix, and therefore, by Theorem 2.1, C is a DTC(3) matrix.

Conversely, suppose that C is a DTC(3) matrix. Let $\hat{C}$ and C^0 be its reduced, (i.e., n-reduced) matrix and submatrix, respectively. Then by Theorem 2.1, C^0 is a DPC(3) matrix. Hence, it satisfies condition (i) or condition (ii) of Theorem 3.11. If C^0 is an elementary DTC(3) matrix, then we have the desired result with $r = n$. Else, suppose C^0 is a DTC(1) matrix with nondiagonal elements $\{0, x, -x\}$ for some $x > 0$. Then C^0 has the structure specified in Corollary 2.2 for some proper nonempty subset S of $\{1, 2, \ldots, n-1\}$. It is easy to see that for any $r \in \{1, 2, \ldots, n-1\}$, the r-reduced matrix of $\hat{C}$ (which is the same as the r-reduced matrix of C) is an elementary DTC(3) matrix. This proves the result. ∎

4. Special asymmetric TSPs. In this section we consider some special ATSPs that can be solved as one or more symmetric TSPs of the same size. The digraph G in this section is not necessarily complete. Note that this means that Theorem 1.1 does not hold, as it requires G to be a complete digraph, as is shown below, for a general digraph testing if an associated cost matrix C is DTC(1) is NP-hard.

Two cost matrices C and D associated with the same digraph G are said to be *tour value equivalent* if and only if $C(H) = D(H)$ for every tour H in G. Clearly, C

and D are tour value equivalent if and only if $C - D$ is a DTC(1) matrix with all tours costs equal to zero.

For arbitrary graphs, testing tour value equivalence of two cost matrices is NP-hard. To see this, suppose a polynomial time oracle exists which, with input C, D and the associated digraph G, tells us "yes" if the matrices are tour value equivalent and "no" if they are not. For a given digraph G, let C and D be associated cost matrices with $c_{ij} = 0$ and $d_{ij} = 1$ for each $(i, j) \in E(G)$. Invoke the oracle with C, D, and G as input. If the oracle answers "yes," then G has no Hamiltonian tours, else G contains at least one tour. Since testing Hamiltonicity of a digraph is NP-hard, testing tour value equivalence is also NP-hard. It may be noted that, given two cost matrices C and D and a digraph G, the decision problem "Are C and D tour value equivalent?" is not known to be in NP. But this problem belongs to the class co-NP.

In spite of this negative result, for a large class of (di)graphs, called SC-Hamiltonian graphs [12], tour value equivalence can be tested in polynomial time. A digraph G, not necessarily complete, is said to be *separable constant Hamiltonian (SC-Hamiltonian)* if and only if it is Hamiltonian and, for any DTC(1) matrix associated with it, there exist mappings $a, b : V(G) \to \Re$ such that $c_{ij} = a_i + b_j$ for all (i, j) in $E(G)$. An undirected graph G is said to be SC-Hamiltonian if and only if it is Hamiltonian and for any associated DTC(1) matrix C there exists a mapping $a : V(G) \to \Re$ such that $c_{ij} = a_i + a_j$ for all (i, j) in $E(G)$. Obviously, testing whether a given (di)graph is SC-Hamiltonian is NP-hard. Kabadi and Punnen [12] identified a large class of graphs and digraphs that are SC-Hamiltonian. This class includes complete (di)graphs, complete bipartite (di)graphs, etc.

THEOREM 4.1. *Two distance matrices C and D associated with an SC-Hamiltonian digraph are tour value equivalent if and only if there exist mappings $a, b : V(G) \to \Re$ such that $c_{ij} - d_{ij} = a_i + b_j$ for all (i, j) in $E(G)$, and $\sum_{i=1}^{n}(a_i + b_i) = 0$.*

Proof. By definition, C and D are tour value equivalent if and only if $C(H) = D(H)$ for every tour H in G, which is true if and only if $(C - D)(H) = 0$ for every tour H in G. Let $Q = C - D$. Since G is SC-Hamiltonian, $Q(H) = 0$ for every tour H in G if and only if there exist mappings $a, b : V(G) \to \Re$ such that $q_{ij} = a_i + b_j$ and $\sum_{i=1}^{n}(a_i + b_i) = 0$. □

Since a_i's and b_j's of Theorem 4.1 can be obtained in $O(n^2)$ time, the tour value equivalence of two cost matrices associated with an SC-Hamiltonian graph can be verified in $O(n^2)$ time.

It is easy to show that tour value equivalence is reflexive, symmetric, and transitive and hence an equivalence relation. Thus tour value equivalence partitions the space of cost matrices of a digraph G into equivalence classes. It can be verified that each of these equivalence classes is a convex set.

Let C be a cost matrix associated with a symmetrical digraph G. Recall from Section 1 that C is *Hamiltonian symmetrical* if and only if $C(H) = C(H^*)$ for every tour H in G, where H^* is the reversal of H. Since $C(H^*) = C^T(H)$, C is Hamiltonian symmetrical if and only if C and C^T are tour cost equivalent. Recently, Halskau [9] showed that when G is complete, C is Hamiltonian symmetrical if and only if there exist mappings $a, b : V(G) \to \Re$ such that $c_{ij} = a_i + b_j + d_{ij}$, where $D = (d_{ij})$ is a symmetric matrix. It can be verified that this characterization extends to all symmetrical SC-Hamiltonian digraphs. Halskau [9] provided other characterizations of Hamiltonian symmetrical cost matrices for a complete digraph, as given in the following theorem.

THEOREM 4.2 (see [9]). *Let C be any $n \times n$ cost matrix associated with a complete digraph G. Then the following statements are equivalent:*

(1) *C is Hamiltonian symmetrical*
(2) *$C = K + D$, where K is a $DTC(1)$ matrix and D is a symmetric matrix.*
(3) *$S^k(C)$ is symmetrical for any node k, where $S^k(C)$ is the savings matrix associated with C and the (i,j)th element of $S^k(C)$ is $c_{ik} + c_{kj} - c_{ij}$.*
(4) *$c_{ij} - c_{ji} = \frac{1}{n}(R_i(C) - K_i(C)) - (R_j(C) - K_j(C))$ for all $i, j, i \neq j$, where $R_i(C)$ is the ith row sum of C and $K_i(C)$ is the ith column sum on C.*

It may be noted that $S^k(C)$, the savings matrix associated with C, is the negative of k-reduced matrix of C discussed in section 2. We now give another simple characterization of Hamiltonian symmetrical matrices.

THEOREM 4.3. *Let C be a cost matrix associated with a symmetrical digraph G. Then C is Hamiltonian symmetrical if and only if C and $A = \frac{1}{2}(C + C^T)$ are tour value equivalent. If, in addition, G is SC-Hamiltonian, then the Hamiltonian symmetry of C can be tested in $O(n^2)$ time.*

Proof. Note that C is Hamiltonian symmetrical if and only if $C(H) = C^T(H)$ for all tours H in G. Since any matrix C can be written as $C = 1/2(C + C^T) + 1/2(C - C^T)$, the proof of the first part of the theorem follows. Since for symmetrical SC-Hamiltonian graphs tour value equivalence can be tested in $O(n^2)$ time, the proof of the second part follows. □

The characterization above is valid for all symmetrical digraphs. But for digraphs that are not SC-Hamiltonian, verification of the condition above is difficult. In fact, for arbitrary symmetrical digraphs, it can be shown that this verification is NP-hard. The characterization of Theorem 4.3 has important applications in approximation algorithms.

The arc weights of G are said to satisfy the τ-triangle inequality if and only if for any three nodes i, j, and k of G, $\tau(c_{ij} + c_{jk}) \geq c_{ik}$ [1]. When $\tau = 1$, τ-triangle inequality reduces to the triangle inequality. For $1/2 \leq \tau < 1$, τ-triangle inequality is a restriction of the triangle inequality, and for $\tau > 1$ it is a relaxation of the triangle inequality. We now consider a further relaxation of the τ-triangle inequality. A matrix C satisfies *weak τ-triangle inequality* if and only if, for any triplet (i, j, k) with $i \neq j \neq k$,

$$\tau(c_{ij} + c_{ji} + c_{jk} + c_{kj}) \geq c_{ki} + c_{ik}. \tag{4.1}$$

In the above definition, if $\tau = 1$, we say that C satisfies *weak triangle inequality.*

LEMMA 4.4. *Let C be a cost matrix and $A = (C + C^T)/2$.*
(1) *The matrix A satisfies τ-triangle inequality if and only if the matrix C satisfies weak τ-triangle inequality.*
(2) *If C satisfies τ-triangle inequality, then it satisfies weak τ-triangle inequality.*
(3) *If C is Hamiltonian symmetrical, then C satisfies weak triangle inequality if and only if C satisfies triangle inequality.*

Proof. Consider the triplet (i, j, k) corresponding to three nodes of G. Then A satisfies τ-triangle inequality if and only

$$\begin{aligned}\tau(a_{ij} + a_{jk}) \geq a_{ik} &\Leftrightarrow \tau\left(\frac{c_{ij} + c_{ji}}{2} + \frac{c_{jk} + c_{kj}}{2}\right) \geq \frac{c_{ik} + c_{ki}}{2} \\ &\Leftrightarrow \tau(c_{ij} + c_{ji} + c_{jk} + c_{kj}) \geq c_{ki} + c_{ik}.\end{aligned}$$

This completes the proof of part (1). If C satisfies τ-triangle inequality, then

$$\tau(c_{ij} + c_{jk}) \geq c_{ik} \tag{4.2}$$

and

$$\tau(c_{kj} + c_{ji}) \geq c_{ki}. \tag{4.3}$$

Adding inequalities (4.2) and (4.3), we get the proof of (2).

Let us now prove part (3). Since C is Hamiltonian symmetrical, we have $C = A + X$, where $X = (C - C^T)/2$ is a DTC(1) skew-symmetric matrix. Thus there exist $a_1, a_2, \ldots, a_n$ such that $x_{ij} = a_i - a_j$. (See the discussion after Theorem 1.1.) Hence the elements of X satisfy the triangle *equality*. Now suppose that C satisfies the weak triangle inequality. Then by part (1), A satisfies the triangle inequality. Thus $A + X$ (and hence C) satisfies triangle inequality. The converse of part (3) follows from part (2) of the lemma. □

It may be noted that from Lemma 4.4, if C satisfies τ-triangle inequality, then $A = (C + C^T)/2$ also satisfies τ-triangle inequality. Further, part (3) of the above lemma says for Hamiltonian symmetrical matrices, the τ-triangular inequality and weak τ-triangular inequality are equivalent if $\tau = 1$. But we like to point out that for $\tau \neq 1$ this equivalence need not hold even for Hamiltonian symmetrical matrices.

The best known performance bound for a polynomial time ϵ-approximation algorithm for the metric ATSP is $\epsilon = 4/3 \log_3 n \approx 0.842 \log_2 n$ [13]. When C is Hamiltonian symmetrical and satisfies triangle inequality, we can obtain a 3/2-approximate solution for the ATSP by applying Christofides algorithm on the cost matrix $(C+C^T)/2$. Thus Lemma 4.4 and Theorem 4.3 extend the applicability of the Christofides bound beyond the class of symmetric matrices. It may be noted that Lemma 4.4 need not hold for the symmetric matrix D obtained by Halskau [9] if we want to use D in place of A. When the edge costs satisfy the τ-triangle inequality, Bender and Chekuri [2] obtained a 4τ-approximation algorithm, and Andreae and Bandelt [1] obtained a $(3\tau^2/2 + \tau/2)$-approximation algorithm for the STSP. Thus in view of Theorem 4.3 and Lemma 4.4, these results extend to Hamiltonian symmetric matrices satisfying a weak τ-triangle inequality.

Another application of Theorem 4.3 is when the arc costs satisfy the weak range inequality, i.e.,

$$\max_{ij}\{c_{ij} + c_{ji}\} \leq \tau \min_{ij}\{c_{ij} + c_{ji}\}, \tag{4.4}$$

where $\tau > 1$. The concept of weak range inequality is a generalization of the range inequality considered by Kumar and Rangan, which is given by

$$\max_{ij} c_{ij} \leq \tau \min_{ij} c_{ij}. \tag{4.5}$$

Kumar and Rangan [17] introduced the above inequality for $\tau = 2 + \epsilon$, $\epsilon \geq 0$. For STSP satisfying range inequality with $\tau = 2 + \epsilon$ and $\epsilon \geq 0$, they showed that the Christofides algorithm produces a 4/3-optimal solution when $\epsilon = 0$, and the cycle cover algorithm [4] produces a $\frac{4+\epsilon}{3}$ solution for all $\epsilon \geq 0$.

LEMMA 4.5. *Let C be a cost matrix and $A = (C + C^T)/2$.*

(1) *The matrix A satisfies range inequality if and only if C satisfies weak range inequality.*

(2) *If C satisfies range inequality, then it satisfies weak range inequality.*

Proof. The proof of part (1) follows from the definition. Proof of part (2) follows from the inequality:

$$\max_{ij}\{c_{ij} + c_{ji}\} \leq 2 \max_{ij}\{c_{ij}\} \leq 2\tau \min_{ij}\{c_{ij}\} \leq \tau \min_{ij}\{c_{ij} + c_{ji}\}. \qquad \square$$

Thus by applying the cycle cover algorithm [4] on the cost matrix $A = (C+C^T)/2$, we get a $\frac{4+\epsilon}{3}$-approximate solution for ATSP when the cost matrix C is Hamiltonian symmetrical and satisfies the weak range inequality for $\tau = 2+\epsilon$, and $\epsilon \geq 0$. Again, it may be noted that not all symmetric matrices D obtained in [9] satisfy the weak range inequality even if C satisfies the weak range inequality.

For the maximization version of the TSP, the best known polynomial time approximation algorithm has a performance ratio of 2/3 for the ATSP [13] and 3/4 bound for the STSP [3]. Thus from Theorem 4.3, the maximization TSP when C is Hamiltonian symmetrical can be approximated by the 3/4-approximation algorithm given in [25] for the STSP. In addition, if C satisfies the triangle inequality, this bound can be improved to 7/8 by using the algorithm of Hassin and Rubinstein [10] on the matrix $(C + C^T)/2$. Several polynomially solvable cases of the symmetric TSP can be exploited (for both maximization and minimization versions) to solve the ATSP with cost matrix C whenever $1/2(C + C^T)$ satisfies the required conditions [3, 11].

The discussion above exploits properties of DTC(1) matrices. The next theorem takes advantage of our characterization of skew-symmetric DTC(3) matrices. We first state a lemma that can be easily proved.

LEMMA 4.6. *Let C be an $n \times n$ cost matrix associated with a symmetrical digraph G. Then $D = C - C^T$ is a cost matrix associated with G, and for any $0 \leq \alpha_1 < \alpha_2 < \cdots < \alpha_k$ for some positive integer k the following statements are equivalent:*

(1) *For any tour H in G, $|C(H) - C(H^*)| = \alpha_i$ for some $i \in \{1, 2, \ldots, k\}$, and for any $i \in \{1, 2, \ldots, k\}$ there exists a tour H in G such that $|C(H) - C(H^*)| = \alpha_i$.*

(2) *The set of distinct values of costs of tours in G with respect to cost matrix D is precisely $\{\pm\alpha_1, \pm\alpha_2, \ldots, \pm\alpha_k\}$. Thus the number of distinct tour values in G with respect to D is $2k$ if $\alpha_1 > 0$ and $(2k-1)$ if $\alpha_1 = 0$.*

THEOREM 4.7. *Let C be an $n \times n$ cost matrix associated with the complete digraph G on node set $N = \{1, 2, \ldots, n\}$ such that, for some $x > 0$,* (i) *every tour H in G satisfies $|C(H) - C^T(H)| = 0$ or x;* (ii) *there exists at least one tour H_1 in G with $C(H_1) - C^T(H_1) = 0$; and* (iii) *there exists at least one tour H_2 in G with $|C(H_2) - C^T(H_2)| = x$. Then a minimum (maximum) cost tour in G can be identified by solving at the most $n/2$ symmetric TSPs on n nodes.*

Proof. Suppose that C satisfies the condition of the theorem. Let $D = C - C^T$. Then by Lemma 4.6, D is a DTC(3) matrix with tour costs $0, \pm x$. By Corollary 3.12, it follows that there exist distinct $r, \ell \in N$ and $S \subset N - \{r, \ell\}$ such that the nondiagonal elements of the r-reduced matrix D^0 of D satisfy

$$d^0_{ij} = \begin{cases} \alpha & \text{if } i \in S \text{ and } j = \ell, \\ -\alpha & \text{if } i = \ell \text{ and } j \in S, \\ 0 & \text{otherwise} \end{cases}$$

for $\alpha \in \{x, -x\}$. Since D is skew-symmetric, $D(H) = D^0(H)$ for any tour H in G. Also $D^0(H) = \pm x$ if and only if the subpath $j - \ell - i$ of H satisfies $|\{i, j\} \cap S| = 1$, and these are precisely the tours that satisfy $|C(H) - C^T(H)| = x$. We call such a tour a type I tour, and the remaining tours type II tours. Thus for every type II tour H, $C(H) - C^T(H) = 0$. Let $A = \frac{C+C^T}{2}$. Then for every type I tour H, $A(H) = C(H) \pm x/2$, and for every type II tour $A(H) = C(H)$.

Suppose we want to find a tour $\bar{H}$ in G such that $C(\bar{H})$ is minimum. Find a tour $\hat{H}$ such that $A(\hat{H})$ is minimum. If the tour is of type I, then the tour $\bar{H} \in \{\hat{H}, \hat{H}^*\}$ such that $C(\bar{H}) = A(\hat{H}) - x/2$ is the desired tour. (It may be recalled that $\hat{H}^*$ is the

reverse of $\hat{H}$.) If the tour $\hat{H}$ is of type II, then we need to find a minimum cost type I tour. This can be done as follows.

For each $u \in S$, define the matrix A^u with nondiagonal elements as follows:

$$a^u_{ij} = \begin{cases} a_{ij} - M/n & \text{if } i = \ell, j \in N - S - \{\ell\} \text{ or } i \in N - S - \{\ell\}, j = \ell, \\ -M & \text{if } i = \ell, j = u \text{ or } i = u, j = \ell, \\ a_{ij} & \text{otherwise.} \end{cases}$$

Let H^u be the minimum cost tour in G with respect to cost matrix A^u. Then it is easy to see that H^u is a minimum cost type I tour with respect to A containing the arc (ℓ, u). Choose v such that $A(H^v) = \min\{A(H^u) : u \in S\}$. Then, H^v is a type I tour with minimum cost with respect to A. If $C(\hat{H}) \leq A(H^v) - x/2$, then $\hat{H}$ is a minimum cost tour in G with respect to C. Else $\tilde{H} \in \{H^v, H^{v*}\}$ with $C(\tilde{H}) = A(H^v) - x/2$ is a minimum cost tour in G with respect to C.

The maximization case can be proved analogously. □

5. Conclusion. We have obtained an alternative characterization of DTC(2) matrices, and our proof of validity is relatively easy. Complete characterizations of DTC(2) matrices, skew-symmetric DTC(3) matrices, and skew-symmetric DPC(3) are given. These characterizations leads to new polynomially solvable special cases of the TSP. Our characterization of skew-symmetric matrices can be used to solve ATSPs with special structures as a sequence of at most $n/2$ closely related STSPs. We also identified special classes of ATSPs for which polynomial ϵ-approximation algorithms exist for constant ϵ.

An interesting and challenging question is to study characterization of general DTC(k) and DPC(k) matrices for $k \geq 3$, and we leave this question open.

Acknowledgment. We thank the referees for their constructive comments.

REFERENCES

[1] T. Andreae and H.-J. Bandelt, *Performance guarantees for approximation algorithms depending on parametrized triangle inequalities*, SIAM J. Discrete Math., 8 (1995), pp. 1–16.

[2] M. A. Bender and C. Chekuri, *Performance guarantees for the TSP with a parameterized triangle inequality*, Inform. Process. Lett., 73 (2000), pp. 17–21.

[3] A. Barvinok, E. Kh. Gimadi, and A. I. Serdyukov, *The maximum TSP*, in The Traveling Salesman Problem and Its Variations, G. Gutin and A. P. Punnen, eds., Kluwer Academic Publishers, Boston, Chapter 12, 2002, pp. 585–604.

[4] H.-J. Böckenhauer, J. Hromkovič, R. Klasing, S. Seibert, and W. Unger, *An improved lower bound on the approximability of metric TSP and approximation algorithms for the TSP with sharpened triangle inequality*, in Proceedings of STACS 2000, Lille, France, Lecture Notes in Comput. Sci. 1770, Springer, New York, 2000, pp. 382–394.

[5] R. Chandrasekaran, *Recognition of Gilmore-Gomory traveling salesman problem*, Discrete Appl. Math., 14 (1986), pp. 231–238.

[6] E. Y. Gabovich, *Constant discrete programming problems on substitution sets*, Translated from Kibernetika, 5 (1976), pp. 128–134 (in Russian).

[7] P. C. Gilmore, E. L. Lawler, and D. B. Shmoys, *Well solved special cases*, in The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization, E. L. Lawler, J. K. Lenstra, A. G. H. Rinnooy Kan, and D. B. Shmoys, eds., Wiley, New York, 1985.

[8] G. Gutin and A. Punnen, eds., *The Traveling Salesman Problem and Its Variations*, Kluwer Academic Publishers, Boston, 2002.

[9] O. Halskau, *Decompositions of Traveling Salesman Problems*, Ph.D. Thesis, Norwegian School of Economics and Business Administration, Bergen, Norway, 2000.

[10] R. Hassin and S. Rubinstein, *A 7/8-approximation algorithm for metric max TSP*, Inform. Process. Lett., 81 (2002), pp. 247–251.

[11] S. N. Kabadi, *Polynomially solvable cases of the TSP*, in The Traveling Salesman Problem and Its Variations, G. Gutin and A. P. Punnen, eds., Kluwer Academic Publishers, Boston, 2002, pp. 489–584.

[12] S. N. Kabadi and A. P. Punnen, *Weighted graphs with Hamiltonian cycles of same length*, Discrete Math., 271 (2003), pp. 129–139.

[13] H. Kaplan, M. Lewenstein, N. Shafrir, and M. Sviridenko, *Approximation algorithms for asymmetric TSP by decomposing directed regular multigraphs*, J. ACM, 52 (2005), pp. 602–626.

[14] R. M. Karp, *The fast approximate solution of hard combinatorial problems*, in Proceedings of 6th South Eastern Conference on Combinatorics, Graph Theory and Computing, Utilitas Mathematica, Winnipeg, 1975, pp. 15–21.

[15] A. V. Kostochka and A. I. Serdyukov, *Polynomial algorithms with estimates* 3/4 *and* 5/6 *for the traveling salesman problem of the maximum*, Upravlyaemye Sistemy, 26 (1985), pp. 55–59 (in Russian).

[16] S. Krynski, *Graphs in which all Hamiltonian cycles have the same length*, Discrete Appl. Math., 55 (1994), pp. 87–89.

[17] D. A. Kumar and C. P. Rangan, *Approximation algorithms for the traveling salesman problem with range condition*, Theoret. Inform. Appl., 34 (2000), pp. 173–181.

[18] E. L. Lawler, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart and Winston, New York, 1976.

[19] E. L. Lawler, J. K. Lenstra, A. H. G. Rinnoy Kan, and D. B. Shmoys, eds., *The Traveling Salesman Problem—A Guided Tour of Combinatorial Optimization*, Wiley, Chichester, UK, 1985.

[20] V. K. Leont'ev, *Investigation of an algorithm for solving the travelling salesman problem*, Zh. Vychisl. Mat. Mat. Fiz., 5 (1973).

[21] M. Lewenstein and M. Sviridenko, *A* 5/8 *approximation algorithm for the maximum asymmetric TSP*, SIAM J. Discrete Math., 17 (2003), pp. 237–248.

[22] A. P. Punnen, F. Margot, and S. N. Kabadi, *TSP heuristics: Domination analysis and complexity*, Algorithmica, 35 (2003), pp. 111–127.

[23] M. Queyranne and Y. Wang, *Hamiltonian path and symmetric travelling salesman polytopes*, Math. Program., 58 (1993), pp. 89–110.

[24] V. I. Rublinetskii, *Estimates of the accuracy of procedures in the travelling salesman problem*, Comput. Math. Computers, 4 (1973), pp. 11–15 (in Russian).

[25] A. I. Serdyukov, *An algorithm with an estimate for the traveling salesman problem of the maximum*, Upravlyaemye Sistemy, 25 (1984), pp. 80–86 (in Russian).

[26] S. P. Tarasov, *Properties of the trajectories of the appointments problem and the travelling salesman problem*, U.S.S.R. Comput. Maths. Math. Phys., 21 (1981), pp. 167–174.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 999–1014

TORIC SURFACE CODES AND MINKOWSKI SUMS*

JOHN LITTLE† AND HAL SCHENCK‡

Abstract. Toric codes are evaluation codes obtained from an integral convex polytope $P \subset \mathbb{R}^n$ and finite field $\mathbb{F}_q$. They are, in a sense, a natural extension of Reed–Solomon codes, and have been studied recently in [V. Diaz, C. Guevara, and M. Vath, *Proceedings of Simu Summer Institute*, 2001], [J. Hansen, *Appl. Algebra Engrg. Comm. Comput.*, 13 (2002), pp. 289–300; *Coding Theory, Cryptography and Related Areas* (Guanajuato, 1998), Springer, Berlin, pp. 132–142], and [D. Joyner, *Appl. Algebra Engrg. Comm. Comput.*, 15 (2004), pp. 63–79]. In this paper, we obtain upper and lower bounds on the minimum distance of a toric code constructed from a polygon $P \subset \mathbb{R}^2$ by examining *Minkowski sum* decompositions of subpolygons of P. Our results give a simple and unifying explanation of bounds in Hansen's work and empirical results of Joyner; they also apply to previously unknown cases.

Key words. toric variety, coding theory, Minkowski sum

AMS subject classifications. Primary 14G50; Secondary, 14M25, 94B27

DOI. 10.1137/050637054

1. Introduction.

In [8], J. Hansen introduced the notion of a toric surface code. Let $P \subset \mathbb{R}^2$ be an integral convex polygon and $\mathbb{F}_q$ a finite field such that, after translation, $P \cap \mathbb{Z}^2$ is properly contained in the square $[0, q-2] \times [0, q-2]$ with sides of length $q-1$, which we denote $\square_{q-1}$. Then a code is obtained by evaluating monomials with exponent vector in $P \cap \mathbb{Z}^2$ at some subset (usually all) of the points of $(\mathbb{F}_q^*)^2$. We formalize this as follows.

DEFINITION 1.1. *Let $\mathbb{F}_q$ be a finite field with primitive element ξ. For $0 \le i, j \le q-2$ let $P_{ij} = (\xi^i, \xi^j)$ in $(\mathbb{F}_q^*)^2$. For each $m = (m_1, m_2) \in P \cap \mathbb{Z}^2$, let*

$$e(m)(P_{ij}) = (\xi^i)^{m_1}(\xi^j)^{m_2}.$$

The toric code $C_P(\mathbb{F}_q)$ over the field $\mathbb{F}_q$ associated with P is the linear code of block length $n = (q-1)^2$ spanned by the vectors in $\{(e(m)(P_{ij}))_{0 \le i,j \le q-2} : m \in P \cap \mathbb{Z}^2\}$. If the field is clear from the context, we will often omit it in the notation and simply write C_P.

The properties of these codes are closely tied to the geometry of the toric surface X_P associated with the normal fan Δ_P of the polygon P. For example, intersection theory on X_P can be used to derive information about the minimum distance of toric codes. The monomials $e(m)$ which are evaluated to produce the generating codewords correspond to the lattice points $P \cap \mathbb{Z}^2$ and can be interpreted as sections of a certain line bundle on X_P. In [9], J. Hansen studies several specific families of polygons, depicted in Figure 1 (notice that some families are completely contained in others). The minimum distance for these codes is determined by exhibiting codewords of weight equal to a lower bound obtained from intersection theory.

*Received by the editors May 26, 2005; accepted for publication (in revised form) June 16, 2006; published electronically December 11, 2006.

http://www.siam.org/journals/sidma/20-4/63705.html

†Department of Mathematics and Computer Science, College of the Holy Cross, Worcester, MA 01610 (little@mathcs.holycross.edu).

‡Department of Mathematics, Texas A&M University, College Station, TX 77843-3368 (schenck@math.tamu.edu). The work of this author was partially supported by NSF DMS 03-11142, NSA MDA 904-03-1-0006, and ATP 010366-0103.

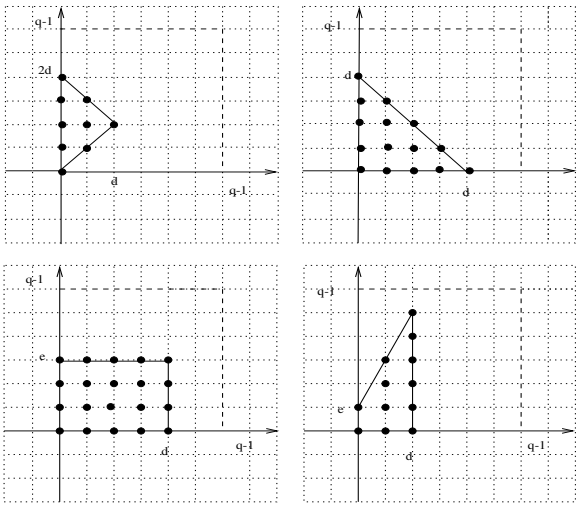


FIG. 1.

In this paper, we give upper and lower bounds on the minimum distance for toric surface codes. Our formulas generalize the results of [9] and also provide theoretical explanations for the some of the values tabulated in [12]. Codewords of small weight come from sections of the corresponding line bundle that have many zeroes in $(\mathbb{F}_q^*)^2$. A natural way to try to obtain these is to consider sections that factor into products of sections of related bundles (we will call these *reducible* sections in the following). Such reducible sections come from polygons $P' \subseteq P$ that decompose as *Minkowski sums* of other smaller polygons. The definition of the Minkowski sum of polytopes will be reviewed in section 2 below. In Proposition 2.3, we derive an upper bound on the minimum distance of a toric surface code when P has a subpolygon that decomposes as a Minkowski sum of other polygons. We then apply these methods in sections 3 and 4 to study the minimum distances of a number of examples, including all toric surface codes from smooth toric surfaces X with $\operatorname{rank}\operatorname{Pic}(X) = 2$ or 3.

In section 5, we derive a statement complementary to the upper bound of Proposition 2.3, giving a lower bound on the minimum distance of toric codes constructed from Minkowski-decomposable polygons. The Hasse–Weil bound on the number of $\mathbb{F}_q$-rational points on a curve shows that for any given polygon P there exists a lower bound on q such that reducible sections of the corresponding line bundle necessarily have more zeroes in $(\mathbb{F}_q^*)^2$ than irreducible sections. For precise statements here, see Proposition 5.2 and Corollary 5.3 below. This leads to our main theorem.

THEOREM 1.2. *Let $\mathbb{F}_q$ be a finite field, and let $P \subset \mathbb{R}^2$ be an integral convex polygon strictly contained in $\square_{q-1}$. Assume that q is sufficiently large (i.e., the bound (1) from Proposition 5.2 applies). Let ℓ be the largest positive integer such that there is some $P' \subseteq P$ that decomposes as a Minkowski sum $P' = P_1 + P_2 + \cdots + P_\ell$ with nontrivial P_i. Then there exists some $P' \subseteq P$ of this form such that*

$$d(C_P(\mathbb{F}_q)) \geq \sum_{i=1}^{\ell} d(C_{P_i}(\mathbb{F}_q)) - (\ell - 1)(q-1)^2.$$

We then apply this result to some additional, less straightforward, examples.

To relate our approach to other previous work, we note that two very general

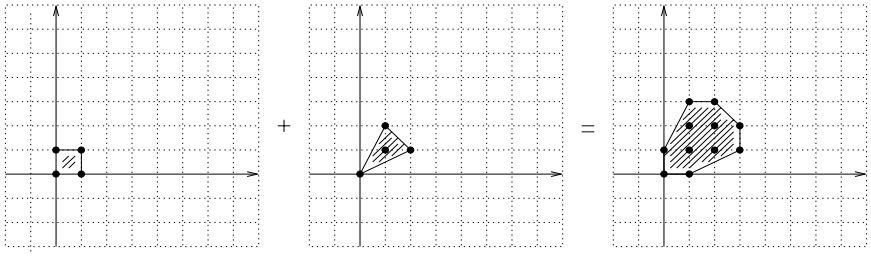


FIG. 2.

methods for obtaining bounds on the minimum distance of codes obtained from a higher dimensional variety X appear in recent work of S. Hansen [10]. The first method requires finding the multipoint Seshadri constant for the line bundle whose sections are evaluated to obtain the code. The second method consists of covering the $\mathbb{F}_q$-rational points of X with curves and then counting points on these curves via inclusion-exclusion; of course, this depends on being able to find "good" curves on X. The methods we introduce here depend on finding sections which factor, so they relate to the second technique.

The methods we use here make use of properties of toric surfaces in an essential way. First, a key fact about *complete* toric varieties is that all the higher cohomology of a globally generated line bundle vanishes. The lattice points in a polygon correspond to global sections of such a line bundle, so the Riemann–Roch theorem provides a relation (see section 5) between lattice points and intersection theory. We also make use of the Hasse–Weil bounds on the number of $\mathbb{F}_q$-rational points of a curve; to apply the formula we need the arithmetic genus of an irreducible section. The adjunction formula [7, p. 91] gives the arithmetic genus in terms of polytopal data.

2. Minkowski sums. In this section, we give a brief discussion of the Minkowski sum operation, referring to Ziegler [17] for more details. For facts on toric varieties, our basic references are Fulton [7] and Sturmfels [16].

DEFINITION 2.1. *Let P and Q be two subsets of $\mathbb{R}^n$. The Minkowski sum is obtained by taking the pointwise sum of P and Q:*

$$P + Q = \{x + y \mid x \in P, y \in Q\}.$$

We write conv to denote the convex hull of a set of points: the set of all convex combinations of the points.

Example 2.2. Let Q be the square $\mathrm{conv}\{(0,0),(1,0),(0,1),(1,1)\}$, and let P be the triangle $\mathrm{conv}\{(0,0),(1,2),(2,1)\}$. Then

$$P + Q = \mathrm{conv}\{(0,0),(1,0),(3,1),(3,2),(2,3),(1,3),(0,1)\},$$

as shown in Figure 2.

If f is a polynomial in two variables,

$$f(x,y) = \sum_{(a,b)\in\mathbb{Z}^2_{\geq 0}} c_{ab}x^a y^b,$$

then

$$NP(f) = \mathrm{conv}\{(a,b) : c_{ab} \neq 0\}$$

is called the *Newton polygon* of f. It is a direct consequence of the definition that if f, g are two polynomials, then $NP(fg) = NP(f) + NP(g)$, where the sum on the right is the Minkowski sum.

Similarly, in the language of toric surfaces, it is easy to see that if P_1 and P_2 are polygons, then the normal fan $\Delta_{P_1+P_2}$ is the common refinement of the fans Δ_{P_1} and Δ_{P_2}. Thus, the lattice points in $P_1 + P_2$ correspond to a basis of the global sections of a certain line bundle $\mathcal{O}(D)$ on the toric surface $X_{P_1+P_2}$, and the lattice points in P_1 and P_2 correspond to bases of global sections for two other line bundles $\mathcal{O}(D_1)$ and $\mathcal{O}(D_2)$ on $X_{P_1+P_2}$ (see [7, p. 67]). If D_1 and D_2 are divisors on the toric surface X corresponding to polygons P_1 and P_2 with $s_1 \in H^0(\mathcal{O}(D_1))$ and $s_2 \in H^0(\mathcal{O}(D_2))$, then

$$s_1 s_2 \in H^0(\mathcal{O}(D_1)) \otimes H^0(\mathcal{O}(D_2)) \subseteq H^0(\mathcal{O}(D_1 + D_2)),$$

which corresponds to the Minkowski sum $P_1 + P_2$. (Indeed, if the D_i are globally generated, then $H^0(\mathcal{O}(D_1)) \otimes H^0(\mathcal{O}(D_2)) = H^0(\mathcal{O}(D_1 + D_2))$; see [7, p. 69].) A good exercise for toric experts is to work this out for the previous example.

A first observation concerning the connection between the minimum distance of C_P and Minkowski sums is the following.

PROPOSITION 2.3. *Let $\sum_{i=1}^{\ell} P_i \subseteq P$, and let X be the toric surface corresponding to P. Let m_i be the maximum number of zeroes in $(\mathbb{F}_q^*)^2$ of a section of the line bundle on X corresponding to P_i, and assume that there exist sections s_i with sets of m_i zeroes that are pairwise disjoint in $(\mathbb{F}_q^*)^2$. Then*

$$d(C_P) \leq \sum_{i=1}^{\ell} d(C_{P_i}) - (\ell - 1)(q-1)^2.$$

Proof. By definition we have $d(C_{P_i}) = (q-1)^2 - m_i$. As noted above, $NP(fg) = NP(f) + NP(g)$, so the product $s = s_1 s_2 \cdots s_\ell$ is a section of the line bundle $\mathcal{O}(D)$ corresponding to $\sum_{i=1}^{\ell} P_i$. Moreover, s has exactly $m = m_1 + \cdots + m_\ell$ zeroes in $(\mathbb{F}_q^*)^2$ by hypothesis. There is a codeword of the toric code C_P with weight

$$w = (q-1)^2 - m = \sum_{i=1}^{\ell} d(C_{P_i}) - (\ell - 1)(q-1)^2,$$

obtained by evaluating s. Hence

$$d(C_P) \leq \sum_{i=1}^{\ell} d(C_{P_i}) - (\ell - 1)(q-1)^2,$$

which is what we wanted to show. □

Of course, the proof of the proposition can be extended to handle the case where pairs of the s_i have common zeroes in $(\mathbb{F}_q^*)^2$. However, the resulting bounds on $d(C_P)$ will involve the inclusion-exclusion principle and are harder to state in that generality. This upper bound also extends immediately to m-dimensional toric codes for all $m \geq 2$ (that is, toric codes constructed from polytopes $P \subset \mathbb{R}^m$).

3. First results and examples. In this section we will present several results on minimum distances of toric codes via Minkowski sum decompositions. These cases can be handled without using Theorem 1.2, and hence involve no hypothesis on q other than that needed to ensure $P \subset \square_{q-1}$.

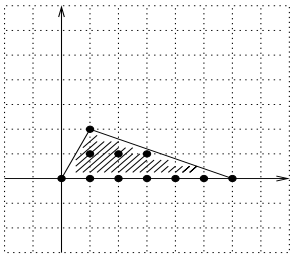


FIG. 3.

PROPOSITION 3.1. *Let* $P = \text{conv}\{(0,0),(a,0)\}$ *be a line segment* (*a polygon of dimension one*). *Then for all* $q > a+1$, $d(C_P(\mathbb{F}_q)) = (q-1)^2 - a(q-1)$.

Proof. The corresponding codes C_P are products of $q-1$ copies of a Reed–Solomon code, and the formula for the minimum distance follows directly. Note that P is also a Minkowski sum of a line segments of length 1. □

Our next result deals with the codes C_P from triangles, as in Figure 3.

PROPOSITION 3.2. *Let* P *be the integral triangle* $P = \text{conv}\{(0,0),(a,0),(b,c)\}$. *If* $a,b,c \geq 0$ *and* $a \geq b+c$, *then for all* $q > a+1$ (*so* $P \subset \square_{q-1}$),

$$d(C_P(\mathbb{F}_q)) = (q-1)^2 - a(q-1).$$

Proof. Note that C_P may be viewed as a subcode of the code C_{Δ_a}, where

$$\Delta_a = \text{conv}\{(0,0),(a,0),(0,a)\}.$$

The toric surface corresponding to the triangle Δ_a is the a-tuple Veronese embedding of $\mathbb{P}^2$. By a result of Serre [15], for all q the curve of degree a in $\mathbb{P}^2$ having the maximum possible number of $\mathbb{F}_q$-rational points is a reducible curve composed of a concurrent lines. When the point of intersection of the a lines lies at infinity or on one of the coordinate axes in the affine plane, then the corresponding curve has $a(q-1)$ $\mathbb{F}_q$-rational points in $(\mathbb{F}_q^*)^2$. Hence $d(C_P) \geq d(C_{\Delta_a}) = (q-1)^2 - a(q-1)$. Letting $P' = \text{conv}\{(0,0),(a,0)\}$, Proposition 2.3 shows that $d(C_P) \leq (q-1)^2 - a(q-1)$ as well. □

The code $C(\Delta_a)$ is also considered in [9], where the result $d(C_{\Delta_a}) = (q-1)^2 - a(q-1)$ is obtained in a different way.

If P' is any integral triangle obtained from P by a unimodular integer affine transformation (so P and P' are *lattice equivalent* polygons), then the same formula applies to give $d(C_{P'})$. This follows from the observation that if P and P' are lattice equivalent polygons, then C_P and $C_{P'}$ are monomially equivalent codes [13]. Propositions 3.1 and 3.2 give a large collection of "building blocks" to use in constructing other polygons. We illustrate this by considering a standard class of toric surfaces and toric codes studied in [9].

Example 3.3. If $P = \text{conv}\{(0,0),(d,0),(0,e),(d,e+rd)\}$ for some $r \in \mathbb{N}$, then P determines a *Hirzebruch surface*, denoted $\mathcal{H}_r$. We assume $e+dr < q-1$. The polygon P can be written as the Minkowski sum of a line segment $L = \text{conv}\{(0,0),(0,e)\}$ and a triangle $T = \text{conv}\{(0,0),(d,0),(d,rd)\}$; see Figure 4. We now apply our results to this $P = T + L$ to determine the minimum distance of $d(C_P)$. The triangle T is lattice equivalent to $\text{conv}\{(0,0),(rd,0),(0,r)\}$. By Proposition 3.2, for all q, $d(C_T) = (q-1)^2 - rd(q-1)$. (The reducible sections of the line bundle corresponding to T defined by $x^d \prod_{j=1}^{rd}(y-\alpha_j)$, α_j distinct in $\mathbb{F}_q^*$, have exactly $rd(q-1)$ zeroes in

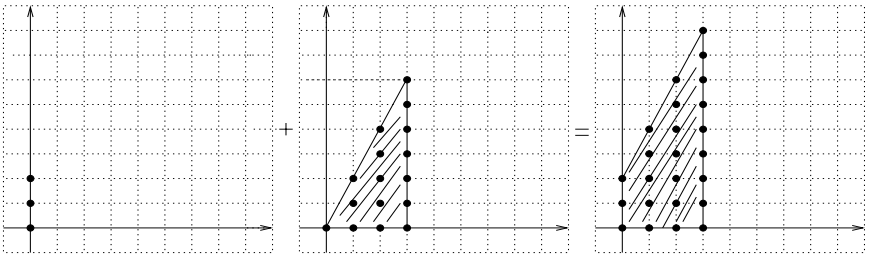


FIG. 4.

$(\mathbb{F}_q^*)^2$. The x^d corresponds to a trivial Minkowski summand and does not contribute to the minimum distance.) Similarly, Proposition 3.1 shows $d(C_L) = (q-1)^2 - e(q-1)$. Thus, by Proposition 2.3,

$$d(C_P) \leq (q-1)^2 - e(q-1) + (q-1)^2 - rd(q-1) - (q-1)^2 = (q-1)^2 - (rd+e)(q-1).$$

The polygon P is a subset of a polygon lattice equivalent to the equilateral triangle Δ_{rd+e}. Hence C_P is monomially equivalent to a subcode of $C_{\Delta_{rd+e}}$. It follows that the opposite inequality also holds, and hence

$$d(C_P) = (q-1)^2 - (rd+e)(q-1).$$

Theorem 1.5 of [9] gives $d(C_P)$ for the codes from the Hirzebruch surfaces $\mathcal{H}_r$ as the minimum of two terms. Since the first term given there is always larger than the second if $r > 0$, the minimum distance we obtain from the Minkowski sum decomposition agrees exactly with the value given in [9]. If $r = 0$, then the triangle T reduces to a horizontal line segment, and the Minkowski sum $T + L$ is a $d \times e$ rectangle. The corresponding toric code has minimum distance

$$d(C_P) = (q-1)^2 - (d+e)(q-1) + de$$

(see [9]). The minimum weight codewords come from evaluating reducible sections

$$\prod_{i=1}^{d}(x - \alpha_i) \prod_{j=1}^{e}(y - \beta_j),$$

where the α_i are distinct and the β_j are distinct in $\mathbb{F}_q^*$. Note that this is one case where the factors have common zeroes, so Proposition 2.3 does not apply directly.

For future reference, we note that by a result of Arkinstall [1], the only lattice polygons with no interior lattice points are triangles lattice equivalent to Δ_2 or to $\text{conv}\{(0,0),(p,0),(0,1)\}$ for some $p \geq 1$, or quadrilaterals with two parallel sides. Any such quadrilateral is lattice equivalent to one of the quadrilaterals defining a Hirzebruch surface with $d = 1$, or to a $1 \times e$ rectangle for some $e \geq 1$. Hence by our discussion in Example 3.3, we know $d(C_P)$ for all toric codes from polygons P with no interior lattice points.

4. Further examples: Smooth surfaces with rank $\mathbf{Pic}(X) = 3$. The Hirzebruch surfaces from Example 3.3 satisfy rank $\text{Pic}(\mathcal{H}_r) = 2$ and, up to isomorphism,

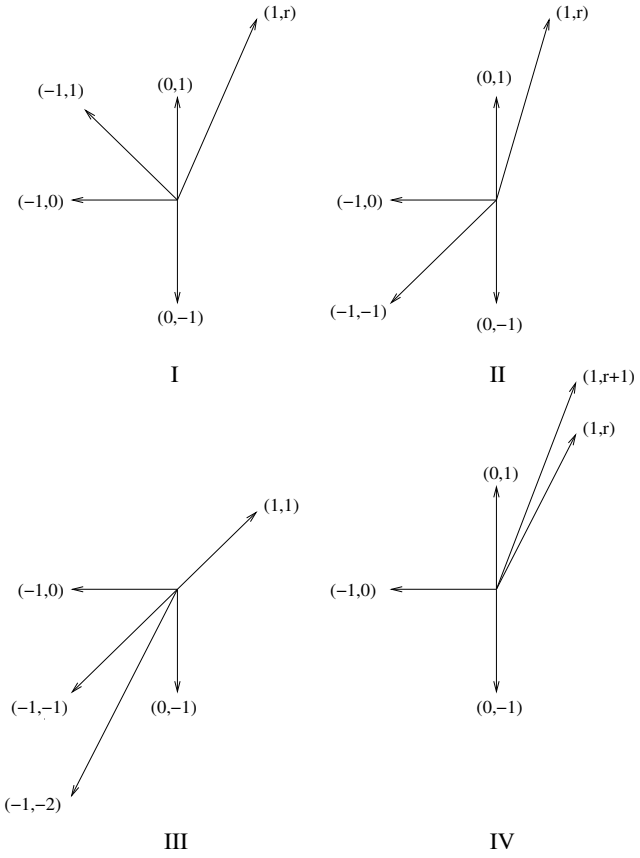


FIG. 5.

account for all smooth toric surfaces with this property. In this section, we work out another extended family of examples and study the toric codes from the next most complicated toric surfaces, those with rank $\mathrm{Pic}(X) = 3$. We will use some facts about toric surfaces, and refer to section 2.5 of [7] for proofs. Recall that any smooth complete toric surface X may be obtained from $\mathbb{P}^2$ or some $\mathcal{H}_r$ by a succession of blow-ups at torus-fixed points. The Picard number of such a surface is $n-2$, where n is the number of one-dimensional cones in the fan defining X. This description makes it reasonably straightforward to write down the fans for all smooth complete toric surfaces with rank $\mathrm{Pic}(X) = 3$; either we add a single ray to the fan of $\mathcal{H}_r$ or a pair of rays to the fan for $\mathbb{P}^2$, in such a way that for any two adjacent rays the determinant of the corresponding two-by-two matrix is ± 1. The possibilities appear in Figure 5.

These fans are the outer normal fans of families of polygons. Polygons with these normal fans can "scale" in size; for example, the fan with rays $\{(\pm 1, 0), (0, \pm 1)\}$ is the normal fan for any rectangle of the form $\mathrm{conv}\{(0,0), (a,0), (0,b), (a,b)\}$. In other words, the polytopes vary with parameters. We will see in a moment that these polygons all have Minkowski sum decompositions as sums of triangles and lines.

For each fan, we want to determine the polygons whose edges are normal to the given rays in the fan. For the fan I pictured in Figure 5, polygons with this outer

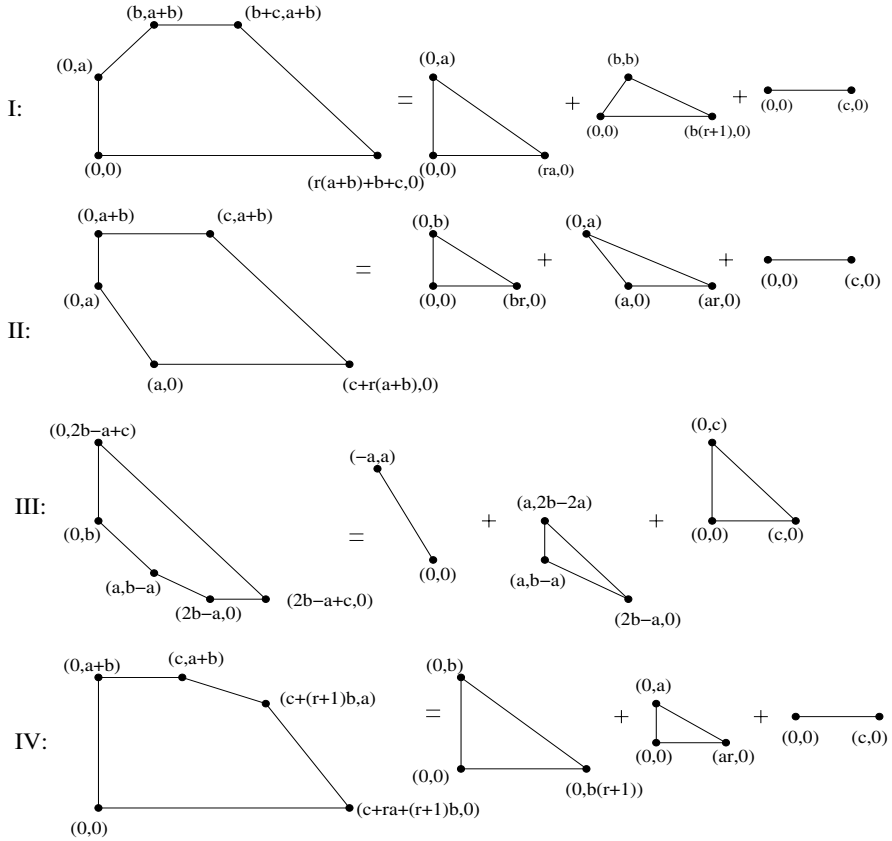


FIG. 6.

normal fan are obtained as the sets of solutions of inequalities as follows:

$$\begin{aligned}(1,r)\cdot(x,y) &\geq \alpha,\\ (0,1)\cdot(x,y) &\geq \beta,\\ (-1,1)\cdot(x,y) &\geq \gamma,\\ (-1,0)\cdot(x,y) &\geq \delta,\\ (0,-1)\cdot(x,y) &\geq \varepsilon,\end{aligned}$$

for some $\alpha, \beta, \gamma, \delta, \varepsilon \geq 0$. Taking $\delta = \varepsilon = 0$, $\gamma = a > 0$, $\beta = a + b$ with $b > 0$, and $\alpha = r(a + b) + b + c$ with $c > 0$, we get a polygon as in Figure 6.

Now we are ready to examine Minkowski sum decompositions for polygons corresponding to the fans in Figure 5. For instance, in case I, we find that the pentagon

$$P = \operatorname{conv}\{(0,0), (r(a+b)+b+c, 0), (b+c, a+b), (b, a+b), (0,a)\}$$

can be decomposed as a Minkowski sum of the triangles

$$P_1 = \operatorname{conv}\{(0,0), (ra,0), (0,a)\}, \quad P_2 = \operatorname{conv}\{(0,0), (b(r+1),0), (b,b)\}$$

and the line segment $P_3 = \operatorname{conv}\{(0,0), (c,0)\}$. There are similar decompositions in each of the other cases as well.

THEOREM 4.1. *Consider toric surface codes corresponding to the families of polygons* I, II, III, IV *above, where* $a, b, c, r \geq 1$ *are integers and* q *is sufficiently large so that the polygon is contained in* $\square_{q-1}$.

1. *In case* I, *for all such* q,
$$d(C_P) = (q-1)^2 - (r(a+b)+b+c)(q-1).$$
2. *In case* II, *for all such* q,
$$d(C_P) = (q-1)^2 - m(q-1),$$
where $m = \max\{a+b, c+(r-1)a+rb\}$.
3. *In case* III, *if* $b > a$ *as in Figure* 6, *then*
$$d(C_P) = (q-1)^2 - (2b+c-a)(q-1).$$
4. *In case* IV, *for all such* q,
$$d(C_P) = (q-1)^2 - (c+ra+(r+1)b)(q-1).$$

Proof. We sketch how the value in case I can be established using the methods presented in sections 2 and 3. We see first that the stated value is an upper bound for $d(C_P)$ using the Minkowski sum decomposition given in Figure 6, Proposition 2.3, Proposition 3.1 for the line segment, and Proposition 3.2 for the triangles. Then, the fact that the given value for $d(C_P)$ is the exact minimum distance follows, as in Example 3.3. The polygon here is contained in the equilateral triangle $\Delta_{r(a+b)+b+c}$. Hence the minimum distance for C_P is bounded below by the minimum distance for the code $C_{\Delta_{r(a+b)+b+c}}$. We leave it as an exercise for the reader to provide detailed proofs for the other parts. In each case, the minimum weight codewords come from evaluation of reducible sections of the corresponding line bundles. For instance, the sections of $\mathcal{O}(D)$ with the maximal number of $\mathbb{F}_q$-rational points in case III are given by $(y-\alpha_1 x)\cdots(y-\alpha_{2b+c-a}x)$ with $\alpha_i \in \mathbb{F}_q^*$ distinct. □

5. Main theorem. In this section we prove our main result, Theorem 1.2. The essential idea is to combine the Minkowski sum construction with the Hasse–Weil bounds on the number of $\mathbb{F}_q$-rational points of a curve: If Y is a smooth, absolutely irreducible curve over $\mathbb{F}_q$, then
$$1+q-2g\sqrt{q} \leq |Y(\mathbb{F}_q)| \leq 1+q+2g\sqrt{q},$$
where g is the genus of Y. In [2], the same inequalities are demonstrated for absolutely irreducible but possibly singular curves, provided that g is interpreted as the *arithmetic genus* of Y.

The intuition behind our results is quite simple: From the Hasse–Weil bound, if P is fixed and q is sufficiently large, then sections which are reducible must have more zeroes than irreducible sections.

We will use the following notation. For a polygon P, $v(P)$ will denote the area (two-dimensional volume) of P, $\#(P) = |P \cap \mathbb{Z}^2|$ will denote the number of lattice points in P, $\partial(P)$ will denote the number of lattice points in the boundary of P, and $I(P) = \#(P) - \partial(P)$ will denote the number of lattice points in the interior of P. Pick's theorem for lattice polygons in $\mathbb{R}^2$ is the equality
$$v(P) = \#(P) - \frac{1}{2}\partial(P) - 1.$$

Recall that we have seen that all polygons P with $I(P) = 0$ correspond to toric surfaces for which the minimum distance of C_P is known by results from sections 2 and 3. Hence, in the following we will assume $I(P) > 0$.

In section 2 we noted that if $\mathcal{O}(D_i),\ i \in \{1, \dots, n\}$, are globally generated line bundles on a toric surface, then the global sections of $\mathcal{O}(\sum D_i)$ correspond to the Minkowski sum of the polygons P_i defined by $H^0(\mathcal{O}(D_i))$. Our starting data is a lattice polygon P, and to find reducible sections, our strategy is to work backwards: We look for Minkowski sums $\sum_{i=1}^{n} P_i = P' \subseteq P$ with n large.

In order to use algebraic geometry, we will first pass to a smooth surface. The toric surface X_Δ defined by the outer normal fan Δ to P need not be smooth. However, we can refine Δ to a fan Δ' such that $X_{\Delta'}$ is smooth, and the line bundle $\mathcal{O}(D)$ on $X_{\Delta'}$ corresponding to P is generated by global sections (see [7, p. 90] or [4]). The numerical invariants D^2 and DK discussed in the next paragraphs have simple interpretations on the smooth surface $X_{\Delta'}$; most importantly, they depend only on P.

Finally, when we deal with subpolygons P_i of P, in order to make the same set-up work, we will refine the fan Δ' to include the outer normals to P_i, and then further subdivide the result (for smoothness) to a fan Δ''. The key point is that ([7, p. 73]) the P_i correspond to globally generated line bundles on the smooth surface $X_{\Delta''}$. So henceforth we will be working with globally generated line bundles on a smooth toric surface.

PROPOSITION 5.1. *Let X be a smooth toric surface, and $K = K_X$ a canonical divisor. Let C be an irreducible curve on X of arithmetic genus g_C such that the corresponding line bundle is globally generated, with P the polytope corresponding to $H^0(\mathcal{O}(C))$. Then the following hold:*

1. $g_C = \frac{C^2+CK}{2} + 1$.
2. $h^0(\mathcal{O}(C)) = \frac{C^2-CK}{2} + 1$.
3. $g_C = 2v(P) + 2 - \#(P) = I(P)$.

Proof. The first formula is simply adjunction; see [11, V.1.5] or [7, p. 91]. Since all the higher cohomology of a globally generated line bundle on a toric variety vanishes, and because a toric surface is rational, if $\mathcal{O}(C)$ is globally generated, then the Riemann–Roch theorem for surfaces ([11, V.1.6]) yields the second formula. Adding the first two formulas shows that $h^0(\mathcal{O}(C)) + g_C = C^2 + 2$. Since $h^0(\mathcal{O}(C)) = \#(P)$ and $C^2 = 2v(P)$ (see [7, p. 111]), the last formula follows from Pick's theorem. ∎

One other fact that will be useful for us is that on a smooth toric surface X, the anticanonical divisor class $-K$ is given by the sum of the divisors corresponding to the one-dimensional cones in the fan defining X [7, p. 85]. Now,

$$(\mathbb{F}_q^*)^2 = X \setminus \bigcup_{\tau \neq \{0\}} V(\tau),$$

where $V(\tau)$ is the closure of the torus orbit of the cone $\tau \subseteq \Delta$; see [7, section 3.1]. In particular, a toric surface decomposes as the union of a two-dimensional torus with a finite set of curves, which correspond exactly to the rays of Δ. Hence, the intersection number $-KC$ accounts for points on C in the complement of the torus in X.

Our first result shows that if q is sufficiently large, then reducible sections with more irreducible components necessarily have more zeroes in $(\mathbb{F}_q^*)^2$ than sections with fewer irreducible components. In what follows, we write $V(s)$ for the zero locus of a section s.

PROPOSITION 5.2. *Let P be a lattice polygon in $\mathbb{R}^2$ with $I(P) > 0$, and let $P' = \sum_{i=1}^{m} P_i'$ and $P'' = \sum_{k=1}^{\ell} P_k''$ (with P_i' and P_k'' nontrivial) be two polygons*

contained in P. Let X be a smooth toric surface obtained by refining the normal fan Δ of P as described above, so that P' and P'' correspond to line bundles $\mathcal{O}(D')$ and $\mathcal{O}(D'')$ on X. Let $s' = s'_1 s'_2 \dots s'_m \in H^0(\mathcal{O}(D'))$ and $s'' = s''_1 s''_2 \dots s''_\ell \in H^0(\mathcal{O}(D''))$ be reducible sections with $V(s'_i)$ and $V(s''_k)$ irreducible. If $m > \ell$ and

$$q \geq (4I(P) + 3)^2, \tag{1}$$

then

$$|V(s') \cap (\mathbb{F}_q^*)^2| > |V(s'') \cap (\mathbb{F}_q^*)^2|.$$

Proof. Let D'_i be the divisor corresponding to $V(s'_i)$, and D''_k be the divisor corresponding to $V(s''_k)$. We write $g_i = g(D'_i)$ and $g''_k = g(D''_k)$. Our starting point is the observation that

$$|V(s') \cap (\mathbb{F}_q^*)^2| \geq \sum_{i=1}^{m} \left((q+1) - 2g'_i\sqrt{q} \right) - \sum_{i<j} D'_i D'_j + D'K.$$

This follows because

$$|V(s') \cap (\mathbb{F}_q^*)^2| = \sum_{i=1}^{m} |V(s'_i)| - T - B,$$

where T is the number of common intersection points of the curves inside the torus $(\mathbb{F}_q^*)^2$ and B is the number of points of D' in the "boundary" $X \setminus (\mathbb{F}_q^*)^2$. Since the number of common intersection points of D'_i and D'_j is the intersection number $D'_i D'_j$, $T \leq \sum_{i<j} D'_i D'_j$. As noted earlier, the number of points of D' outside the torus is $-D'K$, so that $B \leq -D'K$ (note that $D'_i D'_j$ and $-D'K$ do not distinguish $\mathbb{F}_q$ rational points, so they may well overcount). Substituting the Hasse–Weil lower bound $|V(s'_i)| \geq q + 1 - 2g'_i\sqrt{q}$ gives the result. Similarly, by the Hasse–Weil upper bound,

$$\sum_{k=1}^{\ell} \left((q+1) + 2g''_k\sqrt{q} \right) \geq |V(s'') \cap (\mathbb{F}_q^*)^2|.$$

Hence if q satisfies

$$(m-\ell)(q+1) > 2\left(\sum_i g'_i + \sum_k g''_k \right)\sqrt{q} + \sum_{i<j} D'_i D'_j - D'K, \tag{2}$$

then the conclusion of the proposition follows. Write

$$\beta = \frac{1}{m-\ell}\left(\sum_i g'_i + \sum_k g''_k \right).$$

By Proposition 5.1.3, $g'_i = I(P'_i)$, and so

$$\sum_i g'_i = \sum_i I(P'_i) \leq I(P') \leq I(P),$$

and similarly for $\sum_i g''_i$. Because $m - \ell \geq 1$, we see that

$$\beta \leq \frac{2}{m-\ell} I(P) \leq 2I(P). \tag{3}$$

The inequality (2) is quadratic in $\sqrt{q}$, so by the quadratic formula, (2) will hold if

$$\begin{aligned}\sqrt{q} &> \beta + \sqrt{\beta^2 + \sum_{i<j} D_i' D_j' - D'K + 1} \\ &\geq \beta + \sqrt{\beta^2 + \frac{1}{m-\ell}\left(\sum_{i<j} D_i' D_j' - D'K\right) + 1}.\end{aligned}$$

Since $D' = \sum D_i'$, we have

$$\sum_{i<j} D_i' D_j' = \frac{(D')^2 - \sum_i (D_i')^2}{2}. \tag{4}$$

Now we apply (4) and Proposition 5.1:

$$\begin{aligned}\sum_{i<j} D_i' D_j' - D'K + 1 &= \frac{(D')^2 - D'K}{2} - \frac{\sum_i \left((D_i')^2 + D_i'K\right)}{2} + 1 \\ &= h^0(\mathcal{O}(D')) - \sum_i g_i' + m \\ &= (\#(P') - \sum_i I(P_i')) + m \\ &\leq 2\#(P).\end{aligned}$$

The last step follows because $m \leq \#(P)$ (each time we add in a new Minkowski summand, we get at least one new lattice point in the Minkowski sum), and because $(\#(P') - \sum_i I(P_i')) \leq \#(P') \leq \#(P)$. Now we use the theorem of Scott [14],

$$\#(P) \leq 3I(P) + 7,$$

for a lattice polygon P such that $I(P) > 0$. From the above, we see

$$\sum_{i<j} D_i' D_j' - D'K + 1 \leq 6I(P) + 14.$$

Hence if the lower bound (1) holds, since $I(P) > 0$ we have

$$\begin{aligned}\sqrt{q} &\geq 2I(P) + 2I(P) + 3 \\ &= 2I(P) + \sqrt{4I(P)^2 + 12I(P) + 9} \\ &> 2I(P) + \sqrt{4I(P)^2 + 6I(P) + 14} \\ &\geq \beta + \sqrt{\beta^2 + \sum_{i<j} D_i' D_j' - D'K + 1} \quad \text{by (3)},\end{aligned}$$

which is what we wanted to show. □

A number of very crude estimates were used to show that (1) implies the conclusion here. Our lower bound on q will rarely be sharp. Much smaller lower bounds on q can be obtained if we know more about possible factorizations of sections of $\mathcal{O}(D)$. For instance, we have the following statement.

COROLLARY 5.3. *In the situation of Proposition* 5.2*, suppose that* $g_i' = I(P_i') = 0$ *and* $g_k'' = I(P_k'') = 0$ *for all* i, k*. Then the conclusion of Proposition* 5.2 *holds for all* $q > \#(P) + m$*.*

Proof. In this case $\beta = 0$ in the proof of Proposition 5.2. □

Theorem 1.2 follows almost immediately from Proposition 5.2.

Proof of Theorem 1.2. Let $d = d(C_P)$. Given P, the proposition shows that under the hypothesis (1) on q, the number of zeroes of a section can always be increased by finding a reducible section in $H^0(\mathcal{O}(D))$ with more nontrivial factors, if there is one. Hence the sections with the largest number of zeroes in $(\mathbb{F}_q^*)^2$ must come from nontrivial factorizations with the largest possible number of factors. Say $s = s_1 s_2 \cdots s_m$ is a nonzero section with the maximum number of zeroes $(q-1)^2 - d$. Then, counting the number of zeroes,

$$(q-1)^2 - d \le \sum_{i=1}^{m} m_i,$$

where m_i is the number of zeroes of s_i. We have $d(C_{P_i}) \le (q-1)^2 - m_i$ for each i. Hence

$$\sum_{i=1}^{m} m_i \le m(q-1)^2 - \sum_{i=1}^{m} d(C_{P_i}).$$

Rearranging the inequalities gives

$$d \ge \sum_{i=1}^{m} d(C_{P_i}) - (m-1)(q-1)^2,$$

as claimed. □

We have not tried to account for common zeroes of the s_i in the proof of the theorem. Moreover, in applying this statement, it is important to realize that there may be several different subpolygons with the maximal number of Minkowski summands. The bound in Theorem 1.2 is only guaranteed to hold for the one that minimizes

$$\sum_{i=1}^{m} d(C_{P_i}) - (m-1)(q-1)^2.$$

Example 5.4. Consider the polygon

$$P = Q_1 + Q_2 := \operatorname{conv}\{(0,0),(1,1),(2,1),(1,2)\} + \operatorname{conv}\{(0,0),(1,0)\}.$$

Taking $P' = P$ and

$$P'' = P_1 + P_2 := \operatorname{conv}\{(1,1),(1,2)\} + \operatorname{conv}\{(0,0),(1,0)\} \subset P$$

gives two different Minkowski-decomposable subpolygons of P with the same number $m = 2$ of nontrivial summands. However, since $I(Q_1) = 1$, the sections having Newton polygon equal to Q_1 have arithmetic genus 1 and can have more zeroes in $(\mathbb{F}_q^*)^2$ than the rational curves corresponding to the summands in P''. So in applying Theorem 1.2 to this example, we should use the decomposition $P = Q_1 + Q_2$ rather than $P'' = P_1 + P_2$. In fact, we see this already for fields such as $\mathbb{F}_8$, where q is much

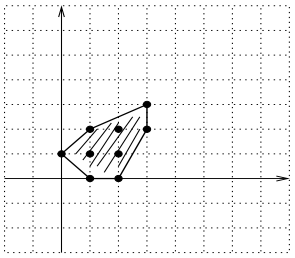


FIG. 7.

smaller than the bound from Proposition 5.2. Indeed, by a Magma computation using the routines from [12],

$$d(C_P(\mathbb{F}_8)) = 33,$$

while $\sum_{i=1}^{2} d(C_{Q_i}(\mathbb{F}_8)) - (q-1)^2 = 33$ and $\sum_{i=1}^{2} d(C_{P_i}(\mathbb{F}_8)) - (q-1)^2 = 35$.

Next we will show that our results shed some additional light on the good examples of toric surface codes tabulated in [12].

Example 5.5. In Example 3.9 of [12], Joyner gives an example of a toric code over $\mathbb{F}_8$ with $k = 11$ and $d = 28$. These parameters were better than any known code in Brouwer's tables [3] at the time his article was written. The convex hull of the integral points is a triangle $P = \text{conv}\{(0,0),(1,4),(4,1)\}$. Note that P contains a translate of the triangle Δ_3. Applying Propositions 2.3 and 3.2, we obtain $d(C_P(\mathbb{F}_q)) \leq (q-1)^2 - 3(q-1)$ for all q, so $d(C_P(\mathbb{F}_8)) \leq 28$. The lower bound $d(C_P(\mathbb{F}_q)) \geq (q-1)^2 - 3(q-1)$ also holds for q sufficiently large, by Theorem 1.2. Joyner's computations show that this bound on d is also valid for $q = 8$, but our general statements are not quite strong enough to prove this.

The following example gives an indication of some additional interesting behavior that can occur for small q.

Example 5.6. Consider the polygon pictured in Figure 7:

$$P = \text{conv}\{(1,0),(2,0),(0,1),(1,2),(3,2),(3,3)\}.$$

Note that $P \subset \square_{q-1}$ for all $q \geq 5$. We see that P contains a pair of Minkowski-decomposable subpolygons: the 1×2 rectangle $P' = \text{conv}\{(1,0),(2,0),(1,2),(2,2)\}$ and the 2×1 parallelogram $P'' = \text{conv}\{(1,0),(1,1),(3,2),(3,3)\}$. P' can be written as the Minkowski sum of two vertical line segments of length 1 and a horizontal line segment of length 1. Each P_i gives $d(C_{P_i}) = (q-1)^2 - (q-1)$. P'' has a similar decomposition with three summands. There are no other Minkowski-decomposable subpolygons of P with more than three Minkowski summands, and there are no Minkowski summands with interior lattice points. Hence we have

$$d(C_P(\mathbb{F}_q)) \geq (q-1)^2 - 3(q-1)$$

for $q > \#(P) + 3 = 12$ by Corollary 5.3.

Both of these subpolygons give rise to reducible sections of the corresponding line bundles. For instance, from P' we obtain reducible sections of the form $s = x(x-a)(y-b)(y-c)$. If $a, b, c \in \mathbb{F}_q^*$ and $b \neq c$, then s has $3(q-1) - 2$ zeroes in $(\mathbb{F}_q^*)^2$. Hence, by reasoning like that used in the proof of Proposition 2.3 (but in the case where the factors do have some common zeroes), we have

$$d(C_P(\mathbb{F}_q)) \leq (q-1)^2 - 3(q-1) + 2.$$

Computations using Magma show that

$$\begin{aligned} d(C_P(\mathbb{F}_5)) &= 6 \text{ vs. } 4^2 - 3 \cdot 4 + 2 = 6, \\ d(C_P(\mathbb{F}_7)) &= 20 \text{ vs. } 6^2 - 3 \cdot 6 + 2 = 20, \\ d(C_P(\mathbb{F}_8)) &= 28 \text{ vs. } 7^2 - 3 \cdot 7 + 2 = 30, \\ d(C_P(\mathbb{F}_9)) &= 42 \text{ vs. } 8^2 - 3 \cdot 8 + 2 = 42, \\ d(C_P(\mathbb{F}_{11})) &= 72 \text{ vs. } 10^2 - 3 \cdot 10 + 2 = 72. \end{aligned}$$

The dimension is $k = \#(P) = 9$ in each case.

The case $q = 8$ is the most interesting one here. We may ask: Where does a section with $49 - 28 = 21$ zeroes in $(\mathbb{F}_8^*)^2$ come from? By examining the minimum weight codewords of this code, we find exactly 49 such words. One of them comes, for instance, from the evaluation of

$$\begin{aligned} x + x^3y^3 + y^2 &\equiv x(1 + x^2y^3 + x^6y^2) \bmod \langle x^7 - 1, y^7 - 1 \rangle \\ &\equiv x(1 + x^2y^3 + (x^2y^3)^3) \bmod \langle x^7 - 1, y^7 - 1 \rangle. \end{aligned}$$

Here $\langle x^7 - 1, y^7 - 1 \rangle$ is the ideal of the $\mathbb{F}_8$-rational points of the two-dimensional torus. So $1 + x^2y^3 + (x^2y^3)^3$ has exactly the same zeroes in $(\mathbb{F}_8^*)^2$ as $x + x^3y^3 + y^2$. Recall that $1 + u + u^3$ is one of the two irreducible polynomials of degree 3 in $\mathbb{F}_2[u]$, and hence $\mathbb{F}_8 \cong \mathbb{F}_2[u]/\langle 1 + u + u^3 \rangle$. Hence if β is a root of $1 + u + u^3 = 0$ in $\mathbb{F}_8$, then

$$1 + x^2y^3 + (x^2y^3)^3 = (x^2y^3 - \beta)(x^2y^3 - \beta^2)(x^2y^3 - \beta^4),$$

and there are exactly $3 \cdot 7 = 21$ points in $(\mathbb{F}_8^*)^2$ where this is zero. It is interesting to note that it is still a sort of reducibility that is producing a section with the largest number of zeroes here, even though the reducibility only appears when we look modulo the ideal $\langle x^7 - 1, y^7 - 1 \rangle$. We also note that these minimum weight codewords come from curves with many rational points over the field $\mathbb{F}_8$ as in the construction used in [5]. Similar phenomena will occur for many other P with q small.

Acknowledgments. This collaboration began while both authors were members of Mathematical Sciences Research Institute during the commutative algebra program in 2003. We also thank the Institute for Scientific Computation at Texas A&M for logistical support, and two anonymous referees for careful readings and suggestions.

REFERENCES

[1] J. Arkinstall, *Minimal requirements for Minkowski's theorem in the plane*, Bull. Austral. Math. Soc., 22 (1980), pp. 259–283.

[2] Y. Aubry and M. Perret, *A Weil theorem for singular curves*, in Arithmetic, Geometry, and Coding Theory, R. Pellikaan, M. Perret, and S. G. Vladut, eds., de Gruyter, Berlin, 1996, pp. 1–7.

[3] A. E. Brouwer, *Bounds on linear codes*, in Handbook of Coding Theory, Elsevier, New York 1998, pp. 295–461; updates online at http://www.win.tue.nl/∼aeb/voorlincod.html.

[4] T. Beck and J. Schicho, *Sparse parametrization of plane curves*, Appl. Algebra Engrg. Comm. Comput., to appear.

[5] P. Beelen and R. Pellikaan, *The Newton polygon of plane curves with many rational points*, Des. Codes Cryptogr., 21 (2000), pp. 41–67.

[6] V. Diaz, C. Guevara, and M. Vath, *Codes from n-Dimensional Polyhedra and n-Dimensional Cyclic Codes*, in Proceedings of the SIMU Summer Institute, 2001.

[7] W. Fulton, *Introduction to Toric Varieties*, Princeton University Press, Princeton, NJ, 1993.

[8] J. Hansen, *Toric surfaces and error-correcting codes*, in Coding theory, Cryptography and Related Areas (Guanajuato, 1998), Springer, Berlin, 2000, pp. 132–142.

[9] J. Hansen, *Toric varieties Hirzebruch surfaces and error-correcting codes*, Appl. Algebra Engrg. Comm. Comput., 13 (2002), pp. 289–300.
[10] S. Hansen, *Error-correcting codes from higher-dimensional varieties*, Finite Fields Appl., 7 (2001), pp. 531–552.
[11] R. Hartshorne, *Algebraic Geometry*, Springer, New York, 1977.
[12] D. Joyner, *Toric codes over finite fields*, Appl. Algebra Engrg. Comm. Comput., 15 (2004), pp. 63–79.
[13] J. Little and R. Schwarz, *On m-dimensional toric codes*, preprint, available online from http://arxiv.org/abs/cs.IT/0506102.
[14] P. R. Scott, *On convex lattice polygons*, Bull. Austral. Math. Soc., 15 (1976), pp. 395–399.
[15] J. P. Serre, *Lettre à M. Tsfasman*, Journées Arithmétiques, 1989 (Luminy, 1989), Astérisque, 198-200 (1991), pp. 351–353.
[16] B. Sturmfels, *Gröbner Bases and Convex Polytopes*, AMS University Lectures Series, Vol. 8, AMS, Providence, RI, 1995.
[17] G. Ziegler, *Lectures on Polytopes*, Springer-Verlag, Berlin, 1995.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 1015–1030

ONLINE BIN PACKING WITH CARDINALITY CONSTRAINTS*

LEAH EPSTEIN†

Abstract. We consider a one-dimensional storage system where each container can store a bounded amount of capacity as well as a bounded number of items $k \geq 2$. This defines the (standard) bin packing problem with cardinality constraints, which is an important version of bin packing. Following previous work on the unbounded space online problem, we establish the *exact* best competitive ratio for bounded space online algorithms for every value of k. This competitive ratio is a strictly increasing function of k which tends to $\Pi_\infty + 1 \approx 2.69103$ for large k. Lee and Lee showed in 1985 [*J. ACM*, 32 (1985), pp. 562–572] that the best possible competitive ratio for online bounded space algorithms for the classical bin packing problem is the sum of a series, and tends to Π_∞ as the allowed space (number of open bins) tends to infinity. We further design optimal online bounded space algorithms for *variable sized bin packing*, where each allowed bin size may have a distinct cardinality constraint, and for the *resource augmentation* model. All algorithms achieve the *exact* best possible competitive ratio possible for the given problem and use constant numbers of open bins. Finally, we introduce unbounded space online algorithms with smaller competitive ratios than the previously known best algorithms for small values of k, for the standard cardinality constrained problem. These are the first algorithms with competitive ratio below 2 for $k = 4, 5, 6$.

Key words. online algorithms, bin packing, cardinality constraints

AMS subject classifications. 68Q25, 68W40

DOI. 10.1137/050639065

1. Introduction.

The classical bin packing problem [20, 5, 3] assumes no limit on the *number* of items which may be packed into a single bin. In practice, many applications require such a bound either due to overheads or additional constraints that are not modeled. For example, a disk cannot keep more than a certain number of files, even if these files are indeed very small. A processor cannot run more than a given number of tasks during a given time, even if all tasks are very short. The problem where there is a given bound $k > 1$ on the number of items which can coexist in one bin is called "bin packing with cardinality constraints" [12, 1]. We consider several versions of this problem.

We first define the classical online bin packing problem. In this problem, we receive a sequence σ of *items* $p_1, p_2 \dots p_n$, arriving one by one. The values p_i are the *sizes* of the items. We have an infinite supply of *bins*, each of which is of unit size. An item must be assigned to a bin upon arrival, so that the sum of items in no bin exceeds 1. A bin is *empty* if no item is assigned to it; otherwise it is *used*. The goal is to minimize the number of bins used. In the *cardinality constrained bin packing problem*, an additional constraint is introduced. A parameter k bounds the number of items that can be assigned to a single bin.

The standard measure of algorithm quality for online bin packing is the *asymptotic competitive ratio*, which we now define. For a given input sequence σ, let $\mathcal{A}(\sigma)$ (or $\mathcal{A}$) be the number of bins used by algorithm $\mathcal{A}$ on σ. Let $OPT(\sigma)$ (or OPT) be the cost of an optimal offline algorithm which knows the complete sequence of items in

*Received by the editors August 28, 2005; accepted for publication (in revised form) June 5, 2006; published electronically December 15, 2006. A preliminary version of this paper appeared in the Proceedings of the 13th European Symposium on Algorithms, Palma de Mallorca, Spain, 2005, Lecture Notes in Comput. Sci. 3669, Springer, New York, 2005, pp. 604–615.
http://www.siam.org/journals/sidma/20-4/63906.html
†Department of Mathematics, University of Haifa, 31905 Haifa, Israel (lea@math.haifa.ac.il).

advance, i.e., the minimum possible number of bins used to pack items in σ. The *asymptotic performance ratio* for an algorithm $\mathcal{A}$ is defined to be

$$\mathcal{R}(\mathcal{A}) = \limsup_{n\to\infty} \sup_{\sigma} \left\{ \frac{\mathcal{A}(\sigma)}{OPT(\sigma)} | OPT(\sigma) = n \right\} .$$

In the *resource augmented* bin packing problem, the online algorithm is supplied with larger bins at its disposal than those of the offline algorithm that it is compared to. The competitive ratio then becomes a function of the bin size. All online bins are of the same size, and all the offline bins are of the same size, but these two sizes are not necessarily the same.

In the *variable-sized* bin packing problem, there is a supply of several bin sizes that can be used to pack the items. The cost of an algorithm is the sum of sizes of used bins. In this problem, the generalization into cardinality constrained packing assumes that each bin size $s_i \leq 1$ is associated with a parameter k_i which bounds the number of items that can be packed into such a bin.

We stress the fact that items arrive *online*; this means that each item must be assigned in turn, without knowledge of the next items. We consider *bounded space* algorithms, which have the property that they have only a constant number of bins available to accept items at any point during processing; these bins are also called "open bins." The bounded space assumption is a quite natural one. Essentially the bounded space restriction guarantees that output of packed bins is steady, and that the packer does not accumulate an enormous backlog of bins which are only output at the end of processing.

Previous results. Cardinality constrained bin packing was studied in the offline environment as early as 1975 by Krause, Shen, and Schwetman [13, 14]. They showed that the performance guarantee of the well-known first fit algorithm is at most $2.7-\frac{12}{5k}$. Additional results were offline approximation algorithms of performance guarantee 2. These results were later improved in two ways. Kellerer and Pferschy [12] designed an improved offline approximation algorithm with performance guarantee 1.5, and finally an APTAS (asymptotic polynomial time approximation scheme) was designed in [2] (for a more general problem). On the other hand, Babel et al. [1] designed a simple *online* algorithm with competitive ratio 2 for any value of k. They also designed improved algorithms for $k = 2, 3$ of competitive ratios $1 + \frac{\sqrt{5}}{5} \approx 1.44721$ and 1.8, respectively. The same paper [1] also proved an almost matching lower bound of $\sqrt{2} \approx 1.41421$ for $k = 2$ and mentioned that the lower bounds of [23, 21] for the classical problem hold for cardinality constrained bin packing as well. The lower bound of 1.5 given by Yao [23] holds for small values of $k > 2$, and the lower bound of 1.5401 given by Van Vliet [21] holds for sufficiently large k. No other lower bounds are known.

For the classical bin packing problem, Lee and Lee [15] presented an algorithm called HARMONIC, which partitions items into $K > 1$ classes and uses bounded space of at most $K - 1$ open bins. For any $\varepsilon > 0$, there is a number K_1 such that the HARMONIC algorithm that uses $K = K_1$ classes has a performance ratio of at most $(1 + \varepsilon)\Pi_\infty$ [15], where $\Pi_\infty \approx 1.69103$ is the sum of a series (see section 2). They also showed there is no bounded space algorithm with a performance ratio below Π_∞. Currently the best known unbounded space upper bound is 1.58889 due to Seiden [18].

The first to investigate the variable-sized bin packing problem were Friesen and Langston [10]. Csirik [4] proposed the VARIABLE HARMONIC algorithm and showed that it has performance ratio at most Π_∞. Seiden [17] showed that this algorithm

is optimal among bounded space algorithms. Unbounded space variable-sized bin packing was studied also in [19].

The resource augmented bin packing problem was studied by Csirik and Woeginger [6]. They showed that the optimal bounded space asymptotic performance ratio is a function $\rho(b)$ of the online bin size b. Unbounded space resource augmented bin packing was studied also in [8].

Our results. We consider bounded space algorithms. We are interested in the best competitive ratio that can be achieved using some constant number of open bins. For every value of k, we find the best competitive ratio of any online bounded space algorithm. We show that $K = k - 1$ bins are sufficient to achieve this best possible competitive ratio (and it is not only achieved in the limit as it is for the classical problem [15]). The competitive ratio is a strictly increasing function of k, and for large enough k it approaches $1 + \Pi_\infty \approx 2.69103$, where Π_∞ is the best competitive ratio shown by [15] for the classical bounded space problem. This is a surprising feature of the problem, since one would expect this value to simply tend to Π_∞ as k grows.

We further consider the resource augmented problem where the online algorithm may use larger bins compared to the optimal offline algorithm. We design optimal online algorithms for this problem as well. For large enough values of k, the competitive ratios again approach values which differ by 1 from the best competitive ratios for the classical resource augmented problem [6]. We show that the competitive ratios for our problem never drop below 1 (unlike the case studied in [6]) and identify the cases where the competitive ratio is exactly 1.

For the variable-sized bin packing problem, we design algorithms of the exact optimal competitive ratios (among bounded space algorithms) for any set of bins and cardinality constraints. An interesting feature is that we prove that the algorithms have optimal competitive ratios, even though we do not know what these ratios are.

A main difference between our results for bounded space algorithms and the results of [15, 6, 17] is that our algorithms have exactly the best possible competitive ratio achievable by bounded space online algorithms. The algorithms for variants of the classical problem have competitive ratios which tend to the best competitive ratio as the number of open bins grows without bound. Our algorithms need just a constant number of open bins to achieve the best competitive ratios. Therefore we need to be very careful in the analysis since, unlike the classical problem, we may not lose any small constants, which depend on the number of open bins, in the analysis.

For small values of k we design several new unbounded space algorithms, based on combination of large and small items together in bins (see [15, 16, 18]), according to sizes of small items. We prove that the competitive ratios of our algorithms for $k = 3, 4, 5, 6$ are $\frac{7}{4} = 1.75$, $\frac{71}{38} \approx 1.86842$, $\frac{771}{398} \approx 1.93719$, $\frac{287}{144} \approx 1.99306$ (respectively). This improves on the bounds $\frac{9}{5} = 1.8$ ($k = 3$) and 2 ($k = 4, 5, 6$) of [1].

2. Optimal algorithms for bounded space packing. In this section we define bounded space algorithms of optimal competitive ratio for each value of $k > 1$. For every $k > 1$, we define an online bounded space algorithm which packs at most k items in each bin and uses at most $k - 1$ open bins. We show that this algorithm is the best possible among bounded space algorithms. We use the well-known sequence π_i, $i \geq 1$, which is often used for bin packing, let $\pi_1 = 2, \pi_{i+1} = \pi_i(\pi_i - 1) + 1$, and let $\Pi_\infty = \sum_{i=1}^{\infty} \frac{1}{\pi_i - 1} \approx 1.69103$.

This sequence was used by Lee and Lee in [15] and by Van Vliet [21]. Adaptations of this sequence were later used in several papers including [6, 19]. The sequence is

constructed in a way that $1-\sum_{i=1}^{j}\frac{1}{\pi_i}=\frac{1}{\pi_{j+1}-1}$ (which can be easily shown by induction using the sequence definition). This means that each time the next value π_i is picked to be an integer, such that all items $\frac{1}{\pi_j}$ for $j\le i$ can fit together in a bin leaving some empty space. Note that Π_∞ is a lower bound on the best competitive ratio for classical bounded space bin packing, and there exists a sequence of bounded space algorithms with an increasing sequence of open bins whose competitive ratios tend to this value [15, 22]. The algorithms in this section are based on the algorithms in [15] with some differences in the construction and proof due to the cardinality constraint (which also increases the competitive ratio by 1 for large values of k). We also would like to achieve the best possible bound for every value of k separately, and not only in the limit.

Let $\mathcal{R}_k=\sum_{i=1}^{k}\max\left\{\frac{1}{\pi_i-1},\frac{1}{k}\right\}$. We show that for every value of k, the best competitive ratio is *exactly* $\mathcal{R}_k$. We start with some properties of $\mathcal{R}_k$ as a function of k.

THEOREM 1. *The value of $\mathcal{R}_k$ is a strictly increasing function of $k\ge 2$ such that $\frac{3}{2}\le\mathcal{R}_k<\Pi_\infty+1$ and $\lim_{k\to\infty}\mathcal{R}_k=\Pi_\infty+1\approx 2.69103$.*

Proof. We first find the value of $\mathcal{R}_2$. Since $\pi_1=2$ and $\pi_2=3$, we have $\mathcal{R}_2=\frac{3}{2}=1.5$. Note also that $\mathcal{R}_3=\frac{11}{6}\approx 1.83333, \mathcal{R}_4=2, \mathcal{R}_5=2.1$, and $\mathcal{R}_6=\frac{13}{6}\approx 2.16666$. Next we show the monotonicity of $\mathcal{R}_k$. For a given k, let $j_k=\min_{1\le j\le k}\{j|\frac{1}{k}\ge\frac{1}{\pi_j-1}\}$. The value j_k exists for all k since $\pi_k-1\ge k$ for all k. Then we have $\mathcal{R}_k=\sum_{i=1}^{j_k-1}\frac{1}{\pi_i-1}+\sum_{i=j_k}^{k}\frac{1}{k}=\sum_{i=1}^{j_k-1}\frac{1}{\pi_i-1}+\frac{k-j_k+1}{k}$. By definition of the values j_i, clearly $j_k\le j_{k+1}$. Therefore $\mathcal{R}_{k+1}-\mathcal{R}_k=\sum_{j_k}^{j_{k+1}-1}\frac{1}{\pi_i-1}+\frac{k-j_{k+1}+2}{k+1}-\frac{k-j_k+1}{k}>\frac{j_{k+1}-j_k}{k+1}+\frac{1-j_{k+1}}{k+1}-\frac{1-j_k}{k}=\frac{j_k-1}{k}-\frac{j_k-1}{k+1}\ge 0$. We deduce the strict inequality above by $\pi_i-1<k+1$, which holds for $i<j_{k+1}$.

An upper bound on $\mathcal{R}_k$ follows from $\mathcal{R}_k=\sum_{i=1}^{k}\max\left\{\frac{1}{\pi_i-1},\frac{1}{k}\right\}<\sum_{i=1}^{k}\frac{1}{\pi_i-1}+\sum_{i=1}^{k}\frac{1}{k}<\Pi_\infty+1$. We next show that $\mathcal{R}_k$ tends to this value. For a given $\varepsilon>0$, let ℓ be a value such that $\sum_{i=1}^{\ell}\frac{1}{\pi_i-1}\ge\Pi_\infty-\frac{\varepsilon}{2}$ and $\ell\ge\frac{2}{\varepsilon}$. Let $k=\ell^2$; then $\mathcal{R}_k\ge\sum_{i=1}^{\ell}\frac{1}{\pi_i-1}+\sum_{i=\ell+1}^{k}\frac{1}{\ell^2}\ge\Pi_\infty-\frac{\varepsilon}{2}+\frac{\ell^2-\ell}{\ell^2}\ge\Pi_\infty-\frac{\varepsilon}{2}+1-\frac{\varepsilon}{2}=\Pi_\infty+1-\varepsilon$. ∎

Next we define the algorithm CARDINALITY CONSTRAINED HARMONIC$_k$ (CCH$_k$), which is an adaptation of the algorithm HARMONIC$_k$ defined originally by Lee and Lee [15]. The fundamental idea of "harmonic-based" algorithms is to first classify items by size and then pack an item according to its class (as opposed to letting the exact size influence packing decisions).

For the classification of items, we partition the interval $(0,1]$ into subintervals. We use $k-1$ subintervals of the form $(\frac{1}{i+1},\frac{1}{i}]$ for $i=1,\ldots,k-1$ and one final subinterval $(0,\frac{1}{k}]$. Each bin will contain items from only one subinterval (type). Items in subinterval i are packed i to a bin for $i=1,\ldots,k-1$, thus keeping the cardinality constraint. The items in interval k are packed k to a bin. A bin which received the full number of items (according to its type) is closed, and therefore at most $k-1$ bins are open simultaneously (one per interval, except for $(\frac{1}{2},1]$).

To prove the upper bound on the competitive ratio, we use a simplified version of Theorem 9 stated in section 5. We use the technique of weighting functions. This technique was originally introduced for one-dimensional bin packing algorithms [20]. The version we use is as follows.

THEOREM 2. *Consider a bin packing algorithm. Let w be a weight measure. Assume that for every input of the algorithm the number of bins used by an algorithm ALG is bounded by $X(\sigma)+c$ for some constant c, where $X(\sigma)$ is the sum of weights of*

all items in the sequence according to weight measure w. Denote by W the supremum amount of weight that can be packed into a single bin of an offline algorithm according to measure w, i.e., the supremum total weight of a set of items whose total size is at most 1. Then the competitive ratio of the algorithm is upper bounded by W.

We define weights as follows. The weight of item x is denoted $w(x)$. The weight of an item in interval $(\frac{1}{i+1}, \frac{1}{i}]$, for $i = 1, \ldots, k-1$, is $\frac{1}{i}$. The weight of an item in interval $(0, \frac{1}{k}]$ is $\frac{1}{k}$. Recall that, except for $k-1$ open bins that may not receive the full number of items, each output bin receives a total weight of 1. A closed bin for items in interval $(\frac{1}{i+1}, \frac{1}{i}]$ receives i items, of weight $\frac{1}{i}$ each. A closed bin for items in interval $(0, \frac{1}{k}]$ receives k items, of weight $\frac{1}{k}$ each. Therefore we get $CCH_k(\sigma) \leq X(\sigma) + k - 1$.

THEOREM 3. *For every $k \geq 2$, the competitive ratio of* CCH_k *is $\mathcal{R}_k$, and no online algorithm which uses bounded space can have a better competitive ratio.*

Proof. We prove the upper bound first. Let $\varepsilon > 0$ a very small constant such that $\varepsilon \ll \frac{1}{k\pi_{k+1}}$ We claim that the maximum weight of a single bin is achieved for the following set of items: $f_1 \geq \cdots \geq f_k$, so that $f_i = \frac{1}{\pi_i} + \varepsilon$. This set of items fits in a single bin according to the definition of the sequence π_j. The sum of their weights is exactly $\mathcal{R}_k = \sum_{i=1}^{k} \max\{\frac{1}{\pi_i - 1}, \frac{1}{k}\}$.

To show that the maximum weight of any bin is indeed $\mathcal{R}_k$, consider an arbitrary set S of $\ell \leq k$ items which fits into one bin. If $\ell \neq k$, we add $k - \ell$ items of size zero and give them weight $\frac{1}{k}$. For this, we extend the packing rules and pack items of size zero together with other items of the first interval. The change may only result in an increase in the sum of weights. Let $g_1 \geq \cdots \geq g_k$ be the sorted list of items. If $g_i \in (\frac{1}{\pi_i}, \frac{1}{\pi_i - 1}]$ holds for all i such that $\pi_i \leq k$, and $g_i \in [0, \frac{1}{k}]$ for all i such that $\pi_i > k$, then the weight of the items of S is exactly $\mathcal{R}_k$. Otherwise let i be the first index of an item that does not satisfy the above. If $w(g_i) = \frac{1}{k}$, we get that $\sum_{j=1}^{k} w(g_j) = \sum_{j=1}^{i-1} w(g_j) + \sum_{j=i}^{k} w(g_j) = \sum_{j=1}^{i-1} w(f_j) + \sum_{j=i}^{k} \frac{1}{k} \leq \sum_{j=1}^{k} w(f_j) = \mathcal{R}_k$.

Otherwise, assume $w(g_i) > \frac{1}{k}$. Due to the greedy construction of the sequence π_j, and since $g_i \notin (\frac{1}{\pi_i}, \frac{1}{\pi_i - 1}]$, we get that $g_i \leq \frac{1}{\pi_i}$ and therefore $w(g_i) < \frac{1}{\pi_i - 1} = w(f_i)$. Let i' be the smallest index such that $w(g_{i'}) = \frac{1}{k}$ (this value exists as mentioned above since $k \leq \pi_k - 1$). If $i' = i + 1$, we get that $w(g_i) < w(f_i)$, and for $j \geq i'$, $\frac{1}{k} = w(g_j) \leq w(f_j)$. In this case we have $\sum_{j=1}^{k} w(g_i) < \sum_{j=1}^{k} w(f_i) = \mathcal{R}_k$. Otherwise consider the values of j such that $i \leq j \leq i' - 1$. We have $g_j \leq \frac{1}{\pi_i}$, and therefore according to the weight definition for $x > \frac{1}{k}$, $\frac{w(g_j)}{g_j} \leq \frac{\pi_i + 1}{\pi_i}$ for $i \leq j \leq i' - 1$. Given that for $j < i$, $g_j \in (\frac{1}{\pi_j}, \frac{1}{\pi_j - 1}]$, we have $\sum_{j=i}^{k} g_j \leq \frac{1}{\pi_i - 1}$ and therefore $\sum_{j=i}^{i'-1} w(g_j) \leq \frac{\pi_i + 1}{\pi_i^2 - \pi_i}$. However, $w(f_i) + w(f_{i+1}) = \frac{1}{\pi_i - 1} + \frac{1}{\pi_{i+1} - 1} = \frac{1}{\pi_i - 1} + \frac{1}{\pi_i^2 - \pi_i} = \frac{\pi_i + 1}{\pi_i^2 - \pi_i}$. Summarizing, we get $\sum_{j=1}^{k} w(g_j) = \sum_{j=1}^{i-1} w(g_j) + \sum_{j=i}^{i'-1} w(g_j) + \sum_{j=i'}^{k} w(g_j) \leq \sum_{j=1}^{i-1} w(f_j) + w(f_i) + w(f_{i+1}) + \sum_{j=i'}^{k} \frac{1}{k} \leq \sum_{j=1}^{k} w(f_i) = \mathcal{R}_k$.

The proof of the lower bound is similar to previously known lower bound proofs for bounded space algorithms; see [15, 6]. To prove the lower bound, let N be a large constant and $\delta > 0$ a very small constant such that $\delta \ll \frac{1}{k\pi_{k+1}}$. We construct the following sequence. The sequence has k phases. Phase i contains N items of size $\frac{1}{\pi_i} + \delta$. Let K be the number of bins that may be open simultaneously. Except for at most K bins, all bins of each phase are closed after the phase. Such bins can be filled by a maximum number of $\min\{\pi_i - 1, k\}$ items. Therefore phase i contributes at least $\frac{N}{\min\{\pi_i - 1, k\}} - K = N \max\{\frac{1}{\pi_i - 1}, \frac{1}{k}\} - K$ closed bins to the output. The optimal packing of the sequence contains N identically packed bins with one item of

each phase per bin. We get that the competitive ratio is at least $\mathcal{R}_k - \frac{kK}{N}$. This approaches $\mathcal{R}_k$ for large enough N. □

3. Extension to resource augmentation. Following the work of [6], which studied resource augmentation for the classical bin packing problem, we show that the algorithms defined in the previous section are optimal in a resource augmented environment as well.

We compare an online algorithm which uses bins of size 1 to an optimal offline algorithm whose bins are of size $\frac{1}{b}$. We assume that all item sizes are bounded by $\frac{1}{b}$. This problem definition is equivalent to the alternative definition where items have sizes in $(0, 1]$, the online algorithm uses bins of size b, and the offline algorithm uses bins of size 1. The competitive ratio for bounded cardinality k is measured as a function of $b > 1$. The best competitive ratios for bounded space algorithms and unrestricted online algorithms are denoted $\mathcal{R}_k(b)$ and $r_k(b)$ (respectively). We note a fundamental difference between the resource augmented problem associated with the classical bin packing problem and the problem studied in this paper. As we show later in this section, the competitive ratio is never below 1 for our problem, whereas the classical problem has a competitive ratio below 1 for $b \geq 2$ [6, 8].

We show that the competitive ratio (even for unbounded space algorithms) cannot actually reach 1 if $b < k$, and is exactly 1 for $b \geq k$.

THEOREM 4. *For all values of b, k such that $b < k$ and $k \geq 2$, we have $\mathcal{R}_k(b) \geq r_k(b) > 1$. For all values of b, k such that $b \geq k$, we have $r_k(b) = \mathcal{R}_k(b) = 1$.*

Proof. It is easy to see that the functions $r_k(b)$ and $\mathcal{R}_k(b)$ are monotonically decreasing in b and that $\mathcal{R}_k(b) \geq r_k(b)$. Therefore, given $b < k$, we can prove the first part for $b' = \max\{b, k - \frac{1}{2}\} \geq b$; i.e., we prove that $r_k(b') > 1$ and therefore $r_k(b) > 1$. Let $x = \frac{b'+k}{2kb'} < \frac{1}{b'}$, and let $\varepsilon = \frac{1-xb'}{b'(k-1)} > 0$. Let N be a large enough integer. The input sequence consists of a first phase with $Nk(k-1)$ items of size ε, possibly followed by a second phase with Nk items of size x. Denote the optimal offline cost after the first phase by OPT_1 and after the second phase by OPT_2. We get that $OPT_1 = N(k-1)$ since $k\varepsilon < \frac{1}{b'}$, and $OPT_2 = Nk$ since $x + (k-1)\varepsilon = \frac{1}{b'}$. Let R be the competitive ratio of an online algorithm $\mathcal{A}$. Let Y_i $(1 \leq i \leq k)$ be the number of bins into which algorithm $\mathcal{A}$ packed exactly i items during the first phase. Note that $\sum_{i=1}^{k} iY_i = Nk(k-1)$. If the sequence stops here, we have $\sum_{i=1}^{k} Y_i \leq R \cdot OPT_1 = RN(k-1)$. If $\sum_{i=1}^{k} Y_i(k-i) > Nk$, we get $\sum_{i=1}^{k} kY_i > Nk(k-1) + Nk = Nk^2$, which gives $R > \frac{k}{k-1}$. Otherwise if the sequence continues, $\sum_{i=1}^{k} Y_i(k-i)$ is exactly the number of larger items that can fit into the existing bins of the online algorithm. Since this number is at most Nk, the other items need to be packed into new bins. Note that $kx = \frac{b'+k}{2b'} > 1$. Therefore the best packing can be with $k-1$ items per bin. This results in a packing of size

$$\sum_{i=1}^{k} Y_i + \frac{Nk - \sum_{i=1}^{k} Y_i(k-i)}{k-1} \leq R \cdot OPT_2 = RNk.$$

We get $\sum_{i=1}^{k}(k-1)Y_i + Nk - \sum_{i=1}^{k} Y_i(k-i) = Nk + \sum_{i=1}^{k}(i-1)Y_i \leq RNk(k-1)$. Combining with $\sum_{i=1}^{k} Y_i \leq RN(k-1)$, we have $Nk^2 = Nk + Nk(k-1) = Nk + \sum_{i=1}^{k} iY_i \leq RN(k^2-1)$ or $R \geq \frac{k^2}{k^2-1} > 1$.

For the second part we simply use the algorithm Next-Fit [11]. Since $\frac{1}{b} \leq \frac{1}{k}$ and all item sizes are at most $\frac{1}{b}$, each bin receives exactly k items. Given a sequence of f items, we get $\lceil \frac{f}{k} \rceil$ packed bins. However, due to the cardinality constraint, $OPT \geq \lceil \frac{f}{k} \rceil$, and

therefore the competitive ratio is at most 1. Consider now a sequence of Nk items of size $\frac{1}{b}$. No algorithm can pack them into less than N bins (no matter how large b is). Since $OPT = N$ as well we get that $r_k(b) = \mathcal{R}_k(b) = 1$ for the case $b \geq k$. □

The algorithms are defined exactly as in the previous section. However, this means that some of the defined classes do not exist if b is large enough. Note that the algorithm for the case $b \geq k$ becomes exactly Next Fit as described in Theorem 4.

To define the competitive ratio, we first define sequences $\pi_i(b)$ and $\Pi_i(b)$, originally defined by [6] as follows: $\Pi_0(b) = 0$, $\pi_1(b) = \lfloor b \rfloor + 1$, $\Pi_1(b) = \frac{1}{\pi_1(b)}$, $\pi_i(b) = \left\lfloor \frac{1}{\frac{1}{b} - \Pi_{i-1}(b)} \right\rfloor + 1$, and $\Pi_i(b) = \Pi_{i-1}(b) + \frac{1}{\pi_i(b)}$. The intuition behind this function is to find a sequence of integers such that the next integer at each point is picked greedily to be minimal and the sum of their reciprocals is less than $\frac{1}{b}$. The values of $\pi_i(b)$ satisfy $\pi_i(b) > b$. We can show that the values are strictly increasing as a function of b. Clearly the values are nondecreasing. If two values are the same, we let $\pi_i(b) = \pi_{i+1}(b) = f$ be these identical values. Then we argue that $\pi_i(b)$ should have been chosen to be at most $f - 1$. To see that, note that $\frac{1}{b} - \Pi_{i-1}(b) > \frac{2}{f} \geq \frac{1}{f-1}$. This holds for all $f \geq 2$.

Csirik and Woeginger [6] introduced the function $\rho(b) = \sum_{i=1}^{\infty} \frac{1}{\pi_i(b)-1}$ and showed that this is the best possible competitive ratio with resource augmentation b for the classical bin packing problem. Note that $\rho(1) = \Pi_\infty \approx 1.69103$. We can prove the following theorems.

THEOREM 5. *For every $k \geq 2$, the competitive ratio of CCH_k (defined in the previous section) is $\mathcal{R}_k(b) = \sum_{i=1}^{k} \max\left\{\frac{1}{\pi_i(b)-1}, \frac{1}{k}\right\}$, and no online algorithm which uses bounded space can have a better competitive ratio.*

THEOREM 6. *The value of $\mathcal{R}_k(b)$ for a fixed value of b is an increasing function of $k \geq 2$ such that $1 \leq \mathcal{R}_k(b) < \rho(b) + 1$ and $\lim_{k\to\infty} \mathcal{R}_k = \rho(b) + 1$.*

Proof of Theorem 5. We again use Theorem 2. The weights are defined as in the previous section. We prove the upper bound first. Let $\varepsilon > 0$ be a very small constant such that $\varepsilon \ll \frac{1}{k\pi_{k+1}(b)}$ We claim that the maximum weight of a single bin is achieved for the following set of items: $f_1 \geq \cdots \geq f_k$, so that $f_i = \frac{1}{\pi_i(b)} + \varepsilon$. This set of items fits in a single bin of size $\frac{1}{b}$ according to the definition of the sequence $\pi_j(b)$. The sum of their weights is exactly $\mathcal{R}_k(b) = \sum_{i=1}^{k} \max\left\{\frac{1}{\pi_i(b)-1}, \frac{1}{k}\right\}$.

To show that the maximum weight of any bin is indeed $\mathcal{R}_k(b)$ consider an arbitrary set S of $\ell \leq k$ items which fits into one bin of size $\frac{1}{b}$. If $\ell \neq k$, we add $k - \ell$ items of size zero and give them weight $\frac{1}{k}$. For this, we extend the packing rules and pack items of size zero together with other items of the first interval. The change may only result in an increase in the sum of weights.

Let $g_1 \geq \cdots \geq g_k$ be the sorted list of items. If $g_i \in (\frac{1}{\pi_i(b)}, \frac{1}{\pi_i(b)-1}]$ holds for all i such that $\pi_i(b) \leq k$, and $g_i \in [0, \frac{1}{k}]$ for all i such that $\pi_i(b) > k$, then the weight of the items of S is exactly $\mathcal{R}_k(b)$.

Otherwise let i be the first index of an item that does not satisfy the above. If $w(g_i) = \frac{1}{k}$, we get that $\sum_{j=1}^{k} w(g_j) = \sum_{j=1}^{i-1} w(g_j) + \sum_{j=i}^{k} w(g_j) = \sum_{j=1}^{i-1} w(f_j) + \sum_{j=i}^{k} \frac{1}{k} \leq \sum_{j=1}^{k} w(f_j) = \mathcal{R}_k(b)$. Otherwise, assume $w(g_i) > \frac{1}{k}$. Due to the greedy construction of the sequence π_j, and since $g_i \notin (\frac{1}{\pi_i(b)}, \frac{1}{\pi_i(b)-1}]$, we get that $g_i \leq \frac{1}{\pi_i(b)}$ and therefore $w(g_i) < \frac{1}{\pi_i(b)-1} = w(f_i)$. Let i' be the smallest index such that $w(g'_i) = \frac{1}{k}$. If such an index does not exist, we let $i' = k + 1$. If $i' = i + 1$, we get that $w(g_i) < w(f_i)$, and for $i' \leq j \leq k$, $\frac{1}{k} = w(g_j) \leq w(f_j)$. In this case we have $\sum_{j=1}^{k} w(g_i) < \sum_{j=1}^{k} w(f_i) = \mathcal{R}_k(b)$. Otherwise consider the values of j such that

$i \le j \le i' - 1$. We have $g_j \le \frac{1}{\pi_i(b)}$, and therefore according to the weight definition for $x > \frac{1}{k}$, $\frac{w(g_j)}{g_j} \le \frac{\pi_i(b)+1}{\pi_i(b)}$. Given that for $j < i$, $g_j \in (\frac{1}{\pi_j(b)}, \frac{1}{\pi_j(b)-1}]$, we have $\sum_{j=i}^{k} g_j \le \frac{1}{b} - \Pi_{i-1}(b) \le \frac{1}{\pi_i(b)-1}$ and $\sum_{j=i+1}^{k} g_j \le \frac{1}{b} - \Pi_i(b) = \frac{1}{b} - \Pi_{i-1}(b) - \frac{1}{\pi_i(b)} \le \frac{1}{\pi_{i+1}(b)-1}$. Using $g_j \le \frac{1}{\pi_i(b)}$ again, we get $\sum_{j=i}^{k} g_j \le \frac{1}{\pi_{i+1}(b)-1} + \frac{1}{\pi_i(b)}$. This gives $\sum_{j=i}^{i'-1} w(g_j) \le \big(\frac{\pi_i(b)+1}{\pi_i(b)}\big)\big(\sum_{j=i+1}^{k} g_j\big) \le \frac{1}{\pi_i(b)(\pi_i(b)-1)} + \frac{1}{\pi_i(b)} + \frac{1}{\pi_{i+1}(b)-1} = \frac{1}{\pi_i(b)-1} + \frac{1}{\pi_{i+1}(b)-1}$. However, we have $w(f_i) + w(f_{i+1}) = \frac{1}{\pi_i(b)-1} + \frac{1}{\pi_{i+1}(b)-1}$. Summarizing, we get $\sum_{j=1}^{k} w(g_j) = \sum_{j=1}^{i-1} w(g_j) + \sum_{j=i}^{i'-1} w(g_j) + \sum_{j=i'}^{k} w(g_j) \le \sum_{j=1}^{i-1} w(f_j) + w(f_i) + w(f_{i+1}) + \sum_{j=i'}^{k} \frac{1}{k} \le \sum_{j=1}^{k} w(f_i) = \mathcal{R}_k(b)$.

To prove the lower bound, let N be a large constant and $\delta > 0$ a very small constant such that $\delta \ll \frac{1}{k\pi_{k+1}(b)}$. We construct the following sequence. The sequence has k phases. Phase i contains N items of size $\frac{1}{\pi_i(b)} + \delta$. Let K be the number of bins that may be open simultaneously. Except for at most K bins, all bins of each phase are closed after the phase. Such bins can be filled by a maximum of $\min\{\pi_i(b) - 1, k\}$ items. Therefore phase i contributes at least $\frac{N}{\min\{\pi_i(b)-1,k\}} - K = N \max\{\frac{1}{\pi_i - 1}, \frac{1}{k}\} - K$ closed bins to the output. The optimal packing of the sequence contains N identically packed bins with one item of each phase per bin. We get that the competitive ratio is at least $\mathcal{R}_k(b) - \frac{kK}{N}$. This approaches $\mathcal{R}_k(b)$ for large enough N. ∎

Proof of Theorem 6. As shown above, the value of $\mathcal{R}_k(b)$ is at least 1. Next we show the monotonicity of $\mathcal{R}_k(b)$ for a fixed value of b as a function of k. If $k \le b$, the value of the function is 1; therefore we need to prove monotonicity for $k > b$. Note that for every k and b, $\pi_k(b) > k$. This holds since if $\pi_k \le k$, we get that $\sum_{t=1}^{k} \frac{1}{\pi_k(b)} \ge 1 > b$. We therefore need to consider the case $\pi_k(b) \ge k + 1$, $\pi_{k+1}(b) \ge k + 2$. For $k' = k, k + 1$, let $j'_k = \min_{1 \le j \le k'}\{j | \frac{1}{k'} \ge \frac{1}{\pi_j(b)-1}\}$. We have $\mathcal{R}_k = \sum_{i=1}^{j_k - 1} \frac{1}{\pi_i(b)-1} + \sum_{i=j_k}^{k} \frac{1}{k} = \sum_{i=1}^{j_k-1} \frac{1}{\pi_i(b)-1} + \frac{k-j_k+1}{k}$. By definition of the values j_i, clearly $j_k \le j_{k+1}$. Therefore $\mathcal{R}_{k+1} - \mathcal{R}_k = \sum_{j_k}^{j_{k+1}-1} \frac{1}{\pi_i(b)-1} + \frac{k-j_{k+1}+2}{k+1} - \frac{k-j_k+1}{k} > \frac{j_{k+1}-j_k}{k+1} + \frac{1-j_{k+1}}{k+1} - \frac{1-j_k}{k} = \frac{j_k-1}{k} - \frac{j_k-1}{k+1} \ge 0$. The strict inequality above follows from $\pi_i(b) - 1 < k + 1$ for $i < j_{k+1}$.

An upper bound on $\mathcal{R}_k$ follows from $\mathcal{R}_k = \sum_{i=1}^{k} \max\big\{\frac{1}{\pi_i(b)-1}, \frac{1}{k}\big\} \le \sum_{i=1}^{k} \frac{1}{\pi_i(b)-1} + \sum_{i=1}^{k} \frac{1}{k} < \rho(b) + 1$. We next show that $\mathcal{R}_k$ tends to this value. For a given $\varepsilon > 0$, let ℓ be a value such that $\sum_{i=1}^{\ell} \frac{1}{\pi_i(b)-1} \ge \rho(b) - \frac{\varepsilon}{2}$ and $\ell \ge \frac{2}{\varepsilon}$. Let $k = \ell^2$; then $\mathcal{R}_k \ge \sum_{i=1}^{\ell} \frac{1}{\pi_i(b)-1} + \sum_{i=\ell+1}^{k} \frac{1}{\ell^2} \ge \rho(b) - \frac{\varepsilon}{2} + \frac{\ell^2-\ell}{\ell^2} \ge \rho(b) - \frac{\varepsilon}{2} + 1 - \frac{\varepsilon}{2} = \rho(b) + 1 - \varepsilon$. ∎

4. Extension to variable-sized bins. Following the work of Seiden [17] we design optimal online bounded space algorithms for the case of variable-sized bins. Similarly to that case and other work on variable-sized bins [7], we design an algorithm for any set of bin sizes, and we prove the optimality of these algorithms; however, we do not know their competitive ratios. Our algorithms are based on the VARIABLE HARMONIC algorithms of Csirik [4]. The optimality of these algorithms among the class of bounded space algorithms was proved in [17]. As in previous sections, the main difference between these algorithms and our algorithms is in the way that small items are packed. As in previous sections, our algorithms have the exact best possible competitive ratio for a given set of bins and cardinality constraints, this with a constant number of open bins that can be easily computed (as a function of the bins sizes and constraints). The algorithms for the classical problem get close to the best

possible competitive ratio as the number of open bins grows without bound.

In order to define our general algorithm Cardinality Constrained Variable Harmonic (CCVH) we use some definitions. Let the bins sizes be $s_1 < \cdots < s_m = 1$. Let their cardinality constraints be $k_1, \ldots, k_m$ (respectively). We define a set of critical sizes for each bin in the following way. Let $T_i = \{\frac{s_i}{j} | 1 \leq j \leq k_i\}$ and $T = \bigcup_{1 \leq i \leq m} T_i$. Let $|T| = M$, and the members of T be $1 = t_1 > t_2 > \cdots > t_M$. The type of a size t_r is defined to a value $i(r)$ such that $t_r \in T_{i(r)}$ (ties are broken arbitrarily). In this case the order of t_r is $\ell(r) \leq k_i$ such that $t_r = \frac{s_{i(r)}}{\ell(r)}$.

We again classify items into intervals whose right endpoint is a critical size. This associates an item with a type and order. Afterwards we pack an item according to its type and order (here as well as in the previous sections, the exact size does not influence packing decisions). Each bin will contain items of a single interval.

Since $M = |T| \leq \sum_{i=1}^{m} k_i$, there is a bounded number of pairs of type and order. For the classification of items, we partition the interval $(0, 1]$ into subintervals. The "small" interval is $(0, t_M]$. The other intervals are $(t_{j+1}, t_j]$ for $j = 1, \ldots, M-1$. Each bin will contain items from only one pair of type and order. Items in the subinterval whose right endpoint is t_r are packed into bins of size $s_{i(r)}$. The items in this interval are packed $\ell(r)$ to a bin, thus keeping the cardinality constraints. Note that at most $M - m$ bins are open simultaneously, since a bin which received the full amount of items (according to its type) is closed.

The differences between our algorithm and algorithms for the classical variable-sized bin packing problem are as follows. The condition for an item to be "small" (i.e., in the "small" interval) is determined by the cardinality constraints. Items cannot be packed using Next Fit, due to these constraints. Moreover, in [17] the smallest items are packed into bins of size 1. In that case it is actually possible to pack the small items into any type of bin. Here the type of bin for the small items must be $s_{i(M)}$. (If there exists another size i' such that $t_M \in T'_i$, that size can be used for the small items as well.)

The following theorem is used in [17] to prove upper bounds on the competitive ratio of algorithms for variable-sized bins.

Theorem 7. *Consider a bin packing algorithm. Let w be a weight measure. Assume that, for every output of the algorithm, the cost of all the bins used by the algorithm ALG is bounded by $X(\sigma) + c$ for some constant c, where $X(\sigma)$ is the sum of weights of all items in the sequence according to weight measure w. Denote by W_i the supremum amount of weight that can be packed into a (valid) single bin of size s_i of an offline algorithm according to measure w. Then the competitive ratio of the algorithm is upper bounded by* $\max_{1 \leq i \leq m} \left\{\frac{W_i}{s_i}\right\}$.

We assign weights to items in the following way. The weight of an item x is again denoted by $w(x)$. An item of interval $(0, t_M]$ receives weight $\frac{s_{i(M)}}{\ell(M)}$ (note that $\ell(M) = k_{i(M)}$). An item of interval $(t_{j+1}, t_j]$ receives weight $\frac{s_{i(j)}}{\ell(j)}$. Each closed bin of interval $(0, t_M]$ is of size $s_{i(M)}$; it receives $\ell(M)$ items, and thus the weight of items packed in it is equal to its size. Each closed bin of interval $(t_{j+1}, t_j]$ is of size $s_{i(j)}$. It receives $\ell(j)$ items, and thus the weight of items packed in it is equal to its size. Therefore the cost of the algorithm differs from the total weight of all items by the cost of all open bins, which is clearly bounded by $M - m$.

We can now use Theorem 7 to prove the following theorem.

Theorem 8. *For a given set of bins sizes and cardinality constraints, the algorithm* CCVH *is an optimal online algorithm.*

Proof. Let $s = s_i$ be the bin size which maximizes the expression $\max_{1 \leq i \leq m} \left\{\frac{W_i}{s_i}\right\}$.

Let $k = k_i$ be the cardinality constraint of this bin size. We allow the bin to contain items of size 0, and we give them the weight $\frac{s_{i(M)}}{\ell(M)}$ as the weight of other very small items. Assume therefore that a bin which contains a maximum amount of weight has exactly k items. Let $b_1, \ldots, b_k$ be their sizes. Let N be a large enough integer. Consider an offline packing with N bins of size s identically packed with items $b_1, \ldots, b_k$. The cost of this algorithm is Ns.

We show that any bounded space online algorithm is forced to have competitive ratio of at least $\left(\sum_{y=1}^{k} w(b_y)\right)/s$. The input sequence is sorted so that it consists of k phases. Phase y has N identical items of size b_y. Let K be the number of bins that can be open simultaneously. For each bin size s_a, we compute the maximum number of items of size b_y that can be packed in a closed bin of size s_a. This number is $Q(y,a) = \min\{k_a, \lfloor \frac{s_a}{b_y} \rfloor\}$. Let $t_{j(y)}$ be the upper bound of the interval for b_y. According to the above weight definitions, $w(b_y) = t_{j(y)}$. For $1 \le a \le m$ such that $s_a \ge b_y$, let $x(y,a)$ be the smallest integer such that $b_y \le t_{x(y,a)}$ and $i(x(y,a)) = a$.

We charge an item of size b_y, which the online algorithm packs in a bin of size s_a, with $\frac{s_a}{Q(y,a)}$. In this way the cost for all items packed in closed bins is exactly the cost of the online algorithm for the closed bins. We claim that for pairs y, a for which $x(y,a)$ is defined, $t_{x(y,a)} = \frac{s_a}{Q(y,a)}$ and $x(y,a) \le j(y)$ hold. If $Q(y,a) = k_a$, then $\frac{s_a}{b_y} \ge k_a$. Therefore $\frac{s_a}{k_a} \ge b_y$ and $t_{x(y,a)} = \frac{s_a}{k_a}$. Otherwise $\frac{s_a}{b_y} - 1 < Q(y,a) \le \frac{s_a}{b_y}$. Therefore $\frac{s_a}{Q(y,a)+1} < b_y$ and $\frac{s_a}{Q(y,a)} \ge b_y$. This is exactly the definition of $t_{x(y,a)}$. Since $j(y)$ is the largest index that satisfies $t_j(y) \ge b_y$, we get that $x(y,a) \le j(y)$. We got that an item of size b_y in a bin of size b_a is charged with $\frac{s_a}{Q(y,a)} = t_{x(y,a)} \ge t_j(y)$. Let $\kappa = \max_{1 \le i \le m}\{k_i\}$. At most $K\kappa$ items are in open bins after phase y; therefore the cost for this phase is at least $Nt_j(y) - K\kappa = Nw(b_y) - K\kappa$. Summing over all phases, we get the cost $\sum_{y=1}^{k}(Nw(b_y) - K\kappa) = N\sum_{y=1}^{k} w(b_y) - Kk\kappa$. The competitive ratio is therefore at least

$$\frac{\sum_{y=1}^{k} w(b_y)}{s} - \frac{Kk\kappa}{sN}.$$

This value approaches

$$\frac{\sum_{y=1}^{k} w(b_y)}{s}$$

for large enough N. □

5. Improved unbounded space algorithms for small values of k.

5.1. $k = 3$. In this section we design an algorithm for $k = 3$. Already the algorithm of [1] has a competitive ratio lower than the best bounded space algorithm ($\frac{9}{5} = 1.8$, which is smaller than $\frac{11}{6}$). We design an algorithm that uses a more careful partition into classes and has competitive ratio $\frac{7}{4} = 1.75$. The algorithm is based on the idea of the HARMONIC algorithm and its generalizations (see [15, 16, 18, 9]). In these generalizations, items of two intervals are combined together in the same bins. We would like to use a similar approach; however, the boundaries of intervals are chosen in accordance with the cardinality constraints.

We use the following five intervals: $A = (\frac{2}{3}, 1]$, $B = (\frac{1}{2}, \frac{2}{3}]$, $C = (\frac{1}{3}, \frac{1}{2}]$, $D = (\frac{1}{6}, \frac{1}{3}]$, $E = (0, \frac{1}{6}]$. Items which belong to an interval I are called items of type I, type I items, or simply I items. Items of types A, C, and D are packed independently of any other items, at one, two, and three items per bin, respectively. Note that it is always

possible to combine one item of type B with two items of type E. Therefore, each item of type E receives a color upon arrival, white or red. White items are packed in separate bins (three per bin), whereas red items are packed two per bin and combined with one type B item. If there exists such an open bin, the red type E items are added there. Otherwise once a type B item arrives later, it is added to a bin with two type E items. The colors are assigned so that an α fraction of the type E items are red. We use $\alpha = \frac{1}{4}$. Therefore every fourth type E item is red, and all others are white.

We define a bin as incomplete in the four following packings:

- a bin with a single C item,
- a bin with only one or two D items,
- a bin with one or two white E items,
- a bin with a single red E item (and possibly a B item as well).

At every time, the algorithm can have at most four incomplete bins, one for each combination. Therefore upon termination, except for at most four incomplete bins, every bin is packed according to one of the following options:

- a single A item,
- two C items,
- three D items,
- one B item,
- two red E items,
- three white E items,
- one B item and two red E items.

According to the definition of the algorithm, we never have a situation where one bin has only a B item, and another bin has two red E items. This is true since a new bin is opened for such items only if they cannot join a previously opened bin.

The algorithm is therefore at one of the following two situations: 1. there are no bins with two red E items with no B item; 2. there are no bins with one B item and no E items.

We assign two weights to each item, according to the two scenarios. The weights are assigned according to types of items. We use $w_1(I)$ and $w_2(I)$ to denote the weights of type I items according to the two weight functions. Let

$$w_1(A) = w_2(A) = 1,$$

$$w_1(B) = 1, \qquad w_2(B) = 0,$$

$$w_1(C) = w_2(C) = \frac{1}{2},$$

$$w_1(D) = w_2(D) = \frac{1}{3},$$

$$w_1(E) = \frac{1-\alpha}{3} = \frac{1}{4}, \qquad w_2(E) = \frac{1-\alpha}{3} + \frac{\alpha}{2} = \frac{\alpha+2}{6} = \frac{3}{8}.$$

The weights are defined so that in the first scenario, on average all bins (but at most four) have a total amount of weight of at least 1 packed into them according to the first weight measure, and otherwise the same property holds according to the second weight measure.

We use the following theorem; see Seiden [18].

THEOREM 9. *Consider a bin packing algorithm. Let w_1, w_2 be two weight measures. Assume that for every output of the algorithm, there exists i ($i = 1$ or $i = 2$)*

such that the number of bins used by the algorithm ALG is bounded by $X_i(\sigma) + c$ *for some constant* c*, where* $X_i(\sigma)$ *is the sum of weights of all items in the sequence according to weight measure* w_i*. Denote by* W_i *the supremum amount of weight that can be packed into a (valid) single bin according to measure* w_i *(*$i = 1, 2$*). Then the competitive ratio of the algorithm is upper bounded by* $\max(W_1, W_2)$*.*

Proof. Given an input, let i be the value that satisfies the theorem for this input. Clearly $OPT(\sigma) \geq \frac{X_i(\sigma)}{W_i}$. We get $ALG \leq X_i(\sigma) + c \leq W_i OPT + c$. □

To use the theorem, we need to prove that for every input, $ALG \leq X_i(\sigma) + c$ for some i. We ignore the (at most four) incomplete bins, which adds at most 5 to the constant c. The weight of a bin is the sum of weights of items assigned to it. In both scenarios, bins with one A item have weight 1, bins with two C items have weight 1, and so do bins with three D items.

We remove from the sequence items of incomplete bins. Denote the numbers of B items by $n(B)$, and of E items by $n(E)$. The number of red E items is denoted $n(ER)$, and the number of white E items $n(EW)$ (i.e., $n(E) = n(EW) + n(ER)$). According to the color assignments, and since at most two white items and one red item were removed, $3n(ER) \leq n(EW) \leq 3n(ER) + 6$. In the first scenario, no bins contain red E items only. The total weight of B and E items is $n(B) + \frac{n(E)}{4}$. The number of bins used for these types is $n(B) + \frac{n(EW)}{3} \leq n(B) + \frac{n(E)+2}{4}$ (using $n(EW) \leq 3n(ER) + 6$ which gives $4n(EW) \leq 3n(E) + 6$). In this case we get $ALG < X_1 + 5$. In the second scenario, no bins contain a B item only. The total weight of B and E items is $\frac{3n(E)}{8}$. The number of bins used for these types is $\frac{n(ER)}{2} + \frac{n(EW)}{3} = \frac{n(E)}{3} + \frac{n(ER)}{6} \leq n(E)(\frac{1}{3} + \frac{1}{24}) = \frac{3n(E)}{8}$ (using $3n(ER) \leq n(EW)$, which gives $4n(ER) \leq n(E)$). In this case we get $ALG < X_2 + 4$.

Next we analyze the maximum amount of weight that a bin can contain according to the two weight measures. In both weight measures, if no item has weight 1, the total weight of three items does not exceed $\frac{3}{2}$. Using w_1, the smallest item of weight 1 is slightly larger than $\frac{1}{2}$. If there is a C item, then there can be no D item but only an E item. We therefore get $1 + \frac{1}{2} + \frac{1}{4}$. If there is no C item, the worst case is two extra D items. This gives $1 + \frac{2}{3}$. We therefore get $W_1 = \frac{7}{4} = 1.75$. Using w_2, the smallest item of weight 1 is slightly larger than $\frac{2}{3}$. There can be no B or C items. The worst case is two extra E items, and we get $W_2 = 1 + 2 \cdot \frac{3}{8} = 1.75$.

We have proved the following theorem.

THEOREM 10. *The competitive ratio of the above algorithm for* $k = 3$ *is at most* 1.75*.*

5.2. $k = 4, 5, 6$**.** In this section we introduce a general algorithm and analyze it for three values of k. The algorithm is a generalization of the algorithm for $k = 3$ with additional options. The intervals (also called classes) are defined as follows. The interval of largest items is $A = (1 - \frac{1}{k}, 1]$. The next interval, of smaller large items is $B = (\frac{1}{2}, 1 - \frac{1}{k}]$. Intervals $C_2, \ldots, C_{k-1}$ are $C_i = (\frac{1}{i+1}, \frac{1}{i}]$. Intervals $E_1, \ldots, E_{k-1}$ are defined to be $E_i = (\frac{1}{k(i+1)}, \frac{1}{ki}]$ for $i < k - 1$ and $E_{k-1} = (0, \frac{1}{k(k-1)}]$.

We use parameters α_i for intervals E_i. An α_i fraction of the items of interval E_i are colored red, and all others are colored white. All these values are rational, so if $\alpha_i = \frac{p_i}{q_i}$ is a minimal rational representation of α_i, then the input items of this interval are partitioned into sets of q_i items, out of which the first $q_i - p_i$ are colored white and the next p_i are colored red.

The packing is done as follows. Items of class C_i are packed i per bin. White items of classes E_i are packed k per bin. Red items of class E_i are packed i per bin.

This means that a bin never contains more than $k-1$ red items, and they occupy a space of at most $\frac{1}{k}$. These items can always be combined with type B items. Basically, items of class B are packed one per bin, but when possible, they are combined with one of the types E_i. When we need to open a bin for red E_i items for some i, we first check whether there exists a bin with only a class B item, and if so, the red items are added to that bin. Otherwise a new bin is opened for them. When an item of class B arrives, we try to add it into a bin of red items that still has not received a B item, and open a new bin if it does not exist.

A bin is complete if it received its full number of items, or if it contains a B item, or if it contains the full number of red items (possibly without a B item). We can neglect bins that are not complete, since their number is at most $3k-4$. This amount is caused by at most $k-1$ bins for intervals C_i for $1 \le i \le k-1$, $k-1$ bins for white items of $k-1$ types, and $k-2$ bins for red items of $k-2$ types (a bin with a red E_1 item cannot be incomplete). As in the algorithm for $k=3$, only one of the two situations can occur. Either there are no complete bins with red items without a class B item, or there are no bins with a class B item and no red items.

We define weights as follows. Assign two weights to each item, according to the two scenarios. The weights are assigned according to types of items. We again use $w_1(I)$ and $w_2(I)$ to denote the weights of type I items according to the two weight functions. Let $w_1(A) = w_2(A) = 1,\ w_1(B) = 1,\ w_2(B) = 0,\ w_1(C_i) = w_2(C_i) = \frac{1}{i},\ w_1(E_i) = \frac{1-\alpha_i}{k},\ w_2(E_i) = \frac{1-\alpha_i}{k} + \frac{\alpha_i}{i} = \frac{i+(k-i)\alpha_i}{ik}$.

The weights are defined so that in the first scenario, on average all bins (neglecting the bins which are not complete) have a total weight of at least 1 packed in them according to the first weight measure, and otherwise the same property holds according to the second weight measure.

To use Theorem 9, we need to prove that the conditions of the theorem hold.

LEMMA 11. *For every input σ, $ALG(\sigma) \le X_i(\sigma) + c$ holds for some i.*

Neglecting the incomplete bins (which affect only the constant c), we would like to show that $ALG \le X_i + c$. For both weight measures cases, bins with one A item have weight 1, and bins with i class C items have weight 1. Denote the number of B items by $n(B)$, and of E_i items by $n(E_i)$. The number of red E_i items is denoted $n(ER_i)$, and the number of white E_i items $n(EW_i)$ (i.e., $(n(E_i) = n(EW_i) + n(ER_i))$.

According to the color assignments, let $\alpha_i = \frac{p_i}{q_i}$ (a minimal rational representation of α_i). Then $\alpha_i(n(E_i) - (q_i - p_i)) \le n(ER_i) \le \alpha_i n(E_i)$, and $(1-\alpha_i)n(E_i) \le n(EW_i) \le (1-\alpha_i)n(E_i) + q_i - p_i$. In the first scenario, no complete bins contain red E_i items only. The total weight of B and E_i items for all i is $n(B) + \sum_{i=1}^{k-1} \frac{1-\alpha_i}{k} \cdot n(E_i)$. The number of bins used for these types is $n(B) + \sum_{i=1}^{k-1} \frac{n(EW_i)}{k} \le n(B) + \cdot \sum_{i=1}^{k-1} (\frac{1-\alpha_i}{k} n(E_i) + \frac{q_i - p_i}{k})$. In this case we get $ALG < X_1 + c_1$, where c_1 depends on the number of neglected incomplete bins, which is constant (for a given choice of the p_i, q_i values). In the second scenario, no bins contain a B item only. The total weight of B and E items is $\sum_{i=1}^{k-1} \frac{i+(k-i)\alpha_i}{ik} \cdot n(E_i)$. The number of bins used for these types is $\sum_{i=1}^{k-1} \big(\frac{n(EW_i)}{k} + \frac{n(ER_i)}{i}\big) \le \sum_{i=1}^{k-1} \frac{(1-\alpha_i)n(E_i) + q_i - p_i}{k} + \frac{\alpha_i n(E_i)}{i} = \sum_{i=1}^{k-1} n(E_i) \frac{i+(k-i)\alpha_i}{ki} + \frac{q_i - p_i}{k}$. In this case we get $ALG < X_2 + c_2$, where c_2 is a constant which depends on the number of incomplete bins and on the values chosen for q_i, p_i, $1 \le i \le k-1$.

Next we would like to analyze the maximum amount of weight that a bin can contain according to the two weight measures. We do that separately for $k = 4, 5, 6$. We always assume that there are exactly k items in each bin. This is done by allowing items of size 0 that belong to the class E_{k-1}. Note also that we will have ranges of

sizes where weights are fixed to be monotonically nondecreasing functions of size; therefore in these cases, we do not need to consider options where a single item can be replaced by a smaller one.

The case $k = 4$. We are aiming at the competitive ratio $\mathcal{R}(4) = \frac{71}{38} \approx 1.86842$. Define the following values: $\alpha_1 = \frac{1}{19}$, $\alpha_2 = \frac{3}{19}$, $\alpha_3 = \frac{9}{19}$. This implies the weights $w_1(E_1) = \frac{9}{38}$, $w_1(E_2) = \frac{8}{38}$, $w_1(E_3) = \frac{5}{38}$, $w_2(E_i) = \frac{11}{38}$ for $i = 1, 2, 3$.

We compute the maximum amount of weight in a single bin with respect to w_2 first. If no item in the bin is of class A, then the largest weight of any item can be $\frac{1}{2}$. However, a bin can contain at most two such items. All other items have weights of at most $\frac{1}{3}$. This gives a total of at most $\frac{5}{3} < \mathcal{R}(4)$. Next, if a class A item is present, all other others are of classes E_1, E_2, E_3. They all have identical weight. At most three more items can exist, and thus we get the total weight $1 + 3 \cdot \frac{11}{38} = \mathcal{R}(4)$.

Next, we compute the maximum weight with respect to w_1. If no item of weight 1 is present, then all weights are upper bounded by the weights of the same items with respect to w_2, and therefore this case is covered by the calculation done for w_2. Otherwise, an item of weight 1 occupies a space of more than $\frac{1}{2}$. If an item of class C_2 exists, it occupies a space of more than $\frac{1}{3}$, and the two other items are of types E_1, E_2, E_3. Moreover, there is room for only one item of either class E_1 or E_2 (these items are larger than $\frac{1}{12}$). Since weights are monotone for all sizes, the worst case is one item of each class B, C_2, E_1, E_3 whose sum of weights is $1 + \frac{1}{2} + \frac{9}{38} + \frac{5}{38} = \mathcal{R}(4)$. If there is no item of C_2, there are three other items, only one of them can be a C_3 item. This gives the worst case bound $1 + \frac{1}{3} + 2 \cdot \frac{9}{38} < \frac{69}{38} < \mathcal{R}(4)$.

The case $k = 5$. We are aiming at the competitive ratio $\mathcal{R}(5) = \frac{771}{398} \approx 1.93719$. Define the following values: $\alpha_1 = \frac{9}{199}$, $\alpha_2 = \frac{24}{199}$, $\alpha_3 = \frac{54}{199}$, $\alpha_4 = \frac{114}{199}$. This implies the following weights: $w_1(E_1) = \frac{76}{398}$, $w_1(E_2) = \frac{70}{398}$, $w_1(E_3) = \frac{58}{398}$, $w_1(E_4) = \frac{34}{398}$, $w_2(E_i) = \frac{94}{398}$ for $i = 1, 2, 3$ and $w_2(E_4) = \frac{91}{398}$.

We compute the maximum amount of weight in a single bin with respect to w_2 first. If no item in the bin is of class A, then the largest weight of any item can be $\frac{1}{2}$. However, a bin can contain at most two such items, and at most three items larger than $\frac{1}{4}$ (two of which may be of size larger than $\frac{1}{3}$). The weight of three items larger then $\frac{1}{4}$ is therefore at most $2 \cdot \frac{1}{2} + \frac{1}{3}$. All other items have weight of at most $\frac{1}{4}$, which gives a total of at most $\frac{4}{3} + 2 \cdot \frac{1}{4} = \frac{11}{6} < \mathcal{R}(5)$. Next, if a class A item is present, all others are of classes E_1, E_2, E_3, E_4. Four more items are present, but at least one of them must be in class E_4. Items in E_1, E_2, E_3 all have weight $\frac{47}{199}$, and thus we get the total weight of at most $1 + 3 \cdot \frac{47}{199} + \frac{91}{398} = \mathcal{R}(5)$.

Next, we compute the maximum weight with respect to w_1. If no item of weight 1 is present, then again all weights are bounded from above by the weights of the same items with respect to w_2, and therefore this case is covered by the calculation done for w_2. Otherwise, an item of weight 1 occupies a space of more than $\frac{1}{2}$.

If an item of class C_2 exists, it occupies a space of more than $\frac{1}{3}$, and the three other items are of types E_1, E_2, E_3, E_4. Moreover, if there is a class E_1 item, then there is no class E_2 item and at most one class E_3 item. This gives a total weight of $1 + \frac{1}{2} + \frac{38}{199} + \frac{29}{199} + \frac{17}{199} = \frac{765}{398} < \mathcal{R}(5)$. If there is no class E_1 item, then if we have a class E_2 item, we can have another item of either class E_2 or E_3 and a class E_4 item, which gives the weight of at most $1 + \frac{1}{2} + 2 \cdot \frac{35}{199} + \frac{17}{199} = \mathcal{R}(5)$. Finally if there are no E_1 and E_2 items, then the weight is at most $1 + \frac{1}{2} + 3 \cdot \frac{29}{199} = \mathcal{R}(5)$.

If no item of class C_2 exists but there is a class C_3 item, we have the following options. If there is a C_4 item as well, then the occupied area is already more than 0.95, so the other two items are of classes E_4, and this gives weight of at most

$1+\frac{1}{3}+\frac{1}{4}+2\cdot\frac{17}{199}<\frac{699}{398}<\mathcal{R}(5)$. If there is no C_4 item, then the largest weight of the additional three items can be $\frac{38}{199}$ each, which bounds the weight by $1+\frac{1}{3}+3\cdot\frac{38}{199}<\frac{759}{398}<\mathcal{R}(5)$.

If no items of classes C_2, C_3 exist, there can be at most two C_4 items, and other items have weight at most $\frac{38}{199}$, which together bounds the weight by $1+2\cdot\frac{1}{4}+2\cdot\frac{38}{199}<\frac{749}{398}<\mathcal{R}(5)$.

The case $k=6$. We are aiming at the competitive ratio $\mathcal{R}=\frac{287}{144}\approx 1.99306$. Define the following values: $\alpha_1=\frac{2}{48}=\frac{1}{24}$, $\alpha_2=\frac{5}{48}$, $\alpha_3=\frac{10}{48}=\frac{5}{24}$, $\alpha_4=\frac{20}{48}=\frac{5}{12}$, $\alpha_5=\frac{30}{48}=\frac{5}{8}$. This implies the following weights: $w_1(E_1)=\frac{46}{288}$, $w_1(E_2)=\frac{43}{288}$, $w_1(E_3)=\frac{38}{288}$, $w_1(E_4)=\frac{28}{288}$, $w_1(E_5)=\frac{18}{288}$, $w_2(E_i)=\frac{58}{288}$ for $i=1,2,3,4$ and $w_2(E_5)=\frac{54}{288}=\frac{3}{16}$.

We compute the maximum amount of weight in a single bin with respect to w_2 first. If no item in the bin is of class A, then a bin can contain at most two such items larger than $\frac{1}{3}$, or at most three items larger than $\frac{1}{4}$, or at most four items larger than $\frac{1}{5}$. The worst case gives two items of class C_2, one of C_3 and one of C_4. The two other items have weight of at most $\frac{29}{144}$ (since $\frac{1}{5}<\frac{29}{144}$), which gives a total of at most $2\cdot\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+2\cdot\frac{29}{144}=\frac{286}{144}<\mathcal{R}(6)$. Next, if a class A item is present, all others are of classes E_i, $1\le i\le 5$. Five more items are present, but at least one of them must be in class E_5. Items in E_1, E_2, E_3, E_4 all have weight $\frac{29}{144}$, and thus we get the total weight of at most $1+4\cdot\frac{29}{144}+\frac{27}{144}=\mathcal{R}(6)$.

Next, we compute the maximum weight with respect to w_1. If no item of weight 1 is present, then again all weights are upper bounded by the weights of the same items with respect to w_2, and therefore this case is covered by the calculation done for w_2. Otherwise, an item of weight 1 occupies a space of more than $\frac{1}{2}$. Consider the other contents of the bin. We replace an item of class C_i with an item of size $\frac{1}{i+1}$ (without changing its weight). Similarly we replace an item of class E_i with an item of size $\frac{1}{6(i+1)}$ for $i<k-1$ and with an item of size 0 if $i=5$. We only decrease sizes of items; therefore they all fit into the bin. We define the expansion of an item of size x of weight w to be $r(x,w)=\frac{w-\frac{1}{16}}{x}$, and for $x=0$ the expansion is 0. Note that the weight of a set of i items of total size S and of maximum expansion s is at most $Ss+\frac{i}{16}$.

The expansions for classes $C_2,\dots,C_5$ are $\frac{189}{144}=1.3125, \frac{156}{144}\approx 1.08333, \frac{135}{144}=0.9375, \frac{33}{40}=0.825$ (respectively). The expansions for classes $E_1,\dots,E_5$ are $\frac{7}{6}\approx 1.166667, \frac{25}{16}=1.5625, \frac{5}{3}\approx 1.66667, \frac{150}{144}\approx 1.041667, 0$ (respectively).

Let e_2 and e_3 be the numbers of items of classes E_2 and E_3. If there is no class C_2 item, we can bound the weight as follows. There are $i-e_2-e_3$ other items; therefore the weight is bounded by $1+e_2\frac{43}{288}+e_3\frac{19}{144}+\frac{5-e_1-e_2}{16}+(\frac{1}{2}-\frac{e_2}{18}-\frac{e_3}{24})\cdot\frac{7}{6}=\frac{91}{48}+\frac{19}{864}e_2+\frac{18}{864}e_3$. If $e_2+e_3\le 4$, we get at most $\frac{857}{432}<\mathcal{R}(6)$. Otherwise, if $e_1+e_2=5$, we do not have any other items except for an item of weight 1 and five items of weight $\frac{43}{288}$ or $\frac{38}{288}$, which gives a total of at most $\frac{503}{288}<\mathcal{R}(6)$.

If there is an item of class C_2, the empty space left is less than $\frac{1}{6}$. This means that $i=e_2+e_3\le 3$ and $e_2\le 2$. We get a total weight of at most $1.5+e_2\frac{43}{288}+e_3\frac{19}{144}+\frac{4-e_2-e_3}{16}+(\frac{1}{6}-\frac{e_2}{18}+\frac{e_3}{24})\cdot\frac{7}{6}=\frac{35}{18}+\frac{19}{864}e_2+\frac{18}{864}e_3$. If $e_2+e_3\le 2$, we can bound the weight by $\frac{859}{432}<\mathcal{R}(6)$. We are left with the cases $e_2=2, e_3=1$, $e_2=1, e_3=2$, $e_2=0, e_3=3$. In the first two cases, only an item of class E_5 can be added to the bin. In the last case, an item of class E_4 or E_5 can be added. Therefore we need to consider two cases, where the four small items are of classes E_2, E_2, E_3, E_5 and E_3, E_3, E_3, E_4. We get total weights $1.5+2\cdot\frac{43}{288}+\frac{19}{144}+\frac{1}{16}=\mathcal{R}(6)$ and $1.5+3\cdot\frac{19}{144}+\frac{14}{144}=\mathcal{R}(6)$.

We summarize with the following theorem.

THEOREM 12. *The competitive ratios of the above algorithm are at most* $\frac{71}{38} \approx 1.86842$ *for* $k = 4$, $\frac{771}{398} \approx 1.93719$ *for* $k = 5$, *and* $\frac{287}{144} \approx 1.99306$ *for* $k = 6$.

6. Conclusion. The main open question is whether an algorithm with competitive ratio strictly better than 2 can be designed for all values of k. In this paper we showed that such an algorithm cannot be bounded space (unless $k \leq 3$). We note that the methods used in this paper for small values of k cannot be extended for larger k.

REFERENCES

[1] L. BABEL, B. CHEN, H. KELLERER, AND V. KOTOV, *Algorithms for on-line bin-packing problems with cardinality constraints*, Discrete Appl. Math., 143 (2004), pp. 238–251.

[2] A. CAPRARA, H. KELLERER, AND U. PFERSCHY, *Approximation schemes for ordered vector packing problems*, Naval Research Logistics, 92 (2003), pp. 58–69.

[3] E. G. COFFMAN, M. R. GAREY, AND D. S. JOHNSON, *Approximation algorithms for bin packing: A survey*, in Approximation Algorithms, D. Hochbaum, ed., PWS Publishing Company, Boston, 1997, pp. 46–93.

[4] J. CSIRIK, *An online algorithm for variable-sized bin packing*, Acta Inform., 26 (1989), pp. 697–709.

[5] J. CSIRIK AND G. J. WOEGINGER, *On-line packing and covering problems*, in Online Algorithms: The State of the Art, A. Fiat and G. J. Woeginger, eds., Lecture Notes in Comput. Sci. 1442, Springer, NY, 1998, pp. 147–177.

[6] J. CSIRIK AND G. J. WOEGINGER, *Resource augmentation for online bounded space bin packing*, J. Algorithms, 44 (2002), pp. 308–320.

[7] L. EPSTEIN AND R. VAN STEE, *On variable-sized multidimensional packing*, in Proceedings of the 12th Annual European Symposium on Algorithms (ESA2004), Bergen, Norway, 2004, Lecture Notes in Comput. Sci. 3221, Springer, New York, 2004, pp. 287–298.

[8] L. EPSTEIN AND R. VAN STEE, *Online bin packing with resource augmentation*, in Proceedings of the 2nd Workshop on Approximation and Online Algorithms (WAOA 2004), Bergen, Norway, 2004, Lecture Notes in Comput. Sci. 3351, Springer, New York, 2004, pp. 48–60.

[9] L. EPSTEIN AND R. VAN STEE, *Optimal online bounded space multidimensional packing*, in Proceedings of 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'04), New Orleans, LA, 2004, SIAM, Philadelphia, 2004, pp. 207–216.

[10] D. K. FRIESEN AND M. A. LANGSTON, *Variable sized bin packing*, SIAM J. Comput., 15 (1986), pp. 222–230.

[11] D. S. JOHNSON, *Fast algorithms for bin packing*, J. Comput. System Sci., 8 (1974), pp. 272–314.

[12] H. KELLERER AND U. PFERSCHY, *Cardinality constrained bin-packing problems*, Ann. Oper. Res., 92 (1999), pp. 335–348.

[13] K. L. KRAUSE, V. Y. SHEN, AND H. D. SCHWETMAN, *Analysis of several task-scheduling algorithms for a model of multiprogramming computer systems*, J. ACM, 22 (1975), pp. 522–550.

[14] K. L. KRAUSE, V. Y. SHEN, AND H. D. SCHWETMAN, *Errata: "Analysis of several task-scheduling algorithms for a model of multiprogramming computer systems*, J. ACM, 24 (1977), pp. 527–527.

[15] C. C. LEE AND D. T. LEE. *A simple online bin packing algorithm*, J. ACM, 32 (1985), pp. 562–572.

[16] P. RAMANAN, D. J. BROWN, C. C. LEE, AND D. T. LEE, *Online bin packing in linear time*, J. Algorithms, 10 (1989), pp. 305–326.

[17] S. S. SEIDEN, *An optimal online algorithm for bounded space variable-sized bin packing*, SIAM J. Discrete Math., 14 (2001), pp. 458–470.

[18] S. S. SEIDEN, *On the online bin packing problem*, J. ACM, 49 (2002), pp. 640–671.

[19] S. S. SEIDEN, R. VAN STEE, AND L. EPSTEIN, *New bounds for variable-sized online bin packing*, SIAM J. Comput., 32 (2003), pp. 455–469.

[20] J. D. ULLMAN, *The Performance of a Memory Allocation Algorithm*, Technical Report 100, Princeton University, Princeton, NJ, 1971.

[21] A. VAN VLIET, *An improved lower bound for online bin packing algorithms*, Inform. Process. Lett., 43 (1992), pp. 277–284.

[22] G. WOEGINGER, *Improved space for bounded-space, online bin packing*, SIAM J. Discrete Math., 6 (1993), pp. 575–581.

[23] A. C. C. YAO, *New algorithms for bin packing*, J. ACM, 27 (1980), pp. 207–227.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 1031–1041

SET SYSTEMS WITH NO SINGLETON INTERSECTION*

PETER KEEVASH†, DHRUV MUBAYI‡, AND RICHARD M. WILSON†

Abstract. Let $\mathcal{F}$ be a k-uniform set system defined on a ground set of size n with no singleton intersection; i.e., no pair $A, B \in \mathcal{F}$ has $|A \cap B| = 1$. Frankl showed that $|\mathcal{F}| \le \binom{n-2}{k-2}$ for $k \ge 4$ and n sufficiently large, confirming a conjecture of Erdős and Sós. We determine the maximum size of $\mathcal{F}$ for $k = 4$ and all n, and also establish a stability result for general k, showing that any $\mathcal{F}$ with size asymptotic to that of the best construction must be structurally similar to it.

Key words. extremal set theory, restricted intersections

AMS subject classification. 05D05

DOI. 10.1137/050647372

1. Introduction.

Say that a set system $\mathcal{F}$ is *L-intersecting* if for every $A, B \in \mathcal{F}$ we have $|A \cap B| \in L$. Ray-Chaudhuri and Wilson [18] and Frankl and Wilson [10] obtained tight bounds for L-intersecting set systems. They showed that if L is a set of s nonnegative integers and $\mathcal{F}$ is an L-intersecting system on $[n] = \{1, \dots, n\}$, then $|\mathcal{F}|$ is at most $\sum_{i=0}^{s} \binom{n}{i}$, and at most $\binom{n}{s}$ if $\mathcal{F}$ is *k-uniform* for some k, i.e., $|A| = k$ for each $A \in \mathcal{F}$. Equality can hold in these bounds when $L = \{0, \dots, s-1\}$. It is natural to ask for the best possible bound for each specific set L, and in general it is an open problem to even determine the order of magnitude. A detailed compilation of results on the uniform version of this problem can be found in [9].

We will consider the problem of finding the largest k-uniform family with no singleton intersection, i.e., L-intersecting with $L = \{0, 2, 3, \dots, k\}$. One construction of such a family is to take all k-subsets of $[n]$ that contain two specified points; this gives a family of $\binom{n-2}{k-2}$ sets with no singleton intersection, which also happens to have no empty intersection. Erdős and Sós (see [4]) conjectured that this is the maximum number for $k \ge 4$ and sufficiently large n, and this was proved by Frankl [6]. (Note that when $n = 3$ the maximum number is n, which can be achieved when n is divisible by 4 by taking $n/4$ vertex disjoint copies of $K_4^{(3)}$, i.e., the complete triple system on 4 points.)

For a more complete understanding of the problem, one might hope to find the maximum number for all n and to describe the structure of the maximum systems. Our first theorem achieves this when $k = 4$, and our approach gives some additional structural information for general k. Our basic idea is to consider a maximum matching and estimate the rest of the family based on the intersections of its sets with this matching. The same technique has recently been successful for various other extremal problems, such as in [13] and [16].

Before stating our first theorem, we should mention the fundamental intersection theorem of Erdős, Ko, and Rado [5]. Say that a set system $\mathcal{F}$ is *t-intersecting* if

*Received by the editors December 12, 2005; accepted for publication (in revised form) June 5, 2006; published electronically December 15, 2006.
http://www.siam.org/journals/sidma/20-4/64737.html

†Department of Mathematics, Caltech, Pasadena, CA 91125 (keevash@caltech.edu, rmw@caltech.edu). The first author's research was supported in part by NSF grant DMS-0555755.

‡Department of Mathematics, Statistics and Computer Science, University of Illinois, Chicago, IL 60607 (mubayi@math.uic.edu). This author's research was supported in part by NSF grant DMS-0400812 and by an Alfred P. Sloan fellowship.

for every $A, B \in \mathcal{F}$ we have $|A \cap B| \geq t$. They showed that, if $\mathcal{F}$ is k-uniform and t-intersecting on $[n]$ with n sufficiently large, then $|\mathcal{F}| \leq \binom{n-t}{k-t}$. (The case $t = 2$ is pertinent to our current discussion.) Confirming a conjecture of Erdős, Wilson [19] showed that this bound in fact holds for $n \geq (t+1)(k-t+1)$ (which is the best possible strengthening), and furthermore that the unique maximum system consists of all k-sets containing some fixed t-set. To describe the complete solution for all n we need to define the t-intersecting systems $\mathcal{F}_i^{k,t}(n) = \{A \subset [n] : |A| = k, |A \cap [t+2i]| \geq t+i\}$ for $0 \leq i \leq k-t$. The complete intersection theorem, conjectured by Frankl and proved by Ahlswede and Khachatrian [1], is that a maximum size k-uniform t-intersecting family on $[n]$ is isomorphic to $\mathcal{F}_i^{k,t}(n)$, for some i which can easily be computed given n. Note that $\mathcal{F}_0^{k,t}(n)$ is the system of all k-sets containing some fixed t-set. These constructions also appear in our analysis for 4-uniform systems with no singleton intersection.

THEOREM 1.1. *Suppose that $\mathcal{F}$ is a 4-uniform set system on $[n]$ with no pair $A, B \in \mathcal{F}$ satisfying $|A \cap B| = 1$. Then*

$$|\mathcal{F}| \leq \begin{cases} \binom{n}{4}, & n = 4, 5, 6, \\ 15, & n = 7, \\ 17, & n = 8, \\ \binom{n-2}{2}, & n \geq 9. \end{cases}$$

Furthermore, the only cases of equality are $K_n^{(4)}$ for $n = 4, 5$, $\mathcal{F}_2^{4,2}(n) = K_6^{(4)}$ for $n = 6, 7$, $\mathcal{F}_1^{4,2}(8)$ for $n = 8$, and $\mathcal{F}_0^{4,2}(n)$ for $n \geq 9$.

Many extremal problems have a property known as stability, meaning that not only do they have a unique maximizing construction, but also any family with size asymptotic to that of the best construction must be structurally similar to it. Stability theorems can be useful tools for establishing exact results (e.g., [15]) and for enumerating discrete structures (e.g., [3]). They are also interesting in their own right, as they provide information about the problem that is structural, rather than just numerical, and they often motivate new proof techniques where the original ones do not suffice.

A strong stability version of the Erdős–Ko–Rado theorem was obtained by Frankl [7], extending an earlier result of Hilton and Milner [11]. A similar result with different assumptions on the parameters was also obtained by Anstee and Keevash [2]. A simple consequence of Frankl's theorem (which is also easy to prove directly) is that for any k there is $c(k)$ such that, if $\mathcal{F}$ is k-uniform and t-intersecting on $[n]$ with $|\mathcal{F}| > c(k)n^{k-t-1}$ and n sufficiently large, then there is a set of t points that is contained in every set of $\mathcal{F}$.

These stability theorems are stronger than the usual stability paradigm in two senses: first the supposed lower bound on $|\mathcal{F}|$ is of a lower order of magnitude than the maximum possible (rather than asymptotic to it), and second the conclusion is that $\mathcal{F}$ is contained in the best construction (rather than structurally similar to it). An example of a stability theorem for set systems that is not strong was given by Mubayi [17]. Also, a strong stability theorem cannot hold for our problem of having no singleton intersection. To see this, note that if $\mathcal{A}$ and $\mathcal{B}$ are families on disjoint sets X and Y with no singleton intersection, then $\mathcal{A} \cup \mathcal{B}$ is a family on $A \cup Y$ with no singleton intersection. If $X \cup Y = [n]$ and $|Y| = o(n)$, we can take $|\mathcal{A}| \sim \binom{n-2}{k-2}$, but there need not be two points that belong to all of the sets. Our next result is a (normal) stability theorem for systems having no singleton intersection.

THEOREM 1.2. *For any $\epsilon > 0$ there is $\delta > 0$ such that if $\mathcal{F}$ is a k-uniform family on $[n]$ with no singleton intersection and $|\mathcal{F}| \geq (1-\delta)\binom{n-2}{k-2}$, then there are two points x, y so that all but at most ϵn^{k-2} sets of $\mathcal{F}$ contain both x and y.*

A result that is useful in the proof of Theorem 1.2, and is of independent interest, is the following bound, which is slightly suboptimal but has the advantage of being valid for all n.

THEOREM 1.3. *Let $\mathcal{F}$ be a k-uniform family on $[n]$ with no singleton intersection, where $k \geq 3$. Then $|\mathcal{F}| \leq \binom{n}{k-2}$.*

The rest of this paper is organized as follows. We start, in the next section, by quickly deducing Theorem 1.3 from a result of Frankl and Wilson [10]. Then we prove Theorem 1.1 in section 3. Some lemmas used in the proof of Theorem 1.2 are given in section 4, and the proof itself in section 5.

Notation. We write $[n] = \{1, \dots, n\}$. Typically $\mathcal{F}$ is a k-uniform set system (or family, or hypergraph) with ground set $[n]$. Given $A \subset [n]$, the link of $\mathcal{F}$ from A is $\mathcal{F}(A) = \{F \backslash A : A \subset F \in \mathcal{F}\}$. The complete r-uniform hypergraph on s vertices is denoted $K_s^{(r)}$. For $0 \leq i \leq k-t$ we define $\mathcal{F}_i^{k,t}(n) = \{A \subset [n] : |A| = k, |A \cap [t+2i]| \geq t+i\}$.

2. A bound for all n. In this section we prove Theorem 1.3. It is a simple consequence of the following theorem of Frankl and Wilson, implicit in [10]. For the convenience of the reader we briefly reproduce their proof.

THEOREM 2.1. *Suppose that p is prime, $k \in \mathbb{N}$, $L \subset \{0, \dots, k-1\}$, and $f(x)$ is an integer-valued polynomial of degree $d \leq k$ such that $f(\ell) \equiv 0 \bmod p$ for $\ell \in L$ and $f(k) \not\equiv 0 \bmod p$. If $\mathcal{F}$ is a k-uniform L-intersecting set system on $[n]$, then $|\mathcal{F}| \leq \binom{n}{d}$.*

Proof. Let $W_{i,j}$ be the matrix with rows indexed by the i-subsets of $[n]$ and columns by the j-subsets of $[n]$, where, given $|A| = i$ and $|B| = j$, the entry $W_{i,j}(A, B)$ is 1 if $A \subset B$ and 0 if $A \not\subset B$. Let V be the row space of $W_{d,k}$. The identity $W_{i,d}W_{d,k} = \binom{k-i}{d-i}W_{i,k}$ implies that V contains the row space of $W_{i,k}$ for all $i \leq d$. Since f is integer-valued there are integers $a_0, \dots, a_d$ such that $f(x) = \sum_{i=0}^{d} a_i \binom{x}{i}$, where $\binom{x}{i}$ is the polynomial $\frac{1}{i!}x(x-1)\cdots(x-i+1)$. Consider the matrix $M = \sum_{i=0}^{d} a_i W_{i,k}^T W_{i,k}$. The row space of M is contained in V, so rank $M \leq \dim V \leq \binom{n}{d}$. On the other hand, given k-sets A, B, we have $M(A, B) = \sum_{i=0}^{d} a_i \binom{|A \cap B|}{i} = f(|A \cap B|)$. Let M_0 be the submatrix of M consisting of elements $M(A, B)$ with $A, B \in \mathcal{F}$. By our assumptions $M(A, B) \equiv 0 \bmod p$ for $A \neq B$ and $M(A, A) \not\equiv 0 \bmod p$, and so M_0 is nonsingular. Therefore $|\mathcal{F}| = \text{rank } M_0 \leq \text{rank } M \leq \binom{n}{d}$. ∎

Proof of Theorem 1.3. Let p be a prime that divides $k-1$ and $f(x) = \binom{x-2}{k-2}$, a polynomial of degree $k-2$. Then $f(i) = 0$ for $2 \leq i \leq k-1$, $f(0) = (-1)^{k-2}(k-1) \equiv 0 \bmod p$, and $f(k) = 1$. By Theorem 2.1, if $\mathcal{F}$ is a k-uniform family on $[n]$ with no singleton intersection, then $|\mathcal{F}| \leq \binom{n}{k-2}$. ∎

3. Solution for 4-uniform families. Throughout we suppose that $\mathcal{F}$ is a 4-uniform set system on $[n]$ with no singleton intersection; i.e., there is no pair $A, B \in \mathcal{F}$ with $|A \cap B| = 1$. In this section we will prove Theorem 1.1, which describes such families $\mathcal{F}$ of maximum size. We start by discussing the small values of n. Trivially $K_n^{(4)}$ is the maximum family for $n = 4, 5, 6$. Also, when $n = 7$ then $\mathcal{F}$ cannot contain two disjoint sets and so is 2-intersecting, and the complete intersection theorem shows that the maximum family is $\mathcal{F}_2^{4,2}(7) = K_6^{(4)}$. Next suppose that $n = 8$. If $\mathcal{F}$ does not contain two disjoint sets, then as before it is 2-intersecting and so contains at most

17 sets, with equality only for $\mathcal{F}_1^{4,2}(8)$. In fact, this is the maximum family, as shown by the case $t = 2$ of the next lemma.

LEMMA 3.1. *Suppose that $\mathcal{F}$ is a 4-uniform family on $[n]$ with no singleton intersection and contains a perfect matching $A_1, \dots, A_t$, with $t \geq 2$. Then $|\mathcal{F}| \leq 3\binom{2t}{2} - 2t$.*

Proof. We argue by induction on t. First we do the base case, where $t = 2$ and it is required to show that $|\mathcal{F}| \leq 14$. Note that every set in $\mathcal{F}$ other than A_1 or A_2 has two points in each of A_1 and A_2. Given a pair uv in A_1, let $\mathcal{F}(uv)$ be its link in A_2, i.e., the set of pairs xy in A_2 for which $uvxy$ is in $\mathcal{F}$, and write $d(uv) = |\mathcal{F}(uv)|$. Since $\mathcal{F}$ has no singleton intersection the links have the following properties:

(i) If uv and wx are disjoint pairs in A_1 and a, b, c are distinct points of A_2, then we do not have $ab \in \mathcal{F}(uv)$ and $ac \in \mathcal{F}(wx)$.

(ii) If ab and cd are disjoint pairs in A_2 and u, v, w are distinct points of A_1, then we do not have $ab \in \mathcal{F}(uv)$ and $cd \in \mathcal{F}(uw)$.

We consider cases according to the maximum value of $d(uv)$. The above properties imply that if there is a pair uv in A_1 with $d(uv) = 6$, then $d(u'v') = 0$ for all other pairs $u'v'$ in A_1, and if there is a pair uv in A_1 with $d(uv) = 5$, then $d(u'v') \leq 1$ for all other pairs $u'v'$ in A_1. In either case we have $|\mathcal{F}| = 2 + \sum_{u,v \in A_1} d(uv) < 14$. Otherwise, if $d(uv) \leq 4$ for all pairs uv in A_1, we claim that for any two opposite pairs uv, wx in A_1 we have $d(uv) + d(wx) \leq 4$. To see this, we can suppose that, say, $d(uv) \geq 3$. But now, if $ab \in \mathcal{F}(wx)$, by property (i) $\mathcal{F}(uv)$ can contain only ab or $A_2 \setminus ab$, contradicting the assumption that $d(uv) \geq 3$. Therefore $d(wx) = 0$, so $d(uv) + d(wx) \leq 4$. Since K_4 can be decomposed into 3 matchings, $|\mathcal{F}| = 2 + \sum_{u,v \in A_1} d(uv) \leq 2 + 3 \cdot 4 = 14$, as required.

Now suppose $t \geq 3$. By the case $t = 2$, for every $1 \leq i \leq t - 1$ there are at most 12 sets with 2 points in each of A_i and A_t. Thus there are at most $12(t-1) + 1$ sets incident to A_t. By an induction hypothesis there are at most $3\binom{2(t-1)}{2} - 2(t-1)$ sets within $\cup_{i=1}^{t-1} A_i$, so in total we have at most $3\binom{2t}{2} - 2t$. ∎

The heart of the proof of Theorem 1.1 is contained in the following theorem, which in the case when $\mathcal{F}$ is not intersecting gives a stronger bound on its size. We define

$$b_2(n) = 13 + \max\left\{7(n-8), \binom{n-6}{2}\right\} \qquad \text{and}$$

$$b_t(n) = 3\binom{2t}{2} - 2t - 1 + \max\left\{3t(n-4t), \binom{n-4t+2}{2}\right\} \qquad \text{for } t \geq 3.$$

THEOREM 3.2. *Suppose that $\mathcal{F}$ is a 4-uniform family on $[n]$ with no singleton intersection. Let $A_1, \dots, A_t$ be a maximum matching in $\mathcal{F}$, and suppose $t \geq 2$. Then $|\mathcal{F}| \leq b_t(n)$.*

Proof. Let $A = \cup_{i=1}^{t} A_i$ and $B = [n] \setminus A$. By maximality of t there are no sets of $\mathcal{F}$ contained in B. The sets contained within A may be estimated by Lemma 3.1: there are at most $3\binom{2t}{2} - 2t$ of them. The remaining sets intersect both A and B, and since there are no singleton intersections they have two possible types: 2 points in some A_i and 2 in B, or 3 points in some A_i and 1 in B.

Say that a pair xy in B has color i if there is a pair ab in A_i such that $abxy$ is a set of $\mathcal{F}$. Note that a pair may have more than one color or be uncolored. Let M be the set of all pairs xy in B which are colored but do not intersect any other colored pair. Thus M is a perfect matching on some set $D \subset B$. Now if a pair xy has more than

one color, there can be no set of $\mathcal{F}$ that intersects it in one point: this would create a singleton intersection. In this case all sets in $\mathcal{F}$ meeting xy consist of xy together with a pair in some A_i, so there are at most $6t$ such sets. On the other hand, if xy has a unique color i, then all sets meeting it are contained in $A_i \cup \{x, y\}$, so there are at most $\binom{6}{4} - 1 = 14$ such sets. Thus the number of sets of $\mathcal{F}$ meeting a colored pair xy is at most $\max\{6t, 14\}$. Setting $d = |D| = 2|M|$, this gives at most $\max\{3td, 7d\}$ sets of $\mathcal{F}$ meeting D.

All other colored pairs are contained in $B \backslash D$. Let G_i be those of color i and C_i be those vertices contained in some pair of G_i. Note that C_i can be empty. The crucial observation of the proof is that $C_1, \ldots, C_t$ are disjoint (and so the same is true of $G_1, \ldots, G_t$). To see this, suppose to the contrary that $x \in C_i \cap C_j$. Then $xy \in G_i$ and $xz \in G_j$ for some y, z. If $y \neq z$, then we would have a singleton intersection in $\mathcal{F}$. On the other hand, if $y = z$, we note that since $xy \notin M$ there is another colored pair P that intersects it. A color of P is different from at least one of i and j, so again we have a singleton intersection. Thus $C_1, \ldots, C_t$ are disjoint.

Let $C = \cup_{i=1}^{t} C_i$ and $E = B \backslash (C \cup D)$. Any set in $\mathcal{F}$ meeting E has 1 point in E and 3 points in some A_i, which must be uniquely specified to avoid a singleton intersection. Thus there are at most $4e$ such sets, where $e = |E|$. All other sets in $\mathcal{F}$ meet C, so are contained in $A_i \cup C_i$ for some i.

Next we note that the sets in $\mathcal{F}$ within $A_i \cup C_i$ form a 2-intersecting family; for there are no singleton intersections, and if $A_i \cup C_i$ contained two disjoint sets, we could enlarge the matching $A_1, \ldots, A_t$. Let $c_i = |C_i|$, so that $|A_i \cup C_i| = c_i + 4$. Note that c_i is either 0 or ≥ 3, as $c_i = 2$ would correspond to a colored pair that does not intersect any other colored pair, but by definition these pairs belong to D, not C. By the complete intersection theorem, the number of sets within $A_i \cup C_i$ is at most $f(c_i)$, defined to be $\binom{c_i+2}{2}$ for $c_i \geq 5$, 17 for $c_i = 4$, 15 for $c_i = 3$, 1 for $c_i = 0$. Now, given $|C| = c = \sum_{i=1}^{t} c_i$, we claim that $\sum_{i=1}^{t} f(c_i) \leq f(c) + t - 1$, with equality holding when $c_i = c$ for some i, and $c_j = 0$ otherwise. This follows from a variational argument, using the inequalities $f(a+1) + f(b-1) \geq f(a) + f(b)$ for $a \geq b \geq 4$ and $f(a+3) + f(0) \geq f(a) + f(3)$ for $a \geq 3$, which are easy to verify. Excluding the sets $A_1, \ldots, A_t$, we conclude that the number of sets in $\mathcal{F}$ meeting C is at most $f(c) - 1$.

Putting everything together, we have $|\mathcal{F}| \leq 3\binom{2t}{2} - 2t + \max\{3td, 7d\} + 4e + f(c) - 1$, where $n = 4t + c + d + e$. For $t \geq 3$ we can write $|\mathcal{F}| \leq 3\binom{2t}{2} - 2t + 3t(n - 4t - c) + f(c) - 1$. This is a quadratic in c with positive coefficient of c^2 for $5 \leq c \leq n - 4t$, so in this range its maximum occurs at $c = 5$ or $c = n - 4t$. Furthermore, it is easy to see that the value at $c = 0$ is larger than at $c = 3, 4, 5$ (and $c = 2$ is impossible as no c_i equals 1 or 2). Therefore the overall maximum occurs at $c = 0$ or $c = n - 4t$, which gives the stated bound. The bound for $t = 2$ follows in the same way, replacing $3td$ by $7d$ in the upper bound for $\mathcal{F}$. □

Proof of Theorem 1.1. Suppose that $\mathcal{F}$ is a 4-uniform family on $[n]$ with no singleton intersection. If $\mathcal{F}$ is intersecting, then it is 2-intersecting, and so we are done by the complete intersection theorem. Otherwise, suppose that the maximum matching has size $t \geq 2$, so $n \geq 8$. We have an upper bound on $|\mathcal{F}|$ given in Theorem 3.2, and we claim that this is always less than $\binom{n-2}{2}$.

First we consider the case $t = 2$. When $n = 8$ we have $\binom{n-2}{2} = 15$ and $b_2(n) = 14$; when $n = 9$ we have $\binom{n-2}{2} = 21$ and $b_2(n) = 20$. For $n \geq 10$ we note that $\binom{n-2}{2} - \binom{n-1-2}{2} = n - 3$ and $b_2(n) - b_2(n-1) \leq \max\{7, n-6\} \leq n - 3$, where we use the inequality $\max\{a, b\} - \max\{a', b'\} \leq \max\{a - a', b - b'\}$. Therefore $b_2(n) < \binom{n-2}{2}$ for all $n \geq 8$.

For general t, when $n = 4t$ we have $\binom{n-2}{2} - b_t(4t) = (2t-3)(t-1) > 0$. Also, for $n > 4t$ we have $\binom{n-2}{2} - \binom{n-1-2}{2} = n-3$ and $b_t(n) - b_t(n-1) \le \max\{3t, n-4t+1\} \le n-3$, so $b_t(n) < \binom{n-2}{2}$ for all $n \ge 4t$. □

4. Three lemmas. Here we prove some lemmas that will be used in the next section. Our first lemma concerns a multicolored version of our problem, in the sense of [12].

LEMMA 4.1. *Suppose that $\mathcal{F}_1, \dots, \mathcal{F}_c$ are k-uniform families on $[n]$ so that there is no $X \in \mathcal{F}_i$, $Y \in \mathcal{F}_j$ with $|X \cap Y| = 1$ for any $i \ne j$. Then $\sum |\mathcal{F}_i| \le c\binom{n}{k-2} + \binom{n}{k}$.*

Proof. Let $\mathcal{A}$ be the family of sets that occur in more than one $\mathcal{F}_i$, and $\mathcal{B}$ the family of sets that occur in exactly one $\mathcal{F}_i$. Then $\mathcal{A}$ has no singleton intersection, so $|\mathcal{A}| \le \binom{n}{k-2}$ by Theorem 1.3. Therefore $\sum |\mathcal{F}_i| \le c|\mathcal{A}| + |\mathcal{B}| \le c\binom{n}{k-2} + \binom{n}{k}$. □

Remark. By analogy with [14] one might expect that the bound can be improved to $\max\{c\binom{n}{k-2}, \binom{n}{k}\}$, but we do not need such a bound here.

Next we have a lemma on matchings. The argument is similar to one given by Frankl [8, Proposition 11.6]. Here also, it should be possible to replace the summation with a maximum.

LEMMA 4.2. *Suppose that X and Y are disjoint sets with $|X| = x$, $|Y| = y$ and $\mathcal{F}$ is a set system on $X \cup Y$ such that $|F \cap X| = s$, $|F \cap Y| = t$ for every $F \in \mathcal{F}$. If $\mathcal{F}$ contains no matching of size m, then $|\mathcal{F}| < m(\binom{x-1}{s-1}\binom{y}{t} + \binom{y-1}{t-1}\binom{x}{s})$.*

Proof. We argue by induction on s, t, x, y. First we note that in the case $x \le ms$ the number of possible intersections of a set F with X is $\binom{x}{s} \le m\binom{x-1}{s-1}$, so trivially $|\mathcal{F}| \le \binom{x}{s}\binom{y}{t} < m(\binom{x-1}{s-1}\binom{y}{t} + \binom{y-1}{t-1}\binom{x}{s})$. Similarly we are done when $y \le mt$. To complete the base of the induction, note that in the case $s = t = 1$ the system $\mathcal{F}$ is a bipartite graph with no matching of size m, and it is easy to see (e.g., by König's theorem) that $|\mathcal{F}| < m \max\{x, y\} \le m(x+y)$.

For the general case, we use the compression method of Erdős, Ko, and Rado [5]. Define arbitrary linear orders $<_X$ on X and $<_Y$ on Y. Given $a, b \in X$ $a <_X b$, we define the ab-shift S_{ab} by $S_{ab}(\mathcal{F}) = \{S_{ab}(F) : F \in \mathcal{F}\}$, where $S_{ab}(F)$ is equal to $F' = F \setminus \{b\} \cup \{a\}$ if $F' \notin \mathcal{F}$, but equal to F if $F' \in \mathcal{F}$. The same definition applies for $a, b \in Y$ with $a <_Y b$. Clearly $|S_{ab}(\mathcal{F})| = |\mathcal{F}|$. A well-known easy property of the shift is that the maximum matching in $S_{ab}(\mathcal{F})$ is no larger than that in $\mathcal{F}$. Iterating these shifts will eventually produce a family which is invariant with respect to S_{ab}, for any $a, b \in X$ or $a, b \in Y$. We can assume that $\mathcal{F}$ has this property.

Suppose, without loss of generality, that $s > 1$. Let a be the maximal element of X. Consider the systems $\mathcal{F}_0 = \{F : a \notin F \in \mathcal{F}\}$ and $\mathcal{F}_1 = \{F \setminus \{a\} : a \in F \in \mathcal{F}\}$ defined on $X \setminus \{a\} \cup Y$. Since $\mathcal{F}_0$ does not have a matching of size m we have $|\mathcal{F}_0| \le m(\binom{x-2}{s-1}\binom{y}{t} + \binom{y-1}{t-1}\binom{x-1}{s})$ by induction. Also $\mathcal{F}_1$ contains no matching of size m. Suppose that $F_1, \dots, F_m$ are disjoint sets in $\mathcal{F}_1$. Each has $s-1$ points in X, so we can find distinct points $a_1, \dots, a_m$ in $X \setminus \cup_{i=1}^m F_i$. Since $\mathcal{F}$ is invariant with respect to ab-shifts with $a, b \in X$, it contains the sets $F_i \cup \{a_i\}$. However, these form a matching, so indeed $\mathcal{F}_1$ contains no matching of size m. Therefore $|\mathcal{F}_1| \le m\big(\binom{x-2}{s-2}\binom{y}{t} + \binom{y-1}{t-1}\binom{x-1}{s-1}\big)$ by induction.

We conclude that $|\mathcal{F}| = |\mathcal{F}_0| + |\mathcal{F}_1| \le m\big(\binom{x-1}{s-1}\binom{y}{t} + \binom{y-1}{t-1}\binom{x}{s}\big)$. □

Finally, we give a simple optimization lemma concerning sums of binomial coefficients.

LEMMA 4.3. *Consider a function $f(z) = \sum_{j=1}^m c_j\binom{z+s_j}{t_j}$, where $c_j \ge 0$ and s_j, t_j are nonnegative integers with $s_j \ge t_j - 1$ for all j. For any positive integers $x_1, \dots, x_n$, writing $x = \sum_{i=1}^n x_i$, we have $\sum_{i=1}^n f(x_i) \le f(x) + (n-1)f(0)$.*

Proof. Note that $\binom{x_i+1+s_j}{t_j}+\binom{x_{i'}-1+s_j}{t_j}-\left(\binom{x_i+s_j}{t_j}+\binom{x_{i'}+s_j}{t_j}\right)=\binom{x_i+s_j}{t_j-1}-\binom{x_{i'}-1+s_j}{t_j-1}$ ≥ 0 if $x_i \geq x_{i'}-1$. So starting from any sequence $x_1,\ldots,x_n$, we can move to the sequence $x,0,\ldots,0$ without decreasing the function $\sum_{i=1}^n f(x_i)$, and the final value gives the stated upper bound. ∎

5. A stability result. In this section we prove Theorem 1.2, which states the following: for any $\epsilon>0$ there is $\delta>0$ such that if $\mathcal{F}$ is a k-uniform family on $[n]$ with no singleton intersection and $|\mathcal{F}|\geq(1-\delta)\binom{n-2}{k-2}$, then there are two points x,y so that all but at most ϵn^{k-2} sets of $\mathcal{F}$ contain both x and y.

Proof of Theorem 1.2. Suppose that $\mathcal{F}$ is a k-uniform family on $[n]$ with no singleton intersection, and $|\mathcal{F}|\geq(1-\delta)\binom{n-2}{k-2}$. We can suppose in all estimates that δ is sufficiently small and n is sufficiently large (by making δ small). Let $A_1,\ldots,A_t$ be a matching in $\mathcal{F}$ with t as large as possible. If $t=1$, then $\mathcal{F}$ is intersecting, and thus 2-intersecting. As we mentioned in the Introduction, a result of Frankl implies that there is a constant $c(k)$ such that if $\mathcal{F}$ is 2-intersecting and $|\mathcal{F}|>c(k)n^{k-3}$, then there are two points x,y contained in every set of $\mathcal{F}$. Since $|\mathcal{F}|\geq(1-\delta)\binom{n-2}{k-2}>c(k)n^{k-3}$ for large n, we are done in the case $t=1$. Now suppose $t\geq 2$. Let $A=\cup_{i=1}^t A_i$, $B=[n]\backslash A$. Note that all sets in $\mathcal{F}$ meet A, and if they meet any A_i, they meet it in at least 2 points.

Let $\mathcal{F}'\subset\mathcal{F}$ be the family of sets meeting exactly one A_i, i.e.,

$$\mathcal{F}'=\{F\in\mathcal{F}\ :\ \exists 1\leq i(F)\leq t,\ F\cap A_{i(F)}\neq\emptyset,\ F\cap A_j=\emptyset\ \ \forall j\neq i(F)\}.$$

Let $\mathcal{G}=\{F\cap B : F\in\mathcal{F}'\}$. Say that $G\in\mathcal{G}$ has color i if $G=F\cap B$ for some F that meets A_i. (A set can have more than one color.) For $b\in B$ a "flower" on b is a system $\{G_1,\ldots,G_{k-2}\}\subset\mathcal{G}$, so that $G_i\cap G_j=\{b\}$ for every $i\neq j$. The key observation is that if there is a flower on b, then there is a unique i so that all sets in $\mathcal{G}$ containing b have color i and no other color. To see this, first note that all the sets in the flower must have the same color (say i), and no other, to avoid a singleton intersection. Now consider any $G\in\mathcal{G}$ that contains b. Then $|G|\leq k-2$, so there are at most $k-3$ sets in the flower that intersect G in a point other than b. Therefore we can find $1\leq j\leq k-2$ so that $G_j\cap G=b$, and so to avoid a singleton intersection, G_j and G cannot have two different colors; i.e., both have only color i.

Let X_i be the set of all b for which there is a flower of color i on b. It follows from the above observation that $X_1,\ldots,X_t$ are pairwise disjoint. We also note for future reference that there are no two disjoint sets of $\mathcal{F}$ contained in $A_i\cup X_i$ for any i; otherwise we could use them instead of A_i to find a larger matching in $\mathcal{F}$. Since there are no singleton intersections, the sets of $\mathcal{F}$ contained in $A_i\cup X_i$ form a 2-intersecting family. Write $X=\cup_{i=1}^t X_i$, $x=|X|$, $x_i=|X_i|$, $Y=B\backslash X$, $y=|Y|$.

Estimate of $|\mathcal{F}'|$. (1) First we count sets corresponding to those elements of $\mathcal{G}$ contained within X and thus within X_i for some i. By Theorem 1.3, $A_i\cup X_i$ contains at most $\binom{x_i+k}{k-2}$ sets. (In fact, we have noted that these sets form a 2-intersecting family, so we could even obtain a stronger bound from the complete intersection theorem mentioned in the introduction, but this expression will be more convenient.)

(2) Next we count sets corresponding to $\mathcal{J}=\{G : G\in\mathcal{G},\ G\subset Y\}$. Let $\mathcal{J}^s=\{G : G\in\mathcal{J},\ |G|=s\}$, and partition $\mathcal{J}^s=\mathcal{J}_1^s\cup\mathcal{J}_2^s$, where $\mathcal{J}_1^s$ contains those G with exactly one color and $\mathcal{J}_2^s$ those with more than one color. Now $\mathcal{J}_2^s$ has no singleton intersection, or there would be corresponding sets in $\mathcal{F}''$ with singleton intersection, so $|\mathcal{J}_2^s|\leq\binom{y}{s-2}$ by Theorem 1.3. It follows that at most $t\binom{k}{s}\binom{y}{s-2}$ sets in $\mathcal{F}'$ correspond to sets of $\mathcal{J}_2^s$.

Also, for each $a \in Y$, $s \geq 2$ the link $\mathcal{J}_1^s(a)$ is a $(s-1)$-uniform system on Y with no matching of size $k-2$. This is immediate from the definition of X, as if $\mathcal{J}_1^s(a)$ has a matching of size $k-2$, then there is a flower on a; i.e., $a \in X$. By Lemma 4.2, $|\mathcal{J}_1^s(a)| \leq (k-2)\binom{y-1}{s-2}$. Therefore the number of sets in $\mathcal{F}'$ corresponding to sets of $\mathcal{J}_1^s$ is at most $\binom{k}{s}\sum_a |\mathcal{J}_1^s(a)| < \binom{k}{s}y(k-2)\binom{y-1}{s-2}$. (In fact, we are even overestimating by an extra factor of s corresponding to the different choices of a in a set of $\mathcal{J}_1^s$.) For $s=1$ we clearly have $|\mathcal{J}_1^1| \leq y$, corresponding to at most ky sets of $\mathcal{F}'$.

In total, the number of sets in $\mathcal{F}'$ corresponding to elements of $\mathcal{J}$ is at most

$$ky + \sum_{s=2}^{k-2}\left(t\binom{k}{s}\binom{y}{s-2} + \binom{k}{s}y(k-2)\binom{y-1}{s-2}\right) < ky + (t+ky)\binom{y+k}{k-4}\sum_{s=2}^{k-2}\binom{k}{s} < 2^k(ky+t)\binom{y+k}{k-4}.$$

(3) Finally, consider those sets corresponding to $\mathcal{K}$, defined as those $G \in \mathcal{G}$ that meet both X and Y. Such a G is contained in $X_i \cup Y$ for some i and has color i but no other color. For $2 \leq s \leq k-2$, let $\mathcal{K}_i^s = \{G : G \in \mathcal{K},\ |G| = s,\ G \subset X_i \cup Y\}$. As in estimate (2), for each $a \in Y$, $s \geq 2$ the link $\mathcal{K}_i^s(a)$ is a $(s-1)$-uniform system with no matching of size $k-2$. Considering each possible intersection size with X_i and Y separately, we apply Lemma 4.2 to get $|\mathcal{K}_i^s(a)| \leq \sum_{\alpha=1}^{s-2}(k-2)\left(\binom{x_i-1}{\alpha-1}\binom{y-1}{s-\alpha-1} + \binom{y-2}{s-\alpha-2}\binom{x_i}{\alpha}\right)$. Applying Lemma 4.3, we can bound the number of sets in $\mathcal{F}'$ corresponding to elements of $\mathcal{K}$ by

$$\begin{aligned}
\sum_{i=1}^{t}\sum_{s=2}^{k-2}\sum_{a\in Y}\binom{k}{s}|\mathcal{K}_i^s(a)| &= \sum_{s=2}^{k-2}\binom{k}{s}\sum_{a\in Y}\sum_{i=1}^{t}|\mathcal{K}_i^s(a)| \\
&\leq \sum_{s=2}^{k-2}\binom{k}{s}y\sum_{i=1}^{t}\sum_{\alpha=1}^{s-2}(k-2)\left(\binom{x_i}{\alpha-1}\binom{y}{s-\alpha-1} + \binom{y}{s-\alpha-2}\binom{x_i}{\alpha}\right) \\
&\leq \sum_{s=2}^{k-2}\binom{k}{s}y(k-2)\sum_{\alpha=1}^{s-2}\left(\binom{x}{\alpha-1}\binom{y}{s-\alpha-1} + \binom{y}{s-\alpha-2}\binom{x}{\alpha}\right) \\
&\leq \sum_{s=2}^{k-2}\binom{k}{s}y(k-2)\cdot 2\binom{x+y}{s-2} \\
&< 2^{k+1}ky\binom{x+y+k}{k-4}.
\end{aligned}$$

Adding the estimates (1), (2), and (3), we have

$$|\mathcal{F}'| \leq \sum_{i=1}^{t}\binom{x_i+k}{k-2} + 2^k(ky+t)\binom{y+k}{k-4} + 2^{k+1}ky\binom{x+y+k}{k-4}.$$

Estimate of $|\mathcal{F}\backslash\mathcal{F}'|$. Suppose that $2 \leq \alpha \leq t$ and $\beta = (\beta_1, \ldots, \beta_\alpha)$ with $\beta_j \geq 2$ and $\beta^* = \sum_{j=1}^{\alpha}\beta_i \leq k$ are given. Let $\mathcal{H}_\beta$ be the collection of all sets $H \subset \cup_i A_i$ such

that the list $|H \cap A_i|$, $1 \le i \le t$, consists of β and $t-\alpha$ zeroes, in some order. Then $|\mathcal{H}_\beta| \le \alpha! \binom{t}{\alpha} \prod_{j=1}^{\alpha} \binom{k}{\beta_j} < 2^{k^2} t^\alpha$, where we crudely estimate that each product term $\binom{k}{\beta_j}$ is at most 2^k and that there are at most k terms (as $\beta^* \le k$).

We can obtain a matching of size t in $\mathcal{H}_\beta$ as follows. For $1 \le i \le t$ let $A_i^1, \dots, A_i^\alpha$ be disjoint subsets of A_i with $|A_i^j| = \beta_j$ for $1 \le j \le \alpha$. Let $M_\gamma = \cup_{j=1}^{\alpha} A_{\gamma+j}^j$ for $1 \le \gamma \le t$, where $A_{\gamma+j}$ is to be interpreted as $A_{\gamma+j-t}$ for $\gamma + j > t$. Then $\mathcal{M} = \{M_1, \dots, M_t\}$ is a matching in $\mathcal{H}_\beta$. Let $\mathcal{G}_\gamma = \{F \cap B : F \cap \cup_i A_i = M_\gamma\}$. Then $\mathcal{G}_\gamma$ for $1 \le \gamma \le t$ are $(k-\beta^*)$-uniform systems satisfying the hypothesis of Lemma 4.1, which thus have total size

$$\sum_{\gamma=1}^{t} |\mathcal{G}_\gamma| \le t\binom{n-kt}{k-\beta^*-2} + \binom{n-kt}{k-\beta^*}.$$

Now we average this estimate over all possible isomorphic choices of the matching $\mathcal{M}$ in $\mathcal{H}_\beta$. Let m be the number of such matchings, and m' be the number of such matchings that contain some fixed set $M \in \mathcal{H}_\beta$ (this is independent of M). By counting pairs $(M, \mathcal{M})$, where $\mathcal{M}$ is a maximum matching containing a set M, we see that $mt = |\mathcal{H}_\beta| m'$. Writing

$$\mathcal{F}_\beta = \{F \in \mathcal{F} : F \cap \cup_i A_i \in \mathcal{H}_\beta\},$$

we have (recalling that $\mathcal{F}_\beta(M)$ denotes the link of $\mathcal{F}_\beta$ from M)

$$\begin{aligned} |\mathcal{F}_\beta| = \sum_{M \in \mathcal{H}_\beta} |\mathcal{F}_\beta(M)| &= \sum_{M \in \mathcal{H}_\beta} \frac{1}{m'} \sum_{\mathcal{M} \ni M} |\mathcal{F}_\beta(M)| \\ = \frac{1}{m'} \sum_{\mathcal{M}} \sum_{M \in \mathcal{M}} |\mathcal{F}_\beta(M)| &\le \frac{m}{m'} \left(t\binom{n-kt}{k-\beta^*-2} + \binom{n-kt}{k-\beta^*} \right) \\ = |\mathcal{H}_\beta| \left(\binom{n-kt}{k-\beta^*-2} + t^{-1}\binom{n-kt}{k-\beta^*} \right). \end{aligned}$$

Since $\beta^* = \sum_j \beta_j$ satisfies $2\alpha \le \beta^* \le k$ and $\alpha \ge 2$ we have

$$\begin{aligned} |\mathcal{F} \backslash \mathcal{F}'| \le \sum_{\alpha,\beta} |\mathcal{F}_\beta| &\le \sum_{\alpha,\beta} 2^{k^2} t^\alpha \left(\binom{n-kt}{k-\beta^*-2} + t^{-1}\binom{n-kt}{k-\beta^*} \right) \\ &\le 2^{k^2} \max_\alpha t^\alpha (n^{k-2\alpha-2} + t^{-1} n^{k-2\alpha}) \cdot \sum_{\alpha,\beta} 1 \\ &< 2^{3k^2} n^{k-3}, \end{aligned}$$

where we crudely estimate that there are at most $k^{k+1} < 2^{2k^2}$ ways to choose the numbers $\alpha, \beta_1, \dots, \beta_\alpha$.

Adding the estimates for $|\mathcal{F}'|$ and $|\mathcal{F} \backslash \mathcal{F}'|$, we obtain

$$|\mathcal{F}| \le \sum_{i=1}^{t} \binom{x_i + k}{k-2} + \delta\binom{n}{k-2},$$

for n sufficiently large. By the hypothesis of the theorem, this gives $\sum_{i=1}^{t} \binom{x_i+k}{k-2} \ge (1-2\delta)\binom{n-2}{k-2}$.

Suppose, without loss of generality, that $x_1 \geq x_i$ for all i. Now some routine calculations imply that $x_1 > (1-8\delta)n$. For the convenience of the reader we will give the details here, but the casual reader may skip to the last paragraph of the proof. Write $1/(r+1) < x_1/n \leq 1/r$ for some natural number r. It follows easily from Lemma 4.3 and induction that $\sum_{i=1}^{t} \binom{x_i+k}{k-2} \leq r\binom{n/r+k}{k-2} + 1_{t>r}(t-r)\binom{k}{k-2}$. This is less than $(1-2\delta)\binom{n-2}{k-2}$ if $r \geq 2$ (since $k \geq 4$), and so we have $r = 1$. Now Lemma 4.3 gives

$$\sum_{i=1}^{t} \binom{x_i+k}{k-2} \leq \binom{x_1+k}{k-2} + \binom{x-x_1+k}{k-2} + (t-2)\binom{k}{k-2}.$$

From the identity $\binom{a+b}{c} = \sum_i \binom{a}{i}\binom{b}{c-i}$ we have

$$\binom{x_1+k}{k-2} + \binom{x-x_1+k}{k-2} \leq \binom{x+2k}{k-2} - (x-x_1+k)\binom{x_1+k}{k-3},$$

and so

$$(1-3\delta)\binom{n-2}{k-2} \leq \binom{x+2k}{k-2} - (x-x_1+k)\binom{x_1+k}{k-3}.$$

In particular, $(1-3\delta)\binom{n-2}{k-2} \leq \binom{x+2k}{k-2}$, so $x > (1-4\delta)n$. Also, since $\binom{x+2k}{k-2} < (1+\delta)\binom{n-2}{k-2}$, we must have $(x-x_1+k)\binom{x_1+k}{k-3} < 4\delta\binom{n-2}{k-2}$. Now $f(q) = (x-q+k)\binom{q+k}{k-3}$ is a concave function of q; to see this, note that

$$\frac{f(q)^2}{f(q-1)f(q+1)} = \frac{(q+4)(q+k)(x-q+k)^2}{(q+3)(q+k+1)((x-q+k)^2-1)} > \frac{(q+4)(q+k)}{(q+3)(q+k+1)} > 1.$$

If $x_1 \leq (1-8\delta)n$, since $x_1 \geq n/2$, it follows that $f(x_1) \geq \min\{f(n/2), f((1-8\delta)n)\} > 4\delta\binom{n-2}{k-2}$, contradiction. Therefore $x_1 > (1-8\delta)n$, as claimed.

The number of sets of $\mathcal{F}$ not contained in $A_1 \cup X_1$ is at most $\sum_{i=2}^{t}\binom{x_i+k}{k-2} + \delta\binom{n}{k-2} < \binom{8\delta n+k}{k-2} + (t-2)\binom{k}{k-2} + \delta\binom{n}{k-2} < \epsilon n^{k-2}$ for small δ. Also, the sets of $\mathcal{F}$ contained in $A_1 \cup X_1$ form a 2-intersecting family, and as in the first paragraph of the proof, it follows that there are two points x, y such that every set in $A_1 \cup X_1$ contains both x and y. This completes the proof. □

Remarks.

1. The proof shows not only that there are at most ϵn^{k-2} sets that do not contain both x and y, but also that all such sets intersect a set $\overline{A_1 \cup X_1}$ of size at most say ϵn.

2. A more careful analysis of the argument gives a new proof of Frankl's result, and some numerical experiments indicate that the smallest n for which the proof works is considerably smaller than his value; perhaps $n = k^5$ will do, compared with $k^{\Theta(k)}$. We will not attempt to present these calculations here, as the main goal should be to prove the result for all n.

REFERENCES

[1] R. Ahlswede and L. H. Khachatrian, *The complete intersection theorem for systems of finite sets*, European J. Combin., 18 (1997), pp. 125–136.

[2] R. P. Anstee and P. Keevash, *Pairwise intersections and forbidden configurations*, European J. Combin., 27 (2006), pp. 1225–1362.

[3] J. Balogh, B. Bollobás, and M. Simonovits, *The number of graphs without forbidden subgraphs*, J. Combin. Theory Ser. B, 91 (2004), pp. 1–24.
[4] P. Erdős, *Problems and results in graph theory and combinatorial analysis*, in Proceedings of the Fifth British Combinatorial Conference (University Aberdeen, 1975), Congressus Numerantium 15, Utilitas Mathematics, Winnipeg, MB, 1976, pp. 169–192.
[5] P. Erdős, C. Ko, and R. Rado, *Intersection theorems for systems of finite sets*, Quart. J. Math. Oxford Ser., 12 (1961), pp. 313–320.
[6] P. Frankl, *On families of finite sets no two of which intersect in a singleton*, Bull. Austral. Math. Soc., 17 (1977), pp. 125–134.
[7] P. Frankl, *On intersecting families of finite sets*, J. Combin. Theory Ser. A, 24 (1978), pp. 146–161.
[8] P. Frankl, *Extremal set systems*, in Handbook of Combinatorics, Elsevier, Amsterdam, 1995, pp. 1293–1329.
[9] P. Frankl, K. Ota, and N. Tokushige, *Exponents of uniform L-systems*, J. Combin. Theory Ser. A, 75 (1996), pp. 23–43.
[10] P. Frankl and R. M. Wilson, *Intersection theorems with geometric consequences*, Combinatorica, 1 (1981), pp. 357–368.
[11] A.J.W. Hilton and E. C. Milner, *Some intersection theorems for systems of finite sets*, Quart. J. Math. Oxford Ser., 18 (1967), pp. 369–384.
[12] P. Keevash, M. Saks, B. Sudakov, and J. Verstraete, *Multicolour Turan problems*, Adv. Appl. Math., 33 (2004), pp. 238–262.
[13] P. Keevash and B. Sudakov, *Local density in graphs with forbidden subgraphs*, Combin. Probab. Comput., 12 (2003), pp. 139–153.
[14] P. Keevash and B. Sudakov, *Set systems with restricted cross-intersections and the minimum rank of inclusion matrices*, SIAM J. Discrete Math., 18 (2005), pp. 713–727.
[15] P. Keevash and B. Sudakov, *The Turán number of the Fano plane*, Combinatorica, 25 (2005), pp. 561–574.
[16] D. Mubayi, *Erdős-Ko-Rado for three sets*, J. Combin. Theory Ser. A, 113 (2006), pp. 547–550.
[17] D. Mubayi, *Structure and stability of triangle-free set systems*, Trans. Amer. Math. Soc., 359 (2007), pp. 275–291.
[18] D. K. Ray-Chaudhuri and R. M. Wilson, *On t-designs*, Osaka J. Math., 12 (1975), pp. 735–744.
[19] R. M. Wilson, *The exact bound in the Erdős-Ko-Rado theorem*, Combinatorica, 4 (1984), pp. 247–257.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 1042–1045

ROTA'S BASIS CONJECTURE FOR PAVING MATROIDS*

JIM GEELEN† AND PETER J. HUMPHRIES‡

Abstract. Rota conjectured that, given n disjoint bases of a rank-n matroid M, there are n disjoint transversals of these bases that are all bases of M. We prove a stronger statement for the class of paving matroids.

Key words. Rota's basis conjecture, paving matroids

AMS subject classification. 05B35

DOI. 10.1137/060655596

1. Introduction.

We prove the following theorem.

THEOREM 1.1. *Let $B_1, \dots, B_n$ be disjoint sets of size $n \geq 3$, and let $M_1, \dots, M_n$ be rank-n paving matroids on $\bigcup_i B_i$ such that B_i is a basis of M_i for each $i \in \{1, \dots, n\}$. Then there exist n disjoint transversals $A_1, \dots, A_n$ of $(B_1, \dots, B_n)$ such that A_i is a basis of M_i for each $i \in \{1, \dots, n\}$.*

A *paving matroid* M is a matroid in which each circuit has size $r(M)$ or $r(M)+1$, where $r(M)$ is the rank of M. Theorem 1.1 implies Rota's basis conjecture for paving matroids.

CONJECTURE 1.2 (Rota (see [6])). *Given n disjoint bases $B_1, \dots, B_n$ in a rank-n matroid M, there exist n disjoint transversals $A_1, \dots, A_n$ of $(B_1, \dots, B_n)$ that are all bases of M.*

For $n = 2$, Conjecture 1.2 follows immediately from basis exchange in matroids. Chan [2] proved the conjecture for $n = 3$. Wild [9] proved a stronger conjecture for the class of strongly base-orderable matroids, while more recently a slightly weaker result was proved for a general matroid (Ponomarenko [8]). Further partial results may be found in [1], [3], [4], [5], and [9].

Theorem 1.1 fails for both $n = 2$ and matroids in general. When $n = 2$, if we take $\mathcal{B}(M_1) = \{\{e, f\}, \{e, g\}, \{f, h\}, \{g, h\}\}$ and $\mathcal{B}(M_2) = \{\{e, f\}, \{e, h\}, \{f, g\}, \{g, h\}\}$, then $\{e, f\}, \{g, h\}$ is the only pair of disjoint bases. In the second instance, if $r_{M_1}(E - B_1) = 0$, then there are no M_1-independent transversals of $(B_1, \dots, B_n)$.

The remainder of this paper is taken up with the proof of the theorem. In section 2, we prove that Theorem 1.1 holds when $n = 3$. This result is used, in section 3, as the base case of an inductive proof of Theorem 1.1. The induction argument is surprisingly straightforward and can be read independently of section 2.

2. The case $n = 3$.

For basic concepts in matroid theory, the reader is referred to Oxley [7]. We follow the same notation as Oxley throughout this paper.

A closed set in a matroid is commonly known as a flat. We will primarily be interested in rank-2 flats, or *lines*. In the proof of Theorem 2.1, we make frequent use

*Received by the editors March 29, 2006; accepted for publication (in revised form) June 19, 2006; published electronically December 15, 2006. This research was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada and the New Zealand Marsden Fund.

http://www.siam.org/journals/sidma/20-4/65559.html

†Department of Combinatorics and Optimization, University of Waterloo, Waterloo, ON, Canada N2L 3G1.

‡Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand (pjh96@student.canterbury.ac.nz).

of the fact that if $r_M(X) = r_M(Y) = 2$ and $|X \cap Y| \geq 2$, then X and Y are contained in the same line in M.

THEOREM 2.1. *Theorem* 1.1 *holds for* $n = 3$.

Proof. Assume that the theorem is false. Then there exist bases $B_1 = \{a_1, a_2, a_3\}, B_2 = \{b_1, b_2, b_3\}, B_3 = \{c_1, c_2, c_3\}$ of rank-3 paving matroids M_1, M_2, M_3, respectively, with common ground set $E = B_1 \cup B_2 \cup B_3$, that provide a counterexample. The rank of a set X in M_i will be denoted by $r_i(X)$ and the closure by $\mathrm{cl}_i(X)$. A three-element subset of E will be called a *transversal* if it meets each of B_1, B_2, and B_3. Note that we may assume that every nontrivial line in each matroid contains a transversal, since all nontrivial lines not containing a transversal may be relaxed to provide an alternative counterexample (see [7, section 1.5, Exercise 3]).

2.1.1. *Let* $X \subseteq E$ *be a set that meets each of* B_1, B_2, B_3. *If* $r_i(X) = 3$, *then* X *contains an* M_i*-independent transversal.*

Subproof. Let $T \subseteq X$ be a transversal, and suppose that T is M_i-dependent. Then since $r_i(X) = 3$, there is some $e \in X$ such that $e \notin \mathrm{cl}_i(T)$. Without loss of generality, $e \in B_1$, so let f be the unique element in $T \cap B_1$. Then $r_i((T - f) \cup e) = 3$, and we are done. ∎

2.1.2. *If no* M_1*-dependent transversal contains both* a_1 *and* b_1, *then there exists* $e \in B_3$ *such that* $r_2(E - \{a_1, b_1, e\}) = 2$.

Subproof. For each $a \in B_1$ and $b \in B_2$, there exists $c \in B_3$ such that $\{a, b, c\}$ is M_3-independent (since $r_3(B_3) = 3$). In particular, there exist $e, f, g \in B_3$ such that $\{a_2, b_3, e\}$, $\{a_3, b_3, f\}$, and $\{a_2, b_2, g\}$ are M_3-independent. Then, by 2.1.1, $\{a_3, b_2\} \cup (B_3 - \{e\})$, $\{a_2, b_2\} \cup (B_3 - \{f\})$, and $\{a_3, b_3\} \cup (B_3 - \{g\})$ all have rank 2 in M_2 (since otherwise we would find the required partition into transversals). The second and third of these sets both have two points in common with the first, and so they are all contained in a common line in M_2. ∎

Suppose that M_1 has a line L containing at least seven elements. Since $r_1(B_1) = 3$, $|L - B_1| \geq 5$. Up to symmetry, we may assume that $b_1, b_2, c_1, c_2, c_3 \in L$ and that $a_1 \notin \mathrm{cl}_1(L)$. Now neither $\{a_1, b_1\}$ nor $\{a_1, b_2\}$ is in an M_1-dependent transversal. So by 2.1.2, $r_2(\{a_2, a_3, b_2, b_3\}) = r_2(\{a_2, a_3, b_1, b_3\}) = 2$, contradicting the fact that $r_2(B_2) = 3$. Thus none of M_1, M_2, and M_3 contains a line on seven or more elements.

2.1.3. *Every pair* $e \in B_i, f \notin B_i$ *is contained in some* M_i*-dependent transversal.*

Subproof. Suppose that no M_1-dependent transversal contains both a_1 and b_1. Then, by 2.1.2 and symmetry, we may assume that $r_2(E - \{a_1, b_1, c_1\}) = 2$. Let $X = E - \{a_1, b_1, c_1\}$ and $Y = X - B_1$. Each transversal in $\{a_2, a_3, b_2, b_3, c_1\}$ is M_2-independent, for otherwise $E - \{a_1, b_1\}$ is a seven-point line in M_2. Since each transversal in $\{a_1, b_1, c_2, c_3\}$ is M_1-independent, there is no M_3-independent transversal in X; thus $r_3(X) = 2$. Similarly, since each transversal in $\{a_2, a_3, b_1, c_2, c_3\}$ is M_2-independent and each transversal in $\{a_2, a_3, b_2, b_3, c_1\}$ is M_3-independent, we conclude that $r_1(Y \cup \{a_1\}) = 2$. Without loss of generality, $a_2 \notin \mathrm{cl}_1(Y)$, and so both $\{a_2, b_2, c_2\}$ and $\{a_2, b_3, c_3\}$ are M_1-independent. This means that $\{a_1, b_1, c_2\}$ and $\{a_1, b_1, c_3\}$ are M_2-dependent, for otherwise we again have three disjoint transversals that are independent in their respective matroids. Thus $r_2(\{a_1, b_1, c_2, c_3\}) = 2$, and $E - \{c_1\}$ is an eight-point line in M_2, which is a contradiction. ∎

Assume that B_2 is dependent in M_1. Thus, some line L in M_1 contains B_2; we may assume that L also contains a_1 and c_1, since any nontrivial line contains a transversal. There must be some element a_3, say, of B_1 that is not in $\mathrm{cl}_1(L)$, but then no transversal containing both a_3 and c_1 is dependent in M_1, leading to a contradiction by 2.1.3. Thus each of B_1, B_2, and B_3 is independent in all three matroids. This provides

additional symmetry, since we may now permute (B_1, B_2, B_3).

Suppose next that M_1 contains a five- (or six-) point line L. By the conclusion of the last paragraph, we may assume that $a_1, b_1, b_2, c_1, c_2 \in L$ and that $a_3 \notin \mathrm{cl}_1(L)$. Now, since there is an M_1-dependent transversal containing a_3, b_1, we have that $\{a_3, b_1, c_3\}$ must be M_1-dependent. Likewise $\{a_3, b_2, c_3\}$ is M_1-dependent, and thus $r_1(\{a_3, b_1, b_2, c_3\}) = 2$, contradicting the fact that $a_3 \notin \mathrm{cl}_1(L)$. Hence, none of M_1, M_2, and M_3 have lines containing more than four points.

We suppose now that the transversal $\{a_3, b_3, c_3\}$ is M_2-independent and M_3-dependent. Since $r_1(E - \{a_3, b_3, c_3\}) = 3$, we may assume that $\{a_1, b_1, c_1\}$ is M_1-independent, and also that $r_3(\{a_2, b_2, c_2\}) = 2$, for otherwise we have the required disjoint bases. Now, at most one of a_3, b_3, and c_3 may be contained in $\mathrm{cl}_3(\{a_2, b_2, c_2\})$, so without loss of generality both $\{a_2, b_3, c_2\}$ and $\{a_3, b_2, c_2\}$ are M_3-independent. Then $\{a_3, b_2, c_3\}$ and $\{a_2, b_3, c_3\}$ are both M_2-dependent. The transversal $\{a_2, b_2, c_3\}$ must now be M_2-independent, for otherwise we get a line in M_2 containing $\{a_3, b_3, c_3\}$. Thus $r_3(\{a_3, b_3, c_2\}) = 2$, and further $r_3(\{a_3, b_3, c_2, c_3\}) = 2$. Then both of $\{a_2, b_2, c_3\}$ and $\{a_3, b_2, c_3\}$ are M_3-independent, for otherwise there is a line in M_3 that contains $E - \{a_1, b_1, c_1\}$. So we have $r_2(\{a_3, b_3, c_2\}) = r_2(\{a_2, b_3, c_2\}) = 2$. This, together with the dependence of $\{a_3, b_2, c_3\}$ and $\{a_2, b_3, c_3\}$ in M_2, further implies that $\{a_3, b_3, c_3\}$ is M_2-dependent, which is a contradiction.

From now on, we may assume that M_1, M_2, and M_3 are the same matroid M, since they share the same set of independent transverals. Suppose that M contains the four-point line $\{a_3, b_3, c_2, c_3\}$. Without loss of generality, we may assume that $\{a_1, b_1, c_1\}$ is independent in M, but then both $\{a_2, b_3, c_3\}$ and $\{a_3, b_2, c_2\}$ are also independent in M, so we are done.

Thus, the rank-2 flats in M each contain at most three points. Let $\{a_3, b_3, c_3\}$ be a dependent transversal of M. By 2.1.1, the set $\{a_3, b_2, c_1, c_2\}$ contains a transversal that is independent in M. Suppose without loss of generality that $\{a_3, b_2, c_2\}$ is such a transversal. Then, again by 2.1.1, the set $\{a_1, a_2, b_1, c_1\}$ contains an M-independent transversal, $\{a_1, b_1, c_1\}$ say. Finally, $\{a_2, b_3, c_3\}$ is also independent, for otherwise we get a four-point line, and we have the three required transversals. □

3. Proof of Theorem 1.1. Before proving Theorem 1.1, we require two further lemmas. These allow us to apply induction with Theorem 2.1 as the base case. Let $\mathcal{B}(M)$ denote the set of bases of a matroid M.

LEMMA 3.1. *Let $B_1 \in \mathcal{B}(M_1), B_2 \in \mathcal{B}(M_2)$ be disjoint bases of rank-n paving matroids on the same ground set, where $n \geq 3$. Let X be a two-element subset of B_1. Then there is some $x \in X$, $y \in B_2$ such that $(B_1 - x) \cup y \in \mathcal{B}(M_1)$ and $(B_2 - y) \cup x \in \mathcal{B}(M_2)$.*

Proof. Since M_1, M_2 are paving matroids, $(B_1 - X) \cup y$ is M_1-independent for all $y \in B_2$. Suppose that both $(B_1 - x) \cup y$ and $(B_1 - x') \cup y$ are circuits in M_1, where x, x' are distinct elements of X. Then by circuit elimination, B_1 is also a circuit of M_1. Hence for each $y \in B_2$, at least one of $(B_1 - x) \cup y$ and $(B_1 - x') \cup y$ must be a basis of M_1.

Let y_1, y_2, y_3 be distinct elements of B_2. Then without loss of generality $(B_1 - x) \cup y_1, (B_1 - x) \cup y_2 \in \mathcal{B}(M_1)$. Also, one of $(B_2 - y_1) \cup x$ and $(B_2 - y_2) \cup x$ is a basis of M_2, so we are done. □

LEMMA 3.2. *Let $B_1, \ldots, B_n$ be disjoint sets of size $n \geq 3$, and let $M_1, \ldots, M_n$ be rank-n paving matroids on $\bigcup_i B_i$ such that B_i is a basis of M_i for each $i \in \{1, \ldots, n\}$. Then there is an ordering of the elements of B_1 as $a_1, \ldots, a_n$ and a transversal $\{b_2, \ldots, b_n\}$ of $(B_2, \ldots, B_n)$ such that for all $j \in \{2, \ldots, n\}$ the set*

$(B_1 - \{a_2, \ldots, a_j\}) \cup \{b_2, \ldots, b_j\}$ is a basis of M_1, and $(B_j - b_j) \cup a_j$ is a basis of M_j.

Proof. For $j = 2$, the lemma follows immediately from Lemma 3.1. Suppose now that the lemma holds for some $j \in \{2, \ldots, n-1\}$, so that $B' = (B_1 - \{a_2, \ldots, a_j\}) \cup \{b_2, \ldots, b_j\} \in \mathcal{B}(M_1)$. Then $|B_1 \cap B'| \geq 2$, and so by Lemma 3.1 there is some element $a_{j+1} \in B_1 \cap B'$ and some $b_{j+1} \in B_{j+1}$ such that $(B' - a_{j+1}) \cup b_{j+1} \in \mathcal{B}(M_1)$ and $(B_{j+1} - b_{j+1}) \cup a_{j+1} \in \mathcal{B}(M_{j+1})$, thus proving the lemma. □

Lemma 3.2 is stated for $j \in \{2, \ldots, n\}$ to simplify the induction process. We need the result only for $j = n$ to prove main theorem of this paper.

Proof of Theorem 1.1. Assume that the theorem is true for some $m \geq 3$, and take $n = m + 1$. Let $B_1 = \{a_1, \ldots, a_n\}$ and $b_i \in B_i$ for each $i \in \{2, \ldots, n\}$. By Lemma 3.2 we may assume that $A_1 = \{a_1, b_2, \ldots, b_n\}$ is a basis of M_1 and that $B_i' = (B_i - b_i) \cup a_i$ is a basis of M_i for each $i \in \{2, \ldots, n\}$.

Now let $X = E - (B_1 \cup A_1)$ and $M_i' = (M_i/a_i)|X$ for each $i \in \{2, \ldots, n\}$. Then each M_i' is a rank-m paving matroid having $B_i - b_i$ as a basis. By our induction hypothesis, there are disjoint transversals $A_2', \ldots, A_n'$ of these m bases such that A_i' is a basis of M_i'. Hence $A_i = A_i' \cup a_i$ is a basis of M_i for each $i \in \{2, \ldots, n\}$. Moreover, the bases $A_1, \ldots, A_n$ are disjoint transversals of $(B_1, \ldots, B_n)$, as required. □

Acknowledgment. The authors thank the anonymous referees for their helpful comments.

REFERENCES

[1] R. Aharoni and E. Berger, *The intersection of a matroid and a simplicial complex*, Trans. Amer. Math. Soc., 358 (2006), pp. 4895–4917.
[2] W. Chan, *An exchange property of matroid*, Discrete Math., 146 (1995), pp. 299–302.
[3] T. Chow, *On the Dinitz conjecture and related conjectures*, Discrete Math., 145 (1995), pp. 73–82.
[4] A. A. Drisko, *On the number of even and odd Latin squares of order $p + 1$*, Adv. Math., 128 (1997), pp. 20–35.
[5] A. A. Drisko, *Proof of the Alon-Tarsi conjecture for $n = 2^r p$*, Electron. J. Combin., 5 (1998), paper R28.
[6] 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.
[7] J. G. Oxley, *Matroid Theory*, Oxford University Press, New York, 1992.
[8] V. Ponomarenko, *Reduction of jump systems*, Houston J. Math., 30 (2004), pp. 27–33.
[9] M. Wild, *On Rota's problem about n bases in a rank n matroid*, Adv. Math., 108 (1994), pp. 336–345.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 1046–1055

REPRESENTING SMALL IDENTICALLY SELF-DUAL MATROIDS BY SELF-DUAL CODES*

CARLES PADRÓ† AND IGNACIO GRACIA†

Abstract. The matroid associated with a linear code is the representable matroid that is defined by the columns of any generator matrix. The matroid associated with a self-dual code is identically self-dual, but it is not known whether every identically self-dual representable matroid can be represented by a self-dual code. This open problem was proposed in [R. Cramer et al., *Advances in Cryptology*, Lecture Notes in Comput. Sci. 3621, Springer, New York, 2005, pp. 327–343], where it was proved to be equivalent to an open problem on the complexity of multiplicative linear secret sharing schemes. Some contributions to its solution are given in this paper. A new family of identically self-dual matroids that can be represented by self-dual codes is presented. Additionally, we prove that every identically self-dual matroid on at most eight points is representable by a self-dual code.

Key words. identically self-dual matroids, self-dual codes, multiparty computation, multiplicative linear secret sharing schemes

AMS subject classifications. 05B35, 94A62, 94B05

DOI. 10.1137/05064309X

1. Introduction.

1.1. Self-dual codes and identically self-dual matroids. Let $\mathcal{C}$ be an $[n,k]$ linear code over a finite field $\mathbb{K}$, where n and k are, respectively, the *length* and the *dimension* of $\mathcal{C}$. A *generator matrix* of $\mathcal{C}$ is any $k \times n$ matrix M with entries in $\mathbb{K}$ whose rows span the codewords in $\mathcal{C}$. That is, the vectors of the form $\mathbf{x} = \mathbf{u}M \in \mathbb{K}^n$, where $\mathbf{u} \in \mathbb{K}^k$, are precisely the codewords in $\mathcal{C}$. The columns of the matrix M define a $\mathbb{K}$-representable matroid $\mathcal{M}(M)$ on the set of points $Q = \{1, \ldots, n\}$. (Some basic definitions and results on matroid theory are given in section 2.) All generator matrices of the code $\mathcal{C}$ define the same matroid, and hence $\mathcal{M}(M)$ is said to be the *matroid associated with the code* $\mathcal{C}$ and is denoted by $\mathcal{M}(\mathcal{C})$. In addition, we say that the code $\mathcal{C}$ is a $\mathbb{K}$-representation of the matroid $\mathcal{M}$. While a unique matroid is associated with a linear code $\mathcal{C}$, different codes can represent the same matroid.

Greene's theorem [6], which relates the weight enumerator of a code to the Tutte polynomial of its associated matroid, is the most well-known result about that connection between codes and matroids. Several works have appeared more recently on that subject (see, for instance, the list of references in [4]). Among them, the work by Duursma [5] contains ideas that can be useful for finding new results on the open problem studied here.

Let N be a *parity-check matrix* of the code $\mathcal{C}$, that is, any $(n-k) \times n$ matrix N with maximum rank such that $MN^\top = 0$, where $N^\top$ denotes the transpose of N. Then N is the generator matrix of an $[n, n-k]$ linear code that is called the *dual code*

*Received by the editors October 19, 2005; accepted for publication (in revised form) June 19, 2006; published electronically December 15, 2006. This work was partially supported by the Spanish *Ministerio de Ciencia y Tecnología* under project TIC 2003-00866.

http://www.siam.org/journals/sidma/20-4/64309.html

†Department of Applied Mathematics IV, Technical University of Catalonia, Barcelona, Spain (cpadro@ma4.upc.edu, ignacio@ma4.upc.edu). The work of the first author occurred during a sabbatical stay at CWI, Amsterdam. This stay was funded by the *Secretaría de Estado de Educación y Universidades* of the Spanish Ministry of Education.

of $\mathcal{C}$ and is denoted by $\mathcal{C}^\perp$. If $\mathcal{C}^\perp = \mathcal{C}$, we say that $\mathcal{C}$ is a *self-dual code*. Of course, $n = 2k$ in every self-dual code.

It is well known that the matroid associated with the dual code $\mathcal{C}^\perp$ is the dual matroid of the matroid associated with $\mathcal{C}$. Then the matroid associated with a self-dual code is identically self-dual. Nevertheless, it is not known whether every identically self-dual representable matroid can be represented by a self-dual code. Specifically, the following open problem was stated in [4].

OPEN PROBLEM 1.1. *Determine whether every identically self-dual* $\mathbb{K}$*-representable matroid can be represented by a self-dual linear code over some finite extension of* $\mathbb{K}$.

Matroids that are represented by a self-dual code over the field $\mathbb{K}$ will be said to be *self-dually* $\mathbb{K}$*-representable*. Since every $\mathbb{Z}_2$-representable matroid admits a unique code representing it over $\mathbb{Z}_2$, all identically self-dual $\mathbb{Z}_2$-representable matroids are self-dually $\mathbb{Z}_2$-representable. The uniform matroids $U_{k,2k}$ form another family of identically self-dual matroids for which the answer to Open Problem 1.1 is affirmative. Moreover, if $\mathcal{M}_1$ and $\mathcal{M}_2$ are self-dually $\mathbb{K}$-representable matroids, the 2-sum $\mathcal{M} = \mathcal{M}_1 \oplus_2 \mathcal{M}_2$ of these matroids is self-dually $\mathbb{L}$-representable, where $\mathbb{L}$ is a finite extension of $\mathbb{K}$ with $[\mathbb{L} : \mathbb{K}] \leq 2$. As a consequence of this fact and other properties of the 2-sum of matroids, solving Open Problem 1.1 can be restricted to *indecomposable* matroids, that is, those that can not be expressed as the 2-sum of two smaller matroids [4]. Finally, identically self-dual bipartite matroids were proved to be self-dually representable in [4].

1.2. Ideal multiplicative linear secret sharing schemes. The interest of Open Problem 1.1 is increased by its relation to the multiplicative property of linear secret sharing schemes. That property was introduced by Cramer, Damgård, and Maurer [3] in order to construct efficient secure multiparty computation protocols for a general (that is, not necessarily threshold-based) adversary. Readers are referred to [11, 3, 4] for more information about secret sharing, the multiplicative property, and secure multiparty computation.

A $\mathbb{K}$-linear secret sharing scheme Σ with access structure Γ on the set P of players is said to be *multiplicative* if every player $i \in P$ can compute a value c_i from its shares s_i, s'_i corresponding to two shared secret values $s, s' \in \mathbb{K}$ in such a way that the product ss' is a linear combination of the values $(c_i)_{i \in P}$. Such schemes can be constructed if and only if the set of players is not the union of two unqualified subsets [7, 3]. In this case, we say that the *access structure* of the scheme is $\mathcal{Q}_2$. One of the key results in [3] is a method for constructing, from any $\mathbb{K}$-linear secret sharing scheme with $\mathcal{Q}_2$ access structure, a multiplicative $\mathbb{K}$-linear secret sharing scheme with the same access structure and whose complexity is only twice the complexity of the original scheme. One of the main open problems about this topic is to determine for which $\mathcal{Q}_2$ access structures there are multiplicative schemes with the *same complexity* as the best linear schemes. This problem has been studied in [4] for minimally $\mathcal{Q}_2$ access structures that can be realized by an *ideal* linear secret sharing scheme, that is, a scheme in which all shares have the same length as the secret. The next open problem is proposed in that paper, where it is proved to be equivalent to Open Problem 1.1.

OPEN PROBLEM 1.2. *Determine whether there exists, for every minimally* $\mathcal{Q}_2$ *access structure* Γ *that can be realized by an ideal* $\mathbb{K}$*-linear secret sharing scheme, an ideal multiplicative linear secret sharing scheme over some finite extension of* $\mathbb{K}$.

The equivalence between these two problems is due to the close relation between ideal linear secret sharing schemes, linear codes, and matroids. Actually, an ideal linear secret sharing scheme can be identified with a linear code. The access structure

of the scheme is then determined by the matroid associated with the code. The connection between ideal secret sharing schemes and matroids, which applies to nonlinear schemes as well, was discovered by Brickell and Davenport [2] and has been studied afterwards in many other works including [11, 10, 8, 1]. It plays a key role in one of the main open problems in secret sharing: the characterization of the access structures of ideal secret sharing schemes.

In addition, the notion of *duality* that applies to codes and matroids is extended to access structures. Self-dual access structures coincide with the minimally $\mathcal{Q}_2$ ones. Moreover, every self-dual code defines an ideal multiplicative linear secret sharing scheme with self-dual access structure.

1.3. Our results. The aim of this paper is to provide new results towards the solution of Open Problem 1.1.

A new family of indecomposable self-dually representable matroids is presented in section 5. By using some of the matroids in that family and other techniques we get our main result: the answer to Open Problem 1.1 is affirmative for matroids on at most eight points.

THEOREM 1.3. *Let $\mathcal{M}$ be an identically self-dual connected matroid on at most eight points (or, equivalently, with rank at most four). Then $\mathcal{M}$ is representable. Moreover, if $\mathcal{M}$ is $\mathbb{K}$-representable, then $\mathcal{M}$ can be represented by a self-dual linear code over some finite extension of $\mathbb{K}$.*

This is proved by enumerating all nonisomorphic identically self-dual matroids with rank at most four and checking that the result holds for every one of them.

By taking into account the equivalence between Open Problems 1.1 and 1.2, the following result is a direct consequence of Theorem 1.3.

COROLLARY 1.4. *Let Γ be a self-dual access structure on a set P with at most seven players. Assume that Γ can be realized by an ideal secret sharing scheme over a finite field $\mathbb{K}$. Then for some finite extension $\mathbb{L}$ of $\mathbb{K}$ there exists an ideal multiplicative $\mathbb{L}$-linear secret sharing scheme with access structure Γ.*

1.4. Organization of the paper. Some basic definitions and facts about matroid theory are recalled in section 2. Sections 3 and 4 contain some technicalities that are needed in the proofs in the following sections. A new family of self-dually representable matroids is introduced in section 5. Finally, section 6 contains the proof of Theorem 1.3, our main result.

2. Basics on matroid theory. Matroid theory abstracts many concepts from linear algebra, including independent sets, bases, and subspaces. The reader is referred to [9] for a detailed account of this field. There exist many different equivalent definitions of a matroid. The one we present here uses the concept of *basis*.

DEFINITION 2.1. *A* matroid *$\mathcal{M}$ is a finite set Q together with a family $\mathcal{B}$ of subsets of Q such that*

1. *$\mathcal{B}$ is nonempty,*
2. *for every $B_1, B_2 \in \mathcal{B}$ and $i \in B_1 - B_2$, there exists $j \in B_2 - B_1$ such that $(B_1 - \{i\}) \cup \{j\}$ is in $\mathcal{B}$.*

The set Q is the *ground set* of the matroid $\mathcal{M}$, and the sets in $\mathcal{B}$ are called the *bases* of $\mathcal{M}$. All sets in $\mathcal{B}$ have the same number of elements, which is the *rank* of $\mathcal{M}$ and is denoted $r(\mathcal{M})$. The simplest examples of matroids are the uniform ones. The *uniform matroid* $U_{k,n}$ is the matroid on a set Q of n points whose bases are all sets with exactly k points.

A subset $X \subseteq Q$ is said to be *independent* if there exists a basis $B \in \mathcal{B}$ with $X \subseteq B$, while we say that $X \subseteq Q$ is a *spanning subset* if $B \subseteq X$ for some basis $B \in \mathcal{B}$. The *dependent* subsets are those that are not independent. A point $p \in Q$ is called a *loop* if $\{p\}$ is a dependent subset, and a *coloop* is a point $p \in Q$ such that $p \in B$ for every basis $B \in \mathcal{B}$. A *circuit* is a minimally dependent subset, and the maximally independent subsets are the bases. The *rank* of $X \subseteq Q$, which is denoted $r(X)$, is the maximum cardinality of the subsets of X that are independent. Observe that the rank of Q is the rank of the matroid $\mathcal{M}$ that was defined before. A matroid is said to be *connected* if, for every two points $i, j \in Q$, there exists a circuit C with $i, j \in C$.

We say that $X \subseteq Q$ is a *flat* if $r(X \cup \{i\}) > r(X)$ for every $i \notin X$. The flat $\mathrm{cl}(X) = \{i \in Q : r(X \cup \{i\}) = r(X)\}$ is called the *closure* of X. If X is a flat, any maximally independent subset $B \subseteq X$ is called a *basis* of the flat X.

If $\mathcal{M}$ is a matroid on the set Q, with family of bases $\mathcal{B}$, then $\mathcal{B}^* = \{Q - B : B \in \mathcal{B}\}$ is the family of bases of a matroid $\mathcal{M}^*$ on Q, which is called the *dual* of $\mathcal{M}$. A *self-dual* matroid is isomorphic to its dual, while an *identically self-dual* matroid is *equal* to its dual. Observe that $|Q| = 2r(\mathcal{M})$ if the matroid is self-dual.

Let $\mathbb{K}$ be a finite field and M be a $k \times n$ matrix over $\mathbb{K}$ with rank k. A matroid $\mathcal{M}$ on the set $Q = \{1, \ldots, n\}$ is defined from the matrix M by considering that a subset $B = \{i_1, \ldots, i_k\} \subseteq Q$ is a basis if and only if the corresponding columns of M form a basis of $\mathbb{K}^k$. In this situation, we say that the matrix M is a $\mathbb{K}$-*representation* of the matroid $\mathcal{M}$. The matroids that can be defined in this way are called *representable*. As was said before, all generator matrices of a linear code $\mathcal{C}$ define the same matroid $\mathcal{M} = \mathcal{M}(\mathcal{C})$. In this case, we say that $\mathcal{C}$ is a $\mathbb{K}$-*representation* of $\mathcal{M}$, or that $\mathcal{C}$ *represents* $\mathcal{M}$ over $\mathbb{K}$.

3. Almost self-dual codes. We say that a $[2k, k]$ linear code $\mathcal{C}$ with generator matrix M is *almost self-dual* if there exists a nonsingular diagonal matrix $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_{2k})$ such that MD is a parity check matrix. Since the matrices M and MD represent the same matroid, the matroid associated to an almost self-dual code is identically self-dual. By the next proposition, in order to prove that a matroid is self-dually representable, it is enough to prove that it can be represented by an almost self-dual code.

PROPOSITION 3.1. *If an identically self-dual matroid $\mathcal{M}$ can be represented over a finite field $\mathbb{K}$ by an almost self-dual code, then $\mathcal{M}$ can be represented by a self-dual code over some finite field that extends $\mathbb{K}$.*

Proof. Let $\mathcal{C}$ be an almost self-dual code over a finite field $\mathbb{K}$. Let M be a generator matrix and $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_{2k})$ the nonsingular diagonal matrix such that MD is a parity check matrix. Let us consider, in an extension field $\mathbb{L} \supset \mathbb{K}$, the diagonal matrix $D_1 = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_{2k}})$. Then $M_1 = MD_1$ is a generator matrix of a self-dual code $\mathcal{C}_1$. The matroids associated with $\mathcal{C}$ and to $\mathcal{C}_1$ are equal. □

Let $\mathcal{C}$ be an $[n, k]$ linear code with generator matrix M, and let us set $E = \mathbb{K}^k$. In the dual space E^*, that is, the vector space formed by all linear forms $\pi\colon E \to \mathbb{K}$, let us consider the linear forms $\pi_1, \ldots, \pi_n$ such that $\mathbf{u}M = (\pi_1(\mathbf{u}), \ldots, \pi_n(\mathbf{u}))$ for every $\mathbf{u} \in E$. Observe that these linear forms correspond to columns of M; thus, we write $M = (\pi_1, \ldots, \pi_n)$.

If $\pi \in E^*$, then $\pi \otimes \pi$ denotes the symmetric bilinear form $\pi \otimes \pi\colon E \times E \to \mathbb{K}$ defined by $(\pi \otimes \pi)(\mathbf{u}, \mathbf{v}) = \pi(\mathbf{u})\pi(\mathbf{v})$. Let $\mathcal{S}(E)$ denote the symmetric bilinear forms on E. The dimension of $\mathcal{S}(E)$ is $k(k+1)/2$, where $k = \dim E$. The following lemma is proved in [4].

LEMMA 3.2. *Let $M = (\pi_1, \ldots, \pi_{2k})$ be a generator matrix of a $[2k, k]$ linear code*

$\mathcal{C}$, and let $Q = \{1, \dots, 2k\}$. Suppose that the matroid associated with $\mathcal{C}$ is identically self-dual and connected. Then in the space $\mathcal{S}(E)$, the vectors $\{\pi_j \otimes \pi_j : j \in Q - \{i\}\}$ are linearly independent for every $i \in Q$. In addition, the code $\mathcal{C}$ is almost self-dual if and only if the vectors $\{\pi_j \otimes \pi_j : j \in Q\}$ are linearly dependent.

Let $\mathcal{C}$ be a code for which the associated matroid is identically self-dual and connected. We now present a method for proving that $\mathcal{C}$ is almost self-dual. Let $M = (\pi_1, \dots, \pi_{2k})$ be a generator matrix of $\mathcal{C}$. From Lemma 3.2, it is enough to check that the subspace $\langle \pi_1 \otimes \pi_1, \dots, \pi_{2k} \otimes \pi_{2k} \rangle \subseteq \mathcal{S}(E)$ has dimension $2k-1$. Every symmetric bilinear form $\Lambda \in \mathcal{S}(E)$ can be represented by the symmetric $k \times k$ matrix $M(\Lambda) = (\lambda_{ij})$ such that $\Lambda(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 M(\Lambda) \mathbf{x}_2^\top$ for every $\mathbf{x}_1, \mathbf{x}_2 \in E$. One can prove that $\dim\langle \pi_1 \otimes \pi_1, \dots, \pi_{2k} \otimes \pi_{2k} \rangle = 2k-1$ by exhibiting $\dim \mathcal{S}(E) - (2k-1) = (k-1)(k-2)/2$ linearly independent linear equations of the form $\sum_{1 \le i \le j \le k} c_{ij} \lambda_{ij} = 0$ that are satisfied by the coefficients $(\lambda_{ij})_{1 \le i \le j \le k}$ of each of the bilinear forms $\pi_i \otimes \pi_i$. Observe that, if $\pi = (v_1, \dots, v_k) \in E^*$, then the coefficients of the bilinear form $\pi \otimes \pi$ are $\lambda_{ij} = v_i v_j$, so the code $\mathcal{C}$ is almost self-dual if the components of each of the vectors $\pi_1, \dots, \pi_{2k}$ satisfy $(k-1)(k-2)/2$ linearly independent quadratic equations of the form $\sum_{1 \le i \le j \le k} c_{ij}\ v_i v_j = 0$.

To illustrate this method we apply it to proving the well-known result that the uniform matroid $U_{k,2k}$ can be $\mathbb{K}$-represented by an almost self-dual code for every finite field with $|\mathbb{K}| \ge 2k$. Take $2k$ different elements $x_1, \dots, x_{2k} \in \mathbb{K}$ and, for every $\ell = 1, \dots, 2k$, the linear form $\pi_\ell = (1, x_\ell, x_\ell^2, \dots, x_\ell^{k-1}) \in E^*$. From the properties of the Vandermonde matrix, it is clear that the code $\mathcal{C}$ defined by these linear forms is a $\mathbb{K}$-representation of $U_{k,2k}$. Moreover, all vectors π_ℓ satisfy the $(k-1)(k-2)/2$ linearly independent quadratic equations $v_i v_j = v_{i-1} v_{j+1}$, where $2 \le i \le j \le k-1$, and hence the code $\mathcal{C}$ is almost self-dual.

4. Sum of matroids and flat-partitions. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be matroids on the sets Q_1 and Q_2, respectively. Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be their families of bases. Suppose that $Q_1 \cap Q_2 = \{p\}$ and that p is neither a loop nor a coloop of $\mathcal{M}_i$. The 2-*sum of $\mathcal{M}_1$ and $\mathcal{M}_2$ at the point p*, which will be denoted by $\mathcal{M} = \mathcal{M}_1 \oplus_2 \mathcal{M}_2$, is the matroid with ground set $Q = (Q_1 \cup Q_2) - \{p\}$ whose family of bases is $\mathcal{B} = \mathcal{B}_1' \cup \mathcal{B}_2'$, where

- $\mathcal{B}_1' = \{B_1 \cup C_2 \subseteq Q : B_1 \in \mathcal{B}_1, C_2 \cup \{p\} \in \mathcal{B}_2\}$,
- $\mathcal{B}_2' = \{C_1 \cup B_2 \subseteq Q : C_1 \cup \{p\} \in \mathcal{B}_1, B_2 \in \mathcal{B}_2\}$.

It is not difficult to check that $\mathcal{B}$ satisfies the axioms in Definition 2.1, that $\mathcal{M}$ is connected if both $\mathcal{M}_1$ and $\mathcal{M}_2$ are connected, and that $r(\mathcal{M}) = r(\mathcal{M}_1) + r(\mathcal{M}_2) - 1$. Observe that, if $\mathcal{M}_2$ is the uniform matroid $U_{1,2}$, then $\mathcal{M}_1 \oplus_2 U_{1,2} \cong \mathcal{M}_1$. This is said to be a *trivial* 2-sum. A connected matroid is said to be *indecomposable* if it is not isomorphic to any nontrivial 2-sum of matroids. The matroid $\mathcal{M} = \mathcal{M}_1 \oplus_2 \mathcal{M}_2$ is identically self-dual if and only if both $\mathcal{M}_1$ and $\mathcal{M}_2$ are identically self-dual [4]. The next proposition was also proved in [4].

PROPOSITION 4.1. *If the matroids $\mathcal{M}_1$ and $\mathcal{M}_2$ can be represented over a field $\mathbb{K}$ by almost self-dual codes, then so can their 2-sum $\mathcal{M}$. Moreover, if $\mathcal{M}_1$ and $\mathcal{M}_2$ are self-dually $\mathbb{K}$-representable, then $\mathcal{M}$ is self-dually $\mathbb{L}$-representable for some extension $\mathbb{L}$ of $\mathbb{K}$ with $[\mathbb{L} : \mathbb{K}] \le 2$.*

Let $\mathcal{M}$ be a matroid with ground set Q, and let (X_1, X_2) be a partition of Q. We say that (X_1, X_2) is a *flat-partition* of $\mathcal{M}$ if $X_i \ne \emptyset$ and X_1 and X_2 are flats of $\mathcal{M}$. If $\mathcal{M}$ is connected and $\emptyset \ne X \subsetneq Q$, then $r(X) + r(Q - X) > r(\mathcal{M})$ [9, Proposition 4.2.1]. The following lemma is a direct consequence of this fact.

LEMMA 4.2. *Let $\mathcal{M}$ be a connected matroid, and let (X_1, X_2) be a flat-partition of $\mathcal{M}$. Then $r(X_1) + r(X_2) > r(\mathcal{M})$ and $r(X_i) > 1$ for $i = 1, 2$.*

The next proposition, which is a consequence of [9, Theorem 8.3.1], provides a characterization of indecomposable identically self-dual matroids in terms of their flat-partitions.

PROPOSITION 4.3. *Let $\mathcal{M}$ be a connected identically self-dual matroid. Then $\mathcal{M}$ is indecomposable if and only if $r(X_1) + r(X_2) > r(\mathcal{M}) + 1$ for every flat-partition (X_1, X_2) of $\mathcal{M}$. Moreover, if there exists a flat-partition of $\mathcal{M}$ with $r(X_1) + r(X_2) = r(\mathcal{M}) + 1$, then there exist two connected identically self-dual matroids $\mathcal{M}_1$, $\mathcal{M}_2$ with $r(\mathcal{M}_i) = r(X_i)$ and $\mathcal{M} = \mathcal{M}_1 \oplus_2 \mathcal{M}_2$.*

A *cyclic flat* of a matroid is a flat that is a union of circuits. It is easy to show that X is a cyclic flat of a matroid on Q if and only if $Q - X$ is a cyclic flat of the dual matroid [9, Exercise 2.1.13]; also, the closure of any circuit is a cyclic flat. Applying these ideas to identically self-dual matroids gives the following lemma.

LEMMA 4.4. *For a circuit C of an identically self-dual matroid $\mathcal{M}$ on a set Q, if $0 < r(C) < r(\mathcal{M})$, then $\big(\mathrm{cl}(C), Q - \mathrm{cl}(C)\big)$ is a flat-partition of $\mathcal{M}$.*

5. A family of self-dually representable paving matroids. The *girth* $g(\mathcal{M})$ of a matroid $\mathcal{M}$ is defined as the minimum cardinality of the circuits of $\mathcal{M}$. Observe that $g(\mathcal{M}) \leq r(\mathcal{M}) + 1$ and that the uniform matroids $U_{k,n}$ with $n > k$ are the only ones with $g(\mathcal{M}) = r(\mathcal{M})+1$. Matroids with $g(\mathcal{M}) \geq r(\mathcal{M})$ are called *paving matroids*. In this section, we study the identically self-dual matroids with $g(\mathcal{M}) = r(\mathcal{M})$.

Let $\mathcal{M}$ be an identically self-dual matroid on the set Q with $r(\mathcal{M}) = g(\mathcal{M}) = k$. Since $g(\mathcal{M}) = k$, a k-element subset of Q is either a circuit or a basis, and so $\mathcal{M}$ is determined by the set $\mathbf{C}^k$ of its k-element circuits. For $i \in Q$, let $\mathbf{C}^k(i)$ be $\{C \in \mathbf{C}^k : i \in C\}$. Since $\mathcal{M}$ is identically self-dual, $\mathbf{C}^k = \mathbf{C}^k(i) \cup \{Q - C : C \in \mathbf{C}^k(i)\}$. If $k \geq 2$ and $C \in \mathbf{C}^k$, then $\mathrm{cl}(C)$ is a cyclic flat of $\mathcal{M}$ with $\emptyset \subsetneq \mathrm{cl}(C) \subsetneq Q$, and so $Q - \mathrm{cl}(C)$ is a cyclic flat of $\mathcal{M}$ with $\emptyset \subsetneq Q - \mathrm{cl}(C) \subsetneq Q$; it follows that $\mathrm{cl}(C) = C$ since $|Q| = 2k$ and both $\mathrm{cl}(C)$ and $Q - \mathrm{cl}(C)$ contain k-element circuits. Thus, by Lemma 4.4, if $k \geq 2$ and $C \in \mathbf{C}^k$, then $(C, Q - C)$ is a flat-partition of $\mathcal{M}$.

LEMMA 5.1. *For $C_1, C_2 \in \mathbf{C}^k$, if $C_1 \notin \{C_2, Q - C_2\}$, then $2 \leq |C_1 \cap C_2| \leq k - 2$.*

Proof. If $C \in \mathbf{C}^k$ and $x \in C$, then $C = \mathrm{cl}(C - x)$. Thus, $C_1 \neq C_2$ implies $|C_1 \cap C_2| \leq k - 2$; also, $C_1 \neq Q - C_2$ implies $|C_1 \cap (Q - C_2)| \leq k - 2$; that is, $2 \leq |C_1 \cap C_2|$. □

Let $\mathbb{K}$ be a field with $|\mathbb{K}| \geq 2k$, and let $\alpha_1, \alpha_2, \ldots, \alpha_{2k} \in \mathbb{K}$ be different elements such that $\alpha_1 + \cdots + \alpha_{2k} = 0$. Let $Q = \{1, \ldots, 2k\}$. It is not difficult to check that $\mathcal{B}(\alpha_1, \alpha_2, \ldots, \alpha_{2k}) = \{\{i_1, \ldots, i_k\} \subseteq Q : \alpha_{i_1} + \cdots + \alpha_{i_k} \neq 0\}$ is the family of bases of a matroid with ground set Q, which will be denoted by $\mathcal{S}(\alpha_1, \alpha_2, \ldots, \alpha_{2k})$. All matroids of this form are identically self-dual paving matroids. Moreover, as we prove in the next proposition, they are self-dually representable. Observe that $\mathcal{S}(\alpha_1, \alpha_2, \ldots, \alpha_{2k})$ is precisely the uniform matroid $U_{k,2k}$ if $\alpha_{i_1} + \cdots + \alpha_{i_k} \neq 0$ for all distinct indices $i_1, \ldots, i_k$.

PROPOSITION 5.2. *Let $\mathbb{K}$ be a field with $|\mathbb{K}| \geq 2k$, and let $\alpha_1, \alpha_2, \ldots, \alpha_{2k} \in \mathbb{K}$ be different elements such that $\alpha_1 + \cdots + \alpha_{2k} = 0$. The matroid $\mathcal{M} = \mathcal{S}(\alpha_1, \alpha_2, \ldots, \alpha_{2k})$ can be represented over $\mathbb{K}$ by an almost self-dual code, and hence it is self-dually representable over some finite extension of $\mathbb{K}$.*

Proof. If $g(\mathcal{M}) = k + 1$, then $\mathcal{M}$ is the uniform matroid $U_{k,2k}$. Since $|\mathbb{K}| \geq 2k$, there exists an almost self-dual code representing $\mathcal{M}$ over $\mathbb{K}$.

If $g(\mathcal{M}) = k$, we can suppose without loss of generality that $\alpha_1 + \cdots + \alpha_k = 0$. Consider the linear forms $\pi_i = (1, \alpha_i, \alpha_i^2, \ldots, \alpha_i^{k-2}, \alpha_i^k) \in (\mathbb{K}^k)^*$, where $i = 1, \ldots, 2k$, and the matrix $M = (\pi_1, \ldots, \pi_{2k})$. We prove that M is a $\mathbb{K}$-representation of the matroid $\mathcal{M}$ and a generator matrix of an almost self-dual code.

The first claim is proved by showing that k different vectors $\pi_{i_1}, \ldots, \pi_{i_k}$ are linearly dependent if and only if $\alpha_{i_1} + \cdots + \alpha_{i_k} = 0$. Since the linear dependency of the columns of a $k \times k$ matrix is equivalent to the linear dependency of its rows, these vectors are linearly dependent if and only if there exist values $(c_1, \ldots, c_k) \neq (0, \ldots, 0)$ such that $c_1 + c_2\alpha_{i_j} + c_3\alpha_{i_j}^2 + \cdots + c_{k-1}\alpha_{i_j}^{k-2} + c_k\alpha_{i_j}^k = 0$ for every $j = 1, \ldots, k$. This is equivalent to the polynomial $(x - \alpha_{i_1}) \cdots (x - \alpha_{i_k})$ having the form $c'_1 + c'_2x + \cdots + c'_{k-1}x^{k-2} + x^k$, which is equivalent to $\alpha_{i_1} + \cdots + \alpha_{i_k} = 0$.

To prove that the code $\mathcal{C}$ with generator matrix M is almost self-dual, we check that the vectors $\pi_1, \ldots, \pi_{2k}$ satisfy $(k-1)(k-2)/2$ linearly independent quadratic equations $\sum_{1 \leq i \leq j \leq k} c_{ij}^{\ell}\, v_iv_j = 0$, where $\ell = 1, \ldots, (k-1)(k-2)/2$.

Observe that the $(k-2)(k-3)/2$ equations $v_iv_j = v_{i-1}v_{j+1}$, where $2 \leq i \leq j \leq k-2$, are satisfied by those vectors, as are the $k-3$ equations $v_iv_k = v_{i+2}v_{k-1}$, where $1 \leq i \leq k-3$. Only one more equation is needed, which is $(a_0v_1 + \cdots + a_{k-2}v_{k-1} + v_k)(b_0v_1 + \cdots + b_{k-2}v_{k-1} + v_k) = 0$, where $(x - \alpha_1) \cdots (x - \alpha_k) = a_0 + a_1x + a_2x^2 + \cdots + a_{k-2}x^{k-2} + x^k$ and $(x - \alpha_{k+1}) \cdots (x - \alpha_{2k}) = b_0 + b_1x + b_2x^2 + \cdots + b_{k-2}x^{k-2} + x^k$.

To prove that these $(k-1)(k-2)/2$ equations are linearly independent, we check that we can reorder them in such a way that, for every $\ell = 1, \ldots, (k-1)(k-2)/2$, there exists a pair (i, j) such that $c_{ij}^{\ell} \neq 0$ and $c_{ij}^{\ell'} = 0$ for every $\ell' > \ell$. We take first the last equation, which is the only one with $c_{kk}^{\ell} \neq 0$. Afterwards, we take the equations $v_iv_k = v_{i+2}v_{k-1}$, $i = 1, \ldots, k-3$, because each of them will be the last one with $c_{ik}^{\ell} \neq 0$. Next we will find, for every $j = 2, \ldots, k-2$, only one equation with $c_{1j+1}^{\ell} \neq 0$. At this point the same applies to the coefficients c_{2j+1}^{ℓ} for $j = 3, \ldots, k-2$, and so on. □

6. Identically self-dual matroids with rank at most four. This section is devoted to proving Theorem 1.3. We determine all identically self-dual connected matroids with rank at most four, and we prove that each of them is self-dually representable.

Obviously, the uniform matroid $U_{1,2}$ is the only identically self-dual matroid with rank one. Let $\mathcal{M}$ be an identically self-dual connected matroid with $2 \leq r(\mathcal{M}) \leq 4$. By Lemmas 4.2 and 4.4, the connectedness of $\mathcal{M}$ implies that $g(\mathcal{M}) \geq 3$. It follows that $\mathcal{M} = U_{2,4}$ if $r(\mathcal{M}) = 2$. If k is 3 or 4 and $g(\mathcal{M}) = r(\mathcal{M}) + 1 = k + 1$, then $\mathcal{M} = U_{k,2k}$. If $g(\mathcal{M}) = r(\mathcal{M}) = 3$, by Lemma 4.4 there exists a flat-partition (X_1, X_2) of $\mathcal{M}$ with $r(X_1) = r(X_2) = 2$. From Proposition 4.3, $\mathcal{M} = U_{2,4} \oplus_2 U_{2,4}$. If $r(\mathcal{M}) = 4$ and $g(\mathcal{M}) = 3$, we apply again Lemmas 4.2 and 4.4 and Proposition 4.3, and we get that $\mathcal{M} = U_{2,4} \oplus_2 \mathcal{M}_1$, where $\mathcal{M}_1$ is an identically self-dual connected matroid with $r(\mathcal{M}_1) = 3$. Therefore, $\mathcal{M} = U_{2,4} \oplus_2 U_{3,6}$ or $\mathcal{M} = U_{2,4} \oplus_2 U_{2,4} \oplus_2 U_{2,4}$.

Summarizing, if $\mathcal{M}$ is an identically self-dual connected matroid with rank at most three, or if it has rank four and $g(\mathcal{M})$ is 3 or 5, then $\mathcal{M}$ is an uniform matroid or a 2-sum of uniform matroids. Therefore, for every prime p, the matroid $\mathcal{M}$ is self-dually $\mathbb{K}$-representable for some finite field $\mathbb{K}$ with characteristic p.

Let $\mathcal{M}$ be an identically self-dual connected matroid on the set $Q = \{1, 2, \ldots, 8\}$ with $r(\mathcal{M}) = g(\mathcal{M}) = 4$. Consider the set $\mathbf{C}^4$ of the circuits of $\mathcal{M}$ with exactly four points and $\mathbf{C}^4(8) = \{C \in \mathbf{C}^4 : 8 \in C\} = \{C_1, \ldots, C_m\}$, which contains half of the circuits in $\mathbf{C}^4$.

Consider $\mathbf{D} = \{D_1, \ldots, D_m\}$, where $D_i = C_i - \{8\} \subseteq \{1, \ldots, 7\}$. From Lemma 5.1, $|D_i \cap D_j| = 1$ if $i \neq j$. The matroid $\mathcal{M}$ is completely determined by $\mathbf{D}$, which is a family of 3-element subsets of $\{1, \ldots, 7\}$, each pair of which intersects in exactly one point. Moreover, there exists an identically self-dual paving matroid with rank 4 for every such family $\mathbf{D}$.

The projective plane over the finite field $\mathbb{Z}_2$, which is called the *Fano plane*, consists of 7 points and 7 lines and every line has exactly 3 points. Of course, any two lines intersect in a single point. One can check by case analysis that every family $\mathbf{D}$ with the properties described above can be completed to a family $\{R_1, \ldots, R_7\}$, the set of the lines of some Fano plane defined on the set of points $\{1, \ldots, 7\}$.

If we identify every point in $Q - \{8\} = \{1, \ldots, 7\}$ with the point in $\mathbb{Z}_2^3 - \{(0,0,0)\}$ corresponding to its binary representation, we obtain a Fano plane whose lines are $R_1 = \{2,4,6\}$, $R_2 = \{1,4,5\}$, $R_3 = \{3,4,7\}$, $R_4 = \{1,2,3\}$, $R_5 = \{2,5,7\}$, $R_6 = \{1,6,7\}$, $R_7 = \{3,5,6\}$. Therefore, up to isomorphism, the only identically self-dual matroids with both rank and girth equal to 4 are the matroids $\mathcal{M}_i$, where $i = 1, \ldots, 9$, determined by $\mathbf{D}_1 = \{R_1\}$, $\mathbf{D}_2 = \{R_1, R_2\}$, $\mathbf{D}_3 = \{R_1, R_2, R_3\}$ (three lines intersecting in one point), $\mathbf{D}_4 = \{R_1, R_2, R_4\}$ (three lines without any common point), $\mathbf{D}_5 = \{R_1, R_2, R_4, R_7\}$ (the other three lines intersect in one point), $\mathbf{D}_6 = \{R_1, R_2, R_3, R_4\}$ (the other three lines do not have any common point), $\mathbf{D}_7 = \{R_1, R_2, R_3, R_4, R_5\}$, $\mathbf{D}_8 = \{R_1, R_2, R_3, R_4, R_5, R_6\}$, $\mathbf{D}_9 = \{R_1, R_2, R_3, R_4, R_5, R_6, R_7\}$.

The proof of Theorem 1.3 is concluded by proving that, for every $i = 1, \ldots, 9$, the matroid $\mathcal{M}_i$ is representable and that, for every finite field $\mathbb{K}$ such that $\mathcal{M}_i$ is $\mathbb{K}$-representable, there exists an almost self-dual code that is an $\mathbb{L}$-representation of $\mathcal{M}_i$ for some algebraic extension $\mathbb{L}$ of $\mathbb{K}$. This is done in Propositions 6.1, 6.2, 6.3, and 6.4. For every $i = 1, \ldots, 7$ we take $C_i = R_i \cup \{8\} \subseteq Q$.

PROPOSITION 6.1. *For $i = 1$ and $i = 3$ and for every prime p, and for $i = 2$ and for every prime $p \neq 2$, there exists a finite field $\mathbb{K}$ with characteristic p and 8 elements $\alpha_1, \ldots, \alpha_8 \in \mathbb{K}$ such that $\mathcal{M}_i = \mathcal{S}(\alpha_1, \ldots, \alpha_8)$, and hence $\mathcal{M}_i$ can be represented by an almost self-dual code over the field $\mathbb{K}$.*

Proof. From the definition of $\mathcal{S}(\alpha_1, \ldots, \alpha_{2k})$, it follows that the 4-element circuits of $\mathcal{S}(\alpha_1, \ldots, \alpha_8)$ are the sets $\{i_1, i_2, i_3, i_4\}$ for which $\alpha_{i_1} + \alpha_{i_2} + \alpha_{i_3} + \alpha_{i_4} = 0$. Thus, $\mathcal{M}_1$, in which the only 4-element circuits are $\{1,3,5,7\}$ and $\{2,4,6,8\}$, is $\mathcal{S}(1 + \beta^2 + \beta^4, \beta + \beta^3 + \beta^5, -1, -\beta, -\beta^2, -\beta^3, -\beta^4, -\beta^5)$, where β is any element of degree at least 6 over the prime field $\mathbb{Z}_p$ (that is, β is not a root of any polynomial over $\mathbb{Z}_p$ with degree smaller than 6). The 4-element circuits of $\mathcal{M}_3$ are precisely all unions of pairs of sets among $\{1,5\}$, $\{2,6\}$, $\{3,7\}$, and $\{4,8\}$. Therefore, if β is any element of degree at least 4 over $\mathbb{Z}_p$, the matroid $\mathcal{M}_3$ is $\mathcal{S}(1, \beta, \beta^2, \beta^3, -1, -\beta, -\beta^2, -\beta^3)$ if $p \neq 2$ and is $\mathcal{S}(1+\beta, \beta+\beta^2, 1+\beta^2, 1, \beta^2+\beta^3, 1+\beta^3, \beta+\beta^3, \beta+\beta^2+\beta^3)$ if $p = 2$. Finally, $\mathcal{M}_2$, in which the 4-element circuits are $\{1,3,5,7\}$, $\{1,4,5,8\}$, $\{2,3,6,7\}$, and $\{2,4,6,8\}$, is $\mathcal{S}(-\beta, -\beta-\beta^2, 1+\beta+\beta^2+\beta^3, \beta+\beta^2, -\beta^2-\beta^3, -\beta^3, -1, \beta^3)$, where β is any element of degree at least 4 over $\mathbb{Z}_p$ and the prime p is not 2. □

PROPOSITION 6.2. *The matroid $\mathcal{M}_2$ can be represented by an almost self-dual code over some finite field $\mathbb{K}$ with characteristic 2.*

Proof. In the corresponding algebraic extension $\mathbb{K}$ of $\mathbb{Z}_2$, take $\omega \in \mathbb{K}$ with $\omega^{13} = 1$ and $\omega \neq 1$. Then the matrix

$$M = M(\pi_1, \ldots, \pi_8) = \begin{pmatrix} \omega & 0 & \omega^3 & 0 & \omega^{-1} & 0 & \omega^{-3} & 0 \\ 0 & \omega^2 & 1 & 0 & 0 & \omega^{-2} & 1 & 0 \\ 0 & 1 & 0 & \omega^5 & 0 & 1 & 0 & \omega^{-5} \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

is the generator matrix of an almost self-dual code $\mathcal{C}$ that represents the matroid $\mathcal{M}_2$ over $\mathbb{K}$. This can be proved by using a simple computer program to check that $\det(\pi_{i_1}, \pi_{i_2}, \pi_{i_3}, \pi_{i_4}) = 0$ if and only if $\{i_1, i_2, i_3, i_4\} = C_1, Q - C_1, C_2, Q - C_2$ and that $\dim\langle \pi_1 \otimes \pi_1, \ldots, \pi_8 \otimes \pi_8 \rangle = 7$. □

PROPOSITION 6.3. *For every finite field $\mathbb{K}$ and for every $i = 4, \dots, 9$, if a code $\mathcal{C}$ is a $\mathbb{K}$-representation of the matroid $\mathcal{M}_i$, then $\mathcal{C}$ is almost self-dual.*

Proof. Let $\mathcal{M}$ be one of the matroids $\mathcal{M}_4, \dots, \mathcal{M}_9$, and let $M = (\pi_1, \dots, \pi_8)$ be such that the code $\mathcal{C}$ with generator matrix M is a $\mathbb{K}$-representation of $\mathcal{M}$. Since $\{R_1, R_2, R_4\} \subseteq \mathbf{D}_i$ for every $i = 4, \dots, 9$, we have that $C_1 = \{2, 4, 6, 8\}$, $C_2 = \{1, 4, 5, 8\}$, and $C_4 = \{1, 2, 3, 8\}$, and their complements are circuits of $\mathcal{M}$. For every $i = 1, 2, 4$, let $a_1^i v_1 + a_2^i v_2 + a_3^i v_3 + a_4^i v_4 = 0$ and $b_1^i v_1 + b_2^i v_2 + b_3^i v_3 + b_4^i v_4 = 0$ be, respectively, the equations of the hyperplanes $V_i = \langle \pi_j : j \in C_i \rangle$ and $W_i = \langle \pi_j : j \in Q - C_i \rangle$. Therefore, the three quadratic equations

$$(a_1^i v_1 + a_2^i v_2 + a_3^i v_3 + a_4^i v_4)(b_1^i v_1 + b_2^i v_2 + b_3^i v_3 + b_4^i v_4) = 0,$$

where $i = 1, 2, 4$, are satisfied by each vector π_j. We have to prove only that these quadratic equations are linearly independent. Let $Q_1, Q_2, Q_4 \subseteq \mathbb{K}^4$ be the quadrics defined by these equations. Observe that $Q_i = V_i \cup W_i$. By symmetry, it is enough to prove that $Q_1 \cap Q_2 \not\subseteq Q_4$. This is clear by taking into account that $Q_1 \cap Q_2 = \langle \pi_4, \pi_8 \rangle \cup \langle \pi_2, \pi_6 \rangle \cup \langle \pi_1, \pi_5 \rangle \cup \langle \pi_3, \pi_7 \rangle$ and $Q_4 = \langle \pi_1, \pi_2, \pi_3, \pi_8 \rangle \cup \langle \pi_4, \pi_5, \pi_6, \pi_7 \rangle$. ∎

To conclude the proof of Theorem 1.3, it is enough to prove that the matroids $\mathcal{M}_4, \dots, \mathcal{M}_9$ are representable. This is done in the next proposition and, even though it is not necessary, we determine, for completeness, the characteristics of the fields over which these matroids admit representations.

PROPOSITION 6.4. *For every $i = 4, \dots, 7$ and for every prime p the matroid $\mathcal{M}_i$ is $\mathbb{K}$-representable for some finite field $\mathbb{K}$ with characteristic p. The matroid $\mathcal{M}_8$ is $\mathbb{K}$-representable if and only if the characteristic of $\mathbb{K}$ is not 2. Finally, the matroid $\mathcal{M}_9$ is $\mathbb{K}$-representable if and only if the characteristic of $\mathbb{K}$ is 2.*

Proof. Let p be a prime. Take a prime q with $q \geq 5$ and $q \neq p$. Let $\mathbb{K}$ be a finite field of characteristic p that contains a primitive q-root of unity $\omega \in \mathbb{K}$. Then the matrix

$$M_4 = \begin{pmatrix} \omega^3 & 0 & \omega^2 & 0 & \omega^4 & 0 & \omega & 0 \\ 0 & \omega & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & \omega^3 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

is a $\mathbb{K}$-representation of $\mathcal{M}_4$. The matrix

$$M_5 = \begin{pmatrix} ab & 0 & a & 0 & 1 & 0 & 1 & 0 \\ 0 & b & 1 & 0 & 0 & a^{-1} & 1 & 0 \\ 0 & 1 & 0 & a & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

provides a representation of the matroid $\mathcal{M}_5$ if $a, b \neq 0, 1$ and $b \neq a^{-1}$. A representation of the matroid $\mathcal{M}_6$ is given by the matrix

$$M(a, b) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & a & b & a+b-1 & 0 \end{pmatrix}$$

if $a, b \neq 0, 1$ and $a \neq b$ and $a + b \neq 1$. The code with generator matrix $M(a, 1)$ represents $\mathcal{M}_7$ if $a \neq 0, 1, -1$. Therefore, $\mathcal{M}_5$, $\mathcal{M}_6$, and $\mathcal{M}_7$ are $\mathbb{K}$-representable for

every finite field with $|\mathbb{K}| \geq 5$, and hence they can be represented over fields of every characteristic. Moreover, the matrix $M(1,1)$ is a representation of $\mathcal{M}_8$ for every finite field with characteristic different from 2, and it provides a $\mathbb{K}$-representation of the matroid $\mathcal{M}_9$ if $\mathbb{K}$ has characteristic 2. Finally, it is well known that $\mathcal{M}_8$ cannot be represented over any field with characteristic 2, while $\mathcal{M}_9$ can be represented only over fields with characteristic 2. See, for instance, the Appendix "Some interesting matroids" in [9], in which $\mathcal{M}_8$ and $\mathcal{M}_9$ appear, respectively, as R_8 and $AG(3,2)$. □

Acknowledgments. We would like to thank the anonymous referees for their many valuable comments and suggestions, which have greatly improved this paper.

REFERENCES

[1] A. Beimel, T. Tassa, and E. Weinreb, *Characterizing ideal weighted threshold secret sharing*, in Theory of Cryptography, Proceedings of the Second Theory of Cryptography Conference (TCC 2005), Lecture Notes in Comput. Sci. 3378, Springer, New York, 2005, pp. 600–619.

[2] E.F. Brickell and D.M. Davenport, *On the classification of ideal secret sharing schemes*, J. Cryptology, 4 (1991), pp. 123–134.

[3] R. Cramer, I. Damgård, and U. Maurer, *General secure multi-party computation from any linear secret-sharing scheme*, in Advances in Cryptology (Eurocrypt 2000), Lecture Notes in Comput. Sci. 1807, Springer, New York, 2000, pp. 316–334.

[4] R. Cramer, V. Daza, I. Gracia, J. Jiménez Urroz, G. Leander, J. Martí-Farré, and C. Padró, *On codes, matroids and secure multi-party computation from linear secret sharing schemes*, in Advances in Cryptology (Crypto 2005), Lecture Notes in Comput. Sci. 3621, Springer, New York, 2005, pp. 327–343.

[5] I.M. Duursma, *Combinatorics of the two-variable zeta function*, in Finite Fields and Applications, Lecture Notes in Comput. Sci. 2948, Springer, New York, 2004, pp. 109–136.

[6] C. Greene, *Weight enumeration and the geometry of linear codes*, Stud. Appl. Math. 55, Elsevier, Amsterdam, 1976, pp. 119–128.

[7] M. Hirt and U. Maurer, *Complete characterization of adversaries tolerable in secure multi-party computation*, in Proceedings of the 16th Annual ACM Symposium on Principles of Distributed Computing (PODC '97), Santa Barbara, CA, 1997, ACM, New York, 1997, pp. 25–34.

[8] F. Matúš, *Matroid representations by partitions*, Discrete Math., 203 (1999), pp. 169–194.

[9] J.G. Oxley, *Matroid Theory*, Oxford Science Publications Clarendon Press Oxford University Press, New York, 1992.

[10] J. Simonis and A. Ashikhmin, *Almost affine codes*, Des. Codes Cryptogr., 14 (1998), pp. 179–197.

[11] D.R. Stinson, *An explication of secret sharing schemes*, Des. Codes Cryptogr., 2 (1992), pp. 357–390.

SIAM J. DISCRETE MATH.
Vol. 20, No. 4, pp. 1056–1078

HIGHER-DIMENSIONAL PACKING WITH ORDER CONSTRAINTS*

SÁNDOR P. FEKETE†, EKKEHARD KÖHLER‡, AND JÜRGEN TEICH§

Abstract. We present a first exact study on higher-dimensional packing problems with order constraints. Problems of this type occur naturally in applications such as logistics or computer architecture and can be interpreted as higher-dimensional generalizations of scheduling problems. Using graph-theoretic structures to describe feasible solutions, we develop a novel exact branch-and-bound algorithm. This extends previous work by Fekete and Schepers; a key tool is a new order-theoretic characterization of feasible extensions of a partial order to a given complementarity graph that is tailor-made for use in a branch-and-bound environment. The usefulness of our approach is validated by computational results.

Key words. higher-dimensional packing, higher-dimensional scheduling, reconfigurable computing, precedence constraints, exact algorithms, modular decomposition

AMS subject classifications. 90C28, 68R99

DOI. 10.1137/060665713

1. Introduction.

Scheduling and packing problems. Scheduling is arguably one of the most important topics in combinatorial optimization. Typically, we are dealing with a one-dimensional set of objects ("jobs") that need to be assigned to a finite set of containers ("machines"). Problems of this type can also be interpreted as (one-dimensional) packing problems, and they are NP-hard in the strong sense, as problems like 3-PARTITION are special cases.

Starting from this basic scenario, there are different generalizations that have been studied. Many scheduling problems have *precedence constraints* on the sequence of jobs. On the other hand, a great deal of practical packing problems consider *higher-dimensional* instances, where objects are axis-aligned boxes instead of intervals. Higher-dimensional packing problems arise in many industries, where steel, glass, wood, or textile materials are cut. The three-dimensional problem is important for practical applications such as container loading.

In this paper, we give the first study of problems that comprise both generalizations: these are higher-dimensional packing problems with order constraints—or, from a slightly different point of view, higher-dimensional scheduling problems. In higher-dimensional packing, these problems arise when dealing with precedence constraints that are present in many container-loading problems. Another practical motivation

*Received by the editors July 28, 2005; accepted for publication (in revised form) July 21, 2006; published electronically December 26, 2006. A preliminary extended abstract version reporting on parts of this paper appeared in [4, 5].

http://www.siam.org/journals/sidma/20-4/66571.html

†Department of Mathematical Optimization, Braunschweig University of Technology, D–38116 Braunschweig, Germany (s.fekete@tu-bs.de). This author was partially supported by Deutsche Forschungsgemeinschaft (DFG) within the special focus program "Reconfigurable Computing" (SPP 1148), project "ReCoNodes," grants Fe407/8-1 and 8-2.

‡Department of Mathematics, Technical University Berlin, D–10623 Berlin, Germany (ekoehler@math.tu-berlin.de).

§Department of Computer Science 12 (Hardware-Software-Co-Design), University of Erlangen-Nuremberg, D–91058 Erlangen, Germany (teich@informatik.uni-erlangen.de). This author was partially supported by Deutsche Forschungsgemeinschaft (DFG) within the special focus program "Reconfigurable Computing" (SPP 1148), project "ReCoNodes," grants Te 163/14-1 and 14-2.

for considering multidimensional scheduling problems arises from optimizing the reconfiguration of a particular type of computer chip called FPGAs—described below.

FPGAs and higher-dimensional scheduling. A particularly interesting class of instances of three-dimensional orthogonal packing arises from a new type of reconfigurable computer chip, called *field-programmable gate arrays* (FPGAs). An FPGA typically consists of a regular rectangular grid of equal configurable cells (logic blocks) that allow the prototyping of simple logic functions together with simple registers and with special routing resources (see Figure 1.1). These chips (see, e.g., [1, 34]) may support several independent or interdependent jobs and designs at a time, and parts of the chip can be reconfigured quickly during run-time. (For more technical details on the underlying architecture, see the previous paper [32] and the more recent abstract [6].) Thus, we are faced with a general class of problems that can be seen both as scheduling and packing problems. In this paper, we develop a set of mathematical tools to deal with these *higher-dimensional scheduling problems*, and we show that our methods are suitable for solving instances of interesting size to optimality.

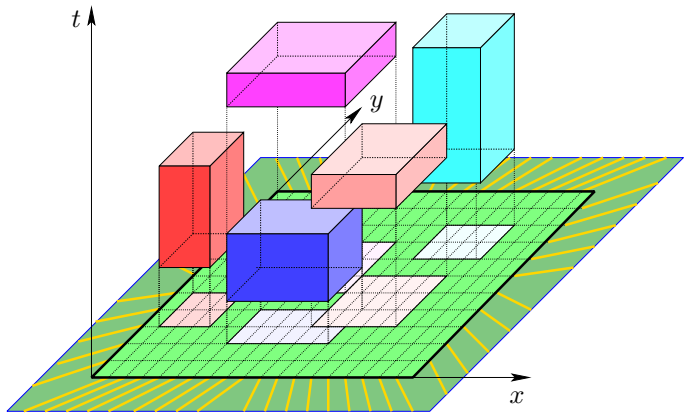


FIG. 1.1. *An FPGA and a set of five jobs, shown as projections in ordinary two-dimensional space and in three-dimensional space-time. Jobs must be placed inside the chip and must not overlap if executed simultaneously on the chip.*

Related work. We are not aware of any exact study of higher-dimensional packing or scheduling problems with order constraints. For a comprehensive survey of classical "one-dimensional" scheduling problems, the reader is referred to [24]. A related problem is dynamic storage allocation, where "processing jobs" means storing them in contiguous blocks of memory from a one-dimensional array. Considering time as the second dimension leads to a two-dimensional packing problem, possibly with order constraints. However, this problem is primarily an online problem; for example, see [25]. In an offline setting, precise starting and ending time values imply order constraints but also provide more information. (See our paper [32] for exact methods for that scenario.)

Closest to our problems is the class of *resource-constrained project scheduling problems* (RCPSP), which can be interpreted as a step towards higher-dimensional packing problems: In addition to a duration t_i and precedence constraints on the temporal order of jobs, each job i may have a number of other "sizes" $x_i^{(1)}, \ldots, x_i^{(k)}$; $x_i^{(j)}$ indicates the amount of resource j required for the processing of job i. The total amount $\sum_i x_i^{(j)}$ of each resource j is limited at any given time. See the book [33] and the

references in the article [28] for an extensive survey of recent work in this area. Even though RCPSPs can be formulated as integer problems, solving resource-constrained scheduling problems is already quite hard for instances of relatively moderate size: The standard benchmark library used in this area consists of instances with 30, 60, 90, and 120 jobs. Virtually all work deals with lower and upper bounds on these instances, and even for instances with 60 jobs, a considerable number has not yet been solved to optimality.

It is easy to see that two-dimensional packing problems (possibly with precedence constraints on the temporal order) can be relaxed to a scheduling problem with one resource constraint, by allowing a noncontiguous use of resources, i.e., the higher-dimensional analogue of preemption. However, the example in Figure 1.2 shows that the converse is not true, even for small instances of two-dimensional packing problems without any precedence constraints: An optimal solution for the corresponding resource-constrained scheduling problem may not correspond to a feasible arrangement of rectangles for the original packing problem. (We leave it to the reader to verify the latter claim.) For $d \geq 2$ the difference becomes more pronounced: The d knapsack constraints for RCSPSP require that for *all* of the d individual resources and every pair of jobs, a disjointness property must be satisfied; on the other hand, the more geometric conditions on d-dimensional packing require that any pair of boxes must be disjoint in *at least one* of their coordinate intervals. Arguably, the disjunctive constraints on $(d+1)$-dimensional packing problems are harder to model.

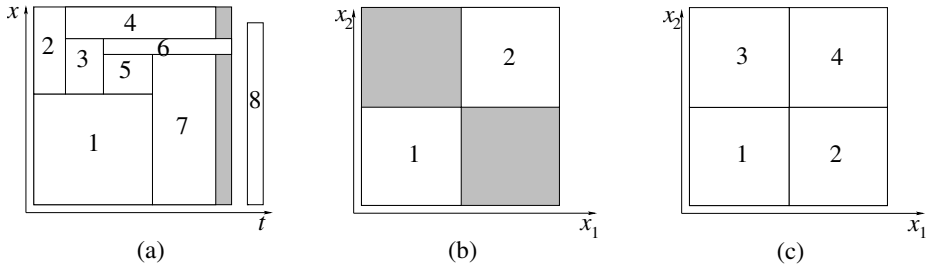


FIG. 1.2. *Differences between RCPSP and packing:* (a) *A set of jobs that is feasible for scheduling with one resource constraint but infeasible for two-dimensional packing: Job 8 does not violate a resource constraint but does not fit as a contiguous rectangle.* (b) *A set of jobs that is just feasible for RCPSP with* $d = 2$ *constraints, i.e., that does not allow any tightening of either constraint without becoming infeasible.* (c) *A set of boxes that is just feasible for packing in* $d = 2$ *dimensions.*

Higher-dimensional packing problems (without order constraints) have been considered by a great number of authors, but only few of them have dealt with the exact solution of general two-dimensional problems. See [7, 10] for an overview. It should be stressed that unlike one-dimensional packing problems, higher-dimensional packing problems allow no straightforward formulation as integer programs: After placing one box in a container, the remaining feasible space will in general not be convex. Moreover, checking whether a given set of boxes fits into a particular container (the so-called *orthogonal packing problem* (OPP)) is trivial in one-dimensional space but NP-hard in higher dimensions.

Nevertheless, attempts have been made to use standard approaches of mathematical programming. Beasley [2] and Hadjiconstantinou and Christofides [18] have used a discretization of the available positions to an underlying grid to get a 0-1 program with a pseudopolynomial number of variables and constraints. Not surprisingly, this approach becomes impractical beyond instances of rather moderate size.

More recently, Padberg [29] gave a *mixed integer programming* formulation for three-dimensional packing problems, similar to the one anticipated by Schepers [30] in his thesis. Padberg expressed the hope that using a number of techniques from branch-and-cut will be useful; however, he did not provide any practical results to support this hope.

In [12, 7, 10, 11, 32], a different approach to characterizing feasible packings and constructing optimal solutions is described. A graph-theoretic characterization of the relative position of the boxes in a feasible packing (by so-called *packing classes*) is used, representing d-dimensional packings by a d-tuple of interval graphs (called *component graphs*) that satisfy two extra conditions. This factors out a great deal of symmetries between different feasible packings, it allows one to make use of a number of elegant graph-theoretic tools, and it reduces the geometric problem to a purely combinatorial one without using brute-force methods like introducing an underlying coordinate grid. Combined with good heuristics for dismissing infeasible sets of boxes [8, 9], a tree search for constructing feasible packings was developed. This exact algorithm has been implemented; it outperforms previous methods by a clear margin.

For the benefit of the reader, a concise description of this approach is contained in section 3.

Graph theory of order constraints. In the context of scheduling with precedence constraints, a natural problem is the following, called *transitive ordering with precedence constraints* (TOP): Consider a partial order $P = (V, \prec)$ of precedence constraints and a (temporal) comparability graph $G = (V, E)$, such that all relations in P are represented by edges in G. Is there a transitive orientation $D = (V, A)$ of G, such that P is contained in D?

Korte and Möhring [21] have given a linear-time algorithm for deciding (TOP), using modified PQ-trees. However, their approach requires knowledge of the full set of edges in G. When running a branch-and-bound algorithm for solving a scheduling problem, these edges of G are known only partially during most of the tree search, but already this partial edge set may prohibit the existence of a feasible solution for a given partial order P. This makes it desirable to come up with structural characterizations that are already useful when only parts of G are known.

Such a set of precedence constraints may be described by a dependency graph; see Figure 1.3.

For a problem instance of this type, we describe a general framework for finding exact solutions to the problem of minimizing the height of a container of given base area, or minimizing the makespan of a higher-dimensional nonpreemptive scheduling problem.

Results of this paper. In this paper, we give the first exact study of higher-dimensional packing with order constraints, which can also be interpreted as *higher-dimensional nonpreemptive scheduling problems*. We develop a general framework for problems of this type by giving a pair of necessary and sufficient conditions for the existence of a solution for the problem TOP on graphs G in terms of forbidden substructures. Using the concept of packing classes, our conditions can be used quite effectively in the context of a branch-and-bound framework, because it can recognize infeasible subtrees at "high" branches of the search tree. In particular, we describe how to find an exact solution to the problem of minimizing the height of a container of given base area. If this third dimension represents time, this amounts to minimizing the makespan of a higher-dimensional scheduling problem. We validate the

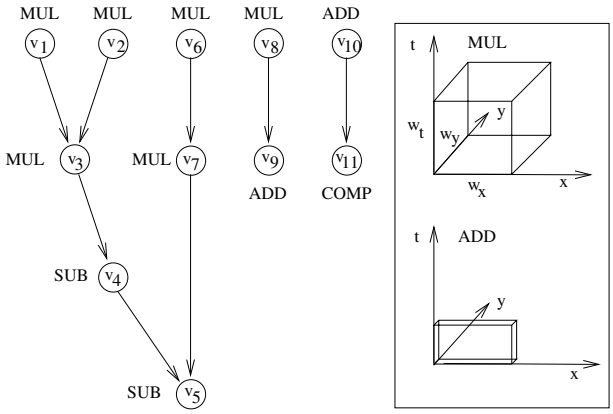


FIG. 1.3. *Dependency graph of jobs and shape of modules (three-dimensional boxes) with the spatial dimensions* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *and the temporal dimension* $\boldsymbol{t}$ *(execution time).*

usefulness of these concepts and results by providing computational results. Other problem versions (like higher-dimensional knapsack or bin packing problems with order constraints) can be treated similarly.

The rest of this paper is organized as follows. In section 2, we describe basic assumptions and some terminology. The notion of packing classes and a solution to packing problems without precedence constraints is summarized in section 3. In section 4, we introduce precedence constraints, describe the mathematical foundations for incorporating them into the search, and explain how to implement the resulting algorithms. Section 5 provides the necessary mathematical foundations for the correctness of our approach. Finally, we present computational results for a number of different benchmarks in section 6.

2. Preliminaries. An FPGA consists of a rectangular grid of identical logic cells. Each job v (or "module") requires a rectangle of size $w_x(v)$ by $w_y(v)$ with fixed axis-parallel orientation and needs to remain available for at least the time $w_t(v)$. Any logic cell that is not occupied by a module may be used by one of the rectangular jobs. As shown in Figure 1.1, we are dealing with a three-dimensional packing problem, possibly with order constraints. In the following, we describe technical as well as mathematical terminology and assumptions.

2.1. Architecture assumptions. The model of having relocatable, rectangular modules is justified by current FPGA technology [1, 34].

Intermodule communication. Intermodule communication is assumed to occur at the end of operation of the sending module (task model). The issuing module may store its result register values into an external memory connected to the FPGA interface (read-out) via a bus interface. Memory is allocated for temporary storage of intermediate results.[1] Afterwards, the receiving module will read the communicated data into its registers via the bus interface. With this communication style, it is justifiable to ignore routing overhead between modules that otherwise might introduce additional placement constraints.

I/O-overhead. The communication time needed for writing out and reading in communicated data may be accounted for by considering this as an offset and being

[1] A static memory allocation may be deduced directly from the static placement.

part of the execution time of a job.

Reconfiguration overhead. The time needed for carrying out reconfigurations may be modeled by a constant (possibly a different number for each job), depending on the target architecture. This may be considered a simplification because the reconfiguration time might depend on the result of the placement. Consider two equal modules with identical placements. A reconfiguration for the second module might not be necessary in case no third module is occupying a (sub)set of cells in the time interval between the execution of the two modules. However, there are many different models for accounting for reconfiguration times, and the particular choice should be adapted individually to the target architecture.

2.2. Mathematical terminology.

Problem instances. We assume that a problem instance is given by a *set V of jobs*. Each job has a *spatial requirement* in the x- and y-directions, denoted by $w_x(v)$ and $w_y(v)$, and a *duration*, denoted by a size $w_t(v)$ along the time axis. The available space H consists of an area of size $h_x \times h_y$. In addition, there may be an overall allowable time h_t for all jobs to be completed. A *schedule* is given by a start time $p_t(v)$ for each job. A schedule is *feasible*, if all jobs can be carried without preemption or overlap of computation jobs in time and space, such that all jobs are within spatial and temporal bounds.

Graphs. Some of our descriptions make use of a number of certain graph-theoretic concepts. An (undirected) graph $G = (V, E)$ is given by a set of vertices V and a set of edges E; each edge describes the adjacency of a pair of vertices, and we write $\{u, w\}$ for an edge between vertices u and w. We consider only graphs without multiple edges and without loops. For a graph G, we obtain the *complement graph* $\overline{G}$ by exchanging the set E of edges with the set $\overline{E}$ of nonedges. In a directed graph $D = (V, A)$, edges are oriented, and we write (u, w) to denote an edge directed from u to w. A graph $G = (V, E)$ is a *comparability graph* if there is a *transitive orientation* for it, i.e., the edges E can be oriented to a set of directed arcs A, such that we get the transitive closure of a partial order. More precisely, this means that $D = (V, A)$ is a cycle-free digraph for which the existence of edges $(u, v) \in A$ and $(v, w) \in A$ for any $u, v, w \in V$ implies the existence of $(u, w) \in A$. Comparability graphs have a variety of nice properties. For our purpose we will make use of the algorithmic result that computing maximum weighted cliques on comparability graphs can be done efficiently (see [17]). A closely related family of graphs, the *interval graphs*, is defined as follows. Given a set of intervals on the real line, every vertex of the graph corresponds to an interval of the set; two vertices are joined by an edge if the corresponding intervals have a nonempty intersection. Interval graphs have been studied intensively in graph theory (see [17, 26]), and, similar to comparability graphs, they have a number of very useful algorithmic properties.

Precedence constraints. Mathematically, a set of precedence constraints is given by a partial order $P = (V, \prec)$ on V. The relations in $\prec$ can be interpreted as a directed acyclic graph $D_P = (V, A_P)$, where A_P is a set of directed arcs corresponding to the relations in $\prec$. In the presence of such a partial order, a feasible schedule is required to satisfy the capacity constraints of the container, as well as these additional constraints.

Packing problems. In the following, we treat jobs as axis-aligned d-dimensional boxes with given orientation, and feasible schedules as arrangements of boxes that satisfy all side constraints. This is implied by the term of a *feasible packing*. There may be different types of objective functions, corresponding to different types of packing problems. The OPP is to decide whether a given set of boxes can be placed within

a given "container" of size $h_x \times h_y \times h_t$. For the *constrained* OPP (COPP), we also have to satisfy a partial order $P = (V, \prec)$ of precedence constraints in the t-dimension. To emphasize the motivation of temporal precedence constraints, we write t to suggest that the time coordinate is constrained, and x and y to imply that the space coordinates are unrestricted. Although our application mainly requires us to consider those temporal constraints, it should be mentioned that our approach works the same way when dealing with spatial restrictions; that is why we are using a generic index i in the mathematical discussion, while some of our benchmark examples consider a temporal dimension t.

There are various optimization problems that have OPP or COPP as their underlying decision problems. The *base minimization problem* (BMP) is to minimize the size h_x for a fixed h_t such that all boxes fit into a container $h_x \times h_x \times h_t$ with quadratic base. This corresponds to minimizing the necessary area to carry out a set of computations within a given time. Because our main motivation arises from dynamic chip reconfigurations, where we want to minimize the overall running time, we focus on the *constrained strip packing problem* (CSPP), which is to minimize the size h_t for a given base size $h_x \times h_y$, such that all boxes fit into the container $h_x \times h_y \times h_t$. Clearly, we can use a similar approach for other objective functions.

3. Solving unconstrained orthogonal packing problems.

3.1. A general framework. If we have an efficient method for solving OPPs, we can also solve BMPs and SPPs by using a binary search. However, deciding the existence of a feasible packing is a hard problem in higher dimensions, and methods proposed by other authors [2, 18] have been of limited success.

Our framework uses a combination of different approaches to overcome these problems:

1. Try to disprove the existence of a packing by classes of lower bounds on the necessary size.
2. In case of failure, try to find a feasible packing by using fast heuristics.
3. If the existence of a packing is still unsettled, start an enumeration scheme in the form of a branch-and-bound tree search.

By developing good new bounds for the first stage, we have been able to achieve a considerable reduction of the number of cases in which a tree search needs to be performed. (Mathematical details for this step are described in [8, 11].) However, it is clear that the efficiency of the third stage is crucial for the overall running time when considering difficult problems. Using a purely geometric enumeration scheme for this step by trying to build a partial arrangement of boxes is easily seen to be immensely time-consuming. In the following, we describe a purely combinatorial characterization of feasible packings that allows us to perform this step more efficiently.

3.2. Packing classes. Consider a feasible packing in d-dimensional space, and project the boxes onto the d coordinate axes. This converts the one d-dimensional arrangement into d one-dimensional ones (see Figure 3.1 for an example in $d = 2$). By disregarding the exact coordinates of the resulting intervals in direction i and considering only their intersection properties, we get the *component graph* $G_i = (V, E_i)$: Two boxes u and v are connected by an edge in G_i iff their projected intervals in direction x_i have a nonempty intersection. By definition, these graphs are *interval graphs*.

Considering sets of d component graphs G_i instead of complicated geometric arrangements has some clear advantages (algorithmic implications for our specific pur-

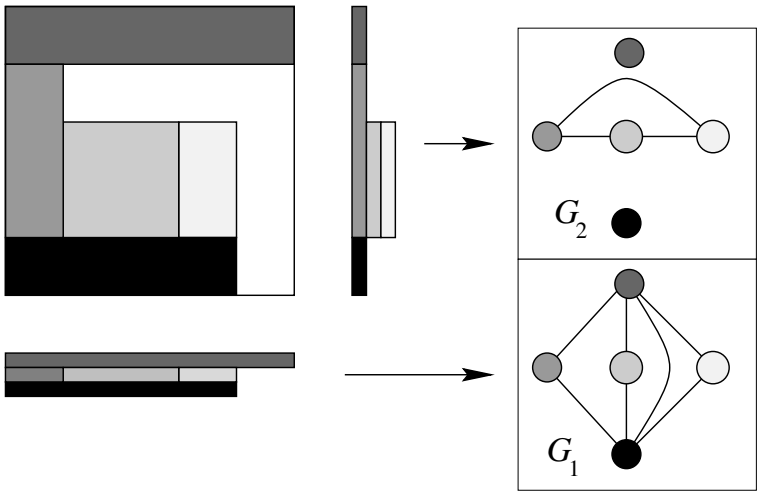


FIG. 3.1. *The projections of the boxes onto the coordinate axes define interval graphs (here in two dimensions:* $\boldsymbol{G_1}$ *and* $\boldsymbol{G_2}$*).*

poses are discussed later). It is not hard to check that the following three conditions must be satisfied by all d-tuples of graphs G_i that are constructed from a feasible packing:

C1: G_i is an interval graph for all $i \in \{1, \dots, d\}$.

C2: Any independent set S of G_i is i-admissible for all $i \in \{1, \cdots, d\}$, i.e., $w_i(S) = \sum_{v \in S} w_i(v) \leq h_i$, because all boxes in S must fit into the container in the ith dimension.

C3: $\cap_{i=1}^{d} E_i = \emptyset$. In other words, there must be at least one dimension in which the corresponding boxes do not overlap.

A d-tuple of component graphs satisfying these necessary conditions is called a *packing class*. The remarkable property (proven in [30, 10]) is that these three conditions are also sufficient for the existence of a feasible packing.

THEOREM 3.1 (Fekete, Schepers). *A set of d-dimensional boxes allows a feasible packing iff there is a packing class, i.e., a d-tuple of graphs $G_i = (V, E_i)$ that satisfies conditions* C1, C2, C3*.*

This allows us to consider only packing classes in order to decide the existence of a feasible packing, and to disregard most of the geometric information.

3.3. Solving OPPs. Our search procedure works on packing classes, i.e., d-tuples of component graphs with the properties C1, C2, C3. Because each packing class represents not only a single packing but a whole family of equivalent packings, we are effectively dealing with more than one possible candidate for an optimal packing at a time. (The reader may check for the example in Figure 3.1 that there are 36 different feasible packings that correspond to the same packing class.)

For finding an optimal packing, we use a branch-and-bound approach. The search tree is traversed by depth first search; see [12, 30] for details. Branching is done by deciding about a single pair of vertices b, c, whether the corresponding edge is contained in E_i or is not contained in E_i, i.e., $\{b, c\} \in E_i$ or $\{b, c\} \notin E_i$. So in fact, there are three classes of edges; those which are fixed to be in E_i, those which are fixed not to be in E_i (nonedges), and those for which it is not decided yet whether or not they

will be contained in E_i. After each branching step, it is checked whether one of the three conditions C1, C2, C3 is violated with respect to the currently fixed edges and nonedges; furthermore it is checked whether a violation can be avoided only by fixing further (formerly undecided) edges or nonedges. Testing for two of the conditions C1–C3 is easy: enforcing C3 is obvious; checking C2 can be done efficiently, since $\overline{G}_i$ is a comparability graph and, as mentioned before, in those graphs maximum weighted cliques can be done efficiently. Note that for this step only nonedges are used, i.e., pairs of vertices for which it has been decided already that they are not contained in E_i. In order to ensure that property C1 is not violated, we use some graph-theoretic characterizations of interval graphs and comparability graphs. These characterizations are based on two forbidden substructures. (Again, see [17] for details; the first condition is based on the classical characterizations by [15, 16]: a graph is an interval graph *iff* its complement has a transitive orientation, and it does not contain any induced chordless cycle of length 4.) In particular, the following configurations have to be avoided:

G1: induced chordless cycles of length 4 in E_i;
G2: so-called 2-chordless odd cycles in the set of edges excluded from E_i (see [12, 17] for details);
G3: infeasible stable sets in E_i.

Each time we detect such a fixed subgraph, we can abandon the search on this node. Furthermore, if we detect a fixed subgraph, except for one unfixed edge, we can fix this edge, such that the forbidden subgraph is avoided.

Our experience shows that in the considered examples these conditions are already useful when only small subsets of edges have been fixed, because by excluding small subconfigurations like induced chordless cycles of length 4, each branching step triggers a cascade of more fixed edges.

4. Packing problems with precedence constraints. As mentioned in the above discussion, a key advantage of considering packing classes is that it makes possible the consideration of packing problems independent of precise geometric placement, and that it allows arbitrary feasible interchanges of placements. However, for most practical instances, we have to satisfy additional constraints for the temporal placement, i.e., for the relative start times of jobs. For our approach, the nature of the data structures may simplify these problems from three-dimensional to purely two-dimensional ones: If the whole schedule is given, all edges E_t in one of the graphs are determined, so we need only to construct the edge sets E_x and E_y of the other graphs. As worked out in detail in [31, 32], this allows it to solve the resulting problems quite efficiently if the arrangement in time is already given.

A more realistic but also more involved situation arises if only a set of precedence constraints is given but not the full schedule. We describe in the following how further mathematical tools in addition to packing classes allow useful algorithms. Note that our method of dealing with order constraints is not restricted to one (the temporal) dimension; in fact, we can also deal with constraints in several dimensions at once, as demonstrated in section 6; see Figure 6.4.

4.1. Packing classes and interval orders. Any edge $\{v_1, v_2\}$ in a component graph G_i corresponds to an intersection between the projections of boxes 1 and 2 onto the x_i-axis. This means that the complement graph $\overline{G_i}$ given by the complement $\overline{E_i}$ of the edge set E_i consists of all pairs of coordinate intervals that are "comparable": Either the first interval is "to the left" of the second, or vice versa.

Any (undirected) graph of this type is a comparability graph. By orienting edges to point from "left" to "right" intervals, we get a partial order of the set V of vertices, a so-called *interval order* [13, 26]. Obviously, this order relation is transitive, inducing a transitive orientation on the (undirected) comparability graph G_i. See Figure 4.1 for a (two-dimensional) example of a packing class, the corresponding comparability graphs, the transitive orientations, and the packing corresponding to the transitive orientations.

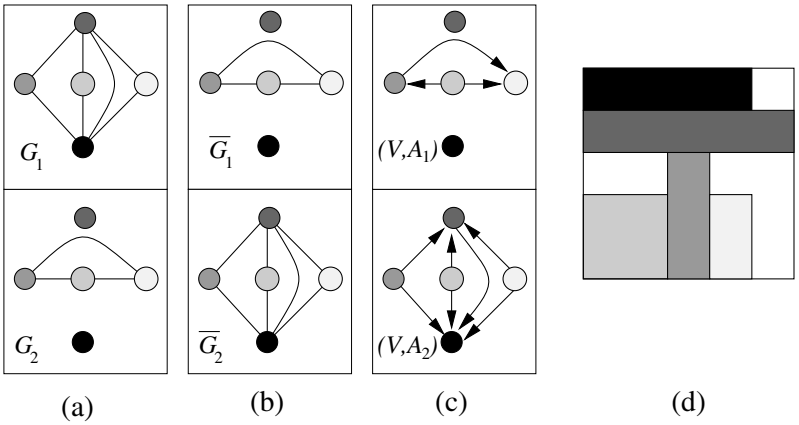


FIG. 4.1. (a) *A two-dimensional packing class.* (b) *The corresponding comparability graphs.* (c) *Two transitive orientations.* (d) *A feasible packing corresponding to the orientation.*

Now consider a situation where we need to satisfy a partial order $P = (V, A_P)$ of precedence constraints in the time dimension. It follows that each arc $a = (u, w) \in A_P$ in this partial order forces the corresponding undirected edge $e = \{u, w\}$ to be excluded from E_i. Thus, we can simply initialize our algorithm for constructing packing classes by fixing all undirected edges corresponding to A_P to be contained in $\overline{E_i}$. After running the original algorithm, we may get additional comparability edges. As the example in Figure 4.2 shows, this causes an additional problem: Even if we know that the graph $\overline{G_i}$ has a transitive orientation, and all arcs $a = (u, w)$ of the precedence order (V, A_P) are contained in $\overline{E_i}$ as $e = \{u, w\}$, it is not clear that there is a transitive orientation that contains all arcs of A_P.

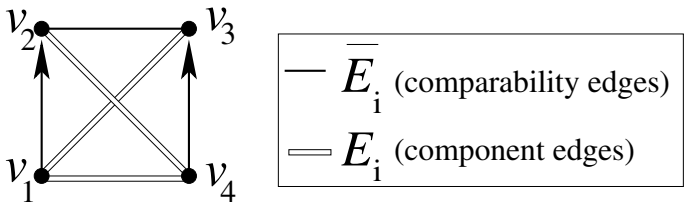


FIG. 4.2. *A comparability graph* $\overline{\boldsymbol{G_i}} = (\boldsymbol{V}, \overline{\boldsymbol{E_i}})$ *with a partial order* $\boldsymbol{P}$ *contained in* $\overline{\boldsymbol{E_i}}$*, such that there is no transitive orientation of* $\overline{\boldsymbol{G_i}}$ *that extends* $\boldsymbol{P}$*.*

4.2. Extending partial suborders. Consider a comparability graph $\overline{G}$ that is the complement of an interval graph G. The problem TOP of deciding whether $\overline{G}$ has a transitive orientation that extends a given partial order P has been studied in the context of scheduling. Korte and Möhring [21] give a linear-time algorithm for

determining a solution, or deciding that none exists. Their approach is based on a very special data structure called *modified PQ-trees*.

In principle it is possible to solve higher-dimensional packing problems with precedence constraints by adding this algorithm as a black box to test the leaves of our search tree for packing classes: In case of failure, backtrack in the tree. However, the resulting method cannot be expected to be reasonably efficient: During the course of our tree search, we are not dealing with one fixed comparability graph but only build it while exploring the search tree. This means that we have to expect spending a considerable amount of time testing similar leaves in the search tree, i.e., comparability graphs that share most of their graph structure. It may be that already a very small part of this structure that is fixed very "high" in the search tree constitutes an obstruction that prevents a feasible orientation of all graphs constructed below it. So a "deep" search may take a long time to get rid of this obstruction. This makes it desirable to use more structural properties of comparability graphs and their orientations to make use of obstructions already "high" in the search tree.

4.3. Implied orientations. As in the basic packing class approach, we consider the component graphs G_i and their complements, the comparability graphs $\overline{G_i}$. This means that we continue to have three basic states for any edge:

1. edges that have been fixed to be in E_i, i.e., *component edges*;
2. edges that have been fixed to be in $\overline{E_i}$, i.e., *comparability edges*;
3. *unassigned edges.*

In order to deal with precedence constraints, we also consider orientations of the comparability edges. This means that during the course of our tree search, we can have three different possible states for each comparability edge:

2a. one possible orientation;
2b. the opposite possible orientation;
2c. no assigned orientation.

A stepping stone for this approach arises from considering the following two configurations; see Figure 4.3. The first configuration (shown in the left part of the

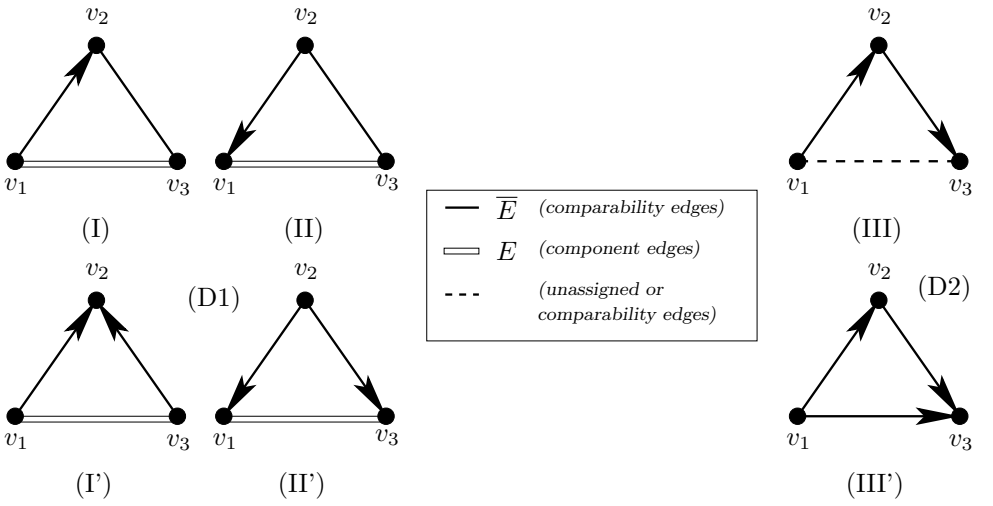


FIG. 4.3. *Implications for edges and their orientations: Above are P_3 implications (D1, left) and transitivity implications (D2, right); below are the forced orientations of edges.*

figure) consists of the two comparability edges $\{v_1, v_2\}$, $\{v_2, v_3\} \in \overline{E_i}$, such that the third edge $\{v_1, v_3\}$ has been fixed to be an edge in the component graph E_i. Now any orientation of just one of the comparability edges forces the orientation of the other comparability edge. In Figure 4.3 the oriented edge in (I) forces the orientation of the second edge as shown in (I'), and similarly for (II) and (II'). Because this configuration corresponds to a partially oriented induced path on three vertices, a P_3 in $\overline{G_i}$, we call this arrangement a P_3 *implication*.

The second configuration (shown in the right part of the figure) consists of two directed comparability edges $(v_1, v_2), (v_2, v_3)$. In this case we know that edge $\{v_1, v_3\}$ must also be a comparability edge, with an orientation of (v_1, v_3). Because this configuration arises directly from transitivity in $\overline{G_i}$, we call this arrangement a *transitivity implication*.

Clearly, any implication arising from one of the above configurations can induce further implications.

In particular, when considering only sequences of P_3 implications, we get a partition of comparability edges into P_3 *implication classes* that will be used in more detail in section 5. Two comparability edges are in the same P_3 implication class, iff there is a sequence of P_3 implications, such that orienting one edge forces the orientation of the other edge. It is not hard to see that the P_3 implication classes form a partition of the comparability edges, because we are dealing with an equivalence relation. For an example, consider the arrangement in Figure 4.2. Here all three comparability edges $\{v_1, v_2\}$, $\{v_2, v_3\}$, and $\{v_3, v_4\}$ are in the same P_3 implication class. Now the orientation of (v_1, v_2) implies the orientation (v_3, v_2), which in turn implies the orientation (v_3, v_4), contradicting the orientation of $\{v_3, v_4\}$ in the given partial order P.

We call a violation of a P_3 implication a P_3 *conflict*.

As the example in Figure 4.4 shows, excluding only P_3 conflicts when recursively carrying out P_3 implications does not suffice to guarantee the existence of a feasible orientation: Working through the queue of P_3 implications, we end up with a directed cycle, which violates a transitivity implication.

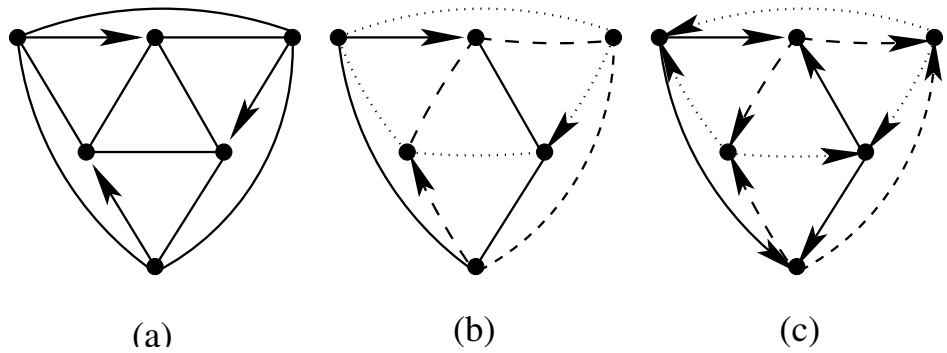


FIG. 4.4. (a) *A graph* $\overline{G_i}$ *with a partial order is formed by three directed edges;* (b) *there are three* P_3 *implication classes that each have one directed arc;* (c) *carrying out* P_3 *implications creates directed cycles, i.e., transitivity conflicts.*

We call a violation of a transitivity implication a *transitivity conflict*.

Summarizing, we have the following necessary conditions for the existence of a transitive orientation that extends a given partial order P:

D1: Any P_3 implication can be carried out without a conflict.

D2: Any transitivity implication can be carried out without a conflict.

These necessary conditions are also sufficient.

THEOREM 4.1. *Let $P = (V, <)$ be a partial order with arc set A_P that is contained in the edge set E of a given comparability graph $G = (V, E)$. A_P can be extended to a transitive orientation of G iff all arising P_3 implications and transitivity implications can be carried out without creating a P_3 conflict or a transitivity conflict.*

A full proof and further mathematical details are described in section 5. This extends previous work by Gallai [14], who extensively studied implication classes of comparability graphs. See Kelly [20], Möhring [26] for helpful surveys on this topic, and Krämer [23] for an application in scheduling theory.

5. Extending partial orientations.

Modular decomposition. The concept of *modular decomposition* of a graph was first introduced by Gallai [14] for studying comparability graphs. This powerful decomposition scheme has a variety of applications in algorithmic graph theory; for further material on this concept and its application the interested reader is referred to [20, 27].

A *module* of a graph $G = (V, E)$ is a vertex set $M \subseteq V$ such that each vertex $v \in V \setminus M$ is either adjacent to all vertices or to no vertex of M in G. (Intuitively speaking, all vertices of a module "look the same" to the other vertices of the graph.) A module is called trivial if $|M| \leq 1$ or $M = V$. A graph G is called *prime* if it contains only trivial modules. Using the concept of modules one can define a decomposition scheme for general graphs by decomposing it recursively into subsets, each of which is a module of G, stopping when all sets are singletons. First of all, observe that every connected component of a given graph G forms a module. It is not hard to see that also every coconnected component of G is a module. If both G and its complement are connected, then the decomposition needs a further idea. Consider the graph in Figure 5.1. Obviously it is connected and coconnected and has a huge number of nontrivial modules. However, if one identifies the *maximal proper submodules* of G, i.e., those modules M that are inclusion-maximal modules of G with $M \neq V$, then one obtains a partition of the vertex set. The corresponding modules of the example G are $M_1 = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, $M_2 = \{20\}$, $M_3 = \{10, 11\}$, $M_4 = \{12, 13, 14, 15, 16, 17, 18, 19\}$.

Gallai [14] showed that any graph G has a particular decomposition (the so-called *canonical decomposition*) of its vertex set into a set of modules with a variety of nice properties. He observed that any graph G is either of *parallel type*, i.e., G is not connected; or G is of *series type*, i.e., $\overline{G}$ is not connected; or G is of *prime type*, i.e., G and $\overline{G}$ are connected. In the first case the canonical decomposition is defined by the set of connected components; in the second case the canonical decomposition is given by the connected components of $\overline{G}$; finally, for prime-type graphs, the canonical decomposition is given by decomposing G into its maximal proper submodules. Gallai also showed that this decomposition is unique.

This recursive decomposition defines a *decomposition tree* $T(G)$ for a given graph G in a very natural way: Create a root vertex of $T(G)$ for the trivial module G itself. Label it series, parallel, or prime, depending on the type of G. For each nonsingleton module of the canonical decomposition of G create a tree vertex, labeled as series-, parallel-, or prime-type node, depending on the type of the module, and make it a child of the vertex corresponding to G; for each singleton module add a tree vertex labeled with the corresponding singleton. Now proceed recursively for each subgraph corresponding to a nontrivial module in the decomposition tree, until all leaves of the

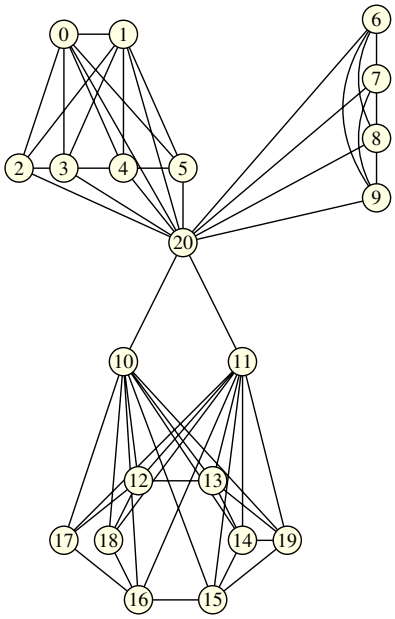


FIG. 5.1. *An example graph G.*

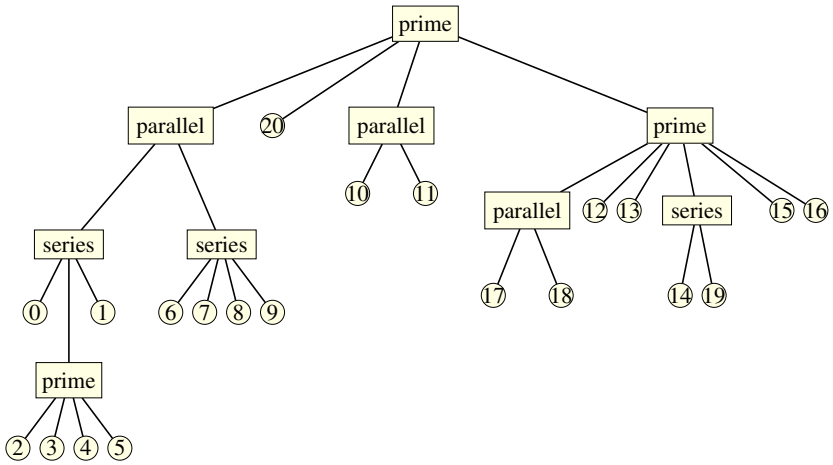


FIG. 5.2. *A modular decomposition tree for the graph G shown in Figure* 5.1.

tree are labeled with singletons. Consequently, the leaves of the tree correspond to the vertices of the graph, while all internal vertices correspond to nontrivial modules of the canonical decomposition of the corresponding parent vertex in $T(G)$. See Figure 5.2 for the decomposition tree of our example.

The *decomposition graph* $G^{\#}$ of a graph G is the quotient of G by the canonical

decomposition into the set of modules $\{A_1, \ldots, A_q\}$, i.e., $V(G^{\#}) = \{A_1, \ldots, A_q\}$, and distinct vertices A_i and A_j are joined by an edge in $G^{\#}$ iff there is an A_iA_j-edge in G. In the following we will look at the decomposition graphs corresponding to internal vertices of $T(G)$ and refer to them as the decomposition graphs of T.

In our example, the decomposition graph $G^{\#}$ of G, i.e., to the root node of $T(G)$, is a path on four vertices, given by

$$G^{\#} = (\{M_1, M_2, M_3, M_4\}, \{\{M_1, M_2\}, \{M_2, M_3\}, \{M_3, M_4\}\}).$$

Modular decomposition and transitive orientations. An important property of the modular decomposition is its close relationship to the concept of P_3 implication classes. Gallai observed the following properties of P_3 implication classes with respect to the modular decomposition.

PROPOSITION 5.1 (see Gallai [14]). *Let $G = (V, E)$ be an undirected graph.*

(1) *If G is not connected and $G_1, \ldots, G_q$ ($q \geq 2$) are the components of G, then the P_3 implication classes of $G_1, \ldots, G_q$ are exactly the P_3 implication classes of G.*
(2) *If $\overline{G}$ is not connected (so that G is connected), $\overline{G_1}, \ldots, \overline{G_q}$ ($q \geq 2$) are the components of $\overline{G}$, and $A_i = V(G_i)$, then A_i and A_j are completely connected to each other whenever $1 \leq i < j \leq q$. Moreover, for all such i and j, the set of A_iA_j-edges form an P_3 implication class E_{ij} of G. The P_3 implication classes of G that are distinct from any E_{ij} are exactly the P_3 implication classes of the graphs $G_i = G[A_i]$ ($i = 1, \ldots, q$).*
(3) *If G and $\overline{G}$ are both connected and have more than one vertex, and the canonical decomposition of G is given by $\{A_1, \ldots, A_q\}$, then we have the following:*
   (a) *If there is one edge between A_i and A_j ($1 \leq i < j \leq q$), then all edges between A_i and A_j are in G.*
   (b) *The set of all edges of G that join different A_i's forms a single P_3 implication class C of G. Every vertex of G is incident with some edge of C (i.e., $V(C) = V(G)$).*
   (c) *The P_3 implication classes of G that are distinct from C are exactly the P_3 implication classes of the graphs $G_i = G[A_i]$ ($1 \leq i \leq q$).*

This strong relationship between P_3 implication classes and the modules in the canonical decomposition of a given graph is a powerful tool for studying graphs having a transitive orientation. Note that the fastest known algorithms for recognizing comparability graphs make extensive use of this relationship. Gallai used the above properties (among others) for proving the following theorem.

THEOREM 5.2 (see Gallai [14]). *Let G be a nonempty graph, let $T = T(G)$ be the tree decomposition of G, and let H be a vertex set corresponding to a node of T.*

(1) *If G is transitively oriented, and A and B are descendants of H in T, then every A, B-edge of G is oriented in the same direction (to or from A). Therefore, $H^{\#}$ receives an induced transitive orientation.*
(2) *Conversely, assuming that $H^{\#}$ is transitively orientable for each $H \in T$, one can choose an arbitrary transitive orientation of each $H^{\#}$ and induce a transitive orientation of G by orienting all A, B-edges (for A and B descendants of H in T) in the same direction that $\{A, B\}$ is oriented in $H^{\#}$.*

It is straightforward to draw the following helpful corollaries from this theorem.

COROLLARY 5.3. *A graph G is a comparability graph iff every decomposition graph in the tree decomposition of G is a comparability graph.*

COROLLARY 5.4. *Let G be a comparability graph and T its tree decomposition. Assigning to each of the decomposition graphs of T a transitive orientation independently results in a transitive orientation of G.*

Furthermore, if only a partial orientation of G is given and we are interested in extending this orientation to a transitive orientation of G, we can formulate the following lemma.

LEMMA 5.5. *Let G be a comparability graph and T its tree decomposition. Furthermore, let P be a partial orientation of G, assigning orientations to some but not all P_3 implication classes of G. P is extendible to a transitive orientation of G iff for each decomposition graph $H^{\#}$ of T the orientation induced on $H^{\#}$ by P is extendible to a transitive orientation on $H^{\#}$.*

Proof. The proof follows immediately from Theorem 5.2(2). □

Now we are ready to prove Theorem 4.1: Conditions D1 and D2 are also sufficient.

Proof of Theorem 4.1. Suppose there is a transitive orientation F of G that contains P. Because F is a transitive orientation, all arcs implied by P_3 or transitivity implications are contained in F. Furthermore, there cannot be any P_3 or transitivity conflict in F, again because F is a transitive orientation. Thus F shows that all arising P_3 and transitivity implications can be carried out without creating a P_3 or transitivity conflict.

Suppose now that D1 and D2 are satisfied, i.e., that there is a directed graph F consisting of all arcs of P together with all orientations of edges of G that are implied by a sequence of P_3 and transitivity implications of arcs of P. In other words, F contains all arcs that are forced by P_3 or transitivity implications together with all their implied arcs; i.e., all arcs that are forced by arcs of F are also contained in F. We show that F can be extended to a transitive orientation of G.

First, observe that, by assumption, there cannot be a P_3 or transitivity conflict in F. In particular, F is an orientation of edges of G and for each P_3 implication class C of G that has at least one edge that is oriented in F, all edges of C are oriented in F and this orientation is conflict-free. By Corollary 5.4, every single conflict-free oriented P_3 implication class of G by itself is extendible to a transitive orientation of G.

Now let T be the decomposition tree of G and consider the decomposition graphs corresponding to T. By the above observation, the orientation of an P_3 implication class C in F implies an orientation of the edge(s) corresponding to this P_3 implication class in the decomposition graphs of T. More precisely, by Proposition 5.1(2), for every series-type node H of T each edge $e = \{AB\}$ of $H^{\#}$ corresponds exactly to one P_3 implication class C_e of G. If C_e is oriented conflict-free in F, this orientation directly induces an orientation of e (see Theorem 5.2). For a prime-type node H the set of edges joining different A_is forms exactly one P_3 implication class C_E of G (see Proposition 5.1(3). Again, if C_E is oriented conflict-free in F, this orientation immediately implies an orientation on $H^{\#}$.

All we have to show now is that for each decomposition graph $H^{\#}$ of T, the partial orientation implied by F can be extended to a transitive orientation of $H^{\#}$. Then, by Corollary 5.4, the implied orientation of G is transitive.

By Corollary 5.4, a *parallel-type node* of T cannot create a contradiction to transitivity—it does not contain any edges.

Also a *prime-type node* of T cannot create a contradiction: All of its edges are contained in only one P_3 implication class and, because all P_3 implication classes of G contained in F are oriented conflict-free, the corresponding orientation induced by

F on this single P_3 implication class has to be transitive.

This leaves the case of *series-type nodes*. Suppose there is a series-type node H of T with decomposition graph $H^{\#}$, for which the partial orientation implied by F cannot be extended to a transitive orientation of $H^{\#}$. Then we claim that this partial orientation has to be cyclic: By definition for each series-type node H of T the decomposition graph $H^{\#}$ is a complete graph and every acyclic partial orientation of a complete graph can be extended to a transitive orientation of this complete graph by taking any topological ordering of the vertices that agrees with the partial orientation. Hence, the partial orientation on $H^{\#}$ has to contain a directed cycle.

However, by the definition of T and the implied orientation of $H^{\#}$ by F, a directed cycle in $H^{\#}$ immediately implies a cyclically oriented cycle in F. Furthermore, with every consecutive pair of oriented edges (x, y), (y, z) of this cycle also the oriented edge (x, z) (which is implied by transitivity) has to be contained in F. Iterating this argument results in an cyclically oriented triangle in F, which is a transitivity conflict. This contradicts our assumption that there are no transitivity conflicts. □

6. Computational experiments.

6.1. Solving problems with precedence constraints. We start by fixing for all arcs $(u, v) \in A_P$ the edge $\{u, v\}$ as an edge in the comparability graph $\overline{G_i}$, and we also fix its orientation to be (u, v). In addition to the tests for enforcing the conditions for unoriented packing classes (C1, C2, C3), we employ the implications suggested by conditions D1 and D2. For this purpose we check directed edges in $\overline{G_i}$ for being part of a triangle that gives rise to either implication. Any newly oriented edge in $\overline{G_i}$ gets added to a queue of unprocessed edges. Like for packing classes, we can again get cascades of fixed edge orientations. If we get an orientation conflict or a cycle conflict, we can abandon the search on this tree node. The correctness of the overall algorithm follows from Theorem 4.1; in particular, the theorem guarantees that we can carry out implications in an arbitrary order. In the following we present our results for different types of instances: The video-codec benchmark described in section 6.3 arises from an actual application to FPGAs. In section 6.4 we give a number of results arising from different geometric packing problems.

Our code was implemented in C++ and was run on a SUN Ultra 10 with 333 MHz.

The first example is a numerical method for solving a *differential equation* (DE) with 11 nodes. The node operations are either multiplications or ALU-type operations. In a second example, a video-codec using the H.261 norm is optimized. These examples are meant to demonstrate the general applicability of our method for practical problems; given other problem instances, or additional constraints, we can easily adapt our algorithm.

6.2. DE benchmark. Let the module library contain two hardware modules (box types): an array-multiplier and a module of type ALU that realizes all other node operations (comparison, addition, subtraction). For a word-length of n=16 bits, we assume a module geometry of 16×1 cells for the ALU module and of 16×16 cells for the multiplier. Furthermore, the execution time of an ALU node takes one clock cycle, while a multiplication requires two clock cycles on our target chip.

The dependency graph is shown in Figure 1.3. First, we compute the transitive closure of all data dependencies to allow our algorithm to find contradictions to feasible packings already in the input.

Next, we solve several instances of the BMP problem for different values of h_t

reported in Table 6.1. Each h_t listed yields a test case for which the container size is minimized (MinA), assuming $h_x = h_y$. Also shown is the CPU-time needed for finding a solution.

TABLE 6.1
Computational results for optimizing reconfigurations for the DE benchmark.

Test	Container sizes			
	h_t	h_x	h_y	CPU-time
1	6	32	32	55.76 s
2	13	17	17	0.04 s
3	14	16	16	0.03 s

The reported optimization times were measured as the CPU-times on a SUN Ultra 10 with 333 MHz.

For the DE benchmark, it turns out that a chip of 32×32 freely programmable cells is necessary to obtain a latency between 6 and 12 clock cycles. As the longest path in the graph has length 6, there does not exist any faster schedule. For 12 and 13 cycles, a chip of size 17×17 is necessary; for $h_t \geq 14$, a chip of size 16×16 cells is sufficient, which is the smallest chip possible to implement the problem, as one multiplication by itself uses the full chip.

The SPP is solved in a similar way. The tradeoff between area size and necessary time is visualized in Figure 6.1, in which the Pareto-optimal points are shown. The figure also shows the Pareto points for the case where no partial order needs to be satisfied (shown dashed).

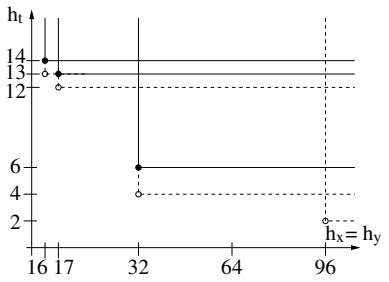


FIG. 6.1. *Pareto-optimal points for minimizing chip area and processing time for the DE benchmark, including partial order constraints (solid lines), and without consideration of partial order constraints (dashed lines).*

6.3. Video-codec benchmark. Figure 6.2 shows a block diagram of the operation of a hybrid image sequence coder/decoder that arises from the FPGA application. The purpose of the coder is to compress video images using the H.261 standard. In this device, transformative and predictive coding techniques are unified. The compression factor can be increased by a predictive method for motion estimates: blocks inside a frame are predicted from blocks of previous images.

The blocks of the operational description shown in the figure possess the granularity of more complex functions. However, this description contains no information corresponding to timing, architecture, and mapping of blocks onto an architecture. The resulting problem graph contains a subgraph for the coder and one subgraph for the decoder.

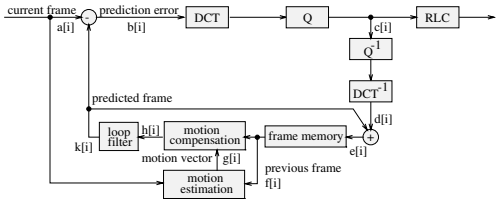


FIG. 6.2. *Block diagram of a video-codec (H.*261*).*

For realizing the device we have a library of three different modules. One is a simple processor core with a (normalized) area requirement of 625 units (25×25 cells, normalized to other modules in order to obtain a coarser grid) called PUM. Second, there are two dedicated special-purpose modules: a block matching module (BMM) that is used for motion estimation and requires $64 \times 64 = 4096$ cells; and a module DCTM for computing DCT/IDCT-computations, requiring $16 \times 16 = 256$ cells. Again, the BMP and the CSPP were considered, and the makespan was minimized for different latency constraints. Here there is only one Pareto point found, shown in Table 6.2.

TABLE 6.2
Optimizing reconfigurations for the video-codec.

Test	Container sizes			
	h_t	h_x	h_y	CPU-time
1	59	64	64	24.87 s

6.4. Geometric instances. We describe computational results for two types of two-dimensional objects. See Table 6.3 for an overview. The first class of instances was constructed from a particularly difficult random instance of the two-dimensional knapsack problem (see [7]). Results are given for order constraints of increasing size. In order to give a better idea of the computational difficulty, we give separate running times for finding an optimal feasible solution and for proving that this solution is best possible.

TABLE 6.3
Optimal packing with order constraints.

Instance	Optimal h_t	h_x	Upper bound	Lower bound
okp17-0	169	100	7.29 s	179 s
okp17-1	172	100	6.73 s	1102 s
okp17-2	182	100	5.39 s	330 s
okp17-3	184	100	236 s	553 s
okp17-4	245	100	0.17 s	0.01 s
square21-no	112	112	84.28 s	0.01 s
square21-mat	117	112	15.12 s	277 s
square21-tri	125	112	107 s	571 s
square21-2mat	[118,120]	[118,120]	346 s	476 s

See Table 6.4 for the exact sizes of the 17 rectangles involved, Table 6.3 for the resuting optimal packings, and Figure 6.3 for their geometric layout. For easier

reference, the boxes in the `okp17` instances are labeled 1–17 in the given order.

The second class of instances arises from the well-known tiling of a 112 × 112 square by 21 squares of different sizes. Again we have added order constraints of various sizes; see Table 6.5 for the exact dimensions. For the instance square21-2mat (with order constraints in two dimensions), we could not close the gap between upper and lower bound. For this instance we report the running times for achieving the best known bounds. Layouts of best solutions are shown in Figure 6.4, with corresponding dimensions listed in Table 6.3.

TABLE 6.4
The problem instances `okp17`.

`okp17`:	base width of container = 100, number of boxes = 17
sizes =	[(8,81),(5,76),(42,19),(6,80),(41,48),(6,86),(58,20),(99,3),(9,52), (100,14),(7,53),(24,54),(23,77),(42,32),(17,30),(11,90),(26,65)]
okp17-0:	no order constraints
okp17-1:	11→8, 11→16
okp17-2:	11→8, 11→16, 8→16
okp17-3:	11→8, 11→16, 8→16, 8→17, 11→7, 16→7
okp17-4:	11→8, 11→16, 8→16, 8→17, 11→7, 16→7, 17→16

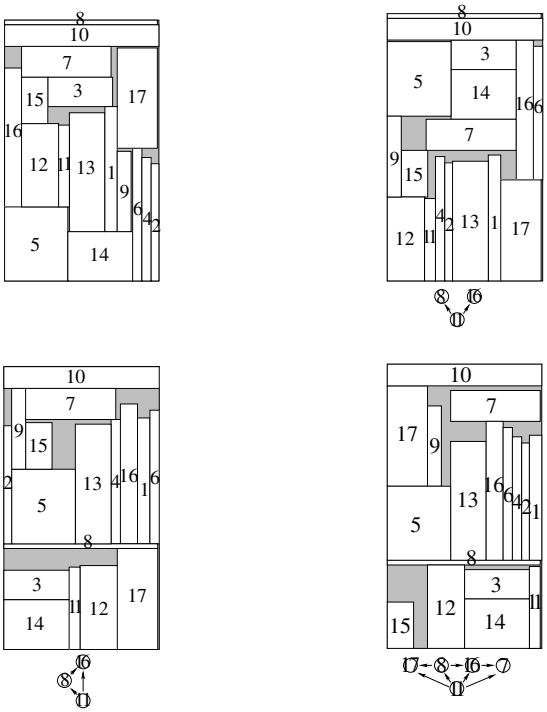


FIG. 6.3. *(top left) An optimal packing of okp*17-0 *of height* 169*; (top right) an optimal packing of okp*17-1 *of height* 172*; (lower left) an optimal packing of okp*17-2 *of height* 182*; (lower right) an optimal packing of okp*17-3 *of height* 184.

TABLE 6.5
The problem instances `square21`.

`square21`:	base width of container = 112, number of boxes = 21
sizes =	[(50,50),(42,42),(37,37),(35,35),(33,33),(29,29),(27,27),(25,25), (24,24),(19,19),(18,18),(17,17),(16,16),(15,15),(11,11),(9,9),(8,8), (7,7),(6,6),(4,4),(2,2)]
square21-0:	no order constraints
square21-mat:	2→4, 6→7, 8→9, 11→15, 16→17, 18→19, 24→25, 27→29, 33→35, 37→42, 2→50, 50→4
square21-tri:	2→15, 15→17, 2→27, 4→16, 16→29, 4→29, 6→17, 17→33, 6→33, 7→18, 18→35, 7→35, 8→19, 19→37, 8→37, 9→24, 24→42, 9→42, 11→25, 25→50, 11→50
square21-2mat:	x-constraints: 2→19, 6→25, 8→29, 11→35, 16→42, 18→4, 24→7, 27→9, 33→15, 37→17, 50→4, 18→50 y-constraints: 2→4, 6→7, 8→9, 11→15, 16→17, 18→19, 24→25, 27→29, 33→35, 37→42, 2→50, 50→4

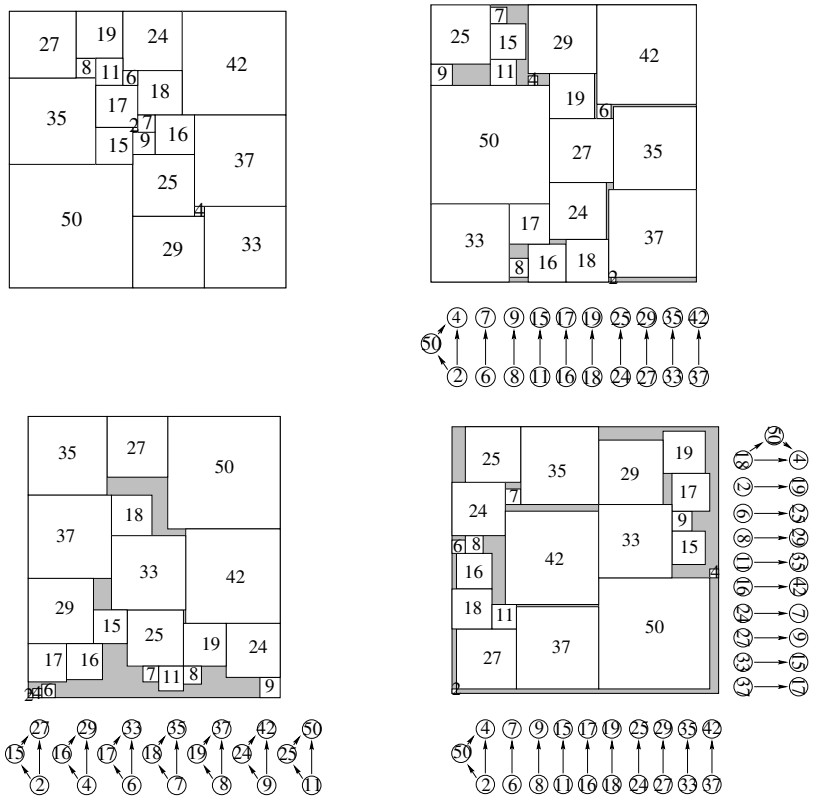


FIG. 6.4. *(top left) An optimal packing of square*21-0 *of height* 112*; (top right) an optimal packing of square*21-*mat of height* 117*; (lower left) an optimal packing of square*21-*tri of height* 125*; (lower right) a packing of square*21-2*mat of size* 120 × 120.

Acknowledgments. We are extremely grateful to Jörg Schepers for letting us continue the work with the packing code that he started as part of his thesis, and for several helpful hints, despite his departure to industry. We thank Nicole Megow for helpful comments, Marc Uetz for a useful discussion on resource-constrained scheduling, and an anonymous referee for a number of helpful suggestions that helped to improve the presentation of this paper.

REFERENCES

[1] Atmel, *AT6000 FPGA Configuration Guide*, Atmel Inc., San Jose, CA, 1997.
[2] J. E. Beasley, *An exact two-dimensional non-guillotine cutting tree search procedure*, Oper. Res., 33 (1985), pp. 49–64.
[3] J. E. Beasley, *OR-Library: Distributing test problems by electronic mail*, J. Oper. Res. Soc., 41 (1990), pp. 1069–1072.
[4] S. P. Fekete, E. Köhler, and J. Teich, *Extending partial suborders*, in Proceedings of the 1st Cologne-Twente Workshop on Graphs and Combinatorial Optimization, J. H. H. Broersma, U. Faigle and S. Pickl, eds., Electron. Notes Discrete Math. 8, Elsevier, Amsterdam, 2001.
[5] S. P. Fekete, E. Köhler, and J. Teich, *Multi-dimensional packing with order constraints*, in Proceedings of the 7th International Workshop on Algorithms and Data Structures, Lecture Notes in Comput. Sci. 2125, Springer-Verlag, Berlin, 2001, pp. 300–312.
[6] S. P. Fekete, E. Köhler, and J. Teich, *Optimal FPGA module placement with temporal precedence constraints*, in Proceedings of Design, Automation and Test in Europe, IEEE Computer Society Press, Los Alamitos, CA, 2001, pp. 658–665.
[7] S. P. Fekete and J. Schepers, *A new exact algorithm for general orthogonal d-dimensional knapsack problems*, in Proceedings of the 5th Annual European Symposium on Algorithms, Lecture Notes in Comput. Sci. 1284, Springer-Verlag, Berlin, 1997, pp. 144–156.
[8] S. P. Fekete and J. Schepers, *New classes of lower bounds for bin packing problems*, in Proceedings of the 6th International Conference on Integer Programming and Combinatorial Optimization, Lecture Notes in Comput. Sci. 1412, Springer-Verlag, Berlin, 1998, pp. 257–270.
[9] S. P. Fekete and J. Schepers, *New classes of lower bounds for the bin packing problem*, Math. Program., 91 (2001), pp. 11–31.
[10] S. P. Fekete and J. Schepers, *A combinatorial characterization of higher-dimensional orthogonal packing*, Math. Oper. Res., 29 (2004), pp. 353–368.
[11] S. P. Fekete and J. Schepers, *A general framework for bounds for higher-dimensional orthogonal packing problems*, Math. Methods Oper. Res., 60 (2004), pp. 311–329.
[12] S. P. Fekete, J. Schepers, and J. v. d. Veen, *An exact algorithm for higher-dimensional orthogonal packing*, Oper. Res., to appear.
[13] P. C. Fishburn, *Interval Orders and Interval Graphs*, John Wiley & Sons, New York, 1985.
[14] T. Gallai, *Transitiv orientierbare Graphen*, Acta Math. Acad. Sci. Hungar., 18 (1967), pp. 25–66.
[15] A. Ghouilà-Houri, *Caractérization des graphes non orientés dont on peut orienter les arrêtes de manière à obtenir le graphe d'une relation d'ordre*, C.R. Acad. Sci. Paris, 254 (1962), pp. 1370–1371.
[16] P. C. Gilmore and A. J. Hoffmann, *A characterization of comparability graphs and of interval graphs*, Canad. J. Math., 16 (1964), pp. 539–548.
[17] M. C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.
[18] E. Hadjiconstantinou and N. Christofides, *An exact algorithm for general, orthogonal, two-dimensional knapsack problems*, European J. Oper. Res., 83 (1995), pp. 39–56.
[19] C.-H. Huang and J.-Y. Juang, *A partial compaction scheme for processor allocation in hypercube multiprocessors*, in Proceedings of the 1990 International Conference on Parallel Processing, 1990, Pennsylvania State University Press, University Press, University Park, PA, pp. 211–217.
[20] D. Kelly, *Comparability graphs*, in Graphs and Order, I. Rival, ed., D. Reidel, Dordrecht, The Netherlands, 1985, pp. 3–40.
[21] N. Korte and R. Möhring, *Transitive orientation of graphs with side constraints*, in Proceedings of the 11th International Workshop on Graph-Theoretic Concepts in Computer Science, H. Noltemeier, ed., Trauner Verlag, Linz, Austria, 1985, pp. 143–160.
[22] N. Korte and R. H. Möhring, *An incremental linear–time algorithm for recognizing interval*

graphs, SIAM J. Comput., 18 (1989), pp. 68–81.

[23] A. Krämer, *Scheduling multiprocessor tasks on dedicated processors*. Doctoral thesis, Fachbereich Mathematik und Informatik, Universität Osnabrück, Osnabrück, Germany, 1995.

[24] E. L. Lawler, J. K. Lenstra, A. H. G. Rinooy Kan, and D. B. Shmoys, *Sequencing and scheduling: Algorithms and complexity*, in Logistics of Production and Inventory, S. C. Graves, A. H. G. Rinnooy Kan, and P. H. Zipkin, eds., Handbooks Oper. Res. Mangmt. Sci. 4, North–Holland, Amsterdam, 1993, pp. 445–522.

[25] M. G. Luby, J. (S.) Naor, and A. Orda, *Tight bounds for dynamic storage allocation*, SIAM J. Discrete Math., 9 (1996), pp. 155–166.

[26] R. H. Möhring, *Algorithmic aspects of comparability graphs and interval graphs*, in Graphs and Order, I. Rival, ed., D. Reidel, Dordrecht, The Netherlands, 1985, pp. 41–101.

[27] R. H. Möhring, *Algorithmic aspects of the substitution decomposition in optimization over relations, set systems, and Boolean functions*, Ann. Oper. Res., 4 (1985), pp. 195–225.

[28] R. H. Möhring, A. S. Schulz, F. Stork, and M. Uetz, *Solving project scheduling problems by minimum cut computations*, Management Sci., 49 (2003), pp. 330–350.

[29] M. Padberg, *Packing small boxes into a big box*, Math. Methods Oper. Res., 52 (2000), pp. 1–21.

[30] J. Schepers, *Exakte Algorithmen für orthogonale Packungsprobleme*, Technical report 97-302, Doctoral thesis, Universität Köln, Köln, Germany, 1997.

[31] J. Teich, S. P. Fekete, and J. Schepers, *Compile-time optimization of dynamic hardware reconfigurations*, in Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, Las Vegas, 1999, pp. 1097–1103.

[32] J. Teich, S. P. Fekete, and J. Schepers, *Optimal hardware reconfiguration techniques*, J. Supercomput., 19 (2001), pp. 57–75.

[33] J. Weglarz, *Project Scheduling. Recent Models, Algorithms and Applications*, Kluwer Academic Publishers, Norwell, MA, 1999.

[34] Xilinx, *XC*6200 *Field Programmable Gate Arrays*, Technical report, Xilinx, Inc., San Jose, CA, 1996.